International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Emotion Recognition in Speech Processing using Fast Fourier Transform

Veerendra Kumar Jammula, Ashok Reddy Gogireddy, Hari Kiran Vege, Kolla Bhanu Prakasha

***Abstract*: The idea of acquiring the state of human emotions from one's speech, we have gathered required data that makes one to understand the concept behind this process. Human emotions can be predicted by his/her facial expressions or by the tone of their voice. Reading the facial expressions is one of the major tasks involved in image processing.**

Likewise, each emotion holds different tone in one's voice. It requires a various emotional tone frequency to calculate and analyse the emotions. We need to fetch approximate frequencies of emotions. It's the challenging task as each speaker has various pitches while speaking and frequencies of the same person varies in his emotion. Another main issue is the noise in the input while a person is speaking, due to less quality recordings or surrounding environment. List of basic emotions are Happy, angry, sad, bored, surprised, disgust, fear.

For this project the prior important concept is speech recognition. The machine must be capable of reading the input in form as speech and must be capable of analysing various contents. The input given is converted into wav format. At the same time machine must be also capable of fetching the frequencies. The calculation is performed using many methodologies that are defined.

Keywords: - Emotions, Emotion Recognition, FFT, Frame rate, Frequency.

I. INTRODUCTION

The process of acquiring sound from an external source as an input in required format is termed as Speech Recognition.[1] Words can be a sentence, command or data entry. [2] It serves as an input to the linguistic process.[3] Speech is the sound that contains an idea or information transferred from human source.[4]

An Automatic Speech Recognition (ASR) considers the useful data that exist in speech signal. [5] In noisy environment Speech recognition is not accurate.[6] One of the most efficient ways to express ourselves is through speaking.[7]

A waveform is used to represent these speech signals which use the short term amplitude spectrum Speech signal, the basic difficulty in speech recognition is the one which varies with the speakers and their rates, content and environment.[8]

A. Speech Emotion

The latest challenge in Speech processing module is Emotion recognition. [9]

Human emotion can be detected through one's speech. [10] Each emotion has a specific tone while speaking, automatic recognition of the emotion is the latest challenge. [11] Emotion recognition is difficult because of the following reasons:

1.Speaker might not express or speak clearly.
2.Noisy environment.
3.Emotion expressing varies with his or her culture & language.
4.Each speaker has his own style of speech and speaking rates.

Therefore, it's difficult to differentiate these portions of utterance. [12] Emotions are classified into two type's long term emotion and transient, this is another problem for the recognizer which does not get a clear picture of emotion. [13]

The evaluation depends upon the level of intensity of the sound waves, which is an input to the system. [14] The speech input given to the system may be real world emotions or acted. [15] It is mostly suggested to use real life situations. [16]

The SER system consists of five main modules of emotional speech

1. Input
2. Feature extraction
3. Feature selection
4. Classification
5. Recognized emotional output

II. EMOTION RECOGNITION

Emotion Recognition through the speech input is used for classification of speech into seven emotions. [17] The speech is taken as recorded one. [18] The model has trained info of speech found in English and wanted database languages. [19] These are compared through Multivariable linear regression classifier. [20] The speech is classified into frequency cepstrum coefficients and modulation spectral features. [21] Emotion recognition through speech is the topic of Human and Computer Interaction and Natural language processing. [22]

Emotion Recognition through speech is extraction of the emotions and feature details of the human emotions.

Manuscript published on November 30, 2019.
* Correspondence Author
Veerendra Kumar Jammula, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, A.P., India.
Ashok Reddy gogireddy, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, A.P., India.
Hari Kiran Vege, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, A.P., India.
Kolla Bhanu Prakash, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, A.P., India.

The linear prediction cepstrum-coefficients,
Mel-Frequency spectral features, energy, pitch, format frequency many research companies are used, those features have the main objects for Emotion Recognition through speech. [23]
The Databases used for the emotional classification are berlin database and Spanish Database. [24]

A. Types of Emotions

1.Anger,
2.Happy,
3.Sadness,
4.Boredom,
5.Neutral,
6.Fear/Scary,
7. Surprised.

B. Classification

The complex task which includes many factors, like distinguishing among emotions where one emotion might start and some other emotion might end. [25]

Robert Plutchik created a new conception of emotions, naming the "wheel of emotions" as it can demonstrate how different emotions can blend into one another and create new emotions. [26] He suggested 8 primary bipolar emotions. [27]

III. FEATURE EXTRACTION

In Emotion recognition through speech, the feature extraction process is the main algorithm in relation between mp3 and wave length. [28] These forms are changing the details from the wave form of extraction for finding the frequency, amplitude to find out the emotion. [29] Every emotion has a certain frequency, certain amplitude and certain pitch. [30] We can find out the Emotion recognition through speech in two ways. They are as follows:

1) Finding out the frequency from the wavelength and frequency is the match condition according to the basic frequency that the basic emotions have. [31]
2) Finding out the pitch which link through the frequency. [32] The relation between pitch and frequency is

$$p=69+12\log(\text{base } 2, \text{power}(f/440))$$

Eq. (III.1)

Note: The frequency is calculated with the basic frequency of 440Hz.

3) The data set which are classified with gender and different emotions

Quick Fourier Transform (FFT) Algorithm, which is consistently viewed as one of the most fundamental numerical calculations of the twentieth century, and method Waveform Audio Files (WAV) containing uncompressed sound encoded the utilization of a straight heartbeat code balance (LPCM) format. [33]
Sound voyages inlongitudinal waves which is a swaying of weight. [34] An individual wave is alwaysdetected by utilizing its length (the separation in time between two high focuses) and adequacy (the total separation vertically from the most elevated point to the absolute bottom). [35] The quality of the wave or its"loudness" can be spoken to by method for its adequacy. [36]

Human ear translates a sound wave by utilizing changing over it into a melodic pitch (or note). [37] Each melodic watch relates to an exact recurrence which is estimated in hertz, or the scope of entire cycles every second of an occasional marvel (for this situation, the sound wave). [38]
Trigonometric Sine work is utilized to symbolize sound information. In time arrangement, t speaks to the time unit, which ascends from zero f speaks to the recurrence. Variablea speaks to a scaling thing (somewhere in the range of zero and 1) to practice to the plentifulness of the wave. Cradle joins the arrangement of 44,100 gliding point esteems speaking to the wave structure. On every cycle, time speaks to the t-organize for which the wave structure is figured as

Eq. (III.2) a*Sin(2*PI*f*t).

When the support is processed, it should be changed over into its relating byte encoding. For 8-piece sound quality, there are 256 elite qualities that can be produced. Normally, 16-piece sound extraordinary can more prominent precisely mannequin a sound wave structure because of the reality it takes into account a total of 65,536 unique qualities.
We are using the first method which gives detailing feature of the speech to wave forms. Through that wave length extraction of frequency and amplitude is shown in graph. To show how the speech mostly goes on. frequency is calculated. That frequency calculation is verified and gives final output as the Emotion.

A. Waveforms

Waveforms are the main aspect in this process of emotion recognition through speech. Because the sound travels in the form of waves.
Through that only the sound travels change according to medium. Through that the speech is changed into the.wav forms. That wav form can predict the frequency and amplitude. The waveforms and frequency are directly related with one and another.

B. Frequency

Frequency is finding the numerical relation between the two frames of the wave in the certain period.

B.A Frequency in the form of wavelength

Frequency is normally in the form of numerical value but for most of the conditions it is represented in the form of wavelength, because it occurs from the waveform in between the time period of the time. The waveform of the frequency gives a clear view of the speech is transformation.

B.B Frequency through amplitude

Frequency is also calculated from amplitude.The amplitude and frequency are directly related. The amplitude is also a graphical representation. Frequency calculation is mainly through the amplitude only for long speech waves.

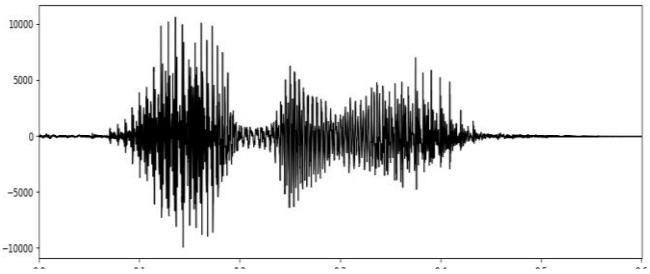

Fig 1: Amplitude

B.C. Average frequency calculation through framerate.

Frequencies may vary from one point to another. So, to find out perfect emotion, the average frequency has to be calculated. The average frequency is calculated by finding out the frame rate.
Frame rate is defined as the frequency for every sound wave or every frame i.e. from starting point to ending point, every millisecond output is calculated and gives
the details in the form of array, With the basic frequency as 440Hz.

C. Emotion extraction

The emotion is extracted through the frequency obtained from the framerate classification and the average frequency is calculated from that framerate.
To calculate the average frequency in python there is function call,.Wav file. Is written as
(".wav", sample rate, basic frequency)

Table 1 Emotion & speech parameters (from Murray and Arnott)

	Anger	Happiness	Sadness	Fear	Disgust
Rate	Slightly faster	Faster or slower	Slightly slower	Much faster	Very much faster
Pitch Average	Very much higher	Much higher	Slightly lower	Very much higher	Very much lower
Pitch Range	Much wider	Much wider	Slightly narrower	Much wider	Slightly wider
Intensity	Higher	Higher	Lower	Normal	Lower
Voice Quality	Breathy, chest	Breathy, blaring tone	Resonant	Irregular voicing	Grumble chest tone
Pitch Changes	Abrupt on stressed	Smooth, upward inflections	Downward inflections	Normal	Wide, downward terminal inflections
Articulation	Tense	Normal	Slurring	Precise	Normal

IV. RESULTS AND DISCUSSION

A. Emotion recognition process through speech

It deals with speech modulation for every change in speech to frequency.
It makes use of tensor flow. The packages used are OS path, Pydub - audio files recognition, matplot - graphs and scipy - wav forms.
Steps for the emotion recognition process through speech:

1) Mp3 to wave length.
2) Wavelength frequency.
3) Fast Fourier Transform.
4) Finding out Amplitude frequency.
5) Framerate Extraction.
6) Average frequency.
7) Distinguishing the Emotion using emotional frequency.

B. Mp3 to wavelength

The basic speech is recorded in mp3 format. Frequency can be calculated using .wav files, so mp3 is converted into that format.

C. Wavelength frequency

Every speech holds a frequency upon there emotion and in the way they speak. It might include some noise. We differentiate noise from exact speech. Result graph is shown.

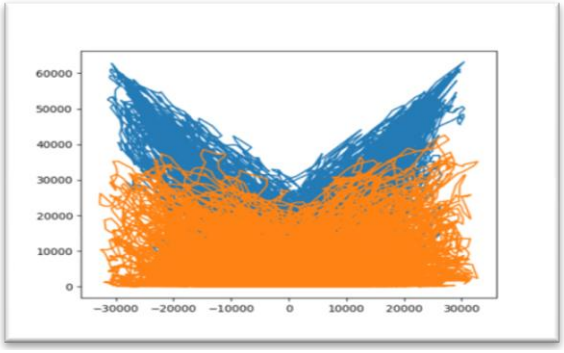

Fig 2: noise in the input

D. Fast Fourier Transform

The algorithm used in the calculation of the Discrete Fourier Transform of an arrangement. It changes over a space or time sign to indication of the recurrence domain.The DFT signal is produced by methods for the conveyance of value arrangements to one of a kind recurrence segment.
Working straightforwardly to change over on Fourier truly change is computationally excessively costly. Along these lines, Fast Fourier change is utilized as it quickly figures through factorizing the DFT lattice as the result of meager elements.

$$F(x) = \int_{-\infty}^{\infty} f(x)e^{-x}\,dt$$

Eq. (D.1)

Subsequently it decreases the DFT calculation intricacy from O(n2) to O(N log N). Furthermore, this is a major qualification when taking a shot at a goliath dataset. Likewise, FFT calculations are exceptionally right as opposed to the DFT definition straightforwardly, within the sight of adjust blunder.
This change is an interpretation from the setup space to recurrence house and this is significant in expressions of investigating every change of clamor and bona fide substance material for increasingly proficient calculation and in investigating the quality range of a sign. This interpretation can be from xn to Xk.
It is changing over spatial or fleeting measurements into the frequencydomain.

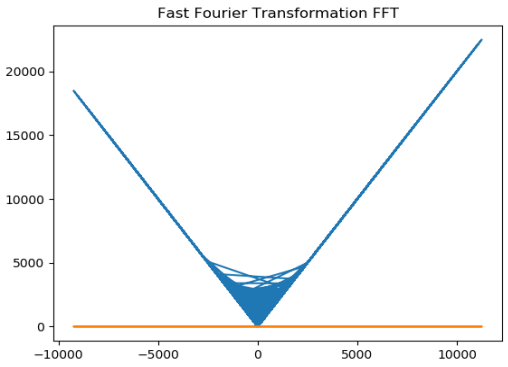

Fig 3 - FFT on audio file (.wav)

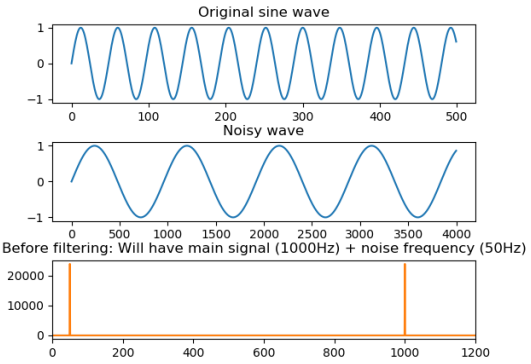

Fig 4 - Representation of Sine waves, noise in the audio

E. Amplitude wavelength

The frequency and amplitude makes difference in the speech recognition. $A=y(t)\sin(2\pi ft+\phi)$ is the relation between the amplitude and frequency.

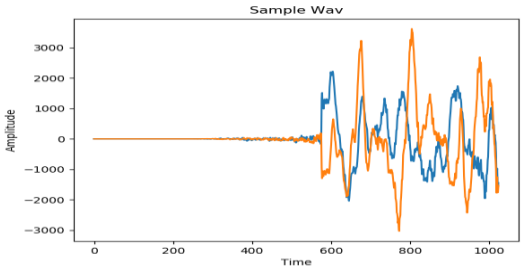

Figure 5: Amplitude with time

F. Framerate extraction

Framerate extraction means finding the frequency rate for every minor detail of the sound, passing by the time to time in an array type, which results in frequency modulation (440hz). First it gives 44100 as of 440Hz and the array shows the details of the framerate.

V. CONCLUSION

In this process,DFT is very slow compared to FFT. The FFT returns all possible frequencies in the signal. And the way it returns is that each index contains a frequency element, such that it can overcome the noise and its more efficient than DFT. During implementation SourceDataLine is used for input as it can acquire large dataset.

REFERENCES

1. Discriminating Emotions in the Valence Dimension from Speech Using Timbre, Features by Anvarjon Tursunov [1], Soonil Kwon [1,*] and Hee-Suk Pang [2]
2. Speech Emotion Recognition Using CNN Zhengwei Huangy, Ming Dongz, Qirong Maoy, Yongzhao Zhany.
3. Speech Emotion Recognition using Convolutional and Recurrent Neural Networks Wootaek Lim, Daeyoung Jang and Taejin Lee.
4. PYAUDIOANALYSIS: AN OPEN-SOURCE PYTHON LIBRARY FOR AUDIO SIGNAL ANALYSIS.
5. Feature extraction of speech signals in emotion identification. M. Morales-Perez ; J. Echeverry-Correa ; A. Orozco-Gutierrez ; G. Castellanos-Dominguez
6. FEATURES EXTRACTION FOR SPEECH EMOTION KAMARUDDIN, NORHASLINDA | WAHAB, ABDUL; *
7. Hidden Markov model-based speech emotion recognition B. Schuller ; G. Rigoll ; M. Lang
8. Speech emotion recognition based on HMM and SVM Yi-Lin Lin ; Gang Wei
9. EMOTION RECOGNITION THROUGH SPEECH USING NEURAL NETWORK PAWAN MISHRA ARTI RAWAT
10. SPEECH EMOTION RECOGNITION FOR PERFORMANCE INTERACTION NIKOLAOS VRYZAS RIGAS KOTSAKIS AIKATERINI LIATSOU CHARALAMPOS DIMOULAS
11. EMOTION DETECTION FROM SPEECH COMPUTER SCIENCE TRIPOS PART II GONVILLE & CAIUS COLLEGE 2009-2010.
12. Kolla, B.P., Dorairangaswamy, M.A. & Rajaraman, A. 2010, "A neuron model for documents containing multilingual Indian texts", 2010 International Conference on Computer and Communication Technology, ICCCT-2010, pp. 451.
13. Kolla, B.P. & Raman, A.R. 2019, "Data Engineered Content Extraction Studies for Indian Web Pages, Advances in Intelligent Systems and Computing, Volume 711, 2019, Pages 505-512.
14. Naga Pawan, Y.V.R. & Prakash, K.B. 2019, "Variants of particle swarm optimization and onus of acceleration coefficients", International Journal of Engineering and Advanced Technology, vol. 8, no. 5, pp. 1527-1538.
15. T. Vijay Muni, G Sai Sri Vidya, N Rini Susan, "Dynamic Modeling of Hybrid Power System with MPPT under Fast Varying of Solar Radiation", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 1 (2017), pp.:530-537.
16. M Srikanth, T. Vijay Muni, M Vishnu Vardhan, D Somesh, "Design and Simulation of PV-Wind Hybrid Energy System", Jour of Adv Research in Dynamical & Control Systems, Vol. 10, 04-Special Issue, 2018, pp: 999-1005
17. S Ilahi, M Ramaiah, T Vijay Muni, K Naidu, " Study the Performance of Solar PV Array under Partial Shadow using DC- DC Converter", Jour of Adv Research in Dynamical & Control Systems, Vol. 10, 04-Special Issue, 2018, pp: 1006-1014.
18. S Moulali, T Vijay Muni, Y Balasubrahmanyam, S Kesav,"A Flying Capacitor Multilevel Topology for PV System with APOD and POD Pulse Width Modulation", Jour of Adv Research in Dynamical & Control Systems, Vol. 10, 02-Special Issue, 2018, pp: 96-101.
19. Prakash, K.B. 2018, "Information extraction in current Indian web documents", International Journal of Engineering and Technology(UAE), vol. 7, no. 2, pp. 68-71.
20. Prakash, K.B. 2017, "Content extraction studies using total distance algorithm", Proceedings of the 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology, iCATccT 2016, pp. 673.
21. Prakash, K.B. 2015, "Mining issues in traditional indian web documents", Indian Journal of Science and Technology, vol. 8, no. 32, pp. 1-11.
22. Prakash, K.B., Ananthan, T.V. & Rajavarman, V.N. 2014, "Neural network framework for multilingual web documents", Proceedings of 2014 International Conference on Contemporary Computing and Informatics, IC3I 2014, pp. 392.
23. Prakash, K.B. & Dorai Rangaswamy, M.A. 2019, "Content extraction studies for multilingual unstructured web documents", Advances in Intelligent Systems and Computing, 749, pp. 653-664.
24. Prakash, K.B. & Dorai Rangaswamy, M.A. 2016, "Content extraction studies using neural network and attribute generation", Indian Journal of Science and Technology, vol. 9, no. 22, pp. 1-10.
25. Prakash, K.B., Dorai Rangaswamy, M.A. & Ananthan, T.V. 2014, "Feature extraction studies in a heterogeneous web world", International Journal of Applied Engineering Research, vol. 9, no. 22, pp. 16571-16579.
26. Prakash, K.B., Dorai Rangaswamy, M.A., Ananthan, T.V. & Rajavarman, V.N. 2015, "Information extraction in unstructured multilingual web documents", Indian Journal of Science and Technology, vol. 8, no. 16.
27. Prakash, K.B., Dorai Rangaswamy, M.A. & Raman, A.R. 2010, "Text studies towards multi-lingual content mining for web communication", Proceedings of the 2nd International Conference on Trendz in Information Sciences and Computing, TISC-2010, pp. 28.
28. Prakash, K.B., Kumar, K.S. & Rao, S.U.M. 2017, "Content extraction issues in online web education", Proceedings of the 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology, iCATccT 2016, pp. 680.
29. Prakash, K.B., Rajaraman, A. & Lakshmi, M. 2017, "Complexities in developing multilingual on-line courses in the Indian context", Proceedings of the 2017 International Conference On Big Data Analytics and Computational Intelligence, ICBDACI 2017, pp. 339.