Implementing Mel-Spectrogram Analysis for Emotion Recognition in Speech

Rishi Ahuja

New Delhi, India

DOI:10.37648/ijrst.v13i04.003

[1]*Received: 22 August 2023; Accepted: 05 October 2023; Published: 15 October 2023*

ABSTRACT

Emotion classification from speech and text is becoming increasingly important in artificial intelligence (AI). A more comprehensive framework for speech emotion recognition must be established to encourage and improve human-machine interaction. Since machines can't now accurately categorize human emotions, models for machine learning development were explicitly developed for this use. Around the world, many researchers are working to increase the accuracy of emotion classification algorithms. To create a speech emotion detectionmodel for this study, two processes are involved: (I) managing and (ii) classifying. Feature selection (FS) was used to find the most relevant feature subset. An extensive range ofdiverse vision-based paradigms were used to meet the increasing demand for precise emotioncategorization throughout the AI technology industry, considering how vital feature selection is. This research approach addresses the difficulty of classifying emotions and the development of machine learning and deep learning techniques. This previously mentioned work focuses on voice expression analysis and offers a paradigm for improving human- computer interaction by developing a prototype cognitive computing system to classifyemotions. The research aims to increase this similar precision, for example, in voice, by utilizing feature selection techniques and, more recently, a variety of deep learning methodologies, most notably TensorFlow. A study further emphasizes how vital component selection is in developing robust machine learning algorithms for the classification of emotions.

INTRODUCTION

Speech emotion detection is one of the most vital study areas, and researchers worldwide are constantly striving to improve the field's capabilities. Davis developed the first algorithms forrecognizing male numerals from 0 to 9 in the Bell Science lab in the United States in 1952 [1].

Our proposed speech emotion classification paradigm in this paper focuses on improving the interaction between humans and machines, as AI Develops, the capacity to accurately classifyemotions in voice and writing becomes increasingly essential. When AI-powered solutions can handle problems with feelings and emotions, they can be significantly more effective in healthcare, education, and service quality.

The two aspects of the suggested methodology are categorization and metadata. Records management involves compiling and organizing a data set of audio recordings, while the classifier uses machine learning techniques to create a model that can create conversational moods.

One of the main components of our approach is the integration of image retrieval algorithms to determine one of the most significant features for template matching. This is important because it helps reduce the data's complexity [2], improve the classifier's efficiency, and increase the findings' capacity for generalization.

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 by using quality evaluation performance measures such as accuracy, precision, and recall. On the whole, this study will add to the constant

[1] How to cite the article: Ahuja R.; October 2023; Implementing Mel-Spectrogram Analysis for Emotion Recognition in Speech; *International Journal of Research in Science and Technology*, Vol 13, Issue 4, 17-22, DOI: http://doi.org/10.37648/ijrst.v13i04.003

attempts of academics around the world to enhance the reliability of emotion categorization systems. A suggested scheme can significantly enhance the results of speech-emotion recognition systems by highlighting the significance of extracted attributes and the use of numerous machine-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 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 speak. 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, teasing, and other expressions are just a few examples. Systems for detecting emotions in speech usea 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.

EXISTING 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 mostapproaches. 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 recognized by one of the other methods, which combines discriminatory feature extraction with the cubic SVM classifier.

ISSUES IN EXISTING SYSTEM

Increasing the number of variables in the model will decrease its accuracy. Only threefeatures can be classified by existing systems (Happy, sad, and neutral) The systems' highly static nature prevents them from performing well in real-time systems. In comparison to correlations 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 to comprehend the audio signal[14]. costly and not upgradeable.

PROPOSED SYSTEM

This feature makes use of the Mel-frequency spectral coefficient and Mel-spectrogram attributes. Voice info is classified into several emotion categories using neural networks and its MFCC characteristic described above. We have the advantage of being able to distinguish 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 and a 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 forclassification because it outperforms all other classification algorithms in terms of performance and also contributes to better results.

The results show that the system can produce an accuracy of up to 90.0% when using theTFD 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.

METHODOLOGY

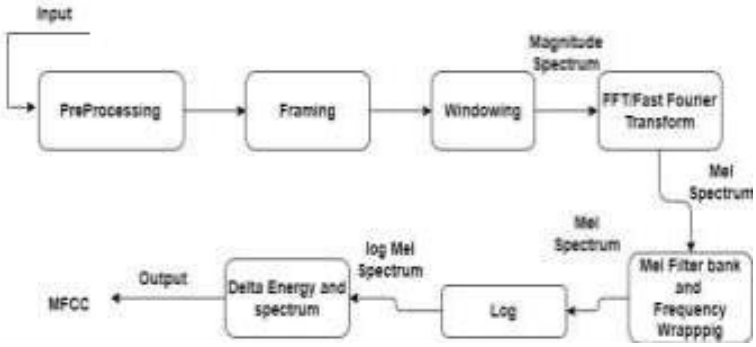

Fig 1: Represents MFCC

In the above fig1 MFCC 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 forapplications requiring a match to human hearing. We can see pics of hidden 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 during 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 evaluating its mel spectrogram. A Mel-frequency wavelet transform serves to divide a resulting scope into multiple frequencybands [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 the algorithm to the output of the filter bank [14] The above depiction grasps the transmitter vibrational text in a concise and informative manner, and it isresistant to signal fluctuations 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 characteristics of a signal in a way that reflects the perception ofthe

human auditory system. They are resistant to changes in speech style and noise and are effective in a variety of language-related tasks such as speech recognition, speaker identification, and emotion recognition [10].

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

Convolutional Neural Networks (CNN)

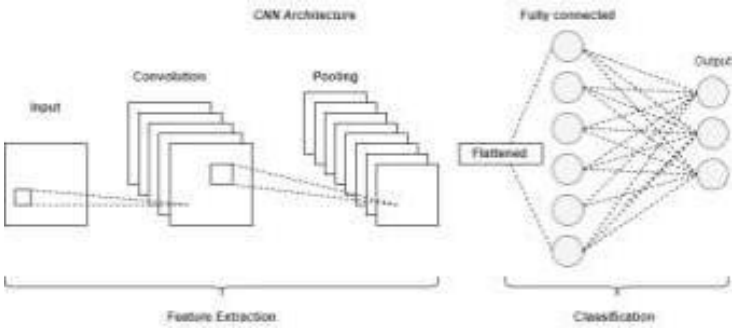

Fig 2: CNN

As shown in Fig 2, CNN employs deep learning to distinguish 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. Unlike previous methods, where filtration had to be manually designed, Convents can learn about such filters and their characteristics. The structure of a Convent was modeled after the visual system and is comparable to the critical neuronal performance parameter seenin the human brain. When specific neurotransmitters are stimulated, the ability to observe andevaluate a small portion of both the occipital lobe [11]. This is one of many overlapping areaswhich make up the entire visual field.

SYSTEM OUTPUT DESIGN

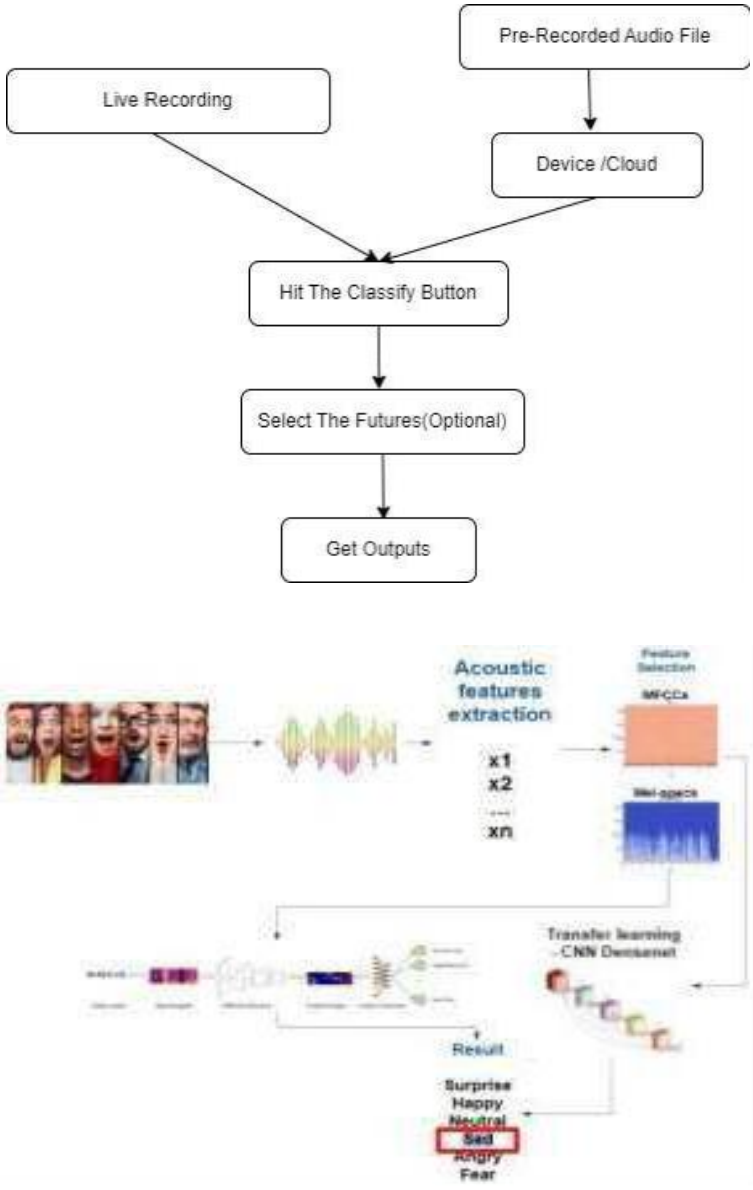

Fig.3. Output design

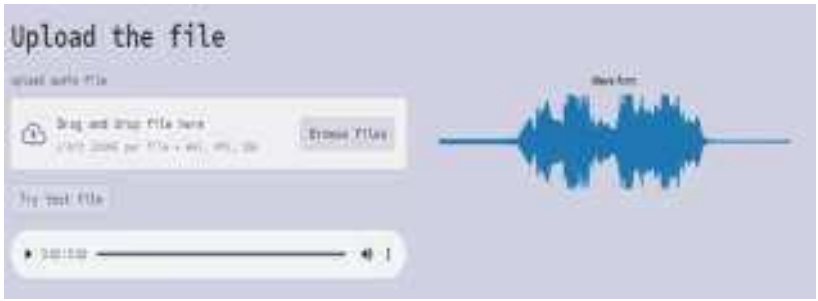

Fig.4.Sound Wave

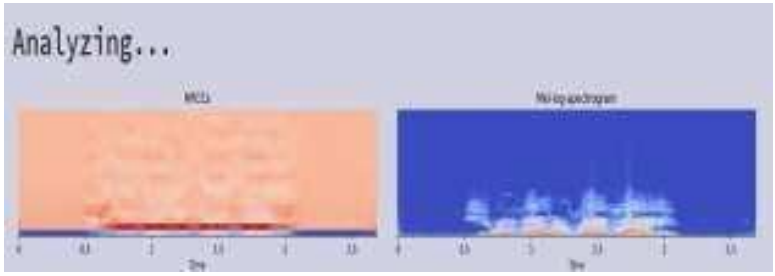

Fig.5.Analyzing the input

RESULT

As shown in Figures 4 and 5, we uploaded the data set and obtained the output sound wave. Itis fed into MFCCs and Mel-log-spectrogram

CONCLUSION

Automatic emotion identification from human speech is becoming more popular since itleads to improved interactions amongst both humans and machines. Combinations of the above approaches can be derived to improve the emotional authentication process. Furthermore, the 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, pages 1215-1221 (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.