Emotion Recognition in Speech Signals using MFCC and Mel-Spectrogram Analysis

Dr.P.Muthuvel
Assistant Professor
Department of computer science and Engineering
Kalasalingam Academy of Research and Education
Krishnankoil-626126
Tamilnadu, India
Muthu.3225@gmail.com

T. Jaswanth
Department Of CSE
Kalasalingam Academy of Research and Education
Krishnankoil-626126
Tamilnadu, India
Jaswanthchowdary172217@gmail.com

Shaik Firoz
Department Of CSE
Kalasalingam Academy of Research And Education
Krishnankoil-626126
Tamilnadu, India
Shaikfiroz123786@gmail.com

S. Navya Sri
Department Of CSE
Kalasalingam Academy of Research and Education
Krishnankoil-626126
Tamilnadu, India
Somanavyasri2000@gmail.com

N. Mukhesh
Department Of CSE
Kalasalingam Academy of Research and Education
Krishnankoil-626126
Tamilnadu, India
Mukheshsunny005@gmail.com

2023 International Conference on Networking and Communications (ICNWC) | 979-8-3503-3600-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICNWC57852.2023.10127355

***Abstract*— In the domain of artificial intelligence, it's becoming more crucial than ever to classify emotions from both text and speech (AI). In order to promote and enhance human-ma- chine interaction, it is essential to establish a broader frame- work for speech emotion recognition. Machines are currently unable to reliably classify human emotions, hence machine learning development models were created for this purpose. Many academics worldwide are attempting to improve the ac- curacy of emotion categorization systems. The two steps of this study's creation of a speech emotion detection model are (I) tasked with managing and (ii) classification. The most pertinent feature subset was discovered using feature selection (FS). A wide variety of different vision - based paradigms were employed to address the growing demand for accurate emotion categorization all across the domain of ai technology, taking into account how crucial feature selection is. This study strategy for both the emotion categorization problem and the establishment of ml algorithms and deep learning methods. This same aforementioned work focuses on speech expression analysis & proposes a paradigm for bettering human-computer interaction through into the construction on prototype cognitive computing that categorizes feelings. The investigation aims to boost this same precision for eg in voice by applying methods for selecting features and now a spectrum different deep learning methodology, notably TensorFlow. A research also high- lights the contribution on component choice mostly in creation of powerful machine-learning algorithms towards feelings categorization.**

I. INTRODUCTION

One of the most important research topics to focus on is speech emotion detection, and academics from all around the world are always working to advance the field's capabil- ities. Male numerals from 0 to 9 may be recognized by the initial methods employ, built by Davis in the Bell Science lab as in The Us in 1952 [1].

The paradigm underlying voice emotion classification we suggest inside this paper aims to enhance human-machine interaction. The ability to appropriately categories emotions in speech and text is becoming more and more crucial as a.i. advances. AI-powered solutions can be far more effective in fields like service quality, healthcare, and education when they are able deal with issues related to feelings and emotions.

The proposed methodology has been divided into two sections metadata and categorization. Records management encompasses assembling but also making preparations a data - set of audio recordings, whilst the classifier employs methods of device going to learn to produce a model which can constructs moods in talk.
The incorporation of image retrieval techniques to identify one of most relevant attributes for template matching is one of the key aspects of our approach. This is critical since it aids in minimizing the complexity of the data [2], strengthen the classifier's effectiveness, and achieve better the generalization ability of the findings.

We use a multitude of deep learning frameworks, including traditional algorithms and deep neural networks, to assess the efficacy of the proposed framework. The designer's effectiveness is evaluated to use quality evaluation performance measures such like accuracy, precision, and recall.

On the whole, this study will add to the constant attempts of academics around the world to enhance the reliability of emotion categorization systems. A suggested scheme has the ability to significantly enhance the results of speech- emotion recognition systems by highlighting the signifi- cance of extracted attributes and the use of numerous ma- chine learning paradigms.

The speaker chose the numbers and said them into a conventional telephone, waiting 350 milliseconds between each word. Using the fundamental ideas of memory as well as matching, Audrey organized the author's input into electric classes that fitted previously defined reference patterns that had been historically created electrically as well as stored in an analog memory. It was evident to watch how Audrey's improper light responded by flashing.

Recognition posed too many obstacles for researchers, such as continuous voice recognition and emotion detection.

Emotions could be understood directly as well as through facial movements. Emotions always are present when people speech. The ability to recognize one's emotions makes emotions important. Speech shows an individual's emotions, including happiness, sadness, etc. As a result, understanding moods through speech has emerged as a new challenge in human-computer interaction (HCI) [3]. HCI needs to be more explicit to understand the basic human feelings. Yelling, sobbing, dancing, laughing, stamping, tease, and other expressions are just a few examples.

Systems for detecting emotions in speech use a variety of feature extraction techniques and classifiers. The three fundamental categories of characteristics are eliciting aspects, idiomatic expressions, and spectral features. For spectral characteristics, various technologies are used, such as MFCC, LPCC [4], and MEDC. Prosodic traits like pitch, intensity, frequency range, loudness, glottal characteristics etc. are examples that can be changed by technology. The Hidden Markov Model (HMM), Gaussian Mixtures Model (GMM), Support Vector Machine (SVM), and Artificial Neural Network (ANN) are several methods for identifying emotions.

II. BACK GROUND RELEATED WORK

Emotion detection from speaking is a complex job in ai. that will have gotten a lot of attention in recent years. The ability to precisely classify feelings in speech can significantly improve the efficacy of AI-powered processes throughout areas such as customer provider, healthcare, but also education. Despite advances in algorithms, the accuracy of emotion detection in speaker remains a difficult problem.

The volatility in how emotions are expressed is a major main challenge in emotion detection from voice. Monologue is a complex message that could also evoke a lot of content, including prosody, tune, fuel, as well as period, where it all may be was using to express feelings. Moreover, moods can really be expressed differently depending on a person's culture, gender, age, and other factors.

Another difficulty is procuring a broad and diverse information for teaching and evaluating recognition system models. The inability to train features that could still generalise well to bilingual voices and moods is hampered by a lack of something like a diverse and extensive dataset.

Emotion monitoring in talk relies on embeddings and great variety. The recognition model's achievement is immediately altered more by depiction of voice signal. Mel-Frequency Based on wavelet transform Co - efficient (MFCCs) and Mel Spectrum analyzer are two techniques for representing features [5].

In recent decades, have numerous studied showing on emotion track from voice. Several researchers concentrated on creating new feature representations, such as MFCCs and Mel Spectrograms, while others concentrated on creating new machine learning algorithms.

Many research findings often use customary machine learning techniques to reported feeling in speaking, such as support vector machines (SVMs) and judgement call trees.

These algorithms have been shown to undertake ok in emotion identification, but teaching needs a vast & different dataset. Learning has been implemented to boost the accuracy of recognition models in current history. Emotions in voice and others have been classified utilising convolutional neu- ral networks (CNNs) & recurrent neural networks (RNNs) [6]. These models have been shown to perform well, but us- ers entail a significant amount of information to train.

Several research on attribute selection technics for emotion recognising too in voice also not only been conducted. Several researchers also selected perhaps the most important variables for emotion detection utilising technics like principal component analyses (PCA) and linear discriminant analyses (LDA).
Overall, the research indicates that there is still some room for enhancement in emotion detection from voice. Such a study seeks to tackle the whole issue by employing technique for feature selection but a variety of machine educational paradigms, incl deep learning, to boost efficiency in emotion detection in speech.

III. EXSISTING SYSTEM

The primary method used by spoken emotion recognition systems to identify emotions is lexical analysis. The three emotions are currently classified as happy, sad, and neutral in most approaches. The degree of correlation between the training and test audio files is used as an integral parameter for identifying a specific emotion type. The maximum cross correlation between audio signal discrete time sequences is calculated [7]. Only the happy, angry, and neutral emotion segments are recognised by one of the other methods, which combines discriminatory feature extraction with the cubic SVM classifier.

IV. ISSUES IN EXISTING SYSTEM

Increasing the number of variables in the model will decrease its accuracy. Only three features can be classified by existing systems (Happy, sad and neutral) The sys- tems' highly static nature prevents them from perform- ing well in real-time systems. In comparison to correla- tions of the entire dataset with just one audio file[13].

The system is incredibly sluggish. Audio files with varying lengths cannot be understood. The model needs to go through several pre-processing processes in order to com- prehend the audio signal[14]. costly and not upgradeable.

V. PROPOSED SYSTEM

This feature makes use of Mel-frequency spectral co - efficient and Mel-spectrogram attributes. Voice info is classified into a number of emotion categories using neural networks and its MFCC characteristic described above. We have the advantage of becoming able to dis- tinguish a wide variety of feelings in real time by using neural networks to process audio signals of varying lengths and lengths.

Real-time continuous improvement precision & combinatorial quantity can be nicely balanced using technology. We use the deep learning algorithm CNN for Mel - frequency cepstral features, & Dense-Network called Densenet for

Mel-spectrogram capabilities. Densenet is a more compact version of CNN.

The advantage of using MFCC and Mel-spectrogram is that they are good at error reduction and can provide a robust feature when the signal is influenced by noise. SVM will be used for classification because it outperforms all other clas- sification algorithms in terms of performance and also con- tributes to better results.

The results show that the system can produce accuracy of up to 90.0% when using the TFD feature and 80.0% when using the MFCC and Mel-spectrogram features. This can be done quickly with any hardware that supports the Python programming language[13]. Processing audio from audio files is very user-friendly and quick. The system can understand audio files of various lengths.

VI. METHOLOGY

MFCC (Mel-frequency cepstral coefficients)

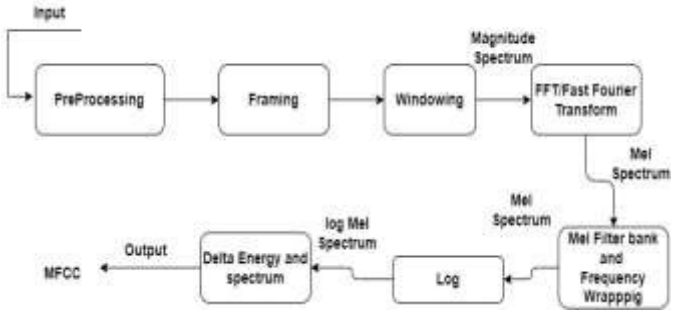


Fig.1.MFCC

In above fig1 MFCC have the Hertz values, like the Mel spectrogram, have been remapped to the Mel scale. Linear audio spectrum analyzers are best suited for situations in which all frequencies are equally important, whereas mel frequency components are ideally adapted for applications requiring a match to human hearing. We can see pics of hid- den audio features using the mel-frequency spectrum. While performing tasks like classification and recognition, CNN models can successfully extract features from images [8].

A mel-spectrogram is now a key point illustration used dur- ing speaker & audio processing one which compactly and informatively captures the spectral content of a sensor[15]. It is produced by performing a series of matrix multiplication on a time-domain signal, such as Fourier transforms, Mel-frequency filter banks, and logarithms.

The Fourier transmogrify serves to convert a time-domain signal into a frequency-domain representation before evalu- ating its mel spectrogram. A Mel-frequency wavelet trans- form serves to divide a resulting scope into multiple fre- quency bands [9]. Such a bit stream is logarithmically spaced and based on an individual's phonemic awareness. That is, at low frequencies, the frequency bands are closer together, while at high frequencies, they are farther apart. This filtration bank's result is a set of expected values, one for each band, representing the effort of the signal within that band. The mel spectrogram is then obtained by applying a logarithm to the output of the filter bank[14] The above depic- tion grasps the transmitter vibrational text inside a concise and informative manner, and it is resistant to signal fluctua- tions generated by various conversation styles, loudness, and some other factors.

Mel sub - bands are widely used within speech & speech synthesis because they can represent the spectral character- istics of a signal in a way that reflects the perception of the human auditory system. They are resistant to changes in speech style and noise and have been shown to be effective in a variety of language-related tasks such as speech recog- nition, speaker identification, and emotion recognition [10].

We propose in this thesis to use Mel - spectrograms as fea- ture representations for voice emotion detection. Mel spec- trograms can be extracted from sound waves and utilized as inputs to a text categorization machine learning module. Standard metrics are employed to evaluate predictive accu- racy, and thus the results are contrasted to those obtained using other feature representations.

Convolutional Neural Networks (CNN)

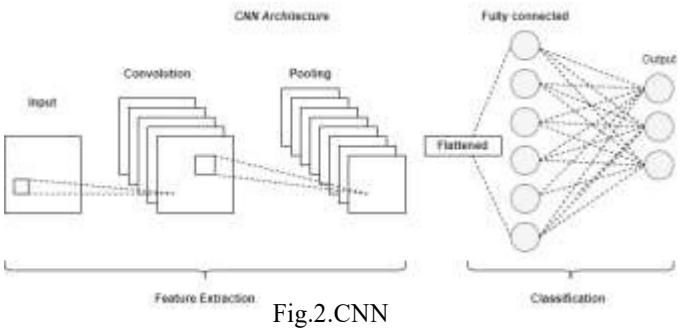


Fig.2.CNN

As shown in fig 2, CNN employs deep learning to distin- guish between different objects in input data by assigning significance to various traits and aspects of the image (teachable biases and weight). A CNN model requires far less pre-processing than traditional separate classifiers. Un- like previous methods, where filtration had to be manually designed, Convents can learn about such filters and their characteristics. The structure of a Convent was modelled af- ter the visual system and is comparable to the critical neu- ronal performance parameter seen in the human brain. When specific neurotransmitters are stimulated, the ability to ob- serve and evaluate a small portion of both the occipital lobe [11]. This is one of many overlapping areas which make up the entire visual field.

Basic architecture of CNN

Thresholding is a convolution layer tool used to split and characterise an object's various study capacities. The feature extraction system is made up of many multi - layer or pool- ing outer layer pairings.

A completely integrated layer which uses a muddled outputs to categorize the image based on the features gathered in previous rounds.

Use a CNN edge detection technique to lower the number of elements in a dataset. The number of functions available is increased by including new functions that are rehashes of old ones within the initial set of attributes. Many CNN parts are depicted in the CNN entity relationship diagram. Con- sider theCNNmodel.

VII.SYSTEM ARCHITECTURE

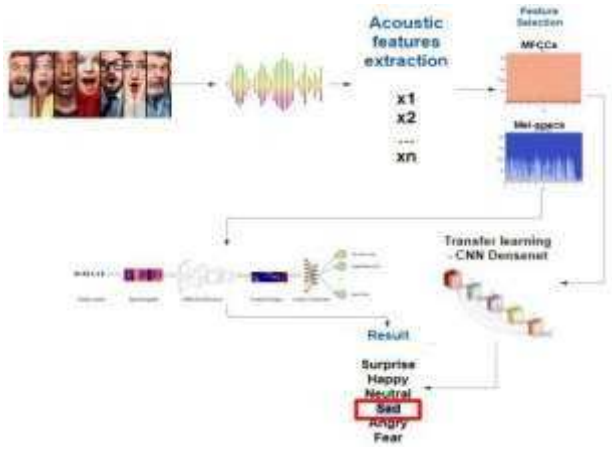


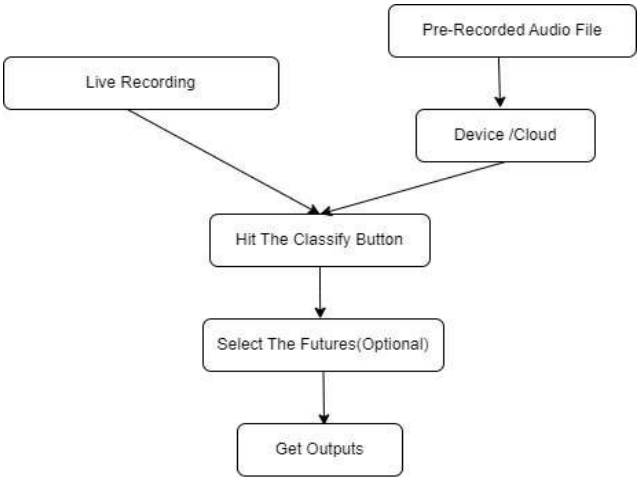
Fig.3.System Architecture

In above fig 3, A data - set, in the type of audio data, is employed in extracting the features. The wav sound files isolated again from dataset were always then utilised to extract. The for-file's values were always saved in matrix form. The stored matrix math format is used to perform two types of feature shortlists.

(I)Mel spectrogram
(II)Cepstral coefficient of Mel (MFCC)

When the signal is impacted by noise, the MFCC and Mel-spectrogram are effective at reducing errors and producing a bullet points. After variable selection, CNN is utilized to ap- ply in order [12], and the final output indicates whether the emotion is happy, sad, or any other. In the MFCC phase, CNN is utilized for Mel - spectrogram extracting features in addition to designation. Below Fig 4 shows the Output De- sign

XIII. SYSTEM OUTPUT DESIGN



Fig,4. Output design

IX. RESULT

As shown in Figures 5 and 6, we uploaded the data set and obtained the output sound wave. It is fed into MFCCs and Mel-log-spectrogram

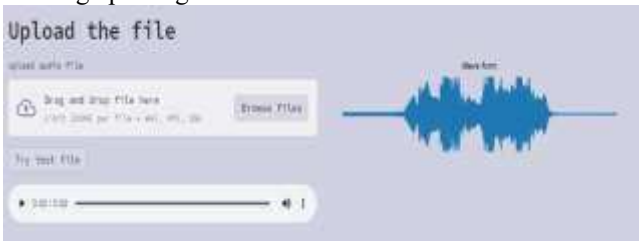


Fig.5.Sound Wave

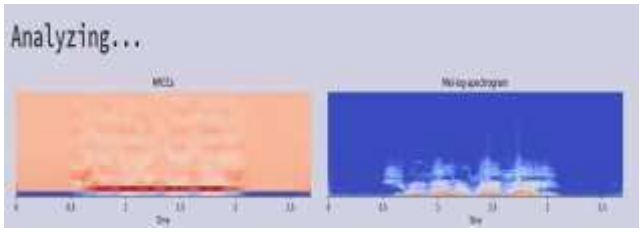


Fig.6.Analyzing the input

X. FUTURE SCOPE AND CONCLUSION

The proposal for voice handling really does have the potential to considerably boost emotion detection efficiency performance. Notwithstanding, further studies is necessary in several areas to wholly explore a benefits of such a approach.

Future Study aims to look ought to take a glance into the application of additional multilayer perceptron for recognizing speech emotions. While MFCCs & Mel - Spectrograms are prevalently utilised, many expresses for empathic sensing can sometimes exist. Characteristics like pitch & fuel, for instance –, could well be essential for emotion classification and ought to be studied further.

Future study should glance in to one of the utilization of other machine educational understandings for emotion detection. Whilst still traditional machine-learning algorithms and supervised learning have proven to be efficacious, those certain paradigms for emotion sensing may exist. Emotion recognition simulations, for instance –, can really gain financially by the use of generative adversary networks (GANs) and Transformers.

Furthermore, investigation remains necessary to comprehend in accuracy for mood tracking technologies under low- resource conditions. It thus includes developing models that can handle less datasets as well as models that can handle noisy or damaged speech.
Finally, it would be fascinating to investigate the prospective pragmatic implementations on our suggested frame- work for identifying speech emotions. This might include developing emotion schemes for customer service, healthcare, and education, as well as exploring a social as sociological concerns of such systems.

Overall, the suggested framework on mood perception is a promising tool at all with the ability for boost emotion sensing system performance. However, rather more study is required too fully grasp its power & utilize it in real-world circumstances.

Vocal mood identification techniques depend upon many classifiers are seen. The signals processing element, which extracts key features form accessible audio input, and a classifier, what detects moods from alongside a voice waveform, are essential challenges to vocal emotion identification. This same precision for many of found in different forms classifications in either a speaker independent system is smaller compared to a sound dependent fashion.

Automatic emotion identification from human speech is becoming more popular since it leads in improved interactions amongst both humans and machines. Combinations of the above approaches can be derived in improve the emotional authentication process. Furthermore, its effectiveness of eeg based emotion identification may be improved by extracting more useful aspects of speech.

Also this work could be extended by making comparisons of the techniques mentioned in this work with accuracy, error or efficiency parameters

REFERENCES

[1] Babak Joze Abbaschian, Daniel Sierra-Soa et.al,"Speech emotion recognition", *MDPI publications*, Sensors , 21(4), 1249 (2021)

[2] Hao Ming, Tianhao Yang et.al "Speech emotion recognition from 3D Log-Mel Spectrograms with Deep Learning Network and with methods", *IEEE publications* , volume 5 , pages1215-1221 (2019)

[3] Wisha Zehra, Abdul Rehman Javed et.al, "Cross corpus multi-lingual speech emotion recognition using ensemble learning", *Springer Nature publications*, volume 7, pages1845–1854 (2021)

[4] Eva Lieskovska, MichalChmulik et.al, "speech emotion recognition using deep learning and attention mechanism", *MDPI publications*, Electronics 10(10), 1163 (2021)

[5] J Ancilin, "Improved speech emotion recognition with Mel frequency magnitude coefficient*", Elsevier publications*, Applied Acoustics 10.1016 108046 (2021)

[6] Ziping Zhao, Qifei Li et.al, "Combining a parallel 2D CNN with a self-attention Dilated Residual Network for CTC-based discrete speech emotion recognition", *Elsevier publications*, Neural Networks 10.1016 (2021)

[7] Prabhav Singh, KPS Rana et.al, "A multimodal hierarchical approach to speech emotion recognition from audio and text", *Elsevier publications*, Knowledge Based Systems 10.1016 107316 (2021)

[8] Youngja Nam, Chankyu lee, "title Cascaded Convolutional Neural Network Architecture for Speech Emotion Recognition in Noisy Conditions", *mdpi publications*, Sensor Networks 21(13), 4399 (2021)

[9] Siddique Latif; Rajib Rana et.al, "Survey of Deep Representation Learning for Speech Emotion Recognition*", IEEE publications*, 10.1109/TAFFC.2021.3114365 (2021)

[10] Mustaqeem, Soonil kwon, "Optimal feature selection speech emotion recognition", *Wiley publications*, 10.1002/int.22505 (2021)

[11] Yuan, Jiahong, Xingyu Cai, Renjie Zheng, Liang Huang, and Kenneth Church. "The role of phonetic units in speech emotion recognition." *arXiv preprint arXiv:2108.01132* (2021).

[12] Ntalampiras, Stavros. "Speech emotion recognition via learning analogies." *Pattern Recognition Letters* 144 (2021): 21-26.

[13] Ali, Hasimah, Muthusamy Hariharan, Sazali Yaacob, and Abdul Hamid Adom. "Facial emotion recognition using empirical mode decomposition." *Expert Systems with Applications* 42, no. 3 (2015): 1261-1277.

[14] Liu, Zhen-Tao, Min Wu, Wei-Hua Cao, Jun-Wei Mao, Jian-Ping Xu, and Guan-Zheng Tan. "Speech emotion recognition based on feature selection and extreme learning machine decision tree." *Neurocomputing* 273 (2018): 271-280.

[15] Ragot, Martin, Nicolas Martin, Sonia Em, Nico Pallamin, and Jean-Marc Diverrez. "Emotion recognition using physiological signals: laboratory vs. wearable sensors." In *Advances in Human Factors in Wearable Technologies and Game Design: Proceedings of the AHFE 2017 International Conference on Advances in Human Factors and Wearable Technologies, July 17-21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA 8*, pp. 15-22. Springer International Publishing, 2018.