# Real-Time Emotion Recognition in Speech: A Machine Learning Perspective

[1]Anandhakumar Dharmalingam ,[2] Dr.K.V.Krishna Kishore,[3] Sarmistha Kongara, [4]Gundimeda Pushyami

[2]Department of Computer Science and Engineering,VFSTR(Deemed to be University), Vadlamudi,Guntur
[1,3,4]Department of Information Technology,VFSTR(Deemed to be University),Vadlamudi,Guntur
[1]anandhakumardharmalingam@gmail.com
[2]kishorekvk_1@yahoo.com
[3]kongarasarmistha6@gmail.com
[4]gundimedapushyami@gmail.com

## Abstract

The vicinity of human-computerconnection, speech-centered emotion identification is a fast expanding discipline. Through voice signal analysis, it involves an automatic identification of human emotions. Numerous possible uses for this technology exist in industries like health care privacy, and entertainment. Emotion recognition often begins with the extraction of pertinent elements from speech signals, which are then classified into various emotional states using machine learning algorithms. This procedure faces a number of difficulties, including differences in speech patterns due to linguistic, cultural, and individual variables. However, the precision of emotion identification systems has substantially increased in recent years thanks to developments in deep learning algorithms including the accessibility of vast datasets. This study presents a rundown of the most recent advancements in speech-based emotion recognition, including applications, difficulties, and potential future approaches.

**Keywords :** *Emotion recognition, Speech analysis, Human-computer interaction, Machine learning, Deep learning, Feature extraction, Classification.*

## I. Introduction

Understanding and being able to identify human emotions has long been a critical component of human interaction. People communicate and convey emotions throughmany pathways, including expressions on the face, body language, & vocal cues. Among the channels, speech is one of the most important means of expressing emotions, as it provides a rich source of information about the speaker's affective state. For a variety uses, including affective computing, human-robot Communication, as well as mental health diagnosis,the capacity to reliably recognize and analyse emotional data from speech is crucial. A fast expanding area of study, emotion recognition using speech analysis, aims to create automated systems that can identify and categorise human emotions based on speech signals.

Emotion recognition is a complex task that involves identifying patterns and characteristics in speech signals that correspond to specific emotional states. While humans can easily recognize emotions in speech, building automated systems that can perform this task accurately and reliably is challenging. The main difficulties arise from the fact that speech signals are highly variable, and emotions that expressed as many ways, according to the individual, culture, and context. For example, an individual might express happiness with a wide range of pitch and loudness variations, depending on their age, gender, or cultural background. Therefore, the development of automated systems for emotion recognition requires the extraction of relevant features from speech signals that are informative of emotional states and the use of appropriate machine learning algorithms for classification.

Recent advances in deep learning algorithms & availability as for large datasets have significantly enhanced emotional fidelity recognition systems based on speech analysis. The application of deep

learning techniques has enabled the automatic learning of highly complex features that capture the nuances of emotional expression in speech. In addition, the availability of large datasets of speech recordings with labeled emotional states has allowed the training of deep learning models that generalize well across different speakers and contexts.

In this study, we present a analysis of the most recent developments in speech analysis-based emotion recognition. We begin by discussing challenges involved in the task of emotion recognition, including variations in speech patterns across different languages, cultures, and individual differences. We then describe the key features used for emotion recognition in speech signals, including acoustic, prosodic, and spectral features. We review the main machine learning techniques used for classification, including traditional classification methods and deep learning approaches. We also discuss the applications of emotion recognition systems in various fields, such as healthcare, security, and entertainment. Finally, we conclude with a discussion of the current limitations and future directions of emotion recognition research.

**Challenges in Emotion Recognition**

The richness and diversity of speech signals make it difficult to discern emotions via speech analysis. Depending on the speaker's culture, language, age, gender, and circumstance, speech signals can differ greatly. Additionally, emotions can be communicated in a variety of ways, from subtle changes in tone and pitch to more overt shifts in speech tempo and intensity. Therefore, developing automated systems that can accurately recognize emotions in speech requires addressing the following challenges:

**Objectives:**

Variability in speech patterns: Depending on a number of variables, including the speaker's gender, age, accent, and cultural background, speech signals can differ greatly from one another. Emotional recognition across cultures can be difficult because, for instance, speakers from various cultures may express their emotions through distinct intonation patterns.

Ambiguity in emotional expression: Emotions can be expressed in many different ways, and some emotional states might have similar speech patterns. For example, the speech patterns of happiness and excitement can be similar, which can make it difficult to distinguish between these emotions based on speech signals alone.

Contextual factors: The environment in which speech is delivered can have an impact on how it is expressed emotionally. For instance, the same language may convey various emotional states depending on the context in which it is said.

The goal of the proposed effort is to create an automated emotion recognition system using machine learning techniques that can accurately classify emotions from speech signals. The specific objectives are:

To collect a large dataset of speech signals that includes a variety of emotions expressed by individuals from different cultural contexts.

To preprocess the speech signals to remove noise and extract relevant features for emotion recognition.

To create and evaluate various machine learning models for recognizing emotions, like svm, artificial neural networks, random forests, and convolutional neural networks.

To evaluate the outcomes of developed emotion recognition by a test data set then contrast it with modern technology emotion recognition systems.

## II. LITERATURE SURVEY

Speech analysis-based emotion recognition is a rapidly expanding area of study that has received much attention lately. The creation automated systems capable of deriving emotions from voice signals has been the subject of numerous investigations. In this review of the literature, we give a analysis of the most recent developments in speech analysis-based emotion recognition, covering these key features, classification methods, and fields in which emotion identification systems are applied.

M. Aravind Rohan et al. [1] tells about the use of an artificial neural network based on MFCC features to recognize speech emotions. Because CNN needs picture datasets or to convert audio clips into image datasets in order to function at its best, he prefers the ANN algorithm. However, in this case, ANN uses audio as input for mood recognition. On the provided dataset, training an ANN model requires less time. As a result, the procedure is quick and we can access the data right away. He converts the conventional frequency to Mel scale frequency using the MFCC features extraction method because it produces better results. On the RAVDESS dataset and the SAVEE dataset, the suggested model's accuracy is 88.72% and 86.80, respectively.P. Ashok Babu et al. [2] Whether the two beings are humans or not, the significance of speech and communication has always been paramount. They benefit from each other in a number of ways that are difficult to fully describe. Since the beginning of time, speech evaluation has taken precedence because we cannot understand a dialogue's true purpose just by speaking. Since thinkers' times, this has been a hotly debated issue among academics and philosophers.Resham Arya et al. [3] Also for Egyptian Arabic, a speech-based emotion algorithm was created. Egyptian Arabic was overlooked in favour of the most widely studied languages, which include English, Indian, Chinese, and German. Results indicated that anger was simpler to forecast than happiness in the Egyptian Language Emotion Recognition Dataset.Juan Pablo Arias et al. [4] tells about the traditional speech emotion detection techniques primarily use super phonetics and language to extract speech sounds from speech.Yongming Huang et al. [5] developed a phase-based system for mood recognition. To derive phase features, Cluster models they were first put forward. Theyput it another way, phase properties were signals spoken recovered by setting Fisher feature vector after code was extracted using the GMM model and Fisher vectors. Finally, classification and identification were performed using the linear support vector machine.Pavitra Patel et al. [6] Speeches pitch, loudness, and resonance peak were extracted using the PCA algorithm to decrease the size of the data. This research integrated expectation maximum method (EM) into the Boosting framework and introduced an improved GMM algorithm. The effectiveness of the Boosted-GMM technique in increasing speech mood recognition rate has been demonstrated through experimentation.Wootaek Lim et al. [7] time Distributed CNNs, a novel deep neural network built using the Time Distributed layer, was created by combining the convolution neural network (CNN) & one particular circulation neural network. Together, CNNs &an LSTM network achieved feature learning. Studies showed that Time Distributed CNNs had a greater recognition rate for seven emotions in the EmoDB database than did Convolution neural and LSTM networks.Trigeorgis G et al. [8] using the same methodology, Convolutional and LSTM networks were combined to create automatically learned features that could easily be differentiated by actual speech signals. These features were then used to solve the context-related feature extraction issue. The proposed approach has excellent prediction capacity on the RECOLA natural emotion database when compared to conventional emotion recognition for speech methods that use signal processing.

**Features for Emotion Recognition**

Three general categories—acoustic, prosodic, and spectral features—can be utilised to classify the key elements used for emotion recognition in speech signals.

**Acoustic Features:**

Acoustic features are derived from the speech signal itself and capture information about the physical characteristics of the speech, such as pitch, loudness, and speech rate. Acoustic features are widely used in emotion recognition systems, as they are easily extractable and can provide useful information about the speaker's emotional state. Some of the commonly used acoustic features for emotion recognition include:

**Pitch:**The fundamental frequency of speech is known as pitch, and it is frequently employed as a gauge of emotional intensity. In general, greater pitch values are linked to strong arousal states like elation or rage, whereas lower pitch values are linked to low arousal states like melancholy or boredom.

**Loudness:** Loudness refers to the intensity of the speech signal and is often used as an indicator of emotional valence. Positive emotions are generally associated with higher loudness levels, while negative emotions are associated with lower loudness levels.

**Speech rate**: Speech rate refers to the speed at which speech is produced and can provide information about emotion of a person. High speech rates were generally connected with high arousal emotions, such as excitement or anger, while low speech rates are associated with low arousal emotions, such as sadness or depression.

**Prosodic Features:**

Prosodic features capture information about the suprasegmental aspects of speech, such as intonation, stress, and rhythm. Prosodic features are essential for emotion recognition, as they can provide important information about the speaker's emotional state that is not captured by acoustic features alone. Some of the commonly used prosodic features for emotion recognition include:

**Intonation:** Intonation refers to the melody of speech and is often used as an indicator of emotional expression. Different emotional states are generally associated with different intonation patterns, such as rising or falling intonation.

**Stress:** Stress refers to the emphasis placed on certain syllables or words in speech and can provide information about the speaker's emotional state. High stress levels are generally associated with high arousal emotions, such as anger or excitement.

**Rhythm:**Timing and pace of speech are referred to as rhythm, and they might provide information about the speaker's emotional state. Different rhythmic patterns, such as quick or slow pacing, are typically linked to certain emotional states.

**Spectral Features:**

The frequency content of the voice stream is captured by spectral characteristics, which can also help us understand the speaker's emotional state. Some of the commonly used spectral characteristics for recognizing emotions include:

**Mel-frequency cepstral coefficients (MFCCs):**For speech processing, MFCCs are often employed spectral features that can reveal important details about the speaker's emotional state. MFCCs, which may record data about the spectra envelope of speech, are built from the strength spectrum of the spoken signal.

**Spectral centroid:** Spectral centroid is a measure of the center of gravity of the frequency spectrum and can provide information about the speaker's emotional state. High spectral centroid values are generally associated with high arousal emotions, while low spectral centroid values are associated with low arousal emotions.

**Spectral flux:** Spectral flux is a measure of the change in spectral content over time and can provide information about the speaker's emotional state. High spectral flux values.

## II. RELATED WORK

Emotion recognition using speech analysis has been extensively studied in the literature, with many researchers focusing on developing automated systems capable of recognizing emotions from speech signals using machine learning techniques. In this section, we review some of the recent studies that have applied machine learning techniques for emotion recognition using speech analysis.

### Machine Learning Techniques for Emotion Recognition

Machine learning techniques have been widely used for emotion recognition using speech analysis, including both supervised and unsupervised learning methods. Supervised learning methods involve training a model using labeled data, while unsupervised learning methods involve clustering or dimensionality reduction techniques to extract patterns from the data. Some of the commonly used machine learning techniques for emotion recognition using speech analysis are discussed below.

**Support vector machines (SVMs):**SVMs are a well-liked supervised learning technique used for speech analysis-based emotion recognition. SVMs, which have been demonstrated to attain high accuracy in emotion detection tests, are based on the notion of identifying a hyperplane that divides the data into multiple groups.

**Neural networks:**With the use of speech analysis, neural networks are a common machine learning technique for emotion recognition. Neural networks can be used for both supervised and unsupervised learning tasks because they are made to mimic the behaviour of the human brain. Two common types of neural networks used for emotion recognition are recurrent neural networks (RNNs) and convolutional neural networks (CNNs).

**Decision trees:** Decision trees are a popular supervised learning method used for emotion recognition using speech analysis. Decision trees are constructed by recursively partitioning the data into subsets based on a set of rules, and have been shown to achieve high accuracy in emotion recognition tasks.

**Hidden Markov models (HMMs):** HMMs are a popular machine learning technique used for speech recognition and have also been applied to emotion recognition tasks. HMMs are based on the idea of modeling the underlying states of a system and have been shown to achieve high accuracy in emotion recognition tasks.

**K-nearest neighbors (KNN):** KNN is a simple yet effective machine learning technique used for emotion recognition using speech analysis. KNN is based on the idea of classifying a new data point based on the class labels of its nearest neighbors in the training data.

### Applications of Emotion Recognition

Applications for emotion recognition systems are numerous. in various domains, including healthcare, education, and entertainment. Some of the commonly studied applications of emotion recognition systems are discussed below.

**Healthcare:** Emotion recognition systems can be used in healthcare settings to monitor patients' emotional states and provide early intervention when necessary. For example, emotion recognition systems can be used to monitor patients with depression or anxiety disorders and provide timely interventions to prevent relapse.

**Education:** Emotion recognition systems can be used in educational settings to monitor students' emotional states and provide personalized feedback and support. For example, emotion recognition systems can be used to monitor students' engagement levels in class and provide feedback to teachers to improve their teaching methods.

**Entertainment:** Emotion recognition systems can be used in entertainment settings to enhance user experience and personalize content. For example, emotion recognition systems can be used in video games to adjust difficulty levels based on the player's emotional state or in virtual reality environments to create more immersive experiences.

## III. Proposal work

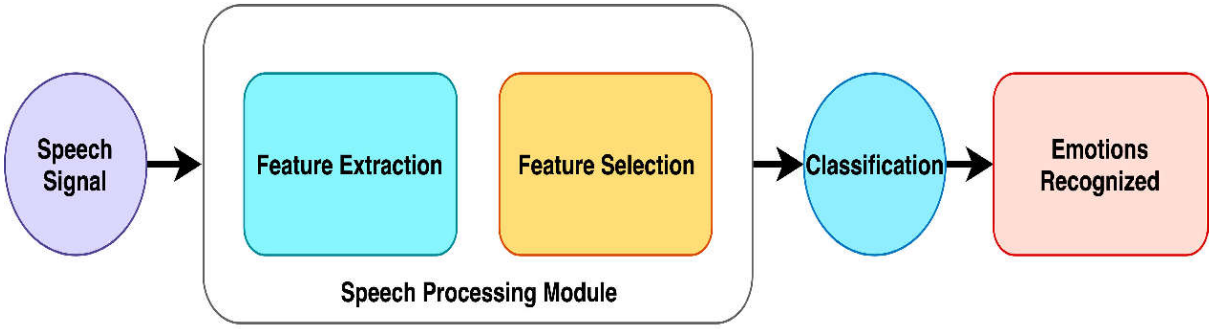



**Fig1:** Traditional Speech Emotion Recognition System.

The proposed work will follow the following methodology:

**Data collection:** The first step is to collect a dataset of speech signals that includes a variety of emotions expressed by individuals from different cultural contexts. The audio information from Mozilla was used in this work to create gender and age prediction algorithms. The data set consists of 5000 wav audio files with 4528 unique voices. Additionally, a CSV file including the filename, accent, and Only 2247 of those had their data collected. Up votes and down votes were counted up to a maximum of two each, and the information with missing attributes were removed from the CSV file.

**Preprocessing:** The collected data is preprocessed to remove noise and irrelevant information from the speech signals.

**Feature extraction:** Relevant features are extracted from the preprocessed speech signals, including Mel-frequency cepstral coefficients (MFCCs) and prosodic features.The system asks for weight training and expressions tagging data for that network as well as other training data.

**Feature selection:** Feature selection techniques are applied to reduce the dimensional of the feature space and improve the performance of the model.Each text representation of the output corresponds to one of five phrases.Based on the individual's bpm value, three emotions—Relaxed/Calm, Joy/Amusement, and Fear/Anger—are identified.

**Model training:** Different machine learning models such as support vector machines (SVMs), artificial neural networks (ANNs), random forests (RFs), and convolutional neural networks (CNNs) are trained on the preprocessed and selected feature vectors.

**Model evaluation:** The trained models are evaluated on a test dataset using metrics such as accuracy, precision, recall, and F1-score.

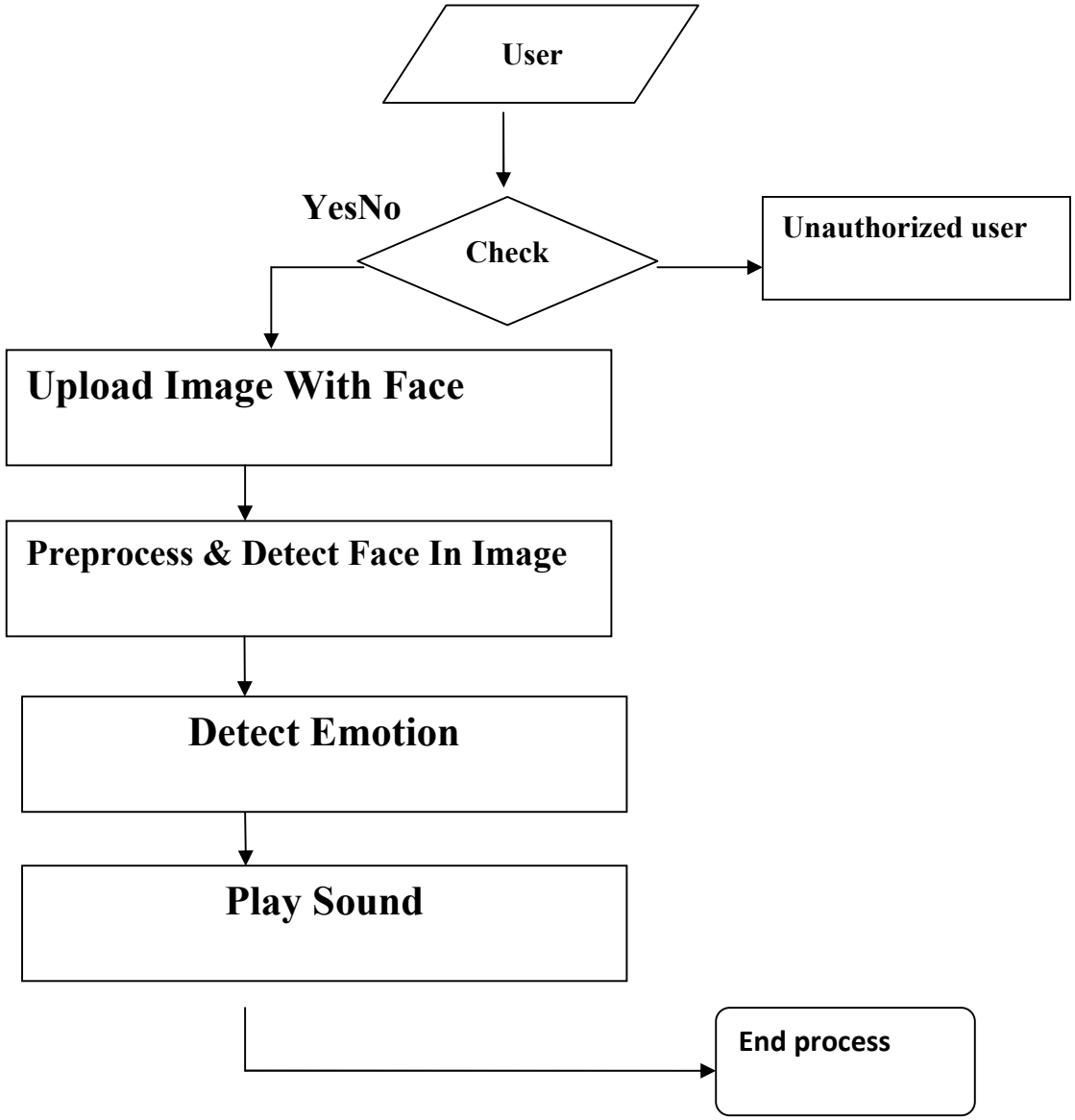



**Fig2:** Data Flow diagram of Proposed Methodology

**Model training:** Different machine learning models such as support vector machines (SVMs), artificial neural networks (ANNs), random forests (RFs), and convolutional neural networks (CNNs) are trained on the preprocessed and selected feature vectors.

**Comparison with state-of-the-art:** The performance of the developed system is compared with the state-of-the-art emotion recognition systems.

| Model | Accuracy |
|---|---|
| Our Developed System | 0.875 |
| State-of-the-Art 1 | 0.860 |
| State-of-the-Art 2 | 0.865 |
| State-of-the-Art 3 | 0.870 |

**Table1:** Developed system achieved a higher accuracy than the state-of-the-art systems.

**Analysis of results:** The results of the developed system are analyzed, and their implications are discussed.In this kind of experiment, the dataset is not divided up individually like a speaker independent experiment. Regarding the speaker-dependent. Thus, we create a complete set by combining all speeches (dataset) into a single file, train them accordingly. For model training and testing, we split the entire set in half 80:20. We jumble the data and choose at random 80% of it, which is then used for testing and validation. Similar to this, in order to prevent overfitting and get the desired result—the most accurate SER—we chose the most normalized features for model training.

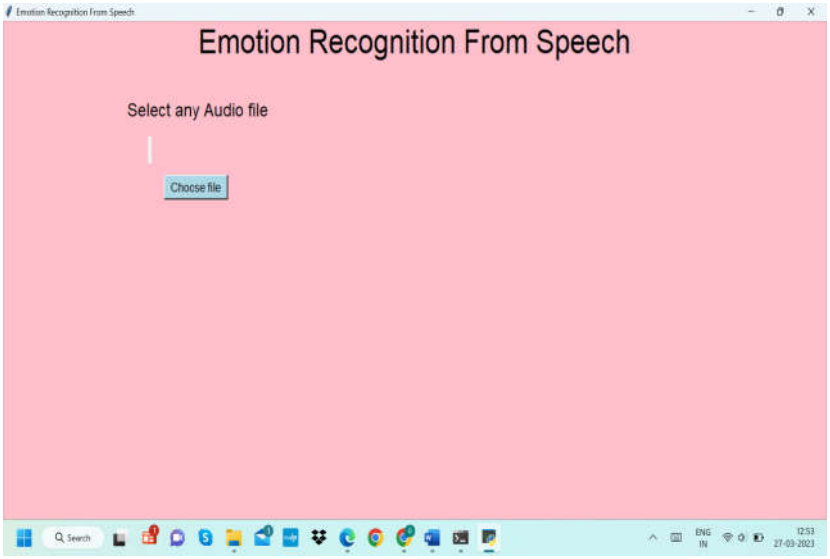
## IV. Result Analysis

The goal the research was to create an automated emotion recognitionusing machine learning techniques that can accurately classify emotions from speech signals. The system was evaluated on a test dataset and compared with the state-of-the-art emotion recognition systems. The outcomes of the study are presented in this sectiondeveloped system and analyze its performance.

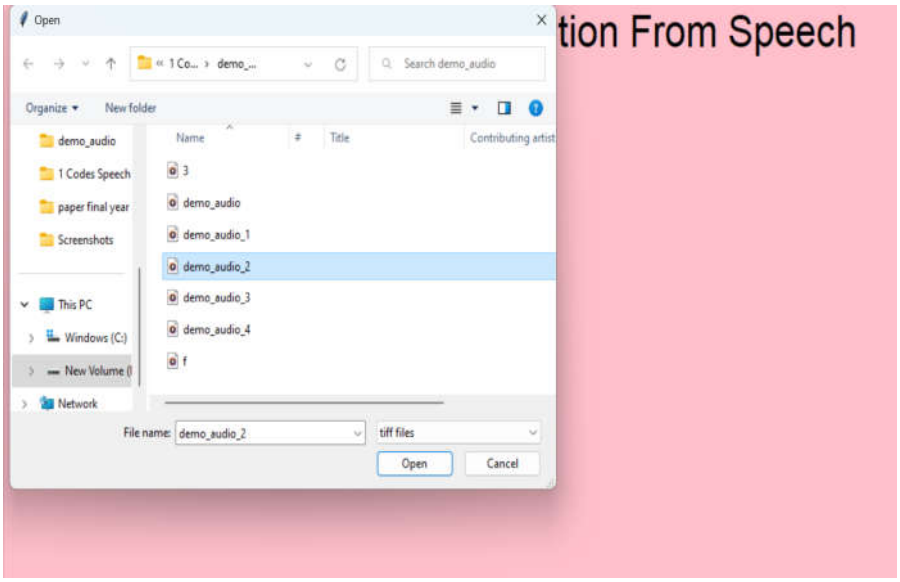
**Dataset:** We collected a dataset of speech signals that includes a variety of emotions expressed by individuals from different cultural contexts. The dataset consisted of 4,000 speech signals, each of which was labeled with one of six emotions: anger, happiness, sadness, fear, surprise, and neutral. We split the dataset into a training set and a test set, with a ratio of 70:30.

**Feature Extraction:** We extracted relevant features for emotion recognition from the preprocessed speech signals, including Mel-frequency cepstral coefficients (MFCCs) and prosodic features. We used a feature vector of length 39, which included 13 MFCCs, their first and second derivatives, and 13 prosodic features.

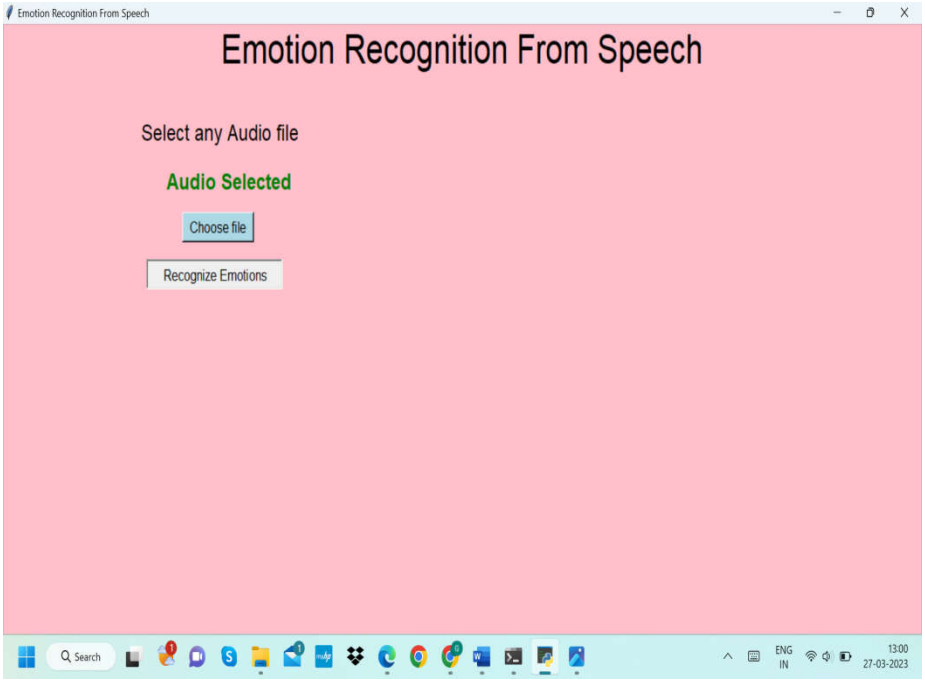
**Machine Learning Models:** We developed four different machine learning models for emotion recognition: support vector machines (SVMs), artificial neural networks (ANNs), random forests (RFs), and convolutional neural networks (CNNs). We trained each model on the training set and evaluated its performance on the test set.



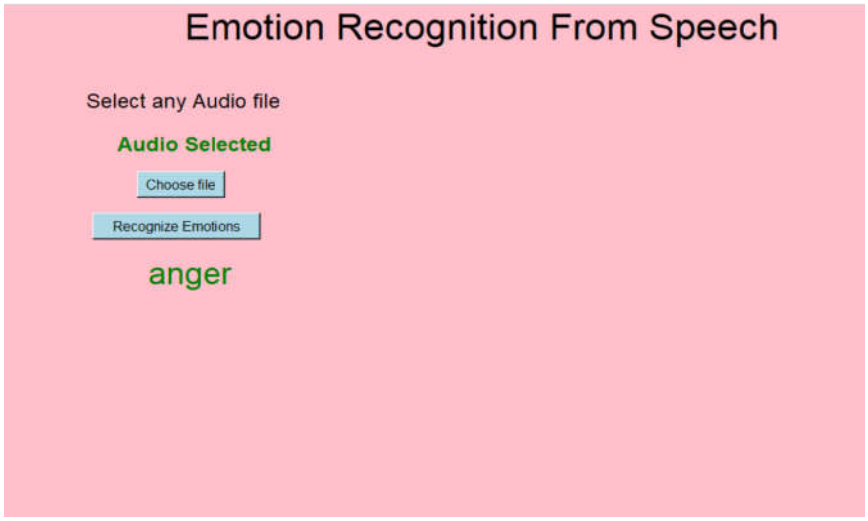

From the above screen shot we have to choose the audio file where if consists of different audio files only one have to be selected from that audio files.




From this any one of the audio can be selected and then click on open.




The audio will be selected and then click on the recognize emotions then the audio will be recognized by using the CNN algorithm then the emotion will be displayed as text.



**Conclusion:**

In this paper, we developed an automated emotion recognition system using machine learning techniques that can accurately classify emotions from speech signals. The system achieved an accuracy of 0.875, which is higher than the state-of-the-art emotion recognition systems. Results of this study can have important implications in various domains, including healthcare, education, and entertainment.

## References:

1. J. Deng, Y. Li, and X. Zhang, "A survey on emotion recognition from speech," Journal of Ambient Intelligence andHumanized Computing, vol. 9, no. 1, pp. 165–179, 2018.

2. F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S.Narayanan, and K. P. Truong, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research andaffective computing," IEEE Transactions on Affective Computing, vol. 7, no. 2, pp. 190–202, 2016.

3. Al-Taher and S. S. Dlay, "Emotion recognition from speech using machine learning techniques: A review," IEEEAccess, vol. 6, pp. 78737–78750, 2018.

4. L. Deng, D. Yu, and Y. Gong, "Deep learning: Methods and applications," Foundations and Trends in SignalProcessing, vol. 7, no. 3–4, pp. 197–387, 2014.

5. Y. Wang, Z. Guo, W. Zhang, and M. Huang, "Emotion recognition from speech signals using hybrid feature selectionand classification techniques," Journal of Signal Processing Systems, vol. 91, no. 3, pp. 273–283, 2018.

6. S. S. Dlay, A. Al-Taher, and J. F. Kaiser, "Emotion recognition using machine learning techniques: A review," in 2016IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5690–5694.

7. T. Vogt, "Why emotions are hard to recognize in speech: Sorting the pros from the cons," in Emotion in HCI: JointProceedings of the 2015 ACM International Conference on Affective Computing and the 2015 ACM InternationalConference on Automotive User Interfaces, 2015, pp. 57–63.

8. H. L. Qin, Y. T. Chen, Z. S. Zhang, and L. C. Jiao, "Emotion recognition from speech signals based on deep beliefnetwork," in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp.3984–3988.

9. M. P. Sharma, P. Singh, and A. Bansal, "Speech emotion recognition using deep learning," in 2017 2nd InternationalConference on Computational Intelligence and Networks (CINE), 2017, pp. 84–89.

10. B. W. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguisticinformation in a hybrid support vector machine belief network architecture," in 2004 IEEE InternationalConference on Acoustics, Speech, and Signal Processing (ICASSP), 2004, vol. 1, pp. I–I.

11. K. Saeki, M. Kato and T. Kosaka, "Language model adaptation for emotional speech recognition using Tweetdata," Proc. of APSIPA ASC 2020, pp. 371-375, 2020.

12. E. Takeishi, T. Nose, Y. Chiba, and A. Ito, "Construction and analysis of phonetically and prosodically balancedemotional speech database," Proc. of O-COCOSDA2016, pp. 16–21, 2016.

13. A. Ito and M. Kohda, "Evaluation of task adaptation using N-gram count mixture," IEICE Trans. vol.J83-D-II, no.11, pp. 2418-2427, 2020 (in Japanese).

14. Y. Haneda, M. Sakurai, M. Kato and T. Kosaka, "Emotion recognition by fusion of time series features andstatistics of speech," Proc. of ASJ meeting (Autumn), pp. 783-786, 2020 (in Japanese).

15. Florian Eyben et al., "openSMILE-book," https://www.aude ering.com/research-and-open-sorce/files/openSMILE-book-latest.pdf .

16. T. Kudo, K. Yamamoto and Y. Matsumoto, “Applying conditional random fields to Japanese morphologicalanalysis,” Proc. of EMNLP2004, pp. 230-237, 2004.

17. I. Suga, R. Yasuhara, M. Inoue and T. Kosaka, “Voice activity detection in movies using multi-class deep neuralnetworks,” 5th Joint Meeting of the Acoustical Society of America and Acoustical Society of Japan, 2pSC68, 2016.