IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Sentiment Analysis on Speech Data using MFCC

Authors - Aman Upadhyay, Arpita Singh, Ananya Gupta, Aparana Sharma, Aditya Pratap Mall, Ms. Pooja Tomar

Abstract

Sentiment analysis is a popular technique these days used to detect emotions or sentiments from the audio. Lot of people have done research in this field bringing out most of the features. Mostly the work is done in speech recognition using the polarities. The functionality can be extended to a plugin system where people can analyze and assess the sentiments of the audience during online meetings and webinars. Here, we tried to work on all most common type of emotions through the best sufficient algorithm. Sentiment analysis gained more recognition over the past years, and largely
focused with text mining techniques around sentiment analysis . However, the research community is still very new to audio sentiment analysis. We implement sentiment analysis on speech transcripts of different speakers in this research to detect emotions of speaker involved in the conversation so it can further be implemented as plugin by various application platforms.

Keywords: Sentiment Analysis, Speech Recognition, MFCC, DTW, Natural Language Processing.

1. INTRODUCTION

Sentiment Analysis involves to detect the sentiments or emotions of people through their speech in conversations. It has various uses. While we talk to a person face-to-face, we can easily examine the people's emotions through their expressions and body language. But it is difficult to do so, when we are talking to someone through online mode or through a telephonic conversation. As we faced the global pandemic, many of us were confined to our homes. All the meetings and classes happened online including staying in touch with close ones. Thus, for such times it was essential to know the other person's sentiments to better understand them. Here, the sentiment analysis's function comes into play. It involves identifying spoken words and then converting them into text using various speech recognition algorithms with help of the machine learning. The resultant text is then analyzed and then categorized as of positive polarity, negative or of neutral polarity. Speech Recognition involves the task of differentiating users based on voice modulation, speech patterns and various other factors. We reviewed the data and use it to gain insights for detecting sentiments of people. Therefore, we first put in place a speaker and speech recognition system, and then we extract data based on this system. Further sentiment analysis process is conducted.

As per the researches, most work has included manually typing the spoken words to examine the human mindset. Many works include neural network concepts such as speech recognition, such as recurrent neural networks (RNNs), Natural Language Processor, and sequence models. In our proposed research, sentiment analysis focuses on identifying the speaker's sentiment by partitioning and categorizing the dataset achieved from the speech recognition algorithms. Some classification algorithms include Naive Bayes, Linear Regression and Support Vector Machines. These are mainly used for selecting and classifying the text for sentiment analysis. The idea is to extract terms out of the text, compare them to a dataset, detect emotions as happiness, sadness, joy, anger, frustration, and fear. The paper offers to explore challenges and methods for implementing audio sentiment analysis while suggesting the functionality of plugin feature for further future use by various platforms. The plugin is a valuable extension that can help provide insights for hosts, participants and organizations. It analyzes the sentiment as well as written communication during online conversations using machine learning and natural language processing methods. But in order to do this, the speech recognition computer must be intelligent enough to distinguish between human voices and background noise.

2. OVERVIEW OF PROCESS

2.1 Sentiment Analysis:

Sentiment Analysis (SA) focuses to detect the sentiment expressed by the speaker in a conversation by assessing it in a text and analyzing whether it conveys a positive or a negative tone. The focus in sentiment analysis research has been mainly on algorithms like Support vector machines, Linear Regression and Naive Bayesian. The text is categorized as objective or subjective through classical machine learning techniques. Once the text is classified into tokens, it becomes easy for the polarity classifier to assess the polarity. Based on this assessment, we declare the sentiment detected. However, it is a time consuming task considering collection of data and classifying them and thus pose as a challenge to the research. In this research, we used the methodologies like Support Vector Machines, Naive Bayes and VADER. Among these, the most efficient one was implemented for the specified purpose leading to the task of sentiment analysis.

2.2 Speech Recognition:

Speech Recognition can referred as the ability to translate spoken words into texts. It is the capability to allow a computer programme to convert the spoken language into the written form. Those written words and phrases assessed from the speech of humans are then converted to a machine readable format for subsequent processing. This research includes various speech recognition tools like Google Speech Recognition and Bing Speech. The signal for speech as a form of communication, combines numerous features that can be employed to extract information that is speaker-specific, linguistic and emotional[8]. Whereas Speech recognition specifically perform the task of extracting and utilizing features that are unique to the speaker from the speech signal.

2.3 Feature Extraction:

It is very important to eliminate the noise and extract the features of different speaker discriminant to achieve heightened accuracy for the process. The accuracy achieved in this phase is crucial for the one that comes after as it provides the input for the next step.
This article uses **MFCC** which is **Mel Frequency Cepstrum Coefficient** that is an audio feature extraction technique to extract speaker specific parameters from speech. It aims to identify performance enhancements. MFCCs are more preferred because they are less susceptible to background noise and channel distortions. It is efficient in capturing the characteristics of sound which are spectral, by emphasizing on human auditory perception. Since human hearing operates on a scale (non-linear), MFCC analyses sounds using a mathematical model. It converts Mel frequencies to actual acoustic frequencies. This distinctiveness aids in the speaker's voice recognition [6].

2.4 Feature Matching:

The **DTW** algorithm- **Dynamic Time Wrapping** is utilized in speech recognition to calculate the distance between speech templates. This is Dynamic Programming technique that calculates the resemblance between two time series. It is necessary because during the time of speech recognition, reference templates can be generated while training the extracted speech. It implements non-linear warping along the temporal axis by varying in size. The warping then obtained is used to measure the similarity between the time series with the implementation methods such as Correlation and Euclidean distance. After classification, the polarity is determined as negative, neutral or positive as it is shown in the figure 1.

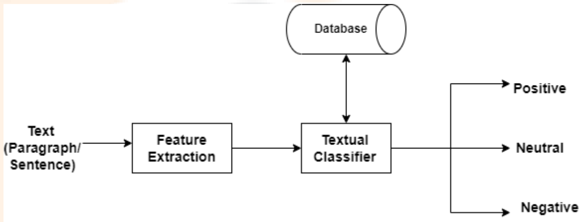

Fig.1.Architecture of Analysis system of Generic Sentiment

3. LITERATURE SURVEY

Tripathi, Anjali, et al.- This paper gives a epitomized review of the ongoing updates in speech emotion recognition[8]. The selection of databases was a significant factor in the excellent quality of the outcome in field of emotion recognition. The most popular machine learning methods are support vector machines used for working Sentiment analysis issue. There are several problems that want to be solved like variety in emotion, recognition of spontaneous emotion and speaker recognition in case of simultaneous verbal exchange. Similarly research is expected to enhance accuracy and increment the wide variety of emotion in an input speech signal.

Wani, Taiba Majid, et al.- Speech databases providing the data for the teaching approach are used in this paper. After preprocessing, feature extraction is finished for the speech sign. Typically, the SER system makes use of features like formant frequencies, speech pace, spectral electricity of speech, and important frequencies. Emotions, traditional classifiers, and deep learning classifiers are all understood through the application of class algorithms[1]. In order to increase accuracy rates and ensure that the system finds the right set of functions, it requires more robust algorithms.

P. Zhou, G. Fortino and M. Chen,- This paper depicts the current Interface of Human Machine system employed according to the different styles of emotion recognition[2]. It provides overview of essential technologies required in the system and defines the emotion communication system. This work provides an emotion communication protocol, which offers a high-position dependable support for emotion dispatches, in order to meet the conditions of communication for both parties when the emotion is conveyed as a type of multimedia information. In the end, this research examines how a spoken emotion transmission is performed in real time and highlights the effectiveness of realizing the emotion communications.

Mehta, Pooja, and Sharnil Pandya- This paper primarily discusses the fundamentals of sentiment mining and its levels[13]. Numerous methods exist along with machine learning techniques available for identifying sentiment in content. Sentiment Analysis, using various classification methods, yields neutral, negative, and positive scores. The research indicates that machine learning techniques like SVM, neural networks and SVM are regarded as the standard learning techniques since they demonstrate the highest accuracy and can be considered as the baseline learning methods. Further research could explore the impact of different combinations of text data and other factors on prediction accuracy. It proposes the need for more future work to enhance performance measures.

de Lope, Javier, and Manuel Graña- In this paper, the research has revolved around SER. SER encompasses some wide range of local small datasets. Many of the most current datasets have not yet been used, and when new datasets are suggested, the previous databases become outdated. Additional validation and analysis are required, as well as an assessment of the real dataset that was utilised for validation and their sensitivity to data sampling. This research includes works that would permit to solve the lack of facts and discover ways to construct more efficient recognizers, which are capable of recognizing feelings throughout the range of functions of the various databases.

4. PROPOSED SYSTEM AND IMPLEMENTATION

4.1 Pre-processing:

First, the tokenization is performed. Tokenization involves breaking down raw text into smaller chunks known as tokens. The process aids in comprehending the evolving model of NLP. Stop words are eliminated from the vocabulary to reduce noise and the feature set's dimensionality [5]. Many APIs are available for applications including sentiment analysis, parts of speech tagging, categorization, and translation through Python's Textblob library [5]. The input analog signal undergoes conversion to format (digital) for handling. After that the digital signal is then undergoes more processing, applying filters to eliminate noise and flatten signals. This process enhances the energy of signals at high frequencies as depicted in Figure 2.

4.2 Speech Model:

In the speech model, the features are extracted using stemming as the technique which extracts the base word by eliminating prefixes and suffixes, thereby revealing the original form of a word. This approach reduces index size and enhances retrieval accuracy [3]. Thus, the features are extracted that can be further utilized for classification. Linear predictive coding (LPC), Mel-Frequency Cepstral Coefficients (MFCC), and dynamic time warping are used to extract features. Google Speech Recognition is utilized as a tool/API in this project. An audio feature extraction method called MFCC is used to extract speaker-specific speech parameters. MFCCs are more preferred because they are less susceptible to background noise and channel distortions. DTW is Dynamic Programming technique that calculates the resemblance between time series with different rates of change or duration.

4.3 Speech-to-Text Conversion:

Speech to text conversion is computational recognition and conversion of speech to text. This is done through Lemmatization. Similar to stemming, lemmatization produces a 'lemma' as output, representing a root word rather than a root stem that has been stemmed. After the process of lemmatization, a valid word conveying the same meaning is obtained. Unlike stemming, lemmatization consistently yields valid words [3]. Thus, the speech is then converted into the text. For speech-to-text conversion, a variety of techniques and algorithms are used, including Hidden Markov Model (HMM) and Cuckoo Search Optimisation (ANN)-based [6].

4.4 Sentiment Analysis:

Sentiment analysis serves the fundamental purpose of comprehending the emotion and opinions expressed in text. The quantification of sentiment analysis results in polarity, a value that can be either positive or negative [15]. The polarity value's sign allows us to deduce if the sentiment is neutral, negative, or positive[4]. Natural Language Processing is used to analyse the sentiment of documents that express opinions within the realm of Computer Science and Artificial Intelligence, deals with interactions between the computers and various human languages. Emotion polarity values [14] for every word in the text are provided by public libraries such as SentiWordNet as shown in given figure 2.

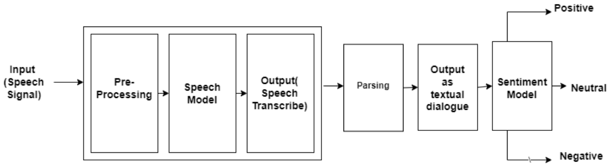

Fig 2: Block Diagram of Sentiment Analysis System

4.5 Flowchart

The flowchart represents the structure implementation of sentiment analysis. The audio input is directed to the Speech Recognition System, where it is segmented into chunks and stored inside the database.

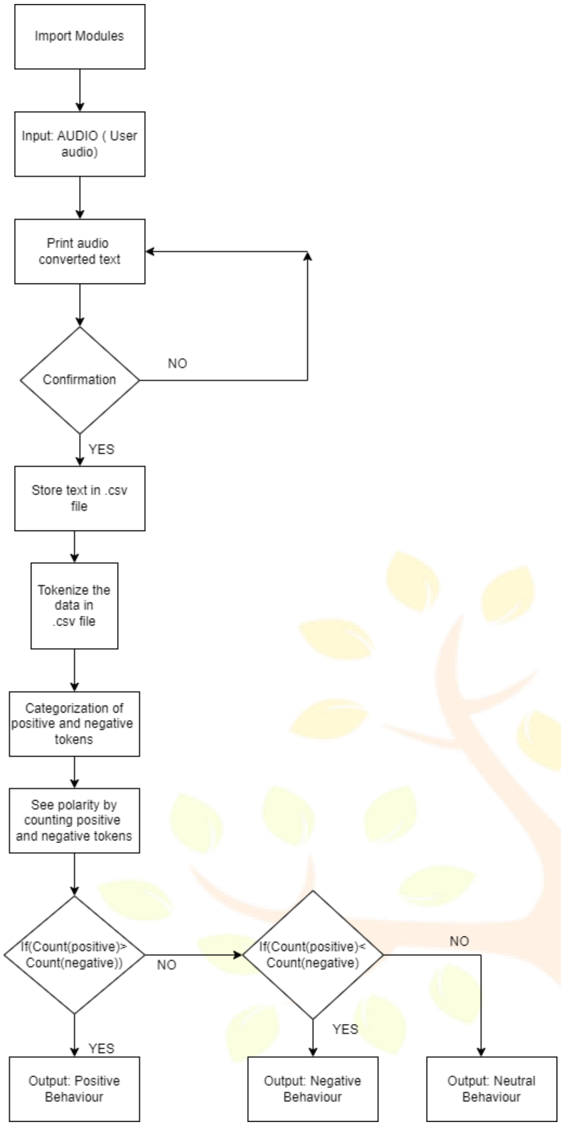
Import Modules
Input: AUDIO (User audio)
Print audio converted text
Confirmation
NO
YES
Store text in .csv file
Tokenize the data in .csv file
Categorization of positive and negative tokens
See polarity by counting positive and negative tokens
If(Count(positive)> Count(negative))
NO
If(Count(positive)< Count(negative)
NO
YES
YES
Output: Positive Behaviour
Output: Negative Behaviour
Output: Neutral Behaviour

Fig 3: Flowchart for the suggested sentiment analysis

After that, these broken chunks are processed and converted into speech from text. Then it is checked, if the speech is fully converted into text. Only after the confirmation the resulting text is saved in CSV file for subsequent sentiment analysis as shown in figure 3. The CSV file undergoes further processing, like tokenization, stemming and lemmatization to refine the data by removing words that aren't necessary [10]. The number of tokens in the sentence determines the polarity of the tokens, which are divided into neutral, negative, and positive attitudes. When there is a positive polarity, the output is positive behaviour. If the polarity is negative, the output is negative behaviour, else the output is neutral behaviour .

5. RESULT ANALYSIS

5.1. Speech Recognition:

Diverse inputs from different speakers are collected, capturing the pitch and loudness of their voices within a controlled environment devoid of external noise. The audio samples are subsequently transformed into text then.

5.2. Sentiment Analysis:

After being transcribed, the text is taken out of the database and put through sentiment analysis. In the process, subjectivity and polarity (expressing personal opinion, judgment, or emotion) are calculated, accompanied by the corresponding processing duration.

Table 1: Accuracy of Sentiments Evaluated

Emotion detected	Accuracy
Happy	86.42%
Disgust	77.77%
Calm	76.51%
Fearful	69.27%

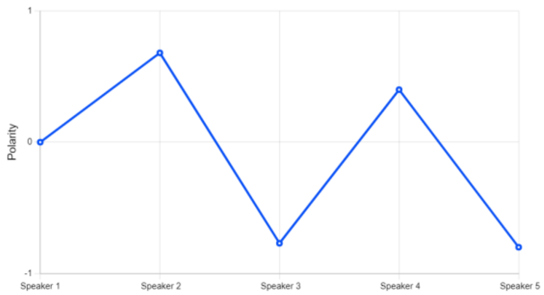

Fig 4: Polarity evaluated with different speakers

6. CONCLUSION

This study introduces a generalized system that utilizes audio input from users, translating spoken language into the written language automatically for analysis. The generated text undergoes sentiment analysis, wherein the polarity and nature of the speech are identified and sentences are tokenized. The model works well with real-time audio inputs, doing sentiment analysis and voice recognition with ease. However, ongoing efforts include the collection of expansive artificial datasets to enhance system efficiency. While the model adeptly converts speech to text, limitations include its capability to manage a single audio input at once and its inability to discern multiple speakers simultaneously. Although the system accurately analyzes the underlying sentiment in speech, opportunities for increased accuracy and customization persist. Future work will address these issues, focusing on refining accuracy, efficiency, and overall system performance.

7. REFERENCES

[1] Wani, Taiba Majid, et al. "A comprehensive review of speech emotion recognition systems." IEEE access 9 (2021): 47795-47814.

[2] M. Chen, P. Zhou and G. Fortino, "Emotion communication system", IEEE Access, vol. 5, pp. 326-337, 2016.

[3] Srinivas Chakravarthy Tokenization for Natural Language processing. Towardsdatascience.com (2020, June)

[4] Natural language toolkit tutorial: stemming and lemmatization. Tutorialspoint.com (2019)

[5] Dipanjan Sarkar emotion sentiment analysis practitioners guide NLP (2018, August)

[6] Ayushi Trivedi Navya Pant, Pinal Shah Simran Soni k and Supriya Agrawal Speech to text and text to speech recognition systems-A review, IOSR Journal of Computer Engineering (IOSR-JCE) (2018).

[7] Maghilnan S, Rajesh Kumar M Sentiment Analysis on Speaker Specific Speech Data (2017 I2C2).

[8] Tripathi, Anjali, et al. "A review on emotion detection and classification using speech." Proceedings of the international conference on innovative computing & communications (ICICC). 2020.

[9] 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. 7, pp. 117327–117345, 2019

[10] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem, ‘‘BAUM-1: A spontaneous audio-visual face database of affective and mental states,’’ IEEE Trans. Affect. Comput., vol. 8, no. 3, pp. 300–313, Jul. 2017

[11] Mustaqeem, M. Sajjad, and S. Kwon, ‘‘Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM,’’ IEEE Access, 2020

[12] M. Neumann and N. T. Vu, ‘‘Improving speech emotion recognition with unsupervised representation learning on unlabeled speech,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., May 2019

[13] Mehta, Pooja, and Sharnil Pandya. "A review on sentiment analysis methodologies, practices and applications." International Journal of Scientific and Technology Research 9.2 (2020): 601-609.

[14] Pang, B., & Lee, L. (2004, July). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd annual meeting on Association for Computational Linguistics (p. 271). Association for Computational Linguistics.

[15] Pang, B., & Lee, L. (2005, June). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Proceedings of the 43rd annual meeting on association for

computational linguistics (pp. 115-124). Association for Computational Linguistics.