

## SPEECH TO TEXT CONVERSION AND SENTIMENT ANALYSIS ON SPEAKER SPECIFIC DATA

Nikhil Jain\*<sup>1</sup>, Manya Goyal\*<sup>2</sup>, Agravi Gupta\*<sup>3</sup>, Vivek Kumar\*<sup>4</sup>

\*<sup>1,2,3,4</sup>Department Of Computer Science And Engineering,

Meerut Institute Of Engineering And Technology, Meerut, India.

### ABSTRACT

Sentiment analysis (SA) is an evolving technique and many people have worked in this field in past years. Most of the work is based on the polarities sentiments through speech recognition and/or audio recognition. In this research paper we tried to work on advance level specifying the actual emotion of speaker through various efficient algorithms. We have also further tried to discriminate the various speakers.

**Keywords:** Speech Recognition, Sentiment Analysis , Natural Language Processor.

### I. INTRODUCTION

Speech recognition (SR) is the technique to identify what the speaker is saying and convert it into text using various speech recognition and conversion algorithms. It is a basic application of machine learning. The text can further be used for various analysis and uses. In this we have used Google Speech Recognition API because of its efficiency.

Speaker recognition is task where we identify different users based on their voice modulation, speech, pitch etc. and store the data. People in research community have worked to transcribe the speeches into text with minimum human help that is typing the speech as we listen to it. People have worked to convert audio, podcasts, debates, speeches, song to text. And the community also worked on audio analysis investigation [1, 2, and 3] to learn and study telephone conversations involving more than one speaker. [4]

A lot of work has used the concepts of neural networks are used in speech recognition like recurrent neural network (RNN), Natural Language Processor (NLP) sequence models etc.

Sentiment Analysis (SA) is the method to identify the speaker's emotion by identification and vectorization of the data set in the text. Overall, Sentiment analysis may involve the following types of classification algorithms: Linear Regression. Naive Bayes. Support Vector Machines.

The basic idea behind this to extract various words from the text and match it with data set and identify the emotion like happy, sad, joy, anger, frustration, fear etc.

### II. PROPOSED SYSTEM

In the past there been not much work in the field of speech to text conversion and sentiment analysis together rather they were operated as two separate modules. Furthermore, the work in sentiment analysis was almost constricted to movie review and twitter sentiment analysis. Here we have combined the both and tried to expand the horizon of sentiment analysis.

#### a. Tokenization:

Tokenization refers to splitting or breaking the raw text into smaller pieces called units or chunks. This breaks the text into tokens which are helpful to understand the developing model of the NLP. Stop words are removed from the vocabulary to reduce noise and to reduce the dimension of the feature set[5]. Textblob is a library in Python which can be used for processing of textual data .It provides various APIs for sentiment analysis, parts of speech tagging, classification, and translation and so on. [6]

#### b. Stemming

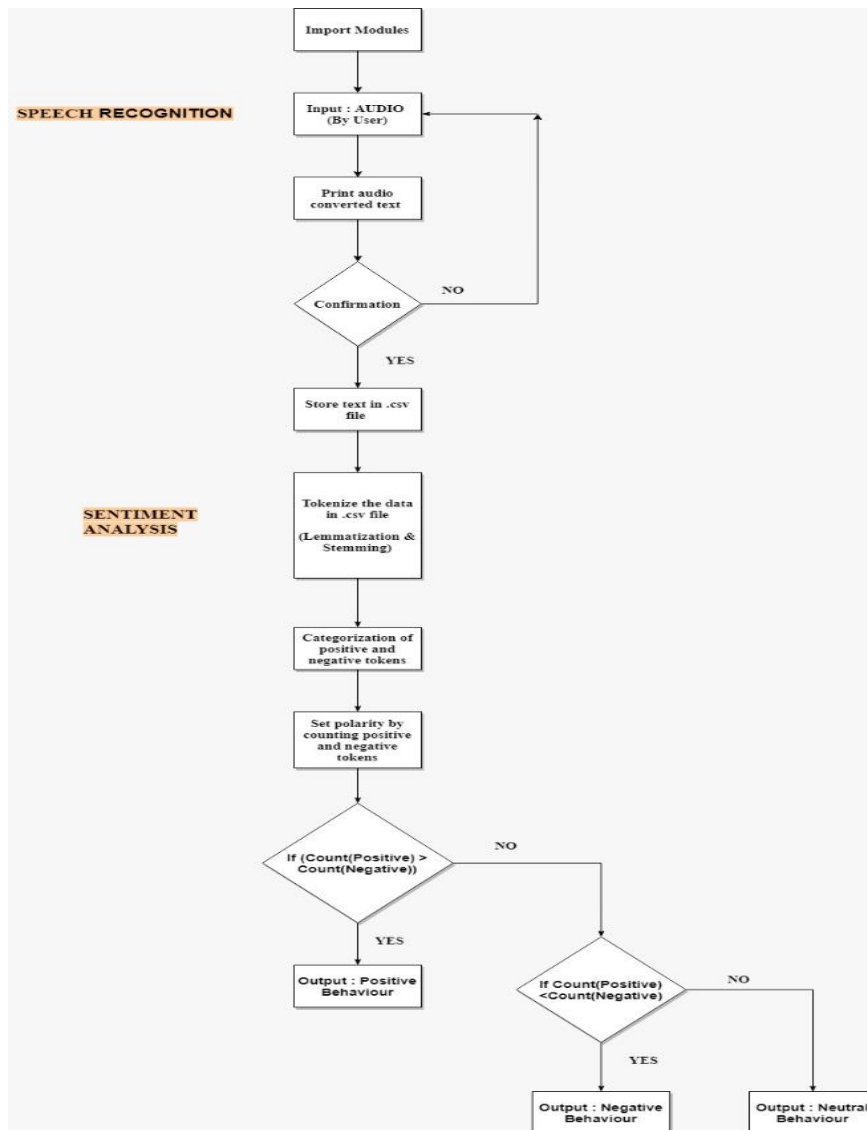
Stemming is a technique that extracts the original form of the word that is the base word by removing prefixes and suffixes from them. .In this way, stemming reduces the size of the index and increases retrieval accuracy [5].

**c. Lemmatization**

Lemmatization technique is like stemming. The output we will get after lemmatization is called 'lemma', which is a root word rather than root stem, the output of stemming. After lemmatization, we will be getting a valid word that means the same thing. It means after applying lemmatization, we will always get a valid word. [7]

**d. Polarity**

The basic and foremost use of sentiment analysis is to find/ analyse the text and understand the emotion and opinions expressed by it. Hence, we quantify the sentiment analysed as positive or negative value known as polarity. We can infer the overall sentiment as a positive, negative or neutral opinion from the sign of the polarity value. [8]



**Fig 1:** Flowchart for the proposed sentiment analysis

**III. DESIGN AND IMPLEMENTATION**

The input voice signal is directed to Speech Recognition System, this then separates them into chunks and then these chunks are stored in a database, these chunks are then sent to a speech recognition algorithm API which recognises the speech and transcribes it into text. The text is saved in a csv file which will then be processed further for sentiment analysis. The data in the csv file is then processed and vectorization is performed. After vectorization the data is then tokenised and additional NLTK, NLP methods are used like lemmatization and stemming to further quantify the data and remove the unnecessary words. The tokens are now categorised into positive and negative, and then further polarity is set to count the number of tokens in the sentence and thus the polarity is displayed with subjectivity and sentiment.

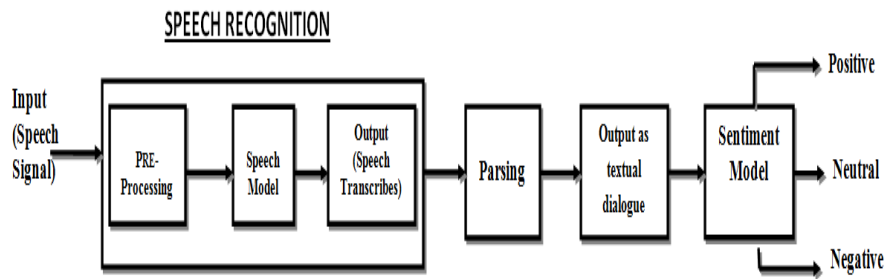


Fig 2: block diagram of the system

**3.1 Pre-processing :** The input analog signal is transformed into digital for processing. The obtained digital signal is then sent to further processing to apply filters to flatten the signals and remove the noise from them. Thus the energy of signals is increased at high frequencies.

**3.2 Speech Model :** Features are extracted through Dynamic Time Warping, Mel-Frequency Cestrum Coefficient (MFCC), Linear Predictive Coding (LPC). Various tools and/or API used. In this project we have used Google Speech Recognition.

**3.3 Speech to text conversion :** It uses many methods/algorithm such as: Artificial Neural Network Classifier (ANN) based Cuckoo Search Optimization [9] and Hidden Markov Model (HMM).

**3.4 Sentiment Analysis :** The approach used for sentiment analysis of opinionated texts is using Natural Language Processing (NLP). In Computer Science and Artificial Intelligence, NLP is concerned with all the interaction between computer and different human languages. There are many public libraries such as SentiWordNet which provides sentiment polarity values of each word in the text.

#### IV. MODULE DESCRIPTION

The proposed system uses speech recognition and sentiment analysis. A detailed analysis for the experiments performed with various algorithms and tools is presented in this research. **Google Speech API, Speech\_Recognition, and pyttsx3 tools** are used for speech recognition and for sentiment analysis **TextBlob** tool is used.

**4.1. Speech Recognition Module :** For initializing the recognizer **Recogniser()** module is used. Then for inputting the audio use of microphone is done through **Microphone()** module. The surrounding noises and sounds level energy threshold is adjusted by **adjust\_for\_ambient\_noise(source, duration)** module. For recognizing audio through Google. **recognize\_google(audio)** is used.

**4.2. Sentiment Analysis Module:** For sentiment analysis **TextBlob, TextBlob. sentiment(), TextBlob. sentiment. polarity()** library functions are used.

**4.3. Details of Experimental Tools:**

##### 4.3.1. Google Search API

The Google Custom Search API allows you to leverage Google's search engine Technology and build powerful search capabilities for your websites and applications.

##### 4.3.2. Python

The first software requirement is Python 2.6, 2.7, or Python 3.3+. This is required to use the library.

##### 4.3.3. Python Speech Recognition module

pip install speechrecognition

Linux users can use the following command :

sudo apt-get install python3-pyaudio

PyAudio can also be installed on windows by users by executing the following command in the command prompt or a terminal

pip install PyAudio

##### 4.3.4. pyttsx3

Python pyttsx3 module:

pip install pyttsx3.

pyttsx3 is a library in Python which is used for text-to-speech conversion. It is compatible with both Python 2 and 3, and it can also work in offline mode .

#### 4.3.5. PyAudio(for Microphone)

PyAudio is necessary for taking an audio input from microphones. Microphone will be undefined if PyAudio is not installed, but the library will still work,

#### 4.3.6. TextBlob

TextBlob is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

## V. DESIGN OF EMOTION IDENTIFIER

The emotion identifier module works on extraction of approximately correct emotion hidden behind the speaker's input sentences .It tries to find tokens related to the words in the speech . Below are the steps for the process of finding the emotion hidden in the speech :

STEP 1 : Making of data set in a .txt file format and maintaining key value word meaning pairs which can be further used to find the emotion of user .

Example :

```
'victimized': 'cheated',  
'accused': 'cheated',  
'acquitted': 'singled out',  
'adorable': 'loved',  
'adored': 'loved'
```

STEP 2 : Extraction of the given text in lower cases and cleansing of text like removing the punctuations and maintaining a tokenised words array using the module `NLTK.word_tokenize()` .

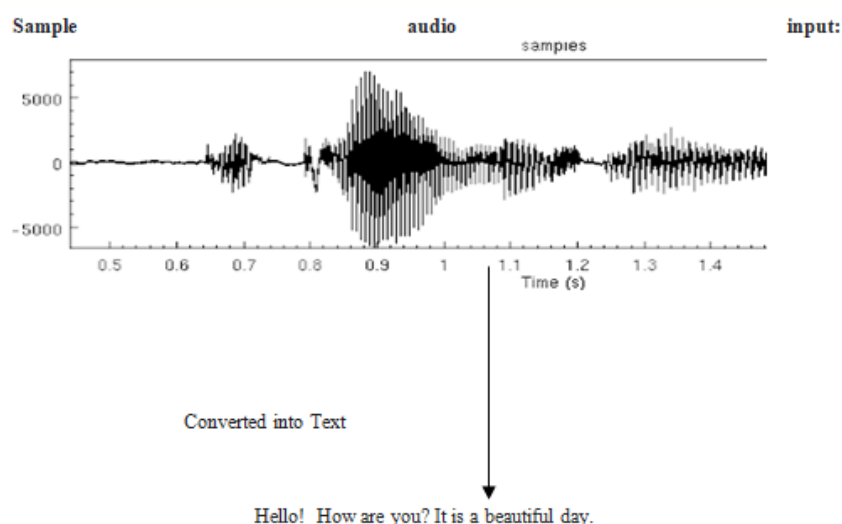
STEP 3 : Downloading of stop words from the natural language toolkit using the module `nltk.download('stopwords')` and Removal of stop words from the text .

STEP 4 : Open the emotions.txt file format in read mode and strip the data line by line and remove extra spaces .

with `open('emotion.txt','r')` as file:

STEP 5 : Check each of the words in the emotions.txt file line by line and compare with the actual text , if the words match then append them to an emotions list which will be printed as per the sentences .






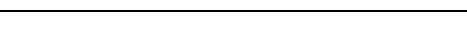
## VI. RESULT ANALYSIS



**6.1. Speech Recognition:**

Inputs from various speakers are taken as shown below. The audio sample shows the pitch, loudness of their voice taken in controlled speakers environment free from outside noise. The samples are further converted into text as shown in the following table along with the time the speaker took to enunciate the sentence (refer Fig 3).

**Table 1: Audio Input**

Speakers	Input : Audio	Output : Text	Time (seconds)
Speaker 1		Hello! How are you?	2.460
Speaker 2		This dress is so beautiful.	2.929
Speaker 3		I am in a bad mood today.	2.820
Speaker 4		She was overwhelmed with joy after she heard the good news.	2.335
Speaker 5		She was grief stricken when she heard the news of her friends' demise.	3.510
Speaker 6		If you take a cruise to Alaska aboard Holland America, you'll stop in Victoria, British Columbia.	4.283

**6.2. Sentiment Analysis:** The transcribed text is fetched from database and sent for sentiment analysis where its polarity and subjectivity (personal opinion, judgement or emotion) are calculated along with the processing time (refer Fig 4).

**Table 2: Polarity and subjectivity**

S No.	Input : Text	Polarity	Subjectivity	Time (seconds)
1.	Hello! How are you?	0.00	0.00	0.004889
2.	This dress is so beautiful.	0.85	1.00	0.008956
3.	I am in a bad mood today.	-0.69	0.66	0.442195
4.	She was overwhelmed with joy after she heard the good news.	0.20	0.20	0.011965

5.	She was grief stricken when she heard the news of her friends' demise.	-0.80	0.20	0.010980
6.	If you take a cruise to Alaska aboard Holland America, you'll stop in Victoria, British Columbia.	0.00	0.00	0.004067

**6.3. Final Result:** The output of the whole experiment is depicted along with the overall sentiment of the inputs provided by the various speakers. The sentiments depicted are positive, negative and neutral in these cases.

**Table 3:** Overall sentiment of the speaker

S. No.	Input : Audio	Text	Sentiment	Emotions
1.		Hello! How are you?	Neutral	[ ]
2.		This dress is so beautiful.	Positive	[pleasant]
3.		I am in a bad mood today.	Negative	[sad]
4.		She was overwhelmed with joy after she heard the good news.	Positive	[happy]
5.		She was grief stricken when she heard the news of her friends' demise.	Negative	[ ]
6.		If you take a cruise to Alaska aboard Holland America, you'll stop in Victoria, British Columbia.	Neutral	[happy]

**6.4. Performance Evaluation :**

The study of above sample data is taken into consideration for calculating accuracy of the algorithm by finding precision, recall, and f-measure for the complete performance evaluation.

Data input size = 6

C = 4 [Speakers 1, 2, 3, 4],

W = 1 [Speaker 5],

M = 1 [Speaker 6]

where C -> Correct Result

W -> Wrong Result

M -> Result not identified

**Precision :**

$$P = C / C+W$$

As per above data shown,

$$P = 4 / 4+1$$

$$P = 4 / 5$$

$$P = 0.8$$

**Recall :**

$$R = W / C+W$$

As per above data shown,

$$R = 1 / 4+1$$

$$R = 1 / 5$$

$$R = 0.2$$

**F-measure :**

$$F = 2PR / P+R$$

As per above data shown,

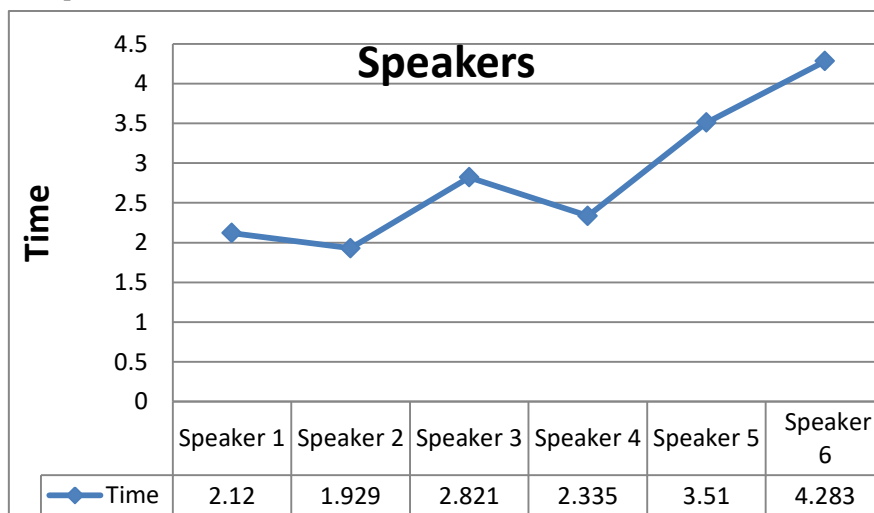
$$F = 2 * 0.8 * 0.2 / 0.8 + 0.2$$

$$F = 0.32 / 1.0$$

$$F = 0.32$$

Note - This value of f-measure is just for a single observation table. When calculated for larger data input size then the f-measure value will be approximately near 1.0 which is considered as a good f-measure.

**6.5. Graphical Representation :**



**Fig 3:** Time Variation in the results according to different inputs



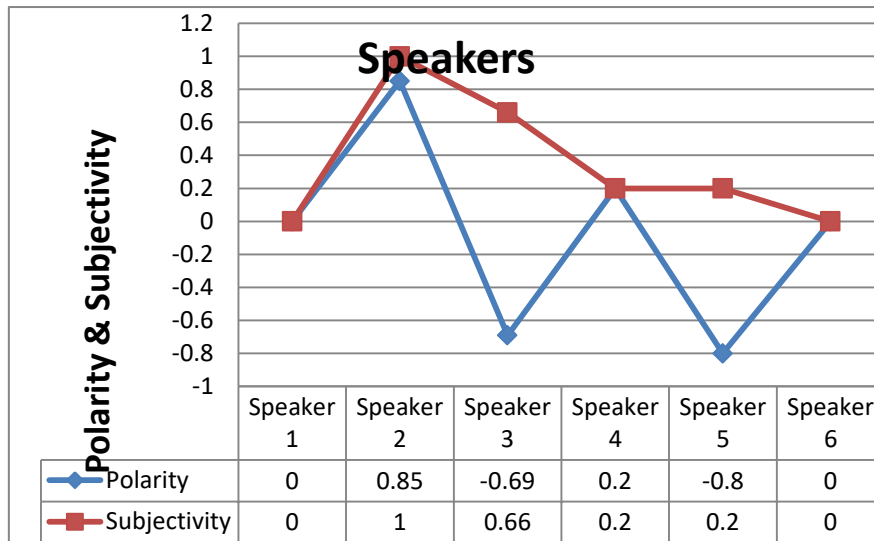


Fig 4: Variation and Contrast in the Polarity and Subjectivity for different inputs

### VII. CONCLUSION

This research puts forward a generalised system that uses an audio input from a user and studies its contents of the spoken words by converting the speech into text automatically. Further the text is directed towards the sentiment analysis model where the sentences are tokenized and accordingly the polarity and nature of the speech is depicted. The model works well with live generated audio inputs and successfully performs speech recognition as well as sentiment analysis on the input but we are working on collecting artificial datasets which are large in size and increasing the efficiency of the system. Though the model works fine and accurately in executing text conversion on the input speech by a user, it still have some drawbacks, the system can smoothly handle only one audio input at a time and cannot understand if two or more speakers speak at the same time. The system is able to analyse the type of thought behind a given speech but its accuracy can still be further increased and more customization can be done. We plan to address these issues in our future work and enhance the accuracy, efficiency and performance of the system.

### VIII. REFERENCES

- [1] 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.
- [2] 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.
- [3] Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics.
- [4] Maghilnan S, Rajesh Kumar M (2017 I2C2). Sentiment Analysis on Speaker Specific Speech Data
- [5] NLTK documentation: <https://www.nltk.org/>
- [6] Srinivas Chakravarthy (2020, June) Tokenization for Natural Language processing. Towardsdatascience.com
- [7] Natural language toolkit tutorial(2019) :stemming and lemmatization. Tutorialspoint.com
- [8] Dipanjan Sarkar (2018, August)emotion sentiment analysis practitioners guide nlp.
- [9] Ayushi Trivedi,Navya Pant, Pinal Shah,Simran Sonik and Supriya Agrawal (2018). Speech to text and text to speech recognition systems-Areview, IOSR Journal of Computer Engineering (IOSR-JCE)