Vocal Vibes: Speech Emotion Recognition using Machine Learning

PROF. DR. MRS. SARITA DESHPANDE[1], SANSKRUTI GAIKWAD[2], PRITESH GADIYA[3], VYANKATESH KHETRI[4], PRERANA PATIL[5]

[1] *Information Technology, P.E.S. Modern College of Engineering Pune, India*

[2, 3, 4, 5] *B.E - Information Technology, P.E.S. Modern College of Engineering Pune, India*

Abstract*— Speech emotion recognition is the process of accurately anticipating a human's emotion from their speech. It improves the way people and computers communicate. Although it is tricky to annotate audio and difficult to forecast a person's sentiment because emotions are subjective, "Speech Emotion Recognition (SER)" makes this possible. Various researchers have created a variety of systems to extract emotions from the speech stream. Speech qualities in particular are more helpful in identifying between various emotions, and if they are unclear, this is the cause of how challenging it is to identify an emotion from a speaker's speech. A variety of the datasets for speech emotions, their modelling, and types are accessible, and they aid in determining the style of speech. After feature extraction, the classification of speech emotions is a crucial component, so in this system proposal, we introduced artificial neural networks (ANNs) that are used to distinguish emotions such as anger, disgust, fear, happiness, neutrality, sadness, and surprise.

Indexed Terms-* Detection, Speech Input, Feature Extraction

I. INTRODUCTION

Speech Emotion Recognition (SER) is a field of research and application within the broader domain of Natural Language Processing (NLP) and Machine Learning. It focuses on the automatic detection and classification of human emotions expressed in spoken language. This technology has a wide range of potential applications, including improving human-computer interaction, mental health diagnostics, customer service, and many more. Emotions are fundamental to human communication and play a crucial role in conveying the speaker's intentions, attitudes, and feelings. Identifying and understanding these emotions from speech can enhance the quality of human-computer interactions and provide valuable insights into the emotional state of individuals. Machine learning algorithms, particularly deep learning models, have made significant progress in this domain, allowing for more accurate and reliable speech emotion recognition. Understanding human emotions through speech has far-reaching implications. It enables machines to interact with humans more effectively, enhancing user experience in applications such as virtual assistants, customer service chat bots, and social robots. It also has immense potential in the fields of psychology, mental health, and sentiment analysis, aiding in the early detection of emotional disorders and improving therapy outcomes. One of the quickest and most normal ways of communicate is to give indications. Emotions can be of different types and can be Expressed in different ways, Facial Expression is one of the most prominent method for recognizing Emotions, Using Facial Expression we can detect Emotions. The Other Way of Communication is through Sign Language. The utilization of sound signs is a quick and effective method for interacting with a human machine. All thoughts are utilized by individuals to all the more likely comprehend the message they are getting. Emotion Detection is a troublesome errand for the machine, however then again, it is typical for people.

II. PROPOSED SYSTEM

The main objective of this project is to detect the emotion of the speaker. Before Speech Emotion Detection was carried out as machine learning (ML). The execution steps are contrasted with other ML undertakings, and better plan processes are improved. The initial step is to accumulate data, which is vital. The model being created depends on the data gave and all choices and reactions to the model being created depend on the information. The subsequent advance, called assembling properties, is to gather countless

assignments that are performed on the gathered information. This methodology resolves many issues connected with data and data quality. The third step is frequently viewed as the way in to a ML project that upholds calculation improvement. This model uses ML algorithms to learn data and figure out how to get all significant data. The last advance is to assess the presentation of the introduced model. Looking at the outcomes will assist you with picking the most suitable ML calculation for the issue. In this paper, we showed speech emotion recognition (SER) utilizing AI calculations to find out feelings. The activity of a feeling acknowledgment structure can altogether change the general activity of the structure in numerous ways and give a greater number of advantages than the activity of the program. This review tells the best way to understand intense sounds, how to work on the current structure as far as data, how to choose elements, and how to group emotional sounds dependent on feelings.

A.SPEECH INPUT MODULE

Input to the system is speech. The digital representation of the given sound through the sound file is currently handled.

B. FEATURE SELECTION AND CLASSIFICATION MODULE

In this process we extract features from the input data. After extracting, selection of required features is done. Those files are used for training and testing of the classifier. We train our classifier with training data and then we proceed for testing data. Using testing data we find out the accuracy and classification report of the trained classifier. Emotions can be detected from audio files because audio files contains different parameters. Parameters can change the emotion information. Voice frequently returns hidden feeling through pitch and tone. The objective of feature extraction is to get useful feature from audio file for feelings. The Audio files contain a lot of information other than emotion detection. Therefore research on how to extract and which parameters to extract are of great important. Features are extracted from the audio file given as information. The features are MFCC, Mel, Chroma, Tonnetz. Emotions can also be recognized by combining the Mel Frequency Cepstral Coefficients (MFCC) with the vibration rate (PITCH) in order to characterize the emotion according to its respective vocal speech signals

C.RECOGNIZED EMOTIONAL OUTPUT

Happy, Disgust, sad, Angry, Surprised, Neutral, Fear, Disgust are primary emotions detected in this speech emotion detection.

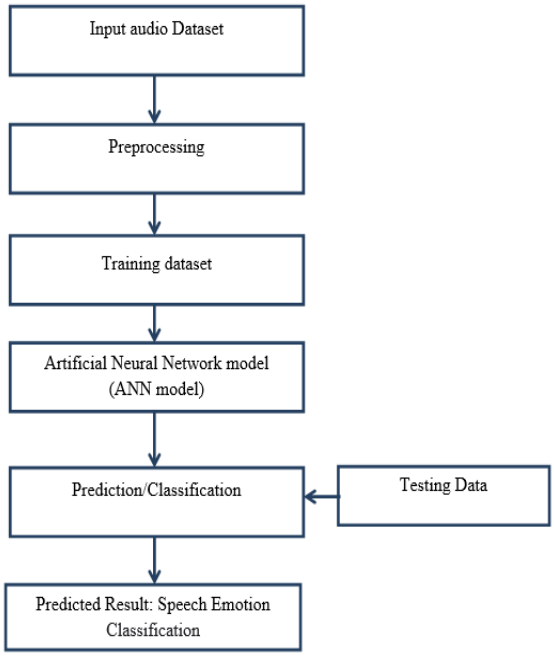
III. LITERATURE SURVEY

Using Machine learning, detecting emotion of the speaker is possible. There are few references to do it. [4] In this paper, they proposed the use of a Gaussian investigation that can separate conversation feeling level dependent on I-vectors demonstrating the dispersion of MFCC usefulness. A review dependent on the IEMOCAP corpus shows that the foundation of the GPLDA surpasses the foundation of the SVM and is less delicate to the I-vector, so the normal level makes it more powerful to change rules during framework improvement. [5] To work on the capacity to continually recognize feelings from conversation, we are upholding an instructing blunder-based way to deal with learning and an organization that recursively creates memory. In such manner, with the assistance of two continuous RNN (Recurrent Neural Networks) strategies, the main model is utilized as a computerized code to recover the first substance and the subsequent model is utilized for passionate forecast. RE (Reconstruction-error-based) of the primary resource is utilized as an extra resource, and is matched to the main resource and put in the subsequent class. The possibility of the framework is that the framework can concentrate on its "shortcomings" in RE. A RECOLA database - based review shows that a given framework is better than a base framework without RE information as far as the Concordas coefficient. [6] Social media correspondence is one of the main parts of compelling correspondence with a Computer. For this sort of association, an unmistakable comprehension of the significance of the word and language arrangement and a familiarity with the feelings contained in the conversation are vital to further develop execution. The language used to communicate feelings passes on Emotions like sadness, fear, joy, and distress. In the main phase of the feeling forecast framework introduced in this article, various sorts of feelings are recognized. The subsequent advance is to utilize a

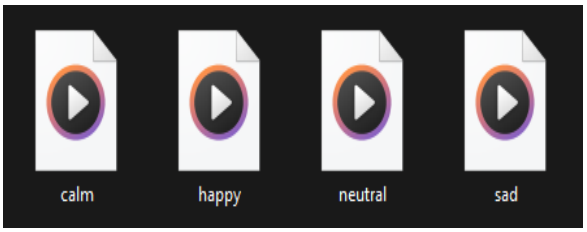
neural organization to anticipate the following series of feelings. Subsequent to joining the diverse discourse images at each point, the grouping is deducted dependent on the communication characters of one tenth of a second. The issue of forecast is that the neural organization is diverted in a nonlinear course, and the change is characterized as the hypothesis of time data. The best gauge of Random Forest accomplished because of anticipating results utilizing the organization is 86.25%. [7] Security (Network protection) is a significant issue today and with trend setting innovation. It is critical that network safety is in excess of a mysterious framework to ensure against cybercrime. Equivalent biometric information and action arranged humanities permit applications to get to explicit or general data. The principle motivation behind this article is to investigate feeling based talk utilizing worldview acknowledgment, and affectionate displaying with neighbors. The five signs are cestrum, Mel frekans cestrum, order, cestrum. The calculation utilizes pockets to prepare the KNN majority. Reconciliation is done straightforwardly. The outcomes show that the greatest presentation gain is accomplished utilizing two distinct KNNs rather than utilizing one KNN. [8] The test is to keep up with the force of the Delete Speech Processing System in the midst of commotion. This page shows a wide data transmission that can work on the responsiveness of the emotion acknowledgment framework when indications are harmed because of chosen clamor. This page shows it as well settling on decisions dependent on explicit prerequisites can give us the greatest aspect to find the best solution

IV. DATA FLOW DIAGRAM

Data Flow Diagram shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.

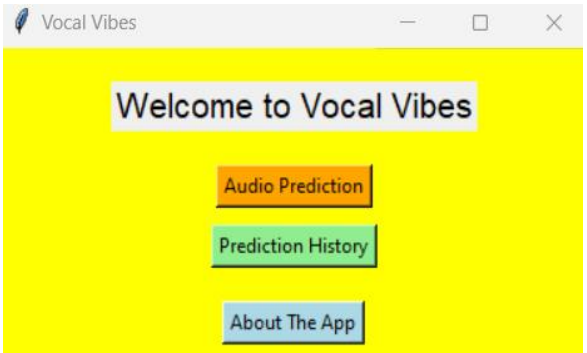


V. DATASETS



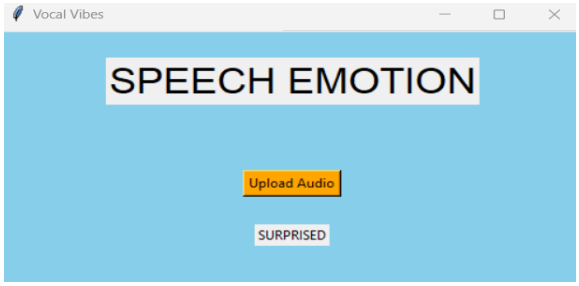
This dataset contains a set of audio files. The emotions are
labelled as follows: 01-'neutral', 02-'calm', 03-'happy', 04-
'sad', 05-'angry', 06-'fearful', 07-'disgust', 08-'surprised'.

VI. RESULTS



In this process we extract features from the input data. After extracting, selection of required features is done. Those files are used for training and testing of the

classifier. We train our classifier with training data and then we proceed for testing data. Using testing data we find out the accuracy and classification report of the trained classifier.



Emotions can be detected from audio files because audio files contains different parameters. Parameters can change the emotion information. Voice frequently returns hidden feeling through pitch and tone.

CONCLUSION

The aim of the paper is to detect the emotions that are elicited by the speaker while speaking. Emotion detection has become an essential task these days. The speech that is in fear, anger, or joy has a higher and wider range in pitch, whereas it has a low range in pitch. The detection of speech is useful in assisting human-machine interactions. These models have been trained to recognize these emotions calm, neutral, surprise, happy, sad, angry, fearful, and disgust.

REFERENCES

[1] Kottursamy, Kottilingam. “A review on finding efficient approach to detect customer emotion analysis using deep learning analysis.” Journal of Trends in Computer Science and Smart Technology 3, no. 2 (2021): 95-113.

[2] Thakur, Amrita, Pujan Budhathoki, Sarmila Upreti, Shirish Shrestha, and Subarna Shakya. “Real Time Sign Language Recognition and Speech Generation.” Journal of Innovative Image Processing 2, no. 2 (2020): 65-76.

[3] Kaur, Jasmeet, and Anil Kumar. “Speech Emotion Recognition Using CNN, k-NN, MLP and Random Forest.” In Computer Networks and Inventive Communication Technologies, pp. 499-509. Springer, Singapore, 2021.

[4] Gamage, Kalani Wataraka, Vidhyasaharan Sethu, Phu Ngoc Le, and Eliathamby Ambikairajah. "An i-vector gplda system for speech based emotion recognition." In 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), pp. 289- 292. IEEE, 2015.

[5] Han, Jing, Zixing Zhang, Fabien Ringeval, and Björn Schuller. "Reconstruction-error-based learning for continuous emotion recognition in speech." In 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 2367-2371. IEEE, 2017.

[6] Akrami, N., F. Noroozi, and G. Anbarjafari. "Speechbased emotion recognition and next reaction prediction." In 25th Signal Processing and Communications Applications Conference, Antalya, pp. 1-6. 2017.

[7] Rieger, S. A., Muraleedharan, R., & Ramachandran, R. P. (2014, September). Speech based emotion recognition using spectral feature extraction and an ensemble of kNN classifiers. In The 9th International Symposium on Chinese Spoken Language Processing (pp. 589-593). IEEE.

[8] Tabatabaei, Talieh S., and Sridhar Krishnan. "Towards robust speechbased emotion recognition." 2010 IEEE International Conference on Systems, Man and Cybernetics. IEEE, 2010

[9] Z. Li, “A study on emotional feature analysis and recognition in speech signal,” Journal of China Institute of Communications, vol. 21, no. 10, pp. 18–24, 2000.

[10] T. L. Nwe, S. W. Foo, and L. C. de Silva, “Speech emotion recognition using hidden Markov models,” Speech Communication, vol. 41, no. 4, pp. 603–623, 2003.

[11] Vaijayanthi, S., and J. Arunnehru. “Synthesis Approach for Emotion Recognition from Cepstral and Pitch Coefficients Using Machine Learning.” In International Conference on Communication, Computing and Electronics Systems, p. 515.

[12] Y. Kim, H. Lee, and E. M. Provost, “Deep learning for robust feature generation in audio-visual emotion recognition,” in Proceedings of the IEEE International Conference on Acoustics,

Speech and Signal Processing (ICASSP '13), Vancouver, Canada, 2013.

[13] Livingstone, Steven R., and Frank A. Russo. "The Ryerson AudioVisual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English." PloS one 13, no. 5 (2018): e0196391

[14] Chourasia, Mayank, Shriya Haral, Srushti Bhatkar, and Smita Kulkarni. “Emotion recognition fromspeech signal using deep learning.” Intelligent Data Communication Technologies and Internet of Things:Proceedings of ICICI 2020 (2021): 471-481.

[15] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review", IEEE Access, vol. 2, no. 7, pp. 117327-117345, 2019.