*Grenze International Journal of Engineering and Technology, June Issue*

# Hidden Markov Model based Human Speech Emotion Recognition

Shivappa M Metagar[1], Nikhil S. Gajjam[2], Mahesh A. Mahant[3] and Farooque R Sayyed[4]
[1-4]Assistant Professor, CSE Department, Walchand Institute of Technology,Solapur, Maharashtra, India
Email: shivametagar@gmail.com, nikhilgajjam@gmail.com, maheshamahant@gmail.com, sayyed.farooque999@gmail.com

***Abstract*—Speech is an attractive interaction medium due to its many characteristics, along with efficiency and naturalness. Speech can be used to convey attitudes and feelings. This paper offers a examine that uses a Hidden Markov model to identify human emotions in speech. a selection of speech characteristics were retrieved so that it will perceive emotion over speech. Thinking about these speech characteristics feelings had been categorized, and the Hidden Markov model's class potential is examined. right here, a variety of emotions, along with neutral, glad, unhappy, bored, irritated, and worried, are recognized. The retrieved features function the basis for the type overall performance. Additionally included are conclusions regarding the effectiveness and boundaries of the speech emotion reputation gadget primarily based at the various classifiers. simplest the speech-based totally prediction of six human feelings is protected in the studies. More human feelings may be predicted with the aid of expanding it. Some of the samples that belonged to the comfortable magnificence and the apprehensive elegance in SVM and RF were incorrectly expected by using the CNN class strategies. By means of extracting more developments to higher differentiate between those lessons, this can be constant. The goal is to noticeably improve the Speech Emotion popularity system's robustness and real-time evaluation.**

***Keywords*— Hidden Markov Model, speech emotion recognition (SER), fear, anger, boredom, sadness, and happiness.**

## I. INTRODUCTION

Ordinary human interactions are appreciably stimulated by way of emotions. this is critical to our cognitive and logical decision-making. via expressing our feelings and providing comments to others, it allows us to relate to and understand their sentiments. Emotion has been shown to have a vast impact on how humans connect to each other. Emotional expressions display a excellent deal approximately a person's mental fitness.

Several inherent advantages make speech signals a very good the affective computing source. as an example, speech alerts are typically less complicated and less expensive to gather than many other organic indicators.

that is why speech emotion recognition (SER) is of interest to maximum Researchers. The purpose of SER is to deduce a speaker's underlying emotional country from her voice. over the past few years, the sector has seen a rise in research interest. The procedure of translating an acoustic signal captured with the aid of a microphone or cellphone right into a string of phrases is known as speech recognition. For packages like statistics access, document practise, and instructions and control, the diagnosed words would possibly serve as a purpose unto themselves. which will comprehend speech, they can also be used as enter for extra linguistic processing.

*Grenze ID: 01.GIJET.11.2.652*

The facts of ideas created inside the speaker's head is contained in speech, that is an auditory signal. The selection of a suitable emotional speech database, the extraction of useful capabilities, and the introduction of a dependable classifier using a system gaining knowledge of set of rules are the three predominant problems that have to be resolved for a SER system to achieve success. virtually, one of the number one issues with the SER gadget is the extraction of emotional capabilities. the character of speech is bimodal. best the acoustic information found in voice indicators is taken into account by automated voice recognition (ASR). it is less correct in loud environments. because it makes use of both visible and auditory statistics observed in speech, Audio visible Speech popularity (AVSR) outperforms ASR.

One of the oldest types of self-expression is speech. these days, biometric recognition structures and system conversation additionally use these voice indicators. these speech indicators are quasi-stationary, which means they change gradually in timing. Its capabilities are quite desk bound whilst analyzed over a quick enough time span (five-one hundred m sec). however, if the signal qualities change over time, it'll mirror the various speech sounds which are being said.

The short-time period amplitude spectrum of the speech wave form simply represents the records within the spoken sign. This allows us to extract information from speech (phonemes) primarily based on their quick-term amplitude spectrum. The number one venture in voice recognition is that speech alerts range substantially among speakers, unbiased of speakme speeds, content material, and acoustic situations.

It has counseled key speech characteristics like energy, pitch, formant frequency, and Mel-frequency cepstral coefficients (MFCC) that carry information approximately emotion. consequently, to rent a feature set that combines a spread of developments that encompass more emotional records. An ASR gadget's feature analysis issue is crucial to the gadget's typical functionality. 3 distinct levels of speech processing are viable. sign degree processing breaks down signals into tiny devices known as frames even as considering the shape of the human auditory device. Speech phonemes are learnt and processed in phoneme degree processing.

## II. Objective of the Paper

This paper's number one objective is to increase a honest yet comprehensive consultant emotion popularity machine for non-stop speech the usage of a speech characteristic extraction technique and the ideal classifier to boom the detection accuracy. This paper's aim is to offer the first-class effects viable within the region of emotion popularity. It is able to also be applied in a spread of industries, along with the inventory market, as an instance. Upstox is an utility that shows market tendencies using machine learning. Fraud profiles can also be diagnosed with it.

Reduced device autonomy, which results in a lack of protection. The largest obstacle we are working to resolve is speech extraction. Learnt machines can take the area of less productive people. Our strategy is to extract the first-rate speech feasible and offer accurate emotion output. This difficulty is extensive to us since it allows us to accomplish all the things that were previously not possible due to the shortage of machine learning era.

## III. Background

### *A. Speech Recognition*

The technology that deals with methods and techniques to perceive speech from speech signals is called speech recognition. Several traits in artificial intelligence and signal processing strategies have made it simpler and sensible to recognize feelings. some other call for it's far "automated Speech reputation." Voice has been identified as the destiny channel for device communique, specially in pc-based totally structures: The want to deduce emotion from spoken phrases is growing at an exponential rate. Because the field of voice popularity has seen terrific advancements. Numerous voice merchandise, which includes Apple home Pod, Google home, and Amazon Alex, were produced and mostly perform thru voice commands. It appears clean that the quality way to communicate with the machines could be via voice.

### *B. Emotion Recognition*

Emotion recognition is basically the look at of determining emotions. Strategies for figuring out emotions can be identified from speech alerts and facial expressions. Numerous techniques, which include signal processing, machine studying, neural networks, and computer vision, have been evolved to pick out feelings. round the world, research and development are being executed on emotion recognition and evaluation. In research, emotion reputation is becoming increasingly famous as a strategy to many issues and a way to make existence less complicated. In artificial intelligence, in which voice signals are the only input for laptop systems, the

number one requirement for emotion popularity from speech is difficult tasks. Moreover, Speech Emotion reputation (SER) is used in a variety of fields, consisting of BPO and call centres to discover consumer pleasure with merchandise, IVR structures to enhance speech interplay, remedy ambiguities in language, and regulate computer systems primarily based on an man or woman's mood and emotion.

### *C. Speech Emotion Recognition*

The goal of the research task of voice emotion recognition is to infer the emotion from speech alerts. in line with numerous surveys, enhancements in emotion popularity will simplify many methods and improve dwelling situations worldwide. The particular use of SER is described later. Emotion popularity is a tough topic because feelings can vary relying at the surroundings and way of life, and a person's facial reaction might produce uncertain effects. Moreover, speech corpora are insufficient to reliably infer feelings, and there are not sufficient voice databases in lots of languages.

Speech is one of the maximum natural methods that humans express themselves. we rely on it so much that we well known its significance while the use of other channels of communique, together with textual content messages and emails, in which we often utilise emotion to convey the emotions which might be linked to communications. Considering emotions are essential to communication, it's far vital to become aware of and examine them inside the digital global of far off communication that exists nowadays. since emotions are subjective, detecting them is a tough challenge. there's no widely wide-spread agreement on the way to classify or degree them. A SER machine is a set of techniques that examine and categorise speech alerts for you to pick out the feelings which can be present in them. programs for any such system are several and encompass caller-agent interplay evaluation and interactive voice-primarily based assistants. in this have a look at, we examine the acoustic characteristics of audio information from recordings that allows you to identify underlying feelings in recorded speech.

## IV. LITERATURE REVIEW

In keeping with the authors RUHUL AMIN KHALIL et al. in [5], emotion reputation from speech alerts is a essential but hard element of human-pc interaction (HCI). Many strategies, consisting of properly-mounted speech analysis and class methodologies, have been used within the literature on speech emotion popularity (SER) to extract emotions from indicators. Nowadays, deep gaining knowledge of techniques were advised instead for conventional SER techniques. further to discussing some recent research that makes use of deep getting to know strategies for speech-based emotion reputation, this paper gives a excessive-level evaluate of those strategies.

The implementation and assessment techniques for speech emotion recognition fashions had been defined by way of Sabrine Dhaouadi et al. in [6]. There are numerous capability uses for speech emotion popularity (SER), a these days evolved area of observe, in each human-computer and human-human interaction contexts. The framework for detecting feelings in voice alerts is pretty confined. Researchers are presently increasing on the capacities that have an effect on speech emotion identity. There is a good sized deal of uncertainty over the excellent set of tips for classifying feelings and which emotions need to be grouped together. we are looking for approaches to cope with those troubles in these paintings.

to categorise opposing feelings, we employ artificial neural networks (ANN) and guide vector machines (SVMs). Human speech may have a variety of spectral and temporal capabilities that can be recognized. Mel Frequency Cepstral Coefficients (MFCCs) are the category algorithms' most efficient inputs, as some distance as we know. in keeping with the type reports derived from the conducted experiments, the ANN model performed better than the SVM at identifying the emotional data in speech for the required parameters.

In [7], Pavol Harár et al. defined how to use a Deep Neural community (DNN) structure with Convolutional, pooling, and fully related layers for Speech Emotion reputation (SER). The German Corpus (Berlin Database of Emotional Speech)'s 3 elegance subsets (indignant, indifferent, and sad) comprised 271 labeled recordings totaling 783 seconds. each audio recording has a unit variance of zero because the primary audio data have been standardized. Each file is split into 20 ms chunks that don't overlap. We separated all statistics into teach (80%), validation (10%), and sorting out (10%) sets and removed silent durations the usage of the Voice activity Detection (VAD) set of criteria. using stochastic gradient descent is optimized in DNN. We used raw information without characteristic choice as input. Within the whole-report class, our educated model accomplished a widespread take a look at accuracy of 97.9%.

One audio document is extracted from the information Set after it's been loaded within the first stage. Following statistics augmentation, an audio document is produced that lacks augmentation, noise, excessive pace, low

speed, stretch, and pitch. Subsequent, each facts-augmented audio report's features are extracted the usage of MFCC. each audio file within the dataset goes through the equal technique. At ultimate, all the files' features had been extracted. Our dataset became divided into a train (80%) and take a look at (20%) set at random.
To guarantee a truthful contrast, the equal split is utilized in each trial. Here, one warm encoding (CNN) and label encoding (SVM, RF) are used for the based variable. After training classification fashions, assessment metrics are recorded.

## V. WORKING OF THE SYSTEM ARCHITECTURE

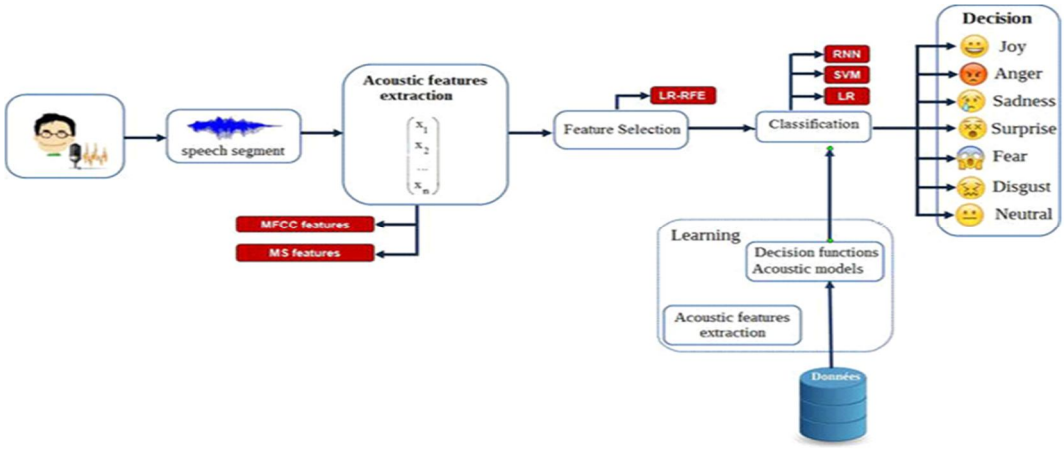


Fig 1: System Architecture

## VI. RESULTS AND DISCUSSIONS

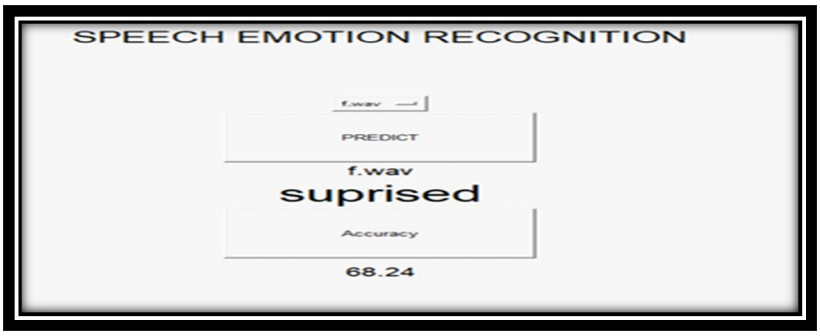


Fig 2: Home page for model emotion

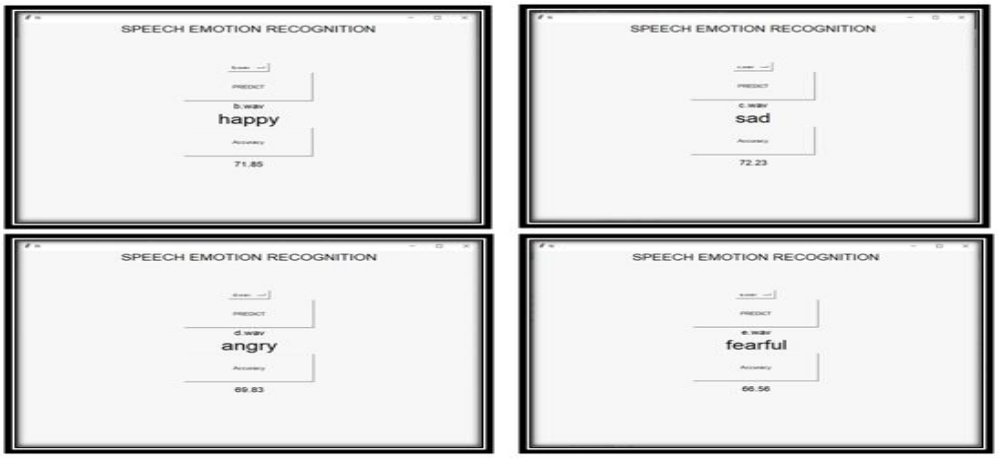


Fig 3: Model Speech Recognition Accuracy.

## VII. Conclusion

The goal of this research is to develop a straightforward yet comprehensive representative emotion recognition system for continuous speech using an appropriate classifier to increase detection accuracy and a speech characteristics extraction method. A thorough discussion of the procedures involved in developing a speech emotion detection system was followed by various experiments to determine the effects of each stage. Initially, it was difficult to deploy a well-trained model due to the small number of publicly accessible speech databases. MFCC was used for feature extraction, and 58 features in total were extracted for emotion prediction. To improve the model's performance, data augmentation was applied to the audio samples in order to diversify the dataset. The RBF kernel was utilized for the SVM classification model. One hundred estimators were employed for the RF classification model. We experimented with several epochs for CNN, and Adam was the optimizer. Categorical cross entropy was the loss function. Accuracy, precision, recall, and F1 score were recorded for the complete model.

## Future Scope

Subsequently, a number of innovative methods for feature extraction had been put out in previous studies, and choosing the most effective one required conducting numerous tests. Lastly, learning about the advantages and disadvantages of each classifying algorithm in terms of emotion recognition was part of the classifier selection process. The experiment's conclusion is that, in comparison to a signal feature, an integrated feature space will yield a higher recognition rate. Employing an ensemble of lexical and acoustic models and taking a lexical features-based approach to SER. Because emotion is sometimes expressed contextually rather than vocally, this will increase the system's accuracy. The emerging growth and development in the field of AI and machine learning have led to the new era.

## References


[1] https://data-flair.training/blogs/python-mini-project-speech-emotion-recognition/

[2] https://www.geeksforgeeks.org/how-to-install-librosa-library-in python/#:~:text=Librosa%20is%20a%20Python%20package,the%20music%20information%20retrieval%20systems.

[3] https://analyticsindiamag.com/a-beginners-guide-to-scikit-learns-mlpclassifier/#:~:text=MLPClassifier%20stands%20for%20Multi%2Dlayer,perform%20the%20task%20of%20classification

[4] https://www.geeksforgeeks.org/best-books-to-learn-machine-learning-for-beginners-and-experts/.

[5] Ruhul Amin Khalil , Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan , Mohammad Haseeb Zafar, And Thamer Alhussain.” Speech Emotion Recognition using Deep Learning Techniques: A Review”, DOI 10.1109/ACCESS.2019.2936124, IEEE Access.

[6] Sabrine Dhaouadi,Hedi Abdelkrim,Slim Ben Saoud, “Speech Emotion Recognition: Models Implementation & Evaluation”. 2019 International Conference on Advanced Systems and Emergent Technologies (IC_ASET).

[7] Pavol Harár, Radim Burget and Malay Kishore Dutta, “Speech Emotion Recognition with Deep Learning”. 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN), 978-1-5090-2797-2/17/$31.00 ©2017 IEEE.