

Speech Emotion Recognition Using Mel-Frequency Cepstral Coefficients

Er. Disha Sharma
Department of CSE
Chandigarh University
Mohali, India
disha.E13241@cumail.in

Suraj Kumar Dev
Department of CSE
Chandigarh University
Mohali, India
surajkumardev121@gmail.com

Ankit Kumar
Department of CSE
Chandigarh University
Mohali, India
subhrajankit@gmail.com

Amanjeet Kumar
Department of CSE
Chandigarh University
Mohali, India
amanjeet233@gmail.com

Rishav Kumar
Department of CSE
Chandigarh University
Mohali, India
rishavkumar0683@gmail.com

Ambuj Shukla
Department of CSE
Chandigarh University
Mohali, India
ambujshukla7982@gmail.com

Abstract—Speech Emotion Recognition (SER) has multiple applications in computational psychology, as well as human-computer interaction, and it also plays a major role in these fields. The accuracy of any machine learning technique is greatly impacted by the extraction as well as the appearance of features. Mel Frequency Cepstrum Coefficient (MFCC) is a frequently used framework across many domains to simulate the characteristics of a signal that contains sound. This study is all about using Mel Frequency Cepstral Coefficients, or MFCCs for short. These are popular tools this use when figuring out emotions in speech signals. The goal here is to see how well MFCCs work in picking up sounds that show different feelings. By using machine learning techniques, like support vector machines (SVM) or even neural networks, this can classify emotions with the MFCC features this have gathered. In this paper, this will look at what others have done in the field of speech emotion recognition. This will also suggest ways to set up systems that detect emotions based on MFCCs. It gives us helpful insights into how Mel frequency features can really help us understand emotions expressed in speech.

Index Terms—Emotion Recognition, Speech Analysis, Feature Extraction, Mel Frequency Cepstral Coefficients (MFCCs), Machine Learning.

I. INTRODUCTION

Emotion recognition from speech signals is a burgeoning field in effective computing, playing a pivotal role in human-computer interaction. This research investigates the application of Mel frequency cepstral coefficients (MFCC) for accurate emotion identification and classification in spoken language. The study aims to bridge gaps in existing emotion recognition techniques as network visualization depicts in Figure 1 by leveraging the distinctive spectral features captured by MFCCs. The objective of researchers in the early stages of pattern recognition has always been to substitute manually

crafted features with trainable multilayer networks. However, the simplicity of this solution remained largely unknown until the mid-1980s [1]. It was revealed that multilayer designs can

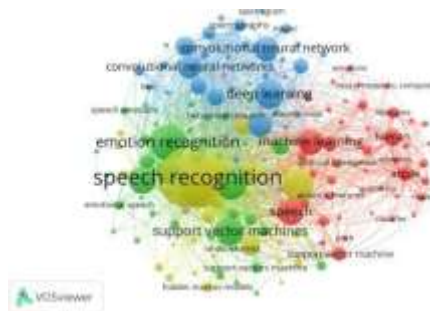


Fig. 1. Network Visualization of Emotion Recognition Techniques in Speech Analysis

be educated through a basic stochastic gradient descent. Deep learning empowers computational models, comprising numerous processing layers, to acquire data representations with varying levels of abstraction. These advances have markedly elevated the state-of-the-art in fields such as voice recognition, visual object identification, object detection, and diverse domains such as drug development and genomics [2]. Deep learning excels in discerning intricate structures within vast datasets, employing the back-propagation technique to guide adjustments in a machine's internal parameters. These parameters are instrumental in deriving the representation in each layer from the previous one. This shift from hand-engineered features to multilayer trainable networks has streamlined the training process, significantly influencing efficient represen-

tation learning. Automation of feature extraction has led to substantial improvements in tasks in various domains, establishing deep learning as a transformative force in contemporary technology [3].

A. Machine Learning in Emotion Recognition

Machine learning is a subset of artificial intelligence that uses algorithms and statistical models to allow computing to perform tasks without explicit instruction. MFCCs are widely used in speech processing due to their ability to capture the nuanced acoustic characteristics of spoken language. This study leverages MFCCs alongside machine learning algorithms such as support vector machines (SVM) and repeat neural networks (RNNs) with long-short-term memory (LSTM) to accurately classify emotions conveyed in speech. The research underscores the potential of MFCC-based systems in enhancing human-computer interaction, healthcare applications, and affective computing, while also addressing the challenges and limitations in current methodologies [4].

When talk about speech, there is this thing called Mel-frequency cepstral coefficients (MFCCs). These are really important! It help pick up the subtle sounds of spoken language. MFCCs show the short-term power spectrum of sounds really well, which helps spot emotional tones and expressions in voices. By changing audio signals into a format that shines a light on key features and overlay visualization given in Figure 2, MFCCs play a huge role in emotion recognition tasks. This study uses MFCCs along with some advanced machine

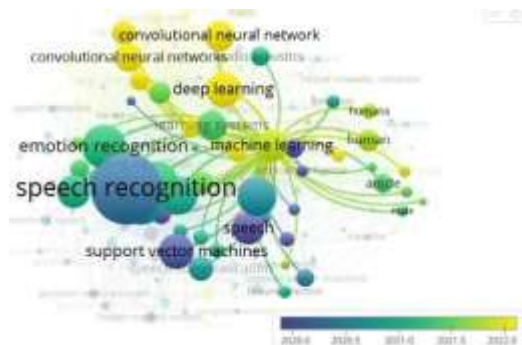


Fig. 2. Emotional Tones and Expressions of MFCC: Overlay Visualization

learning techniques, such as support vector machines (SVM) and recurrent neural networks (RNN) with long short-term memory (LSTM) units. These help in accurately classifying emotions in speech. SVMs work great in high-dimensional spaces and are really good at handling tasks where there is a clear separation between categories. On the other hand, LSTMs are fantastic at processing data sequences and remembering information longer, perfect for understanding the timing in speech [5].

In addition, this research adds to affective computing. That's about making systems that recognize and simulate human emotions, also. By putting emotion recognition into technology, this can build systems that are more caring and responsive, ones that really get what this need as humans.

However, there are also some challenges that this research points out. Speech can vary a lot due to different accents or cultures; individual differences matter too. This can make spotting emotions tricky sometimes. Also, how good the training data is really affects how well the machine learning models work! Having diverse and representative datasets is super important for creating strong emotion recognition systems [6].

To wrap it all up: this study points out the exciting link between machine learning & understanding emotions, especially through using MFCCs & advanced tools like SVM and LSTM. As technology continues to advance, the addition of emotion recognition could lead to more user-friendly applications. Think of all the possibilities, from healthcare to gaming! Future efforts should work hard on those challenges and discover fresh ways to boost the accuracy of emotion-recognition systems [5] [6].

II. RELATED WORK

Mel frequency spectral coefficient (MFCC) systems are used to identify and identify emotional states in spoken language. In the past several years, MFCCs have drawn a lot of interest due to the growing need for computer interactions that are more like human interactions. The research done in this area over the past ten years is compiling in this review of the literature, along with the main methods and strategies applied to MFCCs [7].

This paper introduces something new called Mel Frequency Magnitude Coefficients (MFMC) to help with speech emotion recognition. It deals with some of the problems that older methods like MFCCs face. What is interesting is that MFMC beats MFCC, LFPC and LPCC with precision, especially when using SVM classifiers in six emotional speech databases [8]. Yang and the team (2020) took a step forward in letting smart home assistants understand emotions better! It did this by using MFCC features and SVM for speech emotion recognition (SER), hitting a great accuracy of 92.4%. In addition, SVM is pretty handy for devices with limited resources. Still, there are some drawbacks. It cover only a small range of emotions and rely heavily on certain databases, together with issues around speaker differences and computing efficiency [9].

This 2021 paper by Ancilin and Milton, It explores a new way to extract features called the Mel Frequency Magnitude Coefficient (MFMC) for automatic speech emotion recognition (SER). The authors point out that traditional features like Mel-Frequency Cepstral Coefficients (MFCCs) have limitations because it rely on energy spectrum and Discrete Cosine Transform (DCT) [8] [9].

This paper by Palo, Chandra, and Mohanty dives into different tweaks of a popular method known as Mel-Frequency Cepstral Coefficients (MFCCs) for picking up emotions in human speech. MFCCs are reliable but it can't always capture how emotions shift over time when this speaks [10].

The study by Yang and Huang from 2021 talks about an SER method that mixes MFCCs with BLS. It highlight how important SER is for human-computer interaction while

also noting the struggle to get the best recognition rates possible. However, their focus was only on achieving high accuracy from one specific Chinese emotion database without considering how it affects computing efficiency [11]. Emotion

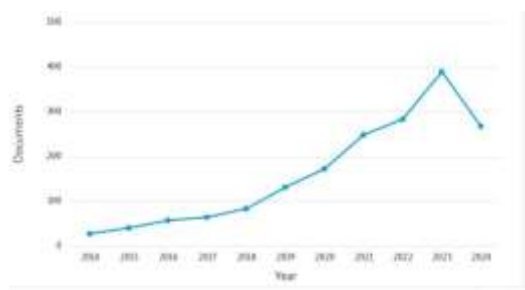


Fig. 3. Speech Emotion Recognition and Human-computer Interaction: Year-wise Publication and citations

recognition from speech using MFCC techniques in March 2017. This paper focuses on Speech Emotion Recognition (SER), which is the key to natural human-computer interaction and year wise publication and citations are depicts in Figure 3. It points out why recognizing emotions in this speech matters, especially for things such as user interfaces, assistive technology, and security systems [6] [7].

This study by Nasrun and Setianingsih investigates how to detect human emotions via speech recognition techniques! Their findings contribute nicely to the growing field focused on mixing emotions into human-computer interaction (HCI) [12].

This method Recognizing emotional speech through CNN and deep learning methods in 2023. Hema and Marquez look into Speech Emotion Recognition Systems and share an approach that uses Convolutional Neural Networks (CNNs), which are cool deep learning tools! CNNs have done really well handling various image tasks and are starting to do great work in audio classification, too, like SER [13].

This paper checks out how this can use Mel-Frequency Cepstral Coefficients (MFCCs) for Speech Emotion Recognition! Emotions are very important when people talk, as it add extra meaning beyond just the words spoken. [10] [12]

This insightful piece by Akcay reviews various elements involved in Speech Emotion Recognition systems—it covers emotional models as well as some speech databases, too! In addition, it goes into detail about feature extraction methods along with processing techniques and classification algorithms. [14]

It dives into the analysis of Mel scale features. Their work focuses on using Mel-Frequency Cepstral Coefficients, or MFCCs, to recognize emotions in speech, a field called Speech Emotion Recognition (SER). What is cool is that it suggests a new way to look at emotions by considering the emotional hints right from the spoken sentences! For example, an exclamation mark might show excitement or anger. On the other hand, question marks can tell us about confusion or curiosity [15]

It also looks at the SER Model. It also investigates how Deep Neural Networks (DNNs) which are like mini-brains in machines—can be used for this task [16].

III. SPEECH EMOTION RECOGNITION

Speech is super important for how this interacts with computers. To better understand emotions in speech signals, this has an SER system that uses different methods. There are many models out there for predicting emotions in speech signals. But guess what? This are focusing on using a recurring neural network (RNN) combined with long-short-term memory (LSTM) to really get a grasp of the audio data in order [17].

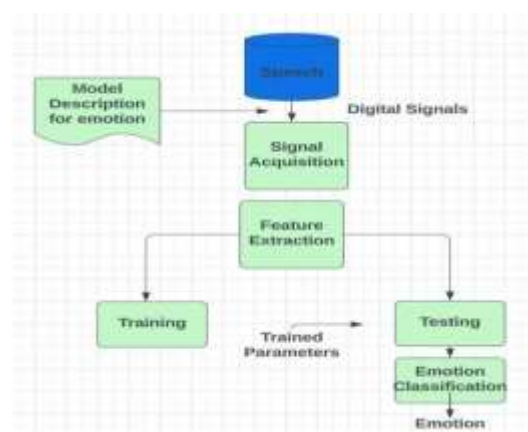


Fig. 4. Speech Emotion Recognition: Machine learning Model

To kick things off, having a solid training dataset is crucial for any machine learning project, and SER is no different! This decided to go with RAVDESS and TESS. These two collections are fantastic for human audio processing and help strengthen this model. Of course, this first clean the data to eliminate noise or long speech gaps [15] [17]. After that, this dive into extracting features! This use those Mel Frequency Cepstral Coefficients (MFCC) as inputs for this model. Finally, this apply an SER algorithm using a deep learning framework of an RNN paired with an LSTM network [18].

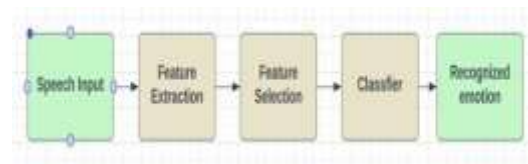


Fig. 5. Speech Emotion Recognition System: Features Extraction

IV. SPEECH EMOTION RECOGNITION : MACHINE LEARNING METHODOLOGY

A. Data Collection

The first step in any machine learning project is to gather data. For Speech Emotion Recognition (SER) using machine

TABLE I
AUDIO'S IN EACH CLASS OF DATASET

Emotion	Count
Angry	400
Disgust	400
Fear	400
Sad Label	400
Neutral	400
Ps	400
Happy	400
Total	2800

learning, this means collecting audio recordings of people showing different emotions with different class given in Table 1. This audio file database can be found in public datasets like the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) or the Toronto Emotional Speech Set (TESS), among a few others. Although there are larger datasets out there, such as IEMOCAP (Interactive Emotional Dyadic Motion Capture) and SAVEE (Surrey Audio-Visual Expressed Emotion), this current model tends to work better with simpler, smaller datasets [18] [19].

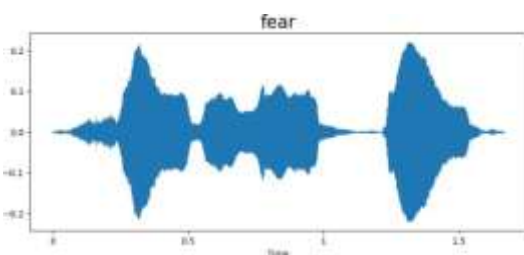


Fig. 6. Audio Recordings Preprocessed Signal of Fear Emotions: Machine Learning

B. Data Labelling

Once this has data, this needs to preprocess those audio files so it can fit into this machine learning model. This process involves several steps: audio segmentation, feature extraction, and normalization. We break the audio files into smaller frames

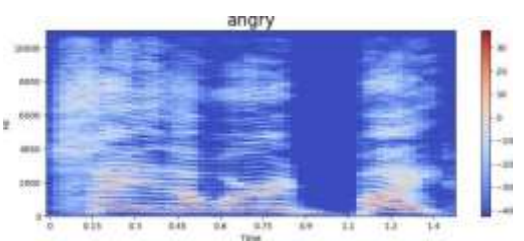


Fig. 7. Audio Recordings Preprocessed Signal of Angry Emotions: Machine learning

and pull out features like Mel-Frequency Cepstral Coefficients (MFCC), pitch, and energy from each frame. To make sure everything is on the same level, normalization techniques like mean-variance normalization are used [19].

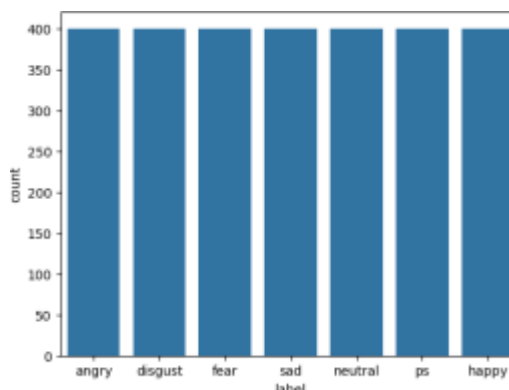


Fig. 8. Data Labeling of Speech Files and Emotions in Machine Learning

C. Feature Selection and Extraction

$$C(x(t)) = F^{-1}[\log(F(x(t)))] \quad (1)$$

Formula to compute the Cepstral

Let's talk about feature selection and extraction now. Speech has many traits that express how a speaker feels. Extracting these emotional signs is a key part of recognizing speech emotions effectively. This divides the features into long-term and short-term categories. It's important to analyze specific sections known as frames in the speech signal. Key elements include prosodic features such as pitch, energy, duration, formant features, Mel Frequency Cepstrum Coefficients (MFCC), and Linear Prediction Cepstrum Coefficient (LPCC). Emotions appear as changes in speech characteristics, things like tempo and pitch, change when someone feels different things. For example, anger often shows higher mean pitch and energy levels; joy is linked to a wider pitch variation range; sadness tends to occur with a slower pace and less energy overall. This use statistical analysis of pitch values and energy to uncover emotional information from these speech signals. The LPCC gives us details on individual speech channel traits that differ with various emotions; it also reduces processing needs while providing sharp vowel characteristics [20].

V. RESULTS AND DISCUSSION

The LSTM model was created with Python's Keras library. It works as a sequential model, which is Eras' way of organizing neural networks in layers. When this model starts with the sequential class, it begins building. Then, it can add layers to make sure everything connects right and functions well. In those LSTM layers, special LSTM cells come into play. It helps capture the temporal connections in the data. Before the LSTM stacks, there are dense layers to help generate results smoothly. After the model is built, it needs to be compiled.

This process is super important because it optimizes how the layers work together for a better matrix transformation. This stuff is key for running on CPUs and relies on how Keras is set up. Then have to set some important details, like which optimizer to use and what error functions are needed before compiling it. Next comes fitting the model to the training

Model: "sequential_1"

Layer (type)	Output Shape	Param #
lstm_1 (LSTM)	(None, 256)	264192
dropout_3 (Dropout)	(None, 256)	0
dense_3 (Dense)	(None, 128)	32896
dropout_4 (Dropout)	(None, 128)	0
dense_4 (Dense)	(None, 64)	8256
dropout_5 (Dropout)	(None, 64)	0
dense_5 (Dense)	(None, 7)	455

Total params: 305,799
 Trainable params: 305,799
 Non-trainable params: 0

Fig. 9. Summary of the Sequential Mode

data. This step aims to tweak its parameters and lower the loss function using methods like stochastic gradient descent (SGD) or Adam optimizer. The training data consists of input (X) and target output (Y), which are essential for changing the model weights over time. A big factor here is the number of training epochs; it really affects how well the model learns without overfitting or underfitting.

After training, it is time to evaluate the model using a separate testing dataset. This step checks how well it performs with metrics like accuracy, precision, recall, and F1-score. These measures give a clear view of how good the model is at making predictions and help compare it with other models so it can help pick the best one. A big part of this LSTM model includes MFCC features taken from audio using the librosa package. These features act as input for the LSTM model and are crucial for figuring out emotions in speech signals and with different combination are given in Table II. The general accuracy of this model is 75.34%, reaching as high as 86.33% at its best also depicts in Table III, showing that it can effectively capture emotional details.

There is also room for improvement! By tweaking data pre-processing steps, such as cutting out random silences from clips and identifying more annotated audio segments, the accuracy and reliability of the model can improve even more.

To wrap things up, this LSTM model made with Keras is a strong tool for recognizing emotions in speech. Its sequential setup with LSTM cells and dense layers helps extract and understand emotional hints from speech signals. With careful data pre-processing, optimization, and evaluation, this LSTM framework shines brightly in real-world uses such as human-computer interaction and psychotherapy.

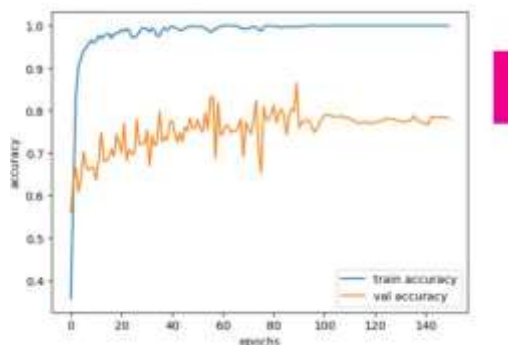


Fig. 10. Emotions in Speech Signals Performance Metrics

TABLE II
EMOTION RECOGNITION ACCURACY OF FEATURE VECTOR

FEATURE SET	ACCURACY
MFCC	75.34%
Frequency Scaled MFCC	42.9%
Cepstrum	71.4%
Frequency Scaled MFCC + Cepstrum	28.6%
MFCC + Cepstrum	57.1%
MFCC + Frequency Scaled MFCC	71.4%

VI. CONCLUSION

Speech emotion recognition is an important and quickly developing field within natural language processing, with its numerous uses in several disciplines. This work presents a brand new deep learning that combines convolutional neural networks with long- and short-term memory networks designed to distinguish feelings expressed through words. The purpose of this article is to identify emotions from speech signals. According to the approach suggested, CNN is used to identify seven distinct emotions using feature extraction techniques, including disgust, normal, and panic. Surprising, happy, rage, and disappointment. The aforementioned feature collection method is frequently employed to analyze voice data and enhance accomplishments. Mel-cepstral coefficients provide improved speech clarity and require less time to shape the spectrum with sufficient data. The overall accuracy of this model that uses Mel Frequency Cepstral Ceptral is 75.34 % and the highest accuracy of this model is 86.43 %. Because of the machine learning model's versatility, it can be used to a variety of fields, such as psychology, and computer-

TABLE III
EMOTIONS IN SPEECH SIGNALS MFCC FEATURES: ACCURACY

Overall Accuracy	75.34 %
Highest Accuracy	86.43%

human interactions. Enhancing the model's sensitivity to subtle emotions and testing its practicality could be the primary focus of future research efforts. The research being conducted represents a significant addition.

REFERENCES

- [1] Ancilin, J. and Milton, A., 2021. Improved speech emotion recognition with Mel frequency magnitude coefficient. *Applied Acoustics*, 179, p.108046.
- [2] Yang, N., Dey, N., Sherratt, RS, and Shi, F., 2020. Recognize basic emotional states in speech using machine learning techniques using features of the Cepstral coefficient of the Mel frequency. *Journal of Intelligent and Fuzzy Systems*, 39(2), pp.1925-1936.
- [3] A.Milton and J.Ancilin (2021). Improved speech emotion recognition with Mel frequency magnitude coefficient 986, pp.108046.
- [4] Palo, H.K., Chandra, M. and Mohanty, M.N., 2018. Recognition of human speech emotion using variants of Mel-frequency cepstral coefficients. *Advances in Systems, Control, and Automation ETAEERE-2016*, pp.491-498.
- [5] Yang, Z., and uang, Y. (2022). Algorithm for speech emotion recognition classification based on Melfrequency cepstral coefficients and broad learning system. *Evolutionary Intelligence*, pp.446-65.
- [6] Likitha, M. S., Sri Raksha R. Gupta, K. Hasitha and A. Upendra Raju. "Speech based human emotion recognition using MFCC." In 2017 international conference on wireless communications, signal processing and networking (WiSPNET), pp. 2257-2260. IEEE, 2017.
- [7] Nasrun, M. and Setianingsih, C., 2021, April. Human emotion detection with speech recognition using Melfrequency cepstral coefficient and support vector machine. In 2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS) (pp. 1-6). IEEE.
- [8] Hema, C. and Marquez, F.P.G., 2023. Emotional speech recognition using CNN and deep learning techniques. *Applied Acoustics*, 211, p.109492.
- [9] Rangra, K. and Kapoor, M., 2021. Exploring the mel scale features using supervised learning classifiers for emotion classification. *International Journal of Applied Pattern Recognition*, 6(3), pp.232-253.
- [10] Akcay, M. B., Ouz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116, 56-76 pp.98656.
- [11] Osadchyy, V., Skuratovskii, R.V., and Williams, A., 2020. Analysis of the Mel scale features using classification of big data and speech signals. *International Journal of Applied Mathematics, Computational Science, and Systems Engineering* pp.456554.
- [12] de Pinto, M.G., Polignano, M., Lops, P. and Semeraro, G., 2020, May. Emotions understanding model from spoken language using deep neural networks and Mel-frequency cepstral coefficients. In 2020 IEEE conference on evolving and adaptive intelligent systems (EAIS) (pp. 1-5). IEEE.
- [13] Nasrun, M. and Setianingsih, C., 2021, April. Human emotion detection with speech recognition using Mel-frequency cepstral coefficient and support vector machine. In 2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS) (pp. 1-6). IEEE.
- [14] Yang, Z. and Huang, Y., 2022. Algorithm for speech emotion recognition classification based on mel-frequency cepstral coefficients and broad learning system. *Evolutionary Intelligence*, 15(4), pp.2485-2494.
- [15] Abdul, Z.K. and Al-Talabani, A.K., 2022. Mel frequency cepstral coefficient and its applications: A review. *IEEE Access*, 10, pp.122136-122158.
- [16] Gou, L., Peng, J., Yang, H., Yu, Y. and Wang, X., 2024, January. Automatic Recognition of Mandarin Speech in Complex Medical Treatment Environments based on Mel-Frequency Cepstral Coefficients Analysis and Deep Learning. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering* (pp. 710-717)Bharti, D. and Kukana, P., 2020, September. A hybrid machine learning model for emotion recognition from speech signals. In 2020 international conference on smart electronics and communication (ICOSEC) (pp. 491-496). IEEE.
- [17] Mukherjee, S., Mundra, S. and Mundra, A., 2023. Speech Emotion Recognition Using Convolutional Neural Networks on Spectrograms and Mel-frequency Cepstral Coefficients Images. In *Information and Communication Technology for Competitive Strategies (ICTCS 2022) Intelligent Strategies for ICT* (pp. 33-41). Singapore: Springer Nature Singapore.
- [18] Mannar Mannan, J., Srinivasan, L., Maithili, K. and Ramya, C., 2023. Human emotion recognize using convolutional neural network (CNN) and Mel frequency cepstral coefficient (MFCC). *Seybold Report Journal*, 18(4), pp.49-61.
- [19] Vijayanthi, S. and Arunnehr, J., 2021. Synthesis approach for emotion recognition from cepstral and pitch coefficients using machine learning. In *International Conference on Communication, Computing and Electronics Systems: Proceedings of ICCCES 2020* (pp. 515-528). Springer Singapore.
- [20] Arya, R., Pandey, D., Kalia, A., Zachariah, B.J., Sandhu, I. and Abrol, D., 2021, October. Speech based emotion recognition using machine learning. In 2021 IEEE Mysore Sub Section International Conference (MysuruCon) (pp. 613-617). IEEE.