Dissertation on (Major Project-II)

"Speech and Pattern Recognition for Emotion Classification"

Submitted in Partial Fulfillment of the Requirement
For the Award of Degree of

Master of Technology

In

Software Technology

By

Vivek Kumar Kohli
University Roll No. 2K16/SWT/519

Under the Esteemed Guidance of

Dr. Ruchika Malhotra
Associate Head & Associate Professor, Department of Computer Science & Software Engineering

2016-2020

DEPARTMENT OF COMPUTER SCIENCE & SOFTWARE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
DELHI – 110042, INDIA

STUDENT UNDERTAKING

Delhi Technological University
(Government of Delhi NCR)
Bawana Road, Delhi- 110042

This is to certify that the thesis entitled **"Speech and Pattern Recognition for Emotion Classification"** done by me for the Major project-II for the achievement of **Master of Technology** Degree in **Software Technology** in the **Department of Computer Science & Engineering**, Delhi Technological University, Delhi is an authentic work carried out by me under the guidance of Dr. Ruchika Malhotra.

Signature: Vivek Kohli

Student Name
Vivek Kumar Kohli
2K16/SWT/519
30/06/2020

Above Statement given by Student is Correct.

Project Guide:
Dr. Ruchika Malhotra
Associate Head & Associate Professor,
Department of Computer Science &
Engineering, DTU

ACKNOWLEDGEMENT

I would like to express sincere thanks and respect towards my guide **Prof. Ruchika Malhotra, Associate Head & Associate Professor, Department of Computer Science & Engineering, Delhi Technological University Delhi.**

I consider myself very fortunate to get the opportunity for work with her and for the guidance I have received from her, while working on this project. Without her support and timely guidance, the completion of the project would have seemed a far. Special thanks for not only providing me necessary project information but also teaching the proper style and techniques of documentation and presentation.

VIVEK KUMAR KOHLI
M. Tech (Software Technology)
2K16/SWT/519
30/06/2020

ABSTRACT

Human speech itself is a very special feature that is used for communication and expression of feelings. Speech analysis is an interesting and developing field for researchers. Physiologists and scholars from around the world are experimenting with speech as a marker for the detection of human mental physiognomies and diseases. Through speech analysis we can identify different human emotions and depressions. In our work we build a speech emotion detection system using convolutional neural network (CNN). Mel-Frequency Cepstral Co-efficient (MFCC) was used for feature extraction and speech recognition. The results showed high accuracies which were overwhelming for the start. Further, we plan to implement different other aspects of machine learning and gender recognition in speech emotion recognition and aid people with difficulty in understanding emotions.

CHAPTER 1

1. INTRODUCTION

1.1 Background

The capacity to adjust vocal sounds and produce speech is one of the highlights which set people apart from other living creatures. The voice of any individual can be portrayed by a few properties, for example, loudness, timbre, pitch, and vocal tone. It has frequently been seen that people express their emotions by changing diverse vocal properties amid speech age. The investigation of the human voice emerged as an imperative zone of concentrate for its different applications in various fields. Speech investigation essentially manages extraction of a few parameters from sound flag for handling of voice in attractive appropriateness by utilizing reasonable systems. These vocal source coordination includes alone, on a solitary held vowel, beat or are equivalent to other capabilities and mirror a noteworthy pressure of the component space (see figure 1).

Altogether, emotions are crucial for people, affecting perception and regular exercises, for example, correspondence, learning and basic leadership. They are communicated through speech, outward appearances, signals and other non-verbal pieces of information. It has been appeared ghastly estimations can recognize emotions from speech, for example, bliss, trouble, fear, outrage, quiet, and so forth. It has additionally been accounted for that one of the qualities of speech motions under an enthusiastic condition is an inconsistency. In the creation of short vowels, the respiratory control framework isn't critical; thus, vowels phonated in a nonstop mold with agreeable dimensions of pitch and loudness are fascinating and valuable in speech design acknowledgment.

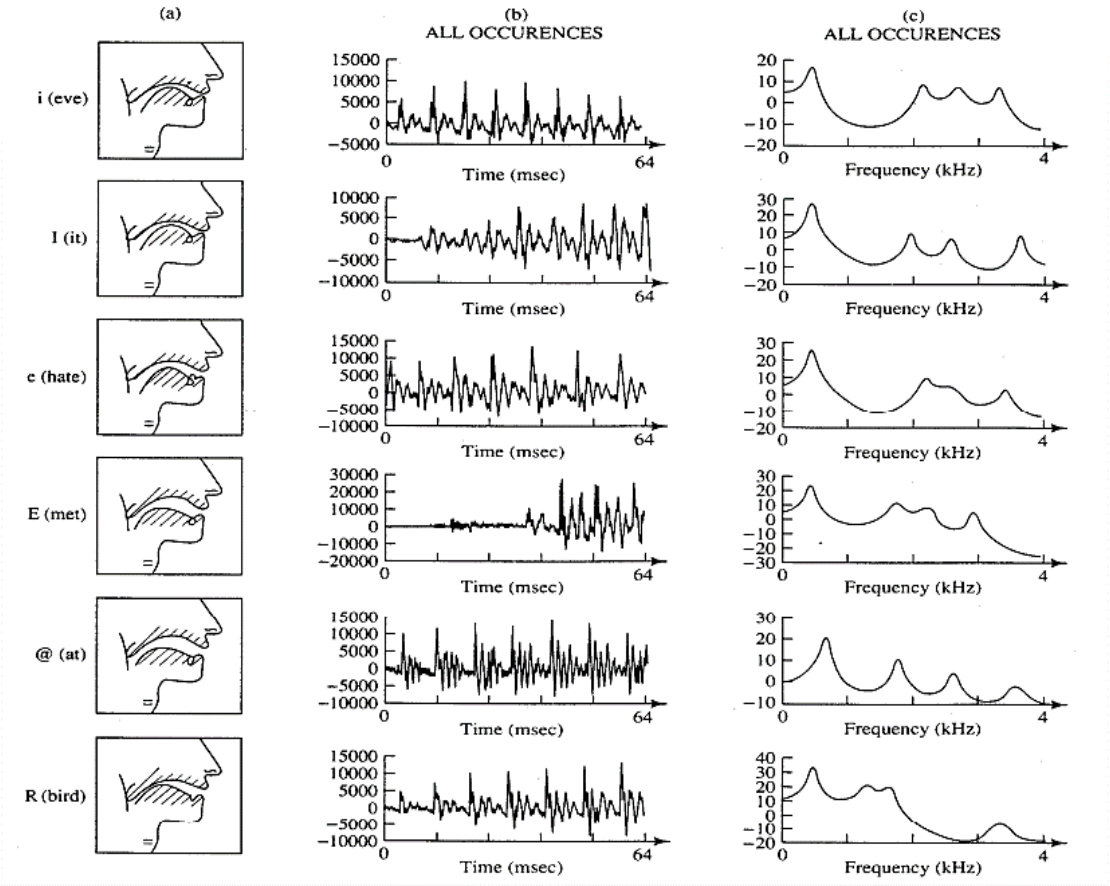

Figure 1: Speech signals for vowels a) waveform and b) frequency modulation.

Speech emotion recognition is only the example acknowledgment framework. This demonstrates the phases that are available in the example acknowledgment framework are likewise present in the Speech feeling acknowledgment framework. The speech feeling acknowledgment framework contains five primary modules passionate speech input, include extraction, high-light choice, grouping, and perceived enthusiastic yield [1].

Subsequently, speech feeling location alludes to dissecting vocal conduct as a marker of affect, with spotlight on the non-verbal parts of speech. Its fundamental suspicion is that there is a lot of impartially quantifiable parameters in voice that mirrors the powerful express an individual is right now communicating. This suspicion is upheld by the way that most full of feeling states include physiological responses which thus change the procedure by which voice is created [2]. For instance, outrage frequently delivers changes in breath and increments solid

pressure, impacting the vibration of the vocal folds and vocal tract shape and influencing the acoustic qualities of the speech [3].

1.2 Human Voice Production

The human voice comprises of sounds created by a person utilizing the vocal folds for doing acoustic exercises, for example, talking, singing, giggling, yelling, and so on. The human voice recurrence is explicitly a piece of the human sound creation instrument in which the vocal lines or overlays are the essential wellspring of produced sounds. For the most part, the instrument for producing the human voice can be subdivided into three sections; the lungs, the vocal overlays inside the larynx, and the articulators [4]. Figure 2 demonstrates the life systems of human voice generation device.

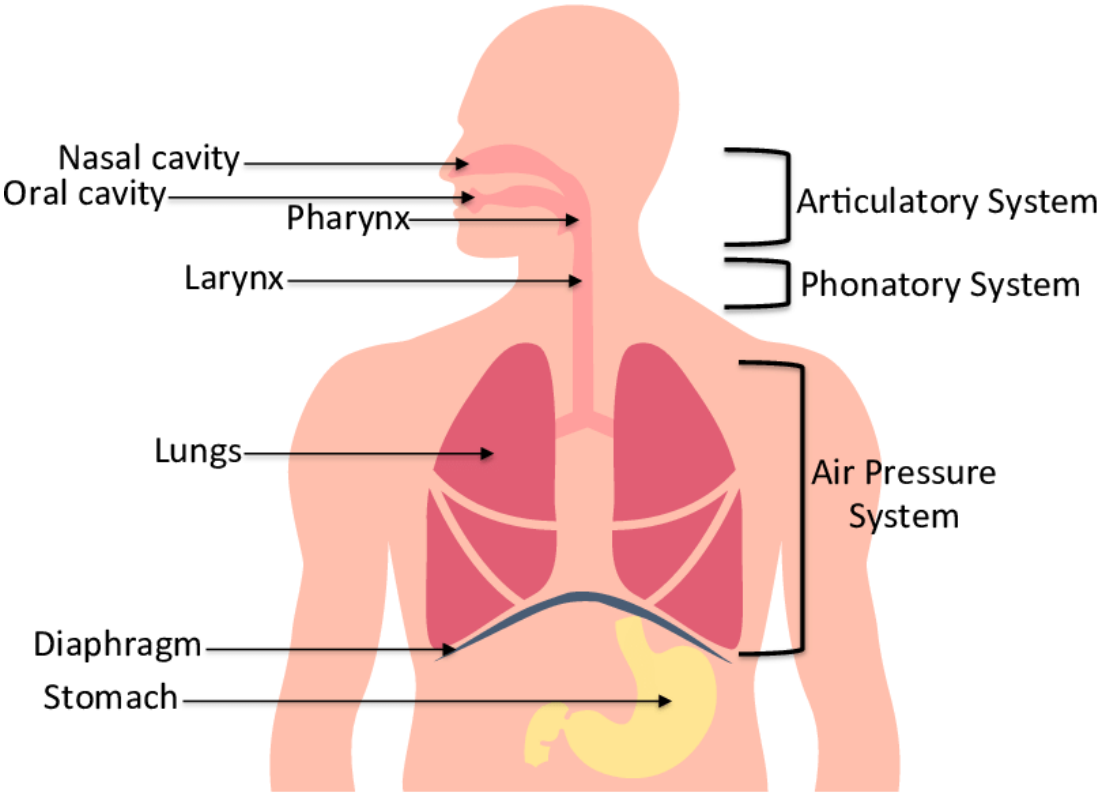

Figure 2: Human Voice apparatus

Also, speech is a perfect biomarker on the grounds that it is reasonable, uncomplicated to gather, easy to oversee, and non-obtrusive to the subject under investigation [5]. Speech highlights have seen introductory accomplishment in numerous examinations and hold guarantee as enlightening biomarkers [6]. Vocal biomarkers have been appeared to be significant in the programmed recognition of numerous states, including emotions, weakness, and subjective load [7]. For instance, [8] utilized pitch and jitter highlights to discover contrasts between temperament states in bipolar patients, while [9] utilized vocal and articulatory highlights, similar to voice quality, laryngeal coordination, and tongue development, to recognize speech of fervor and misery (figure 3).

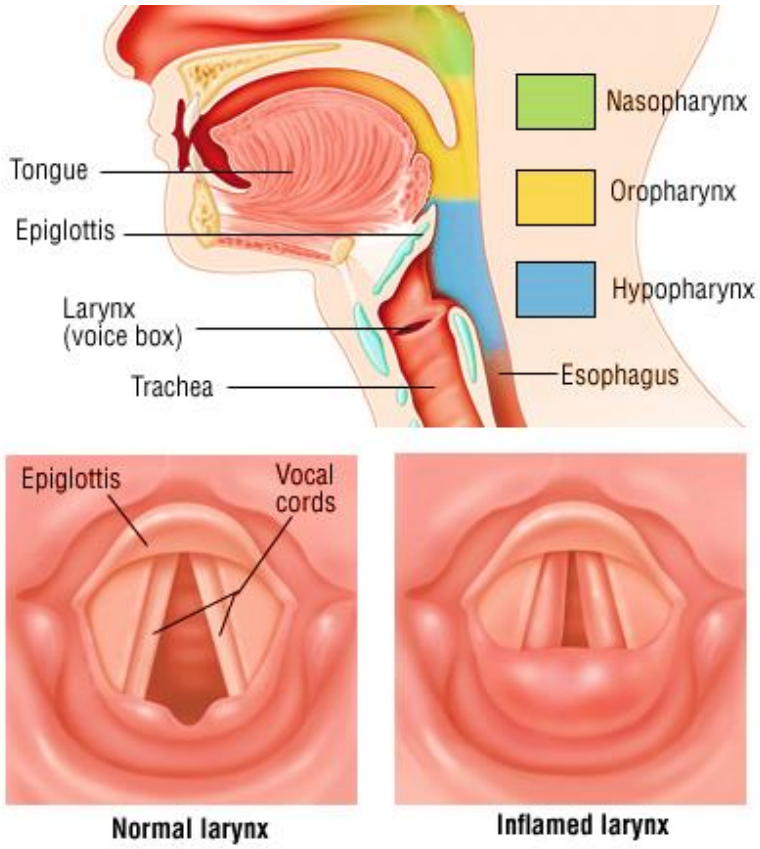

Figure 3: Anatomy of voice box, normal and diseased larynx.

The project is based on the following techniques:

• Artificial Intelligence

• Information Theory and Coding

• Digital Signal Processing

• Natural Language Processing

• Python programming

• Various Machine learning Algorithms

Additionally, a significant amount of knowledge in machine learning was gained throughout, working on the thesis.

Chapter 1 gives a brief introduction to the speech emotion recognition. This chapter also describes aspects of speech and how human voice is produced, which is further elaborated in subsequent chapters. Chapter 2 throws light on various previous related works on speech emotion recognition. Chapter 3 discusses about mechanism of human speech production and employment of speech analysis in various applications. Chapter 4 explains how the speech emotion detector is implemented, including descriptions of some valuable algorithms. Chapter 5 evaluates the proposed speech emotion recognizer in an experimental setup and analyzing its performance and accuracy. Finally, Chapter 6 presents a brief discussion and conclusion with future scopes of the study.

CHAPTER 2

2.1 Literature Review

Speech and facial appearance are the subjects of present examinations in both state identification and feeling location [8] [10]. The two procedures require complex coordination of articulatory muscles that must be definitely planned and executed by neuromotor circuits. Any interruption to engine related zones of the mind because of intellectual, emotive, or pathogenic conditions might be reflected in modifications to or disturbances in the engine yield.

Significant crucial work on speech feeling location was finished by [11], who proposed the utilization of measurable example acknowledgment strategies for feeling identification and set the essential framework engineering still utilized today. From that point forward, the exactness has ceaselessly been enhanced [12] [13].

A few investigations have recently investigated varieties in speech of discouraged individuals demonstrating that stop length and recurrence increment with expanded sadness, showing that increasingly discouraged speech contains progressively in general quietness. Speech rate [14] and phonetic rate [15] [16], which estimates the recurrence of individual sections as opposed to words or syllables, have been appeared to diminish with higher wretchedness seriousness, demonstrating melancholic individuals talk all the more gradually.

Utilizing a parallel machine learning situation, [17] ordered both read and unconstrained speech tests dependent on acoustic highlights like force, vitality, voice quality, and pitch. They found that unconstrained speech was ordered all the more precisely more regularly.

Research dependent on fluffy derivation framework that can recognize the disposition of the individual, communicating ordinary individual, people in rush, upbeat individual and tragic individual) [18]. These emotions were ordered into three classes as per need of voice.

2.2 Problem Statement

Speech and pattern recognition using machine gives a clear and probable insight in the human emotion prediction. These emotions can be then categorized into happy, sad, anger, fear, distress, etc.

Highlights demonstrating distinctive enthusiastic states might cover, and there might be numerous methods for communicating the equivalent passionate state. One system is to register however many highlights as would be prudent. Streamlining calculations would then be able to be connected to choose the highlights contributing most to the segregation while overlooking others, making a smaller feeling code that can be utilized for arrangement. Automatic voice analysis for speech emotion recognition have several aspects:

1) It has quantitative and non-invasive nature.

2) Allows identification and monitoring of the onset of emotional changes.

3) Computer based machine learning models reduce cost and time in addition to classification and improved accuracy.

2.3 Aims and Objectives:

(1) In our study, we will try to identify pattern among types of emotions by speech pattern recognition and to identify acoustic parameter correlated with the emotional expressivity.

(2) To detect emotions among speech signals with software tool and express them as happy, sad, anger, calm, fearful, etc.

(3) To create a model speech recognition system that can efficiently and accurately identify hidden emotions in the audio signal of the given dataset.

Through this paper the application would be able to identify the mood of the subject and classify them based on the learned patterns. Moreover, acoustic analyses have not been frequently performed in studies. Here we aim to perform acoustic and

perceptual evaluations using machine learning within the same study setting, as such studies are not very often reported in emotion recognition using speech.

CHAPTER 3

3.1 Human Speech

Being the normal type of correspondence, speech is the most fundamental and usually utilized correspondence by all the people. For average citizens speech is only the sound waves leaving the human mouth and saw/tuned in through ears. Be that as it may, there is intricate instrument behind its generation. The investigation of human speech generation and perception component is vital and important for the advancement of gadgets for listening devices, cochlear embed, speech acknowledgment, speech improvement, speech reenactment, speech demonstrating and so on.

Typically, individuals talk in dialect as indicated by the district they are raised in. One needn't bother with unique preparing or learning to talk in their primary language. Kids figure out how to talk at an early age of one year by understanding the sound and visual motions. The dialect indications of any dialect can be articulated with the assistance of images called phonemes. Every one of the words with various tones of any dialect can be talked utilizing least arrangement of phonemes [19]. Every one of the dialects talked on the planet have 20 to 60 phonemes. Phonemes of any dialect incorporate relevant impacts, emotions and qualities of the speaker to be articulated which obviously isn't required for composed content of the dialect. These phonemes are essentially structured dependent on the articulatory development of the vocal tract.

Formants

Formants are the recognizing parts of human speech and of singing. The data that a human requires to recognize vowels can be spoken to absolutely quantitatively by the recurrence substance of the vowel sounds. In speech, these are trademark partials that distinguish vowels to the audience. The formant with most minimal recurrence is called f1, the second least called f2, and the third f3. Frequently the initial two formants, f1 and f2, are sufficient to disambiguate a vowel [20]. These two formants decide nature of vowels as far as the open/close and front/back

measurements (which have customarily, however not precisely, been related with position of the tongue). Henceforth, first formant f1 has a higher recurrence for an open vowel, (for example, [a]) and a lower recurrence for a nearby vowel, (for example, [i] or [u]); and the second formant f2 has a higher recurrence for a front vowel, (for example, [i]) and a lower recurrence for a back vowel, (for example, [u]). Figure 4 demonstrates the time-recurrence tweak of an, I, u.

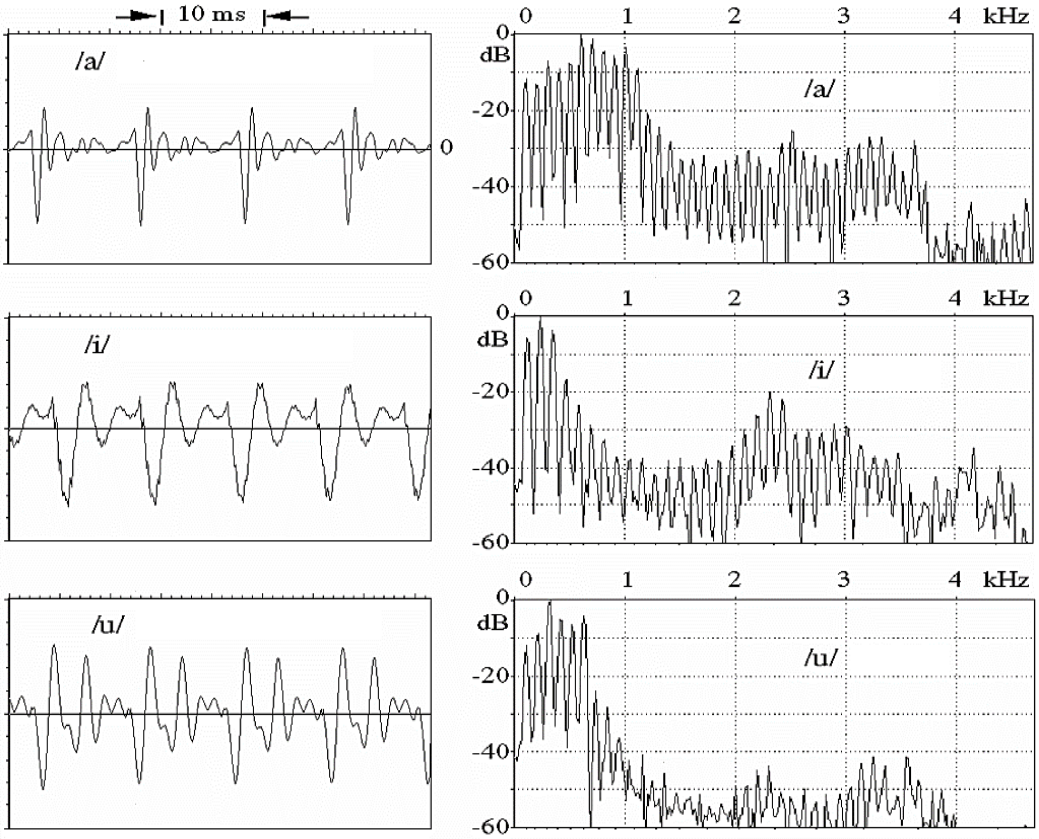

Figure 4: The time and frequency representation for vowels ‘a’, ‘i’, ‘u’.

Vowels will quite often have at least four discernable formants; once in a while there are more than six [21] [22]. Be that as it may, the initial two formants are the most essential in deciding vowel quality, and this is shown regarding a plot of the

first formant against the second formant [23].

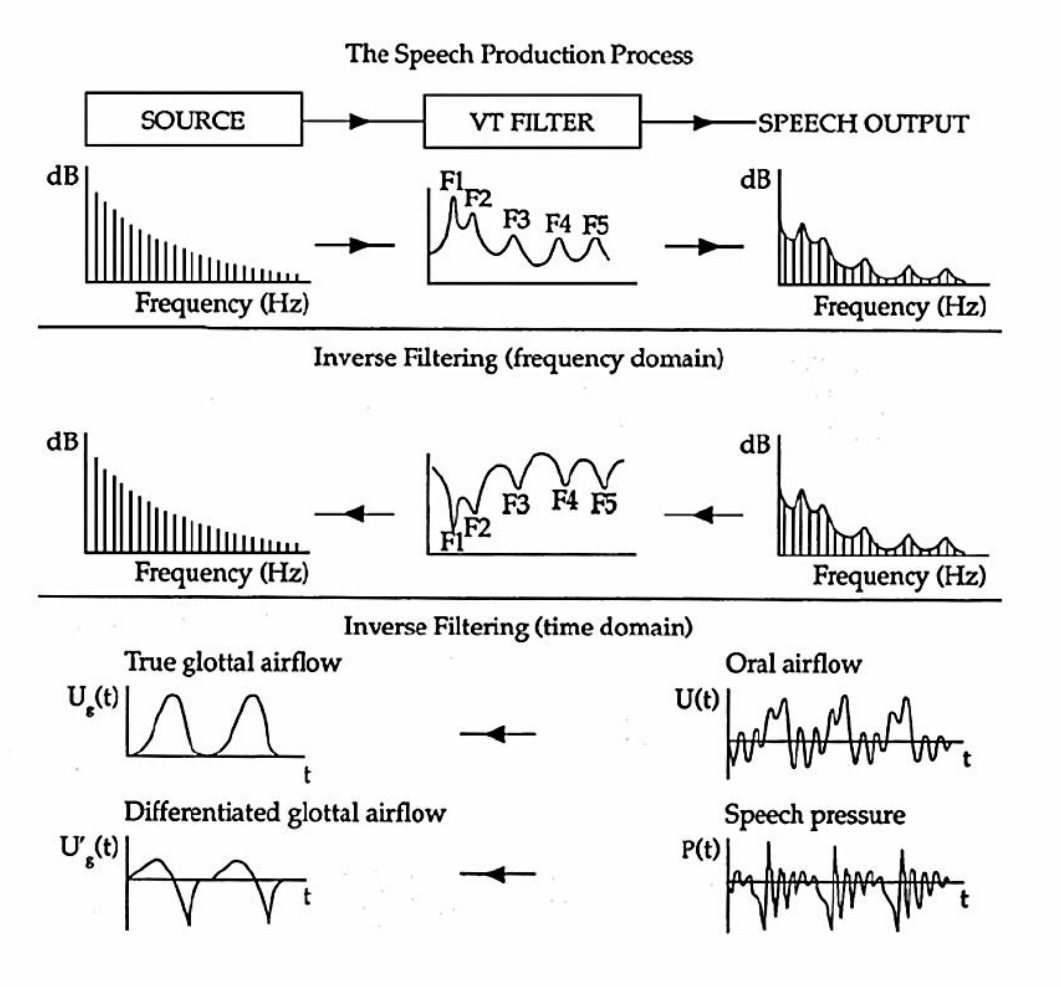

Figure 5: Speech production process

3.2 Physiology of Human Voice

Human voice generation includes the synchronization of ideal glottis situating with the control of the wind current from the lungs to the oropharynx. The vocal folds must be of appropriate consistence to show their dynamic vibratory qualities (figure 5).The normal processes of speech and voice production are overlapping, therefore, the ability to speak can be subdivided into several dimensions, including respiration, phonation, resonance, articulation, and prosody, etc.:

Breath: Using the stomach to rapidly fill the lungs completely, trailed by moderate, controlled exhalation for speech.

Phonation: Using the vocal lines and wind current to deliver voice of differing pitch, loudness, and quality.

Reverberation: Raising and bringing down the delicate sense of taste to guide the voice to resound in the oral or potentially nasal cavities to additionally influence voice quality.

Enunciation: Coordinating snappy, exact developments of the lips, tongue, mandible, and delicate sense of taste for clearness of speech.

Prosody: Combining all components for a characteristic stream of conversational speech, with sufficient loudness, accentuation, and melodic line to upgrade meaning.

In the medical domain, the assessment of pathological voice quality is an important issue, inducing a large amount of research in multidisciplinary domains.

3.2 Speech and Human Emotions

Speech can be the quick and proficient method of communication among human and in specific cases machine additionally [24]. People have the normal capacity to utilize all their accessible faculties for most extreme consciousness of the got message. Through all the accessible faculties individuals in truth sense the enthusiastic condition of their correspondence accomplice. The passionate identification is normal for people yet it is exceptionally troublesome undertaking for machine. Thusly, the motivation behind feeling acknowledgment framework is to utilize feeling related information so that human machine correspondence will be enhanced [25][26].

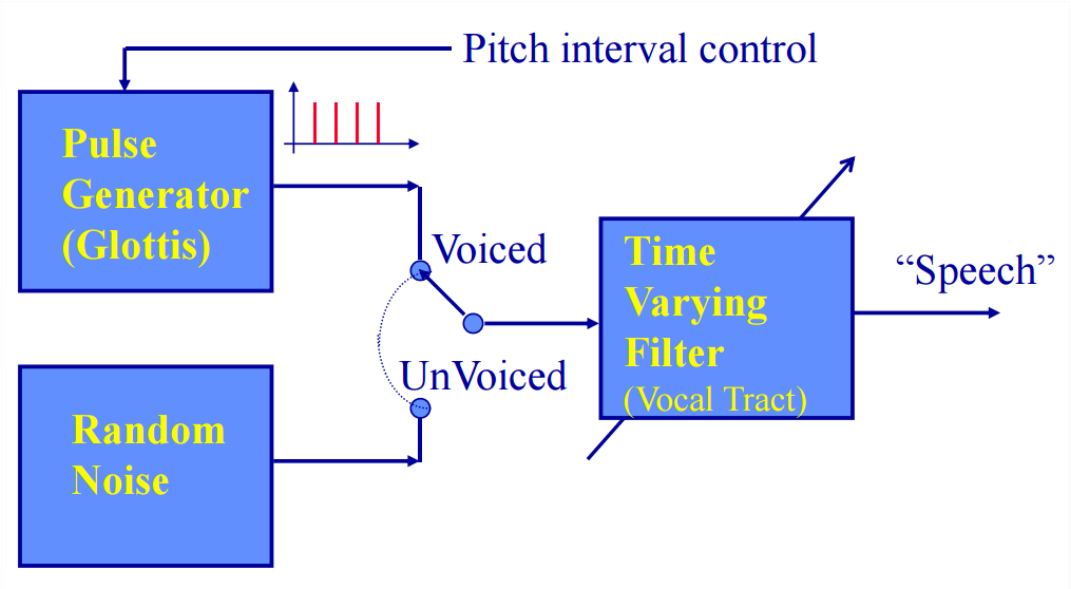

Figure 6: A Speech Production Model

So as to make a characteristic connection, the machine ought to be sufficiently canny to perceive speaker's feeling by examining the acoustics of male/female voices (figure 6 and 7). This has presented a generally new and testing research region with a wide scope of uses in man-machine collaboration, known as Speech Emotion Recognition (SER). SER can improve the performance of automatic speech recognition systems [24]. It is also useful in e-learning, computer games, medicine, psychology and in car-boards [27]. SER is commonly treated as a pattern recognition problem which includes three main stages: feature extraction, feature reduction and classification.

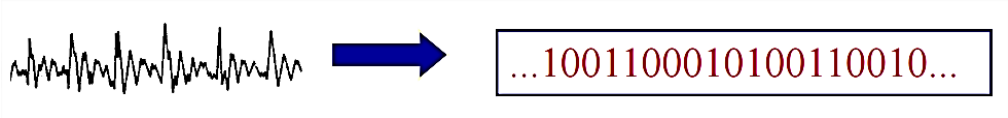

Figure 7: Speech signal to machine translation

3.3 Speech in Different Domains

The utilization of the speech emotion recognition framework incorporate the mental conclusion, savvy toys, lie location, in the Call Center discussions which is the most essential application for the computerized acknowledgment of emotions from the speech, in vehicle board framework where data of the psychological condition of the driver may give to the framework to guarantee his/her security [24].

3.3.1 Dysphonia

Dysphonia is a phonation disorder with the difficulty in the voice production. Dysphonia can be observed with hoarse, harsh, or breathy vowel sounds, as a result of impaired ability of the vocal folds to properly vibrate during exhalation [28]. This can directly relate to the development of underlying diseases that could be diagnosed by the speech evaluations [29] [30]. Factors that influence either the aerodynamic configuration (i.e vocal fold paralysis) or the vibratory property of the glottis (i.e laryngeal cyst) may result in dysphonia [31] [32].

3.3.2 Dysarthria

Dysarthria is a motor speech disorder in which the muscles that are used to produce speech are damaged, weakened or paralyzed. The person with dysarthria cannot control his or her tongue, larynx, vocal cords, and surrounding muscles, which makes it difficult for the person to form and pronounce words [33].

3.3.3 Asthma Diagnosis

Asthma is extremely basic unending provocative muddle of the respiratory routes portrayed by the expanded aviation route hyper-responsiveness, automatic bronchospasm, reversible wind current check, shortness of breath and the related indications are wheezing, hypersensitivities, hacking, sore throat and aggravation around the nasal track. Swollen, aggravated strings don't vibrate proficiently that makes the voice sound dry and consequently voice quality is hindered [34].

3.3.4 The disease of Parkinson

Parkinson's infection is characterized as a turmoil of the focal sensory system that outcomes from the loss of the neurons, basically from the substantia nigra area of the mind. Parkinson's malady can be arranged by huge varieties in voice (Dysphonia) [35]. The trademark side effects of Dysphonia in Parkinson's disease are repetitive, decreased pitch, failure to fluctuate pitch, variable rate, short surges of speech, loose consonants, powerlessness to continue delayed vowel phonation and a hoarse and unforgiving voice. These varieties in voice designs turn out to be more awful as the ailment advances. Research has demonstrated that speech might be a valuable flag for separating PWP (individuals with Parkinson's) from sound people.

3.3.5 Symptom of Aphasia Alzheimer

Alzheimer is the most well-known type of dementia. There is no remedy for the sickness and condition of the patient intensifies as it advances, which in the long run prompts demise. The regular side effects are: memory misfortune, disarray, touchiness, hostility, issue with dialect, and emotional episodes. The most noteworthy impact on practical relational abilities that goes with Alzheimer's illness (AD) is aphasia [36], lost oral informative capacity like breakdowns in semantic preparing, shallow vocabularies and word-discovering challenges prompting the weakening of unconstrained speech.

3.3.6 Depression

Depression is a temperament issue portrayed by discouraged mind-set and loss of intrigue or delight in recently appreciated exercises. With respect to speech explicitly, aggravations as often as possible influence delay times, volume, tone, and prosody. Discouraged individuals wind up restless, weakness, bad tempered, stressed and miserable; confront issues in dozing, retaining, choosing, loses fixation or enthusiasm for exercises that used to be pleasurable, one can endeavor or even attempt suicide.

3.4 Speech Analysis

The mind boggling nature of voice signal that it shifts in time makes it hard to examine. It likewise represents a major test for researchers in making a productive voice examination framework. That is the reason the vast majority of the voice examination methods planned so far go for separating the time-fluctuating highlights of voice signal to streamline the means of assessment, disintegration or adjustment of the signal [37].

The vocal tract is the cavity between the vocal cords and the lips, and acts as a resonator that spectrally shapes the periodic input, much like the cavity of a musical wind instrument.

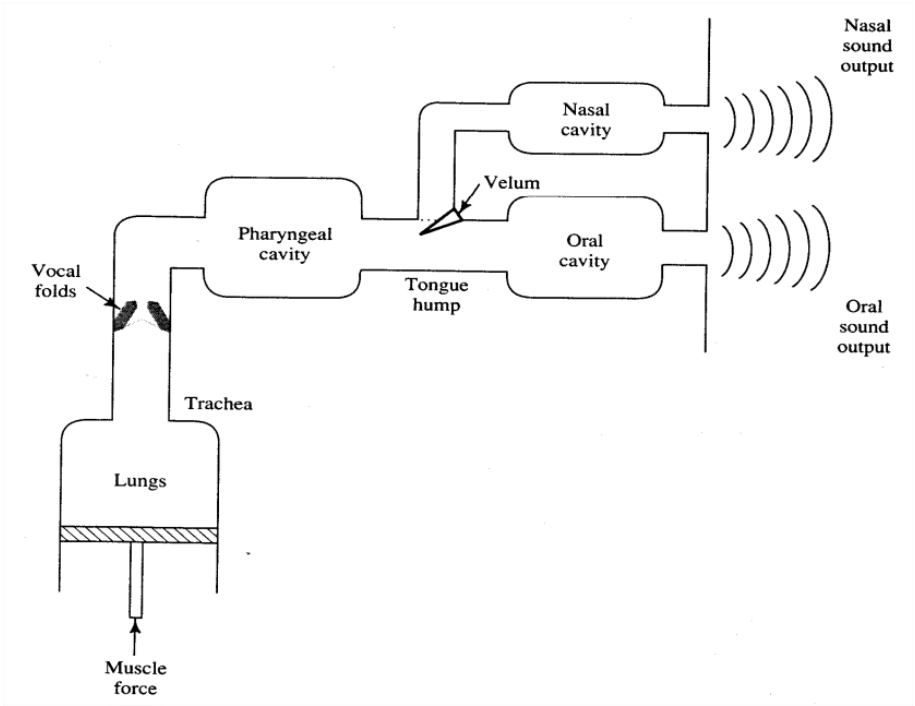

Figure 8: Block diagram of human speech production

Voice signal has voiced portions that relate to periodicity and signal's range vitality and unvoiced fragments of the signal contain non-occasional segments. Research about dependent on Linguistic exhibits that there are different dimensions of portrayal in mapping sound to significance (figure 9). The unmistakable dimensions guessed to frame the reason for speech incorporate 'particular highlights', the littlest building squares of speech that additionally have an acoustic elucidation [38].

Mapping acoustic speech input onto applied and semantic portrayals includes numerous dimensions of calculation and portrayal. These dimensions may incorporate the portrayal of unmistakable highlights, sections (phonemes), syllabic structure, phonological word frames, syntactic highlights and semantic data. The voice signals are fundamentally analysed in the Time and Frequency Domains (figure 10).

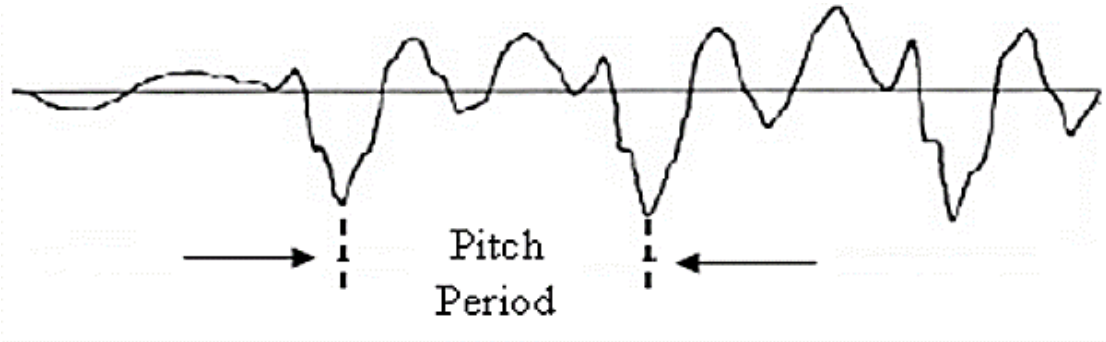

Figure 9: Speech in time-frequency domain

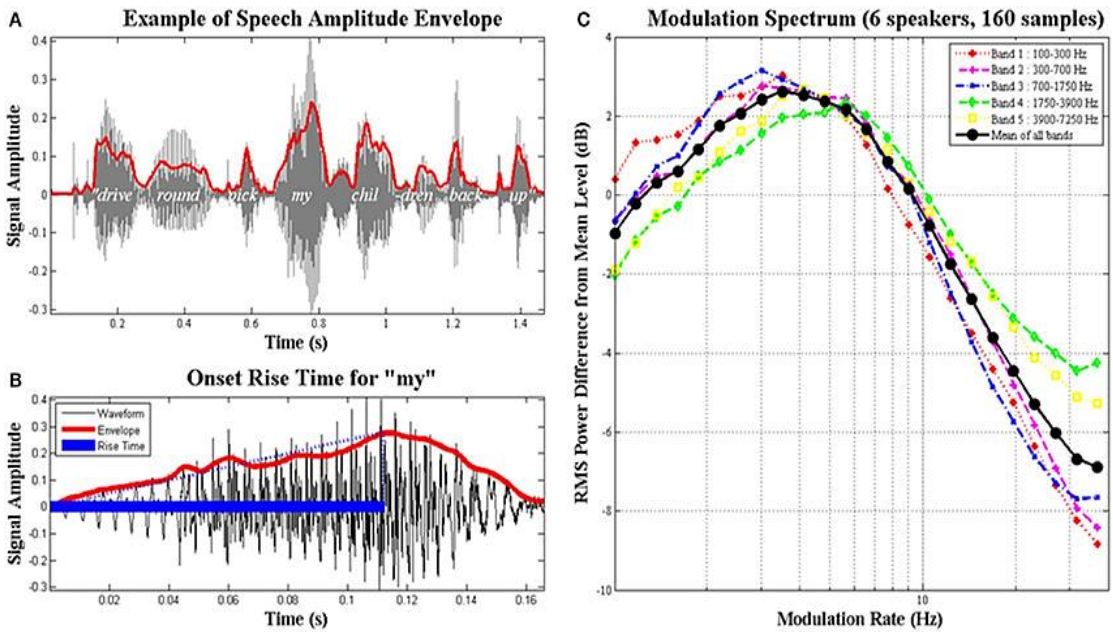

Figure 10: Speech modulation

3.5 Feature Extraction

With several phonatory frequency variability, variation of speech amplitude (shimmer), intensity, and nonlinear dynamics parameters computed from the electroglottographic signals in standard speech tests, it is necessary to select the most discriminant vocal parameters with feature combination methods for further pattern classifications. Feature extraction is the way toward ascertaining the speech signal highlights which are pertinent for speech handling. Since the PC has no feeling of hearing and perception like people, they should be encouraged with these highlights of speech which turn into a deciding variable after order. Highlight extraction includes investigation of speech signal. The MFCC presented by Davis and Mermelstein during the 80's, is the most significant case of a list of capabilities that is broadly utilized in voice preparing [39]. Speech is typically divided in casings of 20 to 30 ms, and the window investigation is moved by 10 ms. Filter methods for feature selection or combination are usually less computationally intensive than the wrapper methods that commonly use a predictive model to score feature subsets [40]. Plenty of statistical feature filter methods are computed based on probability distribution estimations [41].

3.6 Basic model of speech recognition

The standard way to deal with huge vocabulary nonstop speech acknowledgment is to expect a basic probabilistic model of speech creation whereby a predefined word succession, W, delivers an acoustic perception arrangement Y, with likelihood P(W,Y). The objective is then to disentangle the word string, in view of the acoustic perception succession, with the goal that the decoded string has the greatest a osteriori (MAP) probability

$$\hat{P}(W/A) = \arg\max_{W} P(W/A)$$

using Bayesian rule, the above equation can be written as

$$P(W/A)=\frac{P(A/W)P(W)}{P(A)}$$

since *P(A)* is independent of *W*, the MAP decoding rule of above equation can be given by

$$\hat{W}=\text{argmax}_{w}\ P(A/W)P(W)$$

P(A/W), is by and large called the acoustic model, as it evaluates the likelihood of a succession of acoustic perceptions, molded on the word string. Subsequently P(A/W) is registered for extensive vocabulary speech acknowledgment frameworks, it is important to fabricate measurable models for sub word speech units, develop word models from these sub word speech unit models (utilizing a dictionary to depict the structure of words), and after that hypothesize word groupings and assess the acoustic model probabilities by means of standard connection strategies. *P(W)* is called the language model, which describes the probability associated with a postulated sequence of words. Such language models can incorporate both syntactic and semantic constraints of the language and the recognition task.

3.7 Pattern Matching

The pattern-matching methodology [42] includes two basic advances in particular, design preparing and design correlation. The basic element of this methodology is that it utilizes a very much detailed numerical structure and sets up predictable speech design portrayals, for solid example examination, from a lot of named preparing tests through a formal preparing calculation (see figure 11) [43]. A speech design portrayal can be as a speech format or a factual model (e.g., a HIDDEN MARKOV MODEL or HMM) and can be connected to a sound (littler than a word),

a word, or an expression.

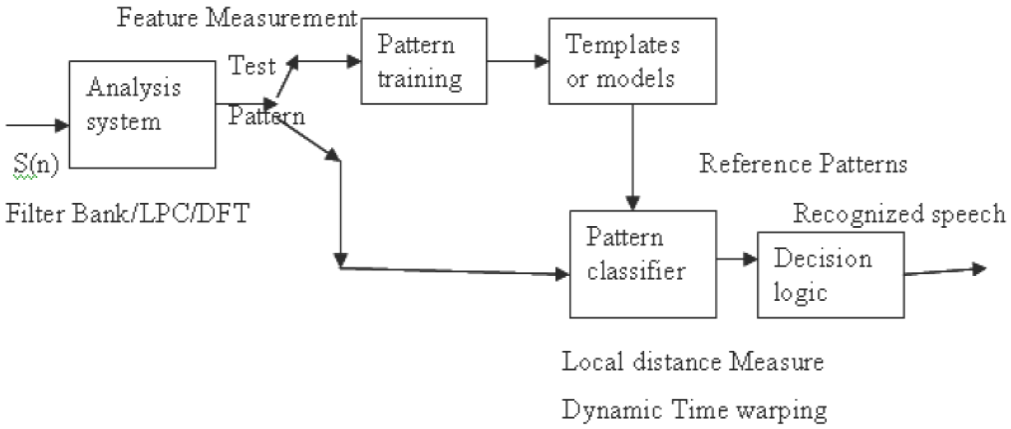

Figure 11: Block diagram of pattern detection model for speech recognition

Isolated Words:

Isolated word recognizers as a rule require every articulation to have calm (absence of a sound signal) on the two sides of the example window. It acknowledges single words or single expression at once. These frameworks have "Tune in/Not-Listen" states, where they require the speaker to hold up between articulations (generally doing preparing amid the stops). Isolated Utterance may be a superior name for this class.

Connected Words:

Connected word frameworks (or all the more accurately 'connected expressions') resemble isolated words, yet enables separate articulations to be 'run-together' with an insignificant interruption between them.

Continuous Speech:

Continuous speech recognizers enable clients to talk normally, while the PC decides the substance. (Fundamentally, it's PC correspondence). Recognizers with

continuous speech abilities are the absolute most hard to make since they use unique strategies to decide articulation limits.

Spontaneous Speech:

At an essential dimension, it tends to be thought of as speech that is characteristic sounding and not practiced. An ASR framework with spontaneous speech capacity ought to have the capacity to deal with an assortment of common speech highlights, for example, words being run together, "ums" and "ahs", and even slight falters.

3.8 State of the art Technologies involved:

Direct analysis and synthesizing the complex voice signal are difficult due to large information contained in the signals. Therefore, the digital signal processes such as Feature Extraction and Feature Matching are introduced to represent the voice signal (figure 12). Several machine learning methods and algorithms such as Mel-frequency cepstral coefficients (MFCC), Linear Predictive Coding (LPC), Hidden Marcov Model (HMM), Dynamic Time warping (DTW), Artificial Neural Network (ANN), Support Vector Machines [44], etc. are evaluated with a view to identify a straight forward and effective method for voice signal. Neural systems have much comparability with Markov Models. Both are factual models spoken to as charts. Where HMMs utilize the likelihood for state advances, neural systems use association qualities and capacities. A key distinction is that neural systems are in a general sense parallel while Markov chains are sequential. Frequencies and speech happen in parallel, while syllable arrangement and words are basically sequential. This implies the two systems are ground-breaking in various angles.

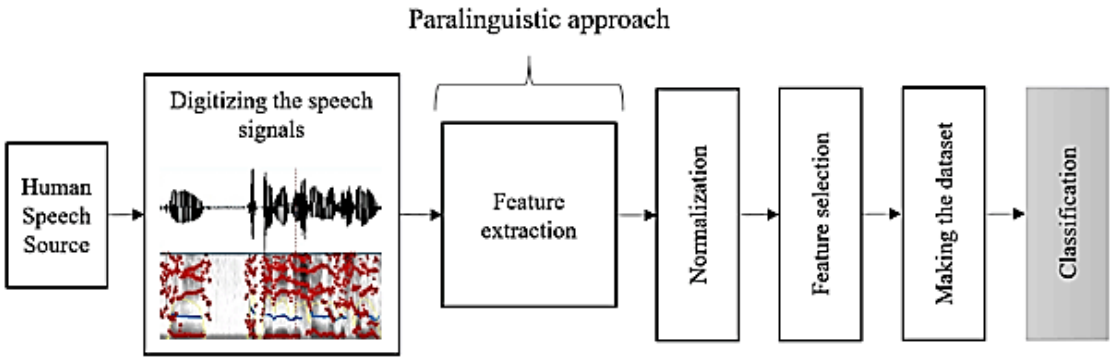

Figure 12: Speech signal processing and classification

3.9 Speech Analysis Algorithms

1. Dynamic Time Warping Algorithm

Dynamic Time twisting (DTW) is a calculation for estimating similitude between two fleeting groupings which may differ in time or speed. For example, similitudes in strolling examples could be distinguished utilizing DTW, regardless of whether one individual was strolling quicker than the other, or if there were increasing speeds and decelerations amid the procedure of a perception. DTW has been connected to fleeting groupings of video, sound, and designs information - to be sure, any information which can be transformed into a straight arrangement can be broke down with DTW. A notable application has been programmed speech acknowledgment, to adapt to various talking speeds. Dynamic time traveling (DTW) is the most clear arrangement and is utilized to tackle precisely this issue in speech acknowledgment applications. Since clearly the voice signal will in general have distinctive worldly rate, the arrangement is imperative to create the better execution.

2. Mel-Frequency Cepstral Coefficients

Mel-Frequency cepstral coefficients (MFCC) have been overwhelmingly utilized in speaker acknowledgment and in speech acknowledgment [45]. This is unreasonable to numerous specialists since speech acknowledgment and speaker acknowledgment look for changed sorts of data from speech, to be specific, phonetic data for speech acknowledgment and speaker data for speaker acknowledgment. MFCC was first proposed for speech acknowledgment and its mel-twisted recurrence scale is to impersonate how human ears process sound. Its ghastly goals progresses toward becoming lower as the recurrence increments. In this way, the data in the higher recurrence locale is down-tested by the mel scale. Notwithstanding, in light of hypothesis in speech creation, speaker qualities related with the structure of the vocal tract, especially the vocal tract length, are reflected more in the high recurrence locale of speech [46].

3. Hidden Morkov Model Algorithm

The Hidden Markov Model have a wide scope of use in worldly example acknowledgment, for example, speech, penmanship, motion acknowledgment, grammatical feature labeling, melodic score following, incomplete releases and bioinformatics. A Hidden Markov Model can be viewed as a speculation of a blend demonstrate where the shrouded factors (or inactive factors), which control the blend part to be chosen for every perception, are connected through a Markov procedure as opposed to autonomous of one another.

a. A very much tuned HMM for the most part gives preferable pressure over a straightforward Markov demonstrate, enabling more groupings to be fundamentally found.

b. The models are coherent (at any rate when drawn as opposed to simply recorded). An excellent model for REPs (Repetitive Extragenic Palindromic successions) as opposed to the checks of the request 8 straightforward Markov display.

c. The low proportion of edges to states implies that substantial parts of the model are basic straight-line successions, which are anything but difficult to attract and to get it.

Speech acknowledgment utilizes a marginally embraced Markov Model. Speech is part into the littlest discernable elements (vowels and consonants as well as conjugated sound like ou,ea,au,…). Every one of these substances are spoken to as states in the Markov Model. As a word enters a Hidden Markov Model it is contrasted with the most appropriate model (element). Well are more incredible than arrangements, since a similar state can be utilized more than once in a way, yet every segment must be utilized once in an arrangement [47].

4. Neural Networks

Neural systems have much closeness with Markov Models. Both are factual models spoken to as diagrams. Where Markov Models utilize the likelihood for state advances, neural systems use association qualities and capacities. A key contrast is that neural systems are on a very basic level parallel while Markov chains are sequential. In the neural systems, the test is to set the proper loads of the associations, the Markov Model difficulties is finding the fitting advances and perceptions conceivable outcomes [48]. In the neural systems, the test is to set the proper loads of the associations, the Markov Model difficulties is finding the fitting changes and perceptions potential outcomes. Figure 13 demonstrates orderly portrayal of a 3 layer neural system.

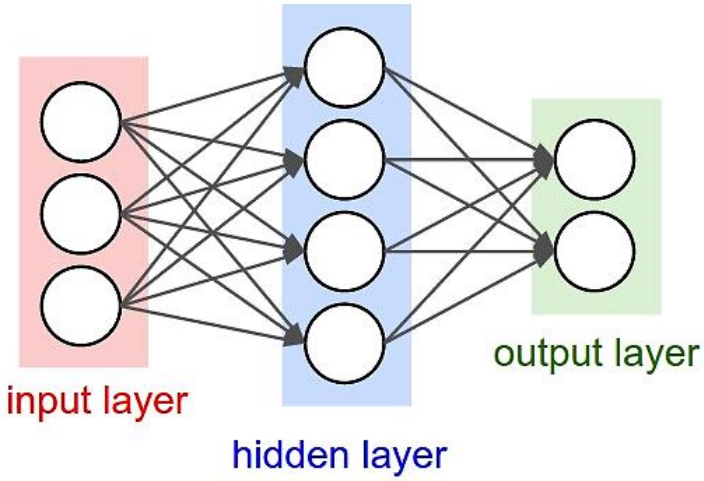

Figure 13: A 3 layer neural network (input, hidden and output)

The Parzen-window method is one of the non-parametric kernel-based PDF modeling techniques, which can be used to establish multimodal PDFs [49]. The Parzen-window method commonly estimates an unknown PDF by averaging the accumulated non-negative kernel functions $\kappa(\cdot)$, the centers of which are located at the vocal pattern data points $x\ i$, written as -

$$\hat{f}(x) = \frac{1}{N h} \sum_{n=1}^{N} \kappa\left(\frac{x - x_i}{h}\right),$$

where N is the number of data points and h represents the kernel bandwidth. In the present study, the Gaussian radial basis function was chosen as the kernel window function. According to [40], the optimal kernel bandwidth of the Gaussian function is given by –

$$h_{\text{opt}} = 1.06 \times \text{SD} \times N^{-1/5},$$

where SD denotes the standard deviation of the data points.

CHAPTER 4

4. METHODOLOGY

Speech is potentially a rich source of markers for detecting and monitoring emotions. Speech markers typically comprise acoustic descriptors extracted from behavioural measures of source, filter, prosodic and linguistic cues. In contrast, in this paper, we extract vocal features based on a model of speech production, reflecting variant parameters that may be more sensitive to individual emotion expressed. These features, which are constrained by expressions, may provide an articulatory complement to acoustic features. Our features represent a mapping from a low-dimensional acoustics-based feature space to a high-dimensional space that captures the underlying emotions, including auditory and subliminal feedback errors.

4.1 Data and Analysis

Aggregating Data is the most troublesome part of any experimental process. Consistently, acoustic speech data for emotion recognition can be collected from different sources, for e.g. [https://zenodo.org/record/1188976]. These databases are equipped with around 500-2000 audio/video samples from various subjects, actors, etc. The datasets are available in zip format. We selected a dataset of 1500 audio files from Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) in ".wav" extension for our analysis. This dataset contains acoustic audios with clear phonetic outputs, specifically meant for speech emotion recognition. 12 males and 12 females where these actors record short audios in 8 different emotions i.e. neutral, calm, happy, sad, angry, fearful, disgust, surprised. Each audio file had a unique identifier in the file name which can be used to determine the emotion the audio file contains. We went to fetch for 5 diverse emotions in the dataset:

1. Happy
2. Sad
3. Angry
4. Fearful
5. Calm

The extraction and matching process is implemented right after the pre-processing or filtering signal is performed. Since it's obvious that the voice signal tends to have different temporal rate, the alignment is important to produce the better performance. Mel-fiequency Cepstral Co-efficient is used for feature extraction and Dynamic time warping (DTW) is used for feature matching in order to detect emotions from the speech. Using machine learning to mine features in the voice, algorithms pick out speech patterns and compare them with speech samples from the learned features.

4.2 Parameters

Jitter (neighborhood): This is the normal supreme distinction between sequential periods, partitioned by the normal time frame. MDVP calls this parameter Jitt, and gives 1.040% as an edge for pathology.

Gleam (neighborhood): This is the normal total contrast between the amplitudes of continuous periods, isolated by the normal abundancy. MDVP calls this parameter Shim, and gives 3.810% as an edge for pathology.

Sounds to-Noise Ratio (HNR): A Harmonicity object speaks to the level of acoustic periodicity, likewise called Harmonics-to-Noise Ratio (HNR). Harmonicity is communicated in dB. Harmonicity can be utilized as a measure for voice quality. For example, a solid speaker can deliver a continued an or I with a harmonicity of around 20 dB, and at around 40 dB; Hoarse speakers will have an a with a harmonicity much lower than 20 dB [50].

Level of Voice Breaks DVB: the proportion of the aggregate length of zones speaking to voice breaks to the season of the entire voiced example; and number of

voice breaks NVB. The criteria for voice break region can be a missing motivation for the present time frame or an outrageous anomaly of the pitch time frame.

Feature Descriptions

For the convenience of voice perturbation feature presentation, average, maximum, and minimum vocal fundamental frequency (in Hz), (as computed by the Kay Pentax Multidimensional voice program (MDVP)) with the abbreviations of MDVP:F0, MDVP:Fhi, and MDVP:Flo, respectively.

The initial step plays out the endpoint location by utilizing momentary fleeting examination. The second step incorporates speech include extraction utilizing MFCC (Mel-Frequency Cepstral Coefficients) parameters and third step is highlight extraction of every one of the characters utilizing Vector Quantization LBG calculation, where we get the yield as characters in mel scale. These characters will be joined to shape an important word. What's more, that word will be contrasted and the pre-arranged educated highlights to get the sound yield.

The accompanying parameters were investigated: F0, F1, F2, F3 Formants recurrence levels, level of voice break in isolated vowels, consistent, key recurrence, Fo (Hz), Jitter (recurrence bother – nearby %), Shimmer (abundancy annoyance – neighborhood %), Harmonic to commotion proportion (HNR – dB), Intensity(dB).

CHAPTER 5

5.1 Experimental Setup

The experiments were performed on a Microsoft© Windows® 10 machine with 8GB RAM, intel Core i5 processor, 32MB cache memory and 2GB Nvidia graphics.

Tools used: Python (v3.x) scripting language, Keras 2.1, matplotlib, sklearn 2.2, Tensorflow 1.7, speechpy, h5py 2.7.1, etc. This library provides most frequently used speech features including MFCCs and filter bank energies alongside with the log-energy of filter banks. It provides the building blocks necessary to create music information retrieval systems. Using the speechpy library we were able to extract features i.e MFCC (Mel Frequency Cepstral Coefficient). MFCC features are widely used in automatic speech or speaker recognition (see figure 14).

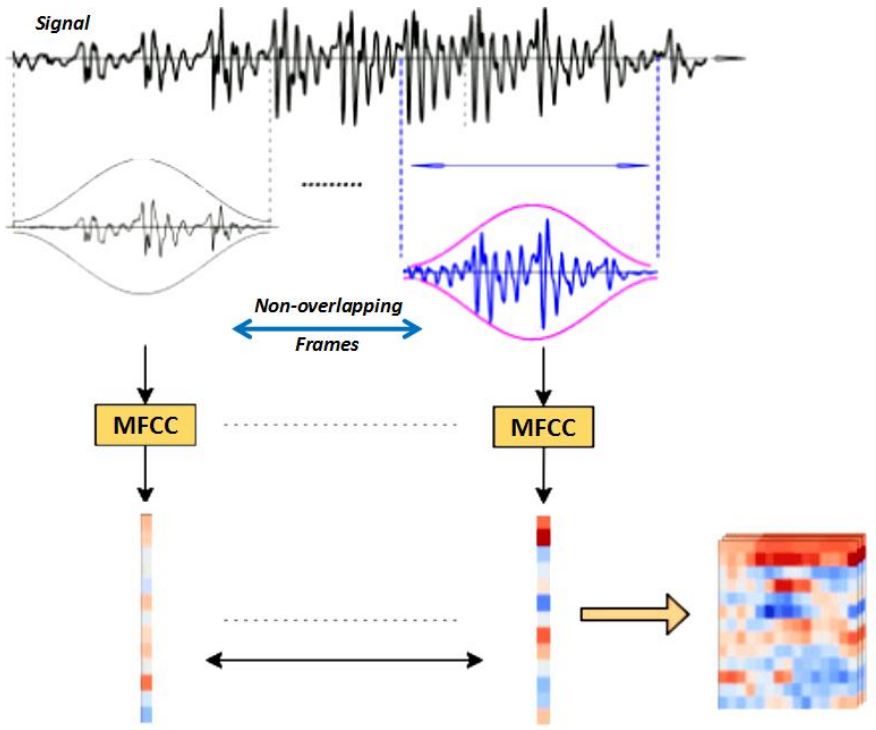

Figure 14: MFCC feature extraction mode

Each audio file gave us several features which were basically array of different values. These features were then appended by the labels which we created earlier.

In this work, we utilized a convolutional neural system classifier in python module. The system utilized in this trial was made out of 3 layers: the information layer, the concealed layer, and the yield layer (figure 15). The information layer takes the elucidating highlight esteems for every articulation. Information highlights were standardized to values in the scope of - 1 to 1. The shrouded layer in beginning stage has 20 hubs, and utilizations a sigmoid exchange work. The quantity of hubs in the yield layer relies upon what number of passionate classifications to perceive.

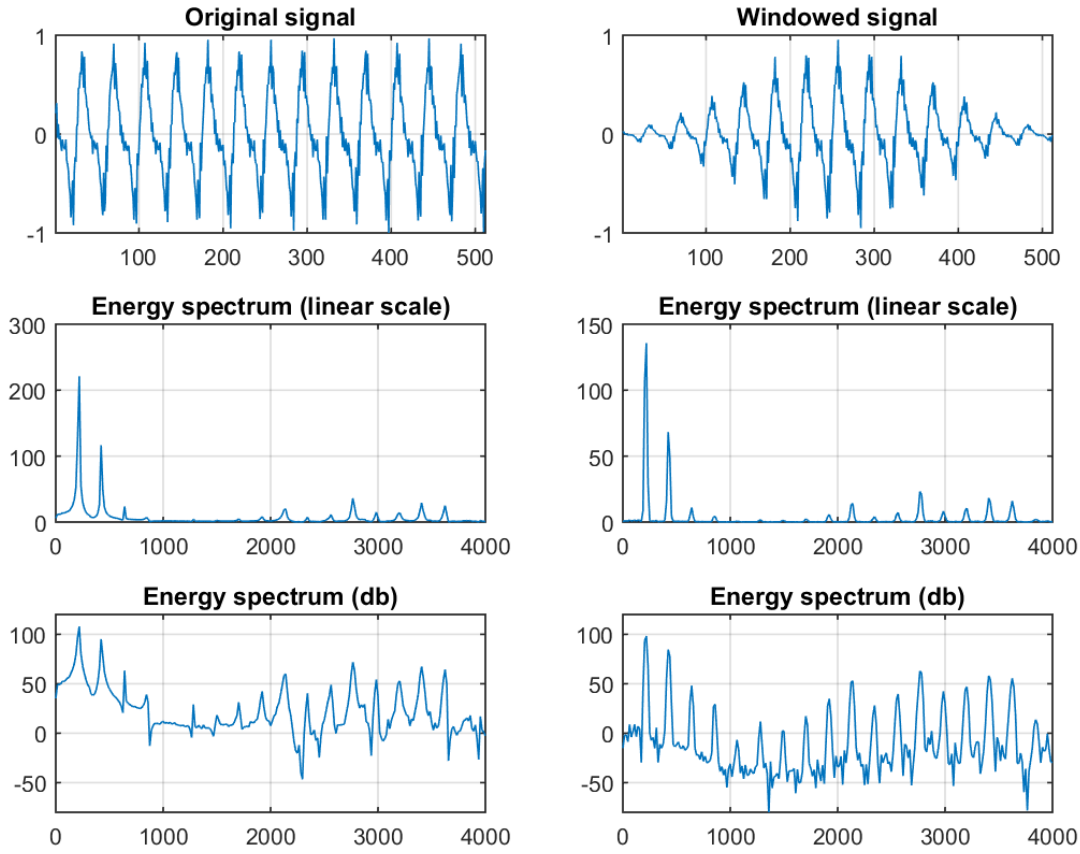

Figure 15: MFCCs energy spectrum

The next step involved dealing with the missing features for some audio files which were shorter in length. We increased the sampling rate by twice to get the unique

features of each emotional speech. We didn't increase the sampling frequency even more since it might collect noise thus affecting the results.

```
# extracting features from the audio files
# pad the signals to have same size if its less than required else slice them
            if s_len < mslen:
                pad_len = mslen - s_len
                pad_rem = pad_len % 2
                pad_len /= 2
                signal = np.pad(signal, (pad_len, pad_len + pad_rem), 'constant'
            else:
                pad_len = s_len - mslen
                pad_rem = pad_len % 2
                pad_len /= 2
                signal = signal[pad_len:pad_len + mslen]
            min_sample = min(len(signal), min_sample)
            mfcc = speechpy.feature.mfcc(signal, fs, num_cepstral=mfcc_len)

# flattening the data
            if flatten:
                mfcc = mfcc.flatten()
            data.append(mfcc)
```

Results

After training numerous models we were able to achieve an overwhelming best validation accuracy of 70% with 19 layers in our first test. Consistent results were obtained with softmax and rmsprop activation functions with batch size of 32 and 1000 epochs.

```
model = Sequential()

model.add(Conv1D(128, 5,padding='same',
                 input_shape=(216,1)))
model.add(Activation('relu'))
model.add(Conv1D(128, 5,padding='same'))
model.add(Activation('relu'))
model.add(Dropout(0.1))
model.add(MaxPooling1D(pool_size=(8)))
model.add(Conv1D(128, 5,padding='same',))
model.add(Activation('relu'))
model.add(Conv1D(128, 5,padding='same',))
model.add(Activation('relu'))
model.add(Conv1D(128, 5,padding='same',))
model.add(Activation('relu'))
model.add(Dropout(0.2))
model.add(Conv1D(128, 5,padding='same',))
model.add(Activation('relu'))
model.add(Flatten())
model.add(Dense(10))
model.add(Activation('softmax'))
opt = keras.optimizers.rmsprop(lr=0.00001, decay=1e-6)
```

```
# training the epochs
    best_acc = 0
    global x_train, y_train, x_test, y_test
    for i in tqdm(range(50)):

# shuffle the data for each epoch in unison
        p = np.random.permutation(len(x_train))
        x_train = x_train[p]
        y_train = y_train[p]
        model.fit(x_train, y_train, batch_size=32, epochs=1)
        loss, acc = model.evaluate(x_test, y_test)
        if acc > best_acc:    # calculate best fit/accuracy
            print ('Updated best accuracy', acc)
            best_acc = acc
            model.save_weights(best_model_path)
    model.load_weights(best_model_path)
    print ('Accuracy = ', model.evaluate(x_test, y_test)[1])
```

After training the model we had to predict the emotions on our test data (Table 1). The confusion matrix (Table 2) of the speech emotions, applying the proposed classifier to classify 5 emotions viz. happy, sad, calm, anger, fear using the combination of all features on the left. Finally, results of the experiments produced an accuracy of 89%, which are remarkably significant. Figure 16 shows the graph of emotion detection through this model. Figure 17 shows a summary of model training. The accuracy graph in figure shows various results of emotions detected from voice based on confusion matrix in table 2.

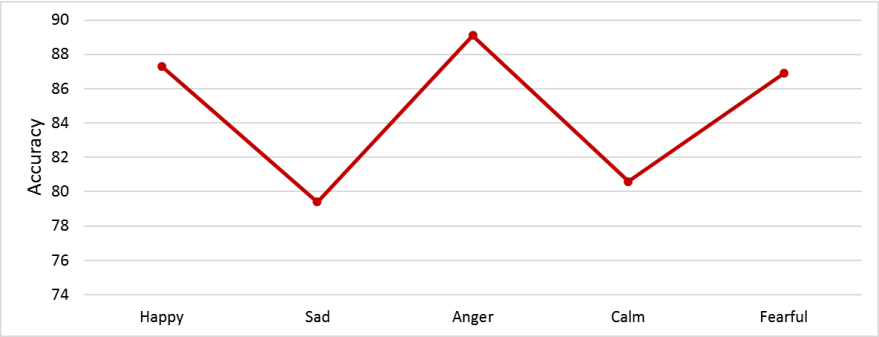

Figure 18: Accuracy graph plotted against the obtained results

Table 1: Results from the trained CNN model

Happy	Sad	Anger	Calm	Fearful
72	69	58	64	47

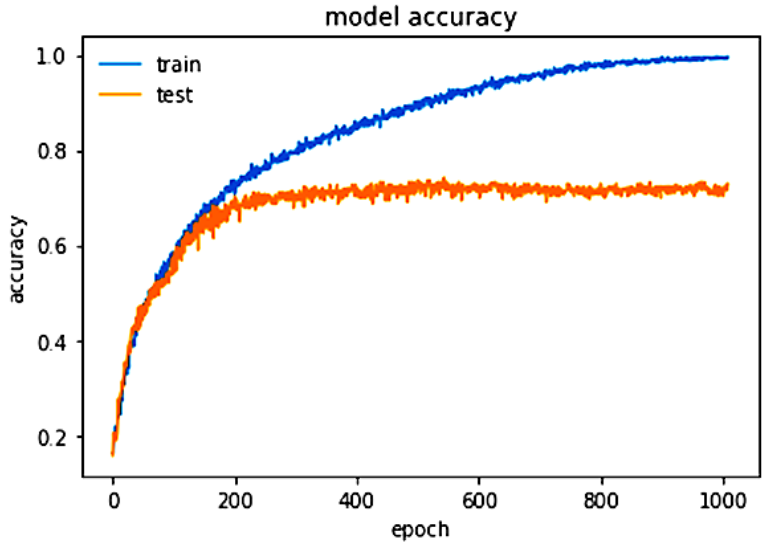

Figure 16: Accuracy graph of training with 1st test.

Table 2: Confusion matrix of the speech emotion analysis

	Happy	**Sad**	**Anger**	**Calm**	**Fearful**
Happy	63	0	13	16	8
Sad	2	55	4	34	5
Anger	10	5	62	13	10
Calm	6	27	1	51	15
Fearful	15	8	23	6	48

Layer (type)	Output Shape	Param #
conv1d_1 (Conv1D)	(None, 216, 128)	768
activation_1 (Activation)	(None, 216, 128)	0
conv1d_2 (Conv1D)	(None, 216, 128)	82048
activation_2 (Activation)	(None, 216, 128)	0
dropout_1 (Dropout)	(None, 216, 128)	0
max_pooling1d_1 (MaxPooling1	(None, 27, 128)	0
conv1d_3 (Conv1D)	(None, 27, 128)	82048
activation_3 (Activation)	(None, 27, 128)	0
conv1d_4 (Conv1D)	(None, 27, 128)	82048
activation_4 (Activation)	(None, 27, 128)	0
conv1d_5 (Conv1D)	(None, 27, 128)	82048
activation_5 (Activation)	(None, 27, 128)	0
dropout_2 (Dropout)	(None, 27, 128)	0
conv1d_6 (Conv1D)	(None, 27, 128)	82048
activation_6 (Activation)	(None, 27, 128)	0
flatten_1 (Flatten)	(None, 3456)	0
dense_1 (Dense)	(None, 10)	34570
activation_7 (Activation)	(None, 10)	0

Total params: 445,578
Trainable params: 445,578
Non-trainable params: 0

Figure 17: Summary of the model training

Since the corpus utilized in our test is generally little, a 10-crease cross approval procedure was connected to expand the unwavering quality of the outcomes. We split the information into two sets; 80% of which were utilized in the instructional meeting, 10% for the preparation approval and the other 10% for the testing reason. We rehash multiple times and utilize diverse one-tenth subsets of the information for testing and take a mean precision. The approval information utilized in preparing is to avert overfitting.

CHAPTER 6

Discussions and Conclusion

Automatic speech emotion recognition is an emerging field and is increasing now a day which it achieved through results in the better human and machine interaction. Deduction of human emotions through voice and speech analysis has a practical plausibility and could potentially be beneficial for improving human conversational and persuasion skills. Small, compact CNN classifiers that were highly optimised using a feature rich library search algorithm were successfully implemented. This paper presented a systematic approach for detection and analysis of human emotions based on the emotions hidden in voice through speech processing. While conducting the proposed study, speech data were sampled at the rate of 10 sample points per second. However, in the beginning of the training process the accuracy were low but as the process continued the test samples achieved an accuracy up to 89% which was quite overwhelming. In order to improve the precision of the mean attribute values, greater number of sampling points can be considered. In conclusion, this project shows that building a fast and efficient speech emotion classifier is a challenging work but can be achieved with proper efforts and little patience.

Future Work and Scope

There is a lot of work on emotional intelligence, and there is likewise independent work on separating other data like age, sexual orientation and so forth. In any case, it has been demonstrated that the voice highlights continue changing by age. Thus, for various sexes the feeling coordinating parameters ought to appear as something else. It tends to be felt effortlessly that when we hear a sound, first thing comes in our mind whether the speaker is male or a female, at that point we gauge the period of individual, at that point we surmise the importance and feeling moving through the voice. There are diverse physiological viewpoints identified with the both sex and comparative is the situation with the time of individual. Every one of these assessments are comprehended inside a flicker of eye inside human cerebrum. In any case, the machines are diverse along these lines, the machine should be prepared

to separate between the sexual orientation and the age gatherings. On the off chance that a woman yells, it indicates outrage of dread, however this equivalent perception can't be connected to the yelling child.

There is a lot of scope of using all the works combined to increase the accuracies of speech emotion detection by the machine. Further, we tend to develop an automated speech recognition system, which will be helpful to people with disorders or find it difficult to understand and identify emotions, for example people suffering from autism.

REFERENCES

1. Yu F, Chang E, Xu YQ, Shum HY. Emotion detection from speech to enrich multimedia content. In Pacific-Rim Conference on Multimedia. Springer, Berlin, Heidelberg, 2001;550-557.
2. Formisano E, De Martino F, Bonte M, Goebel R. Who is saying what? Brain-based decoding of human voice and speech. Science. 2008;322(5903):970-3.
3. Scherer KR. Vocal markers of emotion: Comparing induction and acting elicitation. Computer Speech & Language. 2013;27(1):40-58.
4. Sataloff RT. The human voice. Scientific American. 1992;267(6):108-15.
5. Goberman AM, Blomgren M. Parkinsonian speech disfluencies: effects of L-dopa-related fluctuations. Journal of fluency disorders. 2003;28(1):55-70.
6. Harel BT, Cannizzaro MS, Cohen H, Reilly N, Snyder PJ. Acoustic characteristics of Parkinsonian speech: a potential biomarker of early disease progression and treatment. Journal of Neurolinguistics. 2004;17(6):439-53.
7. Lovestone S, Francis P, Kloszewska I, Mecocci P, Simmons A, Soininen H, Tsolaki M. AddNeuroMed-the European collaboration for the discovery of novel biomarkers for Alzheimer's disease. Annals of the New York Academy of Sciences. 2009;1180(1):36-46.
8. Vanello N, Guidi A, Gentili C, Werner S, Bertschy G, Valenza G. Speech analysis for mood state characterization in bipolar patients. Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2012;2104-2107.
9. Novotný M, Rusz J, Čmejla R, Růžička E. Automatic evaluation of articulatory disorders in Parkinson's disease. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP). 2014;22(9):1366-78.

10. Ringeval F, Chetouani M, Schuller B. Novel metrics of speech rhythm for the assessment of emotion. In Thirteenth Annual Conference of the International Speech Communication Association 2012;2763-2766.

11. Dellaert F, Polzin T, Waibel A. Recognizing emotion in speech. In Fourth International Conference on Spoken Language Processing, 1996.

12. Forbes-Riley K, Litman D. Predicting emotion in spoken dialogue from multiple knowledge sources. In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL, 2004.

13. Steidl S, Levit M, Batliner A, Noth E. Of all things the measure is man automatic classification of emotions and inter-labeler consistency. In IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. 2005;1:I-317.

14. Horwitz R, Quatieri TF, Helfer BS, Yu B, Williamson JR, Mundt J. On the relative importance of vocal source, system, and prosody in human depression. In 2013 IEEE International Conference on Body Sensor Networks (BSN), 2013;1-6).

15. Trevino AC, Quatieri TF, Malyska N. Phonologically-based biomarkers for major depressive disorder. EURASIP Journal on Advances in Signal Processing. 2011;(1):42.

16. Quatieri Jr TF, Malyska N, Trevino AC, inventors; Massachusetts Institute of Technology, assignee. Phonologically-based biomarkers for major depressive disorder. United States patent US 9,936,914. 2018.

17. Alghowinem S, Goecke R, Wagner M, Epps J, Breakspear M, Parker G. Detecting depression: a comparison between spontaneous and read speech. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013;7547-7551.

18. Gharavian D, Sheikhan M, Nazerieh A, Garoucy S. Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy

ARTMAP neural network. Neural Computing and Applications. 2012;21(8):2115-26.

19. Pijanowski BC, Villanueva-Rivera LJ, Dumyahn SL, Farina A, Krause BL, Napoletano BM, Gage SH, Pieretti N. Soundscape ecology: the science of sound in the landscape. BioScience. 2011;61(3):203-16.

20. Graves A, Mohamed AR, Hinton G. Speech recognition with deep recurrent neural networks. In 2013, IEEE international conference on Acoustics, speech and signal processing (ICASSP). 2013;6645-6649.

21. Hollien H, Mendes-Schwartz AP, Nielsen K. Perceptual confusions of high-pitched sung vowels. Journal of Voice. 2000;14(2):287-98.

22. Ladefoged P, Disner SF. Vowels and consonants. John Wiley & Sons; 2012;40.

23. Hawkins S, Midgley J. Formant frequencies of RP monophthongs in four age groups of speakers. Journal of the International Phonetic Association. 2005;35(2):183-99.

24. El Ayadi M, Kamel MS, Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition. 2011;44(3):572-87.

25. Kessous L, Castellano G, Caridakis G. Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. Journal on Multimodal User Interfaces. 2010;3(1-2):33-48.

26. Lalitha S, Madhavan A, Bhushan B, Saketh S. Speech emotion recognition. In 2014, IEEE International Conference on Advances in Electronics, Computers and Communications (ICAECC), 2014;1-4.

27. Schuller B, Rigoll G, Lang M. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. IEEE Proceedings. (ICASSP'04). 2004;1,:I-577.

28. Ma EP, Yiu EM. Multiparametric evaluation of dysphonic severity. Journal of Voice. 2006;20(3):380-90.

29. Lahmiri S. Parkinson’s disease detection based on dysphonia measurements. Physica A: Statistical Mechanics and its Applications. 2017;471:98-105.

30. Dixit VM, Sharma Y. Voice Parameter Analysis for the disease detection. IOSR Journal of Electronics and Communication Engineering, eISSN. 2014:2278-834.

31. Little MA, McSharry PE, Hunter EJ, Spielman J, Ramig LO. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE transactions on biomedical engineering. 2009;56(4):1015-22.

32. Cohen SM, Kim J, Roy N, Asche C, Courey M. Direct health care costs of laryngeal diseases and disorders. The Laryngoscope. 2012;122(7):1582-8.

33. Lihite RJ, Choudhury U, Surender G, Pal B, Lahkar M. An Early Sign of Wilson’s Disease: Dysarthria. Journal of clinical and diagnostic research: JCDR. 2014;8(3):188.

34. Hadjitodorov S, Mitev P. A computer system for acoustic analysis of pathological voices and laryngeal diseases screening. Medical engineering & physics. 2002;24(6):419-29.

35. Shirvan RA, Tahami E. Voice analysis for detecting Parkinson's disease using genetic algorithm and KNN classification method. In18th Iranian conference on biomedical engineering 2011;1416:278-283.

36. Cuetos F, Arango-Lasprilla JC, Uribe C, Valencia C, Lopera F. Linguistic changes in verbal expression: a preclinical marker of Alzheimer's disease. Journal of the International Neuropsychological Society. 2007;13(3):433-9.

37. Murty KS, Yegnanarayana B. Epoch extraction from speech signals. IEEE Transactions on Audio, Speech, and Language Processing. 2008;16(8):1602-13.

38. Hickok G, Poeppel D. The cortical organization of speech processing. Nature Reviews Neuroscience. 2007;8(5):393.

39. Sahidullah M, Saha G. Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. Speech Communication. 2012;54(4):543-65.

40. Hollander M, Wolfe DA, Chicken E. Nonparametric statistical methods. John Wiley & Sons. 2013;751.

41. Yang S, Zheng F, Luo X, Cai S, Wu Y, Liu K, Wu M, Chen J, Krishnan S. Effective dysphonia detection using feature dimension reduction and kernel density estimation for patients with Parkinson's disease. PloS one. 2014;9(2):e88825.

42. Juslin PN, Laukka P. Communication of emotions in vocal expression and music performance: Different channels, same code?. Psychological bulletin. 2003;129(5):770.

43. Bhardwaj P, Debbarma S. A Study of Methods Involved In Voice Emotion Recognition. International Journal of Advanced Research in Computer and Communication Engineering. 2014;3(2):2278-1021.

44. Lin YL, Wei G. Speech emotion recognition based on HMM and SVM. In Proceedings of 2005 International Conference on Machine Learning and Cybernetics, 2005;8:4898-4901.

45. Han W, Chan CF, Choy CS, Pun KP. An efficient MFCC extraction method in speech recognition. In 2006 IEEE International Symposium on Circuits and Systems, (ISCAS 2006). Proceedings. 2006;4.

46. Sapra A, Panwar N, Panwar S. Emotion recognition from speech. International journal of emerging technology and advanced engineering. 2013;3:341-5.

47. Schuller B, Rigoll G, Lang M. Hidden Markov model-based speech emotion recognition. In 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003;2, pp. II-1.

48. Zhao L, Han Z. Speech recognition system based on integrating feature and HMM. In 2010 International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), 2010;3:449-452.

49. Rangayyan RM, Wu Y. Screening of knee-joint vibroarthrographic signals using probability density functions estimated with Parzen windows. Biomedical Signal Processing and Control. 2010;5(1):53-8.

50. Bhuta T, Patrick L, Garnett JD. Perceptual evaluation of voice quality and its correlation with acoustic measurements. Journal of voice. 2004;18(3):299-304.

51. Mundt JC, Snyder PJ, Cannizzaro MS, Chappie K, Geralts DS. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. Journal of neurolinguistics. 2007;20(1):50-64.

52. Alpert M, Pouget ER, Silva RR. Reflections of depression in acoustic measures of the patient's speech. Journal of affective disorders. 2001;66(1):59-69.

53. Lee CM, Narayanan SS. Toward detecting emotions in spoken dialogs. IEEE transactions on speech and audio processing. 2005;13(2):293-303.

54. Ho AK, Iansek R, Marigliani C, Bradshaw JL, Gates S. Speech impairment in a large sample of patients with Parkinson's disease. Behavioural neurology. 1999;11(3):131-7.

55. Duffy JR. Motor Speech Disorders: Substrates, Differential Diagnosis, and Management 2nd edition (St Louis, MO: Mosby). 2005;1-592.

56. Harel B, Cannizzaro M, Snyder PJ. Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: A longitudinal case study. Brain and cognition. 2004;56(1):24-9.

57. López-de-Ipiña K, Alonso JB, Travieso CM, Solé-Casals J, Egiraun H, Faundez-Zanuy M, Ezeiza A, Lizardui UM. On the selection of non-invasive methods based on speech analysis oriented to automatic Alzheimer disease diagnosis. Sensors. 2013;13(5):6730-45.

58. Kiss G, Santen JP, Prud'Hommeaux E, Black LM. Quantitative analysis of pitch in speech of children with neurodevelopmental disorders. InThirteenth Annual Conference of the International Speech Communication Association 2012;1342-1345.

59. Ringeval F, Chetouani M, Schuller B. Novel metrics of speech rhythm for the assessment of emotion. InThirteenth Annual Conference of the International Speech Communication Association 2012;2763-2766.

60. Ringeval F, Chetouani M, Schuller B. Novel metrics of speech rhythm for the assessment of emotion. InThirteenth Annual Conference of the International Speech Communication Association 2012;2763-2766.

61. J. Holmes R, M. Oates J, J. Phyland D, J. Hughes A. Voice characteristics in the progression of Parkinson's disease. International Journal of Language & Communication Disorders. 2000;35(3):407-18.

62. Goberman A, Coelho C, Robb M. Phonatory characteristics of parkinsonian speech before and after morning medication: the ON and OFF states. Journal of Communication Disorders. 2002;35(3):217-39.

63. Van Lancker Sidtis D, Hanson W, Jackson C, Lanto A, Kempler D, Metter EJ. Fundamental frequency (F0) measures comparing speech tasks in aphasia and Parkinson disease. Journal of Medical Speech-Language Pathology. 2004;12(4):207.

64. Rahn III DA, Chou M, Jiang JJ, Zhang Y. Phonatory impairment in Parkinson's disease: evidence from nonlinear dynamic analysis and perturbation analysis. Journal of Voice. 2007;21(1):64-71.

65. Vaziri G, Almasganj F, Behroozmand R. Pathological assessment of patients' speech signals using nonlinear dynamical analysis. Computers in biology and medicine. 2010;40(1):54-63.

66. Elvevåg B, Foltz PW, Weinberger DR, Goldberg TE. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. Schizophrenia research. 2007;93(1-3):304-16.

67. Elvevåg B, Foltz PW, Rosenstein M, DeLisi LE. An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. Journal of neurolinguistics. 2010;23(3):270-84.

68. Bedi G, Carrillo F, Cecchi GA, Slezak DF, Sigman M, Mota NB, Ribeiro S, Javitt DC, Copelli M, Corcoran CM. Automated analysis of free speech predicts psychosis onset in high-risk youths. npj Schizophrenia. 2015;1:15030.

69. Van Santen JP, Prud'hommeaux ET, Black LM. Automated assessment of prosody production. Speech communication. 2009;51(11):1082-97.

70. Little MA, McSharry PE, Hunter EJ, Spielman J, Ramig LO. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE transactions on biomedical engineering. 2009;56(4):1015-22.

71. Bache K, Lichman M. UCI Machine Learning Repository (http://archive. ics. uci. edu/ml), University of California, School of Information and Computer Science. Irvine, CA. 2013.

72. Alghowinem S, Goecke R, Wagner M, Breakspear M, Parker G. A comparative study of different classifiers for detecting depression from spontaneous speech. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013;8022-8026.

73. Low LS, Maddage NC, Lech M, Sheeber L, Allen N. Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents. In 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), 2010;5154-5157.

74. Vogel AP, Fletcher J, Maruff P. Acoustic analysis of the effects of sustained wakefulness on speech. The Journal of the Acoustical Society of America. 2010;128(6):3747-56.

75. Ellgring H, Scherer KR. Vocal indicators of mood change in depression. Journal of Nonverbal Behavior. 1996;20(2):83-110.

76. Mundt JC, Snyder PJ, Cannizzaro MS, Chappie K, Geralts DS. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. Journal of neurolinguistics. 2007;20(1):50-64.

77. Trevino AC, Quatieri TF, Malyska N. Phonologically-based biomarkers for major depressive disorder. EURASIP Journal on Advances in Signal Processing. 2011;2011(1):42.

78. Novotný M, Rusz J, Čmejla R, Růžička E. Automatic evaluation of articulatory disorders in Parkinson's disease. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP). 2014;22(9):1366-78.

79. Skodda S, Schlegel U. Speech rate and rhythm in Parkinson's disease. Movement disorders: official journal of the Movement Disorder Society. 2008;23(7):985-92.