World Journal of Engineering Research and Technology
WJERT
www.wjert.org
SJIF Impact Factor: 4.326

EMOTION SPEECH RECOGNITION SYSTEM FOR ISOLATED WORDS

Dhanashri Rachetti*[1], Shital Sukre[1], Asmita Gajarishi[2], Sagar Shirke[2] and Ritesh Thakur[2]

[1]Department of Computer Engineering, Savitribai Phule Pune University, IOK College of Engineering Pune, India.

[2]Department of Computer Engineering IOK College of Engineering Pune India-412 208.

Article Received on 29/03/2017 Article Revised on 20/04/2017 Article Accepted on 11/05/2017

***Corresponding Author**
Dhanashri Rachetti
Department of Computer Engineering, Savitribai Phule Pune University, IOK College of Engineering Pune, India.

ABSTRACT

Recognition of emotions from speech is one of the most important sub domains in the field of affective computing. Sometimes, a person speaks the sentence while stay in some emotion which makes the tone of speech changes the meaning of the sentence completely. Speech signal consist not only the words and meaning but also it consist of the emotions. The emotion expressed by speech is one of the major influencing factors for the low recognition accuracy achieved during the development of speech based systems. When it comes to human speech emotions affects the tone and the speaking style of the person. The research in this area is needed to overcome these problems of emotion recognition from speech. However the problem is usually deals with the following basic emotion categories: Happy, Sad, Angry, Afraid, Surprise, Neutral. From the literature survey for the proposed study, it is observed that there is no proper emotional speech corpus in any of the Indian languages for carrying out the research on emotional speech processing. No any standard emotional speech database as well as the real life emotional speech database is available in the context of Indian languages. It is also observed from the literature that excitation source information is not thoroughly investigated for the purpose of emotion recognition task. Most of the researchers have used frame-wise spectral features extracted from entire utterance for speech emotion classification. Most of the existing emotion

recognition systems are developed using only gross prosodic features extracted from the entire utterances.

KEYWORDS: Speech recognition system, Emotion recognition system, HTK, Speech Database.

INTRODUCTION

Marathi is an Indo-Aryan language spoken by about 71 million people mainly in the Indian state of Maharashtra and neighbouring states. Marathi is also spoken in Israel and Mauritius. Marathi is thought to be a descendent of Maharashtri, one of the Prakrit languages which developed from Sanskrit. Marathi first appeared in writing during the 11th century in the form of inscriptions on stones and copper plates. From the 13th century until the mid 20th century, it was written with the Modi alphabet. Since 1950 it has been written with the Devanagari alphabet. There are 13 vowels and 36 consonants in Marathi language. The vowels and consonants along with their transliteration and International Phonetics Alphabetic are shown in figure 1. The skill to recognize, interpret and express the emotion referred to emotional intelligent. The emotions recognition and expression are used in human computer interfacing.[1] The speech recognition understands basically what someone speak to a computer, asking a computer to translate speech into its corresponding textual message, where as in speech synthesis, a computer generate artificial spoken dialogs. Speech is the most prominent and natural form of communication between the humans. Speech would thus be a logical choice for man machine communication; hence there is growing interest in developing such machines that can accept speech as input. Given the substantial research efforts in speech recognition worldwide and the steady rate at which computers become faster and smaller.[2,3,4,5,6] The machine that accepts speech as.

अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ओ	औ
A	Āa	I	Ī	U	Ū	ṛ	e	Ai	O	Au
/ə/	/a/	/i/		/u/		/ru/	/e/	/əi/	/o/	/əu/
अं						अः				
aṃ						aḥ				
/ə^{n}/						/əh/				

Fig.1 Vowels in Marathi language along with transliteration and IPA.

क	KA	/KƏ/	ट	ṭa	/ʈə/	प	PA	/PƏ/	ष	ṣa	/ʂə/
ख	KHA	/KʰƏ/	ठ	ṭha	/ʈʰə/	फ	PHA	/FƏ/	स	sa	/sə/
ग	GA	/gƏ/	ड	ḍa	/ɖə/	ब	BA	/BƏ/	ह	ha	/hə/
घ	GHA	/gʱƏ/	ढ	ḍha	/ɖʱə/	भ	BHA	/BʱƏ/	ळ	ḷa	/ɭə/
ङ	Ṅa	/ŋƏ/	ण	ṇa	/ɳə/	म	MA	/MƏ/	क्ष	kṣa	/kʃə/
च	Ca	/tsə/	त	Ta	/t̪ ə/	य	Ya	/jə/	ज्ञ	jña	/ɟɲə/
छ	Cha	/tsʰə/	थ	Tha	/t̪ ʰə/	र	Ra	/rə/			
ज	Ja	/zə/	द	Da	/d̪ ə/	ल	La	/lə/			
झ	Jha	/zʱə/	ध	Dha	/d̪ ʱə/	व	Va	/wə/			
ञ	ṅa	/ŋə/	न	Na	/n̪ ə/	श	Śa	/ʃə/			

Fig 2: consonants in Marathi Language along with Transliteration and IPA.

Input requires generally two stage interfacing. The first step requires an automatic recognition system (ASR) and the second step requires a system for speech understanding. The speech recognition systems that are on the market today are the embodiments of new algorithms that were once the province of the advanced laboratories. Many groups in India have also been engaged in speech related work like Tata Institute of Fundamental Research, Mumbai, Computer Vision and Pattern Recognition Unit at Indian Statistical Institute Kolkata, C-DAC Pune, Indian Institute of Technology, Madras, Indian Institute of Technology, Kanpur and BAMU.[7] There are the different ways to express the emotions by humans. Humans express their emotions by speech and actions like crying, yelling, dancing, laughing, stamping, and many other things.[8] But when it comes to speech human emotions affects the tone and the speaking style of the person. The emotion in speech sound affects the Speech recognition accuracy. The researchers around the globe are taking interest in detection of emotion in the Speech. In human computer interaction, many researchers are finding the depth of the area for emotion detection from speech. During the last few years, the research on speech emotion recognition has got much attention. Many emotional speech databases have been developed and the studies are carried on the developed emotional speech database around the world.[9]

Development of Artificial Emotional Marathi Speech Database

We developed an emotional speech database using Marathi movies. For that purpose, we selected 3 Basic emotions i.e. Happy, Sad, Angry. . It was decided to develop an artificial (from actors' actresses dialogue) emotional speech database using various Marathi movies in which the professional artists simulate the natural emotions. Capturing the natural emotions is not possible. The movies were selected from different genre like drama, comedy & horror.

The audio stream of the movie was sampled at 44 kHz. The selected Words were first viewed and checked for the expressed emotions. The data samples were extracted from the Words using Audacity 2.1.3. As mentioned earlier the audio stream of the Speech was sampled at 16 kHz so we Increase sampled the audio to 44 kHz .Once the data samples were acquired we categorized the samples according to the emotions i.e. Happy, Sad, Anger, . The extracted data samples were saved in .wav file format. We collected speech samples from 30 speakers. The 30 speakers were categorized according to the gender. We have collected the data from 15 males and 15 females in the age group of 21 to 40. The database consists of 150 utterances of each word selected from the list of isolated words. The database consists of in all 2250 utterances of 15 emotional words in Marathi language.

Table 1: Happy words in Marathi Language along with Transliteration and IPA.

Devanagari	Transliterated (Translated In English)	IPA
गप रे	Gapp re (Just Shutup)	/gəpə/ /rəe/
चल निघ	Chal Nigh (Get out)	/tʃələ/ /n̪ əigʱə/
व्हय्घरी	Vhay Ghari (Go to your home)	/ ʋəhəjəgʱərəi/
हट्ट	Hutt (Leave me alone)	/həʈʈə/
मुस्काड फोडीन	Muskaad Fodin (Will Slap your face)	/məusəkaɖə//p ʰəoɖəin̪ ə/

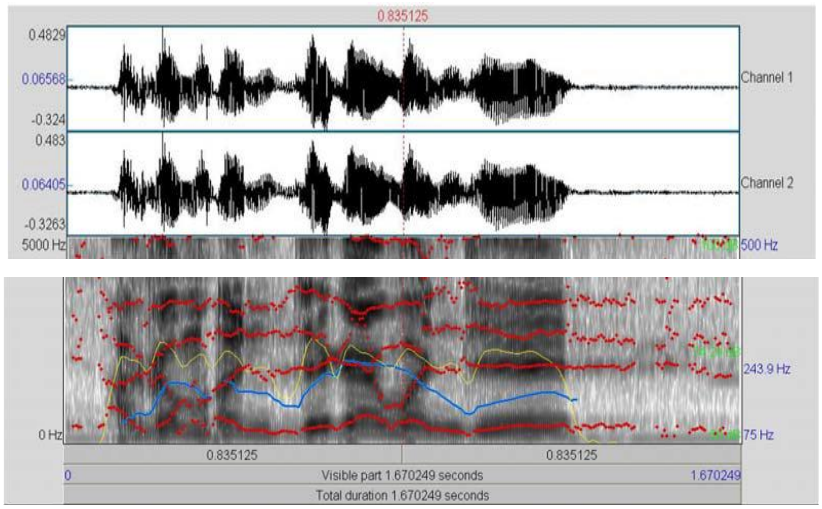
Table 2: Sad words in Marathi Language along with Transliteration and IPA.

Devanagari	Transliterated (Translated In English)	IPA
अरेव्वा	Arewaa (Oh Good)	/ ərevva /
कितीछान	Kiti Chan (How good)	/ kiʈ̪ i/tʃʰəan̪ ə /
कितीगोड	Kiti Goad (How Sweet)	/ kiʈ̪ i/gəoɖə/
मस्त	Mast (Good)	/məsəʈ̪ ə/
खतरनाक	Khatarnaak (Fantastic)	/kʰəʈ̪ ərən̪ əakə/

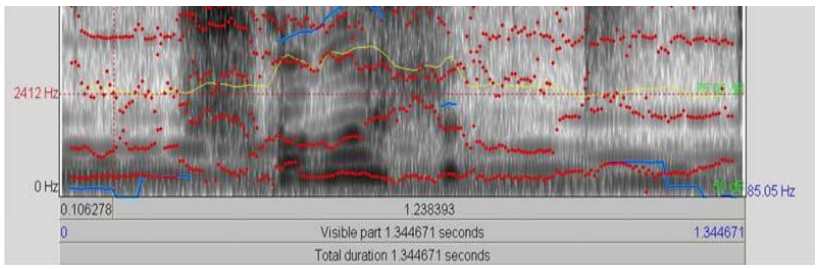
Table 3: Angry words in Marathi Language along with Transliteration and IPA.

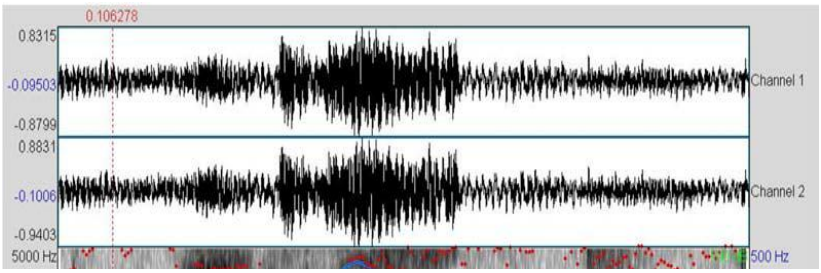
Devanagari	Transliterated (Translated In English)	IPA
आरे देवा	Are Dewa (Oh God)	/əre/ /d̪ eva/
अरेरे	Arere (Ohh)	/ ərere/
अबब	Ababa (Ohh)	/əbəbə/
आई गं	Aai ga (Oh mother)	/ai//gə/
नाही ग जमत	Nahi ga jamat (Its not possible)	/n̪ əahəi//gə/dʒ əməʈ̪ ə/

The table I, II and III represents the words selected for the development of the isolated word emotional speech database along with the respective transliterations and IPA (International Phonetic Alphabet).

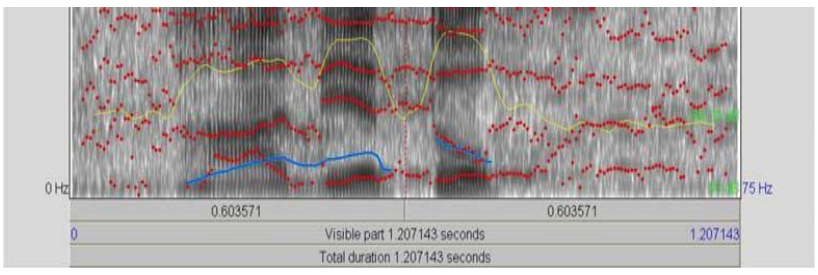


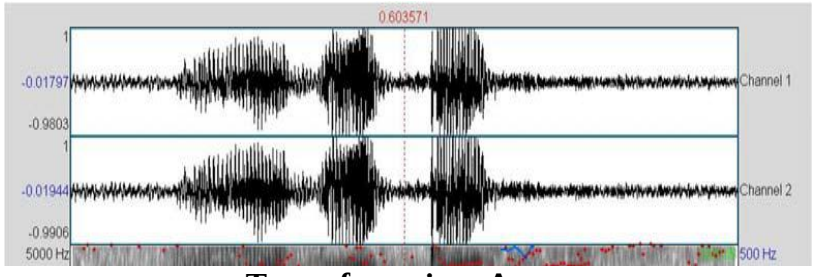
Type of emotion: Happy





Type of emotion: Sad





Type of emotion: Anger

Fig 3: Speech Samples

Emotional Features

Feature extraction is a basic and fundamental pre processing step in pattern recognition and machine learning. It is a special form of dimensionality reduction technique used to reduce the data which is very large to be processed by an algorithm and extraction of specific properties from various features. In feature extraction, the provided input data is transformed into a set of features which provides the relevant information for performing a desired task without the need of the full size data but using the reduced set.[11] The extraction of the best parametric representation of acoustic signals is an important task to produce a better recognition performance. To select suited features carrying information about emotion is necessary for emotion recognition. Studies on emotion of speech indicate that pitch, energy, formant, Feature extraction is process of extracting some valuable parameters for further processing of the input signals. We have studied various features extraction techniques as given below. The prosodic features such as mean, maximum, minimum, standard – deviation of pitch and energy, and audible durations are extracted from four basic classes of emotions namely anger, sadness, happiness and neutral. These features are classified by fuzzy min-max neural network.[12] Kai-Tai Song, et. at. proposed a method of emotion recognition. Firstly, end-point detection and frame setting are done in pre-processing. Secondly, statistical features from pitch and energy are computed. Theses statistical features are classified by using support vector machine.[13] Stevros Ntalampiras and Nikus Fakutakis have used short-term statistics, spectral moments, and autoregressive models as a emotional features. Additionally, they have employed a newly introduced groups of parameters based on wavelet decomposition.[14] It is shown from the above review spectral features are giving more recognition compatibly prosodic features of emotional speech. So we have decided HTKs as features.

Hidden Markov Model ToolKit (HTKs)

Hidden Markov Model (HMM) (Rabiner, 1989) is a doubly stochastic process with one that is not directly observable.This hidden stochastic process can be observed only through another set of stochastic processes that can produce the observation sequence. HMMs are the so far most widely used acoustic models. The reason is just it provides better performance than other methods. HMMs are widely used for both training and recognition of speech system.HMM are statistical frameworks, based on the Markov chain with unknown parameters. Hidden Markov Model is a system which consists of nodes representing hidden states. The nodes are interconnected by links which describes the conditional transition

probabilities between the states. Each hidden state has an associated set of probabilities of emitting particular visible states. HTK is a toolkit for building Hidden Markov Models (HMMs). It is an open source set of modules written in ANSI C which deal with speech recognition using the Hidden Markov Model. HTK mainly runs on the Linux platform. However, to run it on Windows, interfacing package Cygwin (Cygwin, 2011) is used.

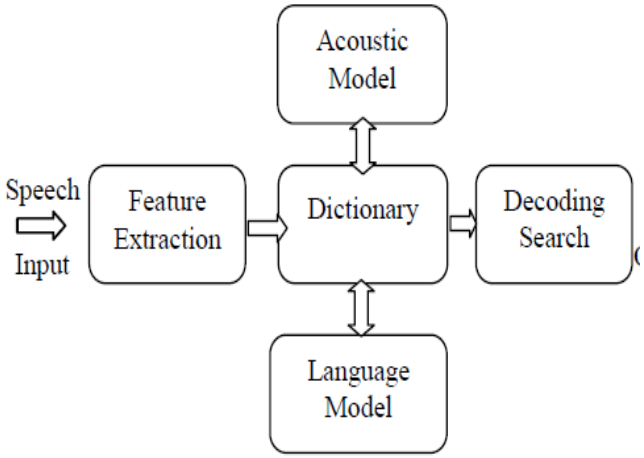
A) *Acoustic Model*

In a statistical framework for speech recognition, the problem is to find the most likely word sequence, which can be described by the equation

$$\hat{W}=\text{argwmax}P(W/X)$$ Applying the Bayes

The term $P(X/W)$ in the above equation can be realized by the Acoustic model. An acoustic model is a file that contains a statistical representation of each distinct sound that makes up a spoken word. It contains the sounds for each word found in the Language model. The speech recognition.

System implemented here uses Hidden Markov Models (HMM) for representing speech sounds. A HMM is a stochastic model. A HMM consists of a number of states, each of which is associated with a probability density function. The model parameters are the set of probability density functions, and a transition matrix that contains the probability of transitions between states. HMM-based recognition algorithms are classified into two types, namely, phoneme level model and word-level model. The word-level HMM has excellent performance at isolated word tasks and is capable of representing speech transitions between phonemes. However, each distinct word has to be represented by a separate model which leads to extremely high computation cost (which is proportional to the number of HMM models). The phoneme model on the other hand can help reproduce a word as a sequence of phonemes. Hence new words can be added to the dictionary without necessitating additional models.[15] Hence phoneme model is considered more suitable in applications with large sized vocabularies and where addition of word is a essential possibility In order to recognize speech, the system usually consists of two phases. They are called pre-processing and post-processing. Pre-processing involves feature extraction and the post-processing stage comprises of building a speech recognition engine. Speech recognition engine usually consists of knowledge about building an acoustic model, dictionary and grammar. Once all these details are given correctly, the recognition engine identifies the most likely match for the given input, and it returns the recognized word.



1) Pre-emphasis: The speech signal s(n) is sent to a high-pass filter,s2(n) = s(n) - a*s(n-1),where s2(n) is the output signal and the value of a is usually between 0.9 and1.0. In our research we used a = 0.98. The goal of pre-emphasis is to compensate the high frequency part that was suppressed during the sound production mechanism of humans. The speech after pre-emphasis sounds became sharper with a smaller volume.

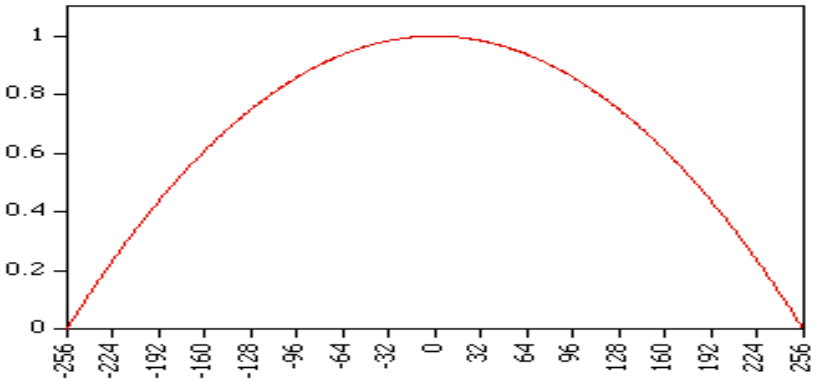
2) ***Frame Blocking:*** The input speech signal is segmented into frames of 20~30 Ms with optional overlap of 1/3~1/2 of the frame size. Usually the frame size (in terms of sample points) is equal power of two in order to facilitate the use of FFT. If this is not the case, we need to do zero padding to the nearest length of power of two.

3) ***Fast Fourier Transform (FFT):*** A Fast Fourier transform (FFT) is an efficient algorithm to compute the Discrete Fourier Transform (DFT) and its inverse. An FFT computes the DFT and produces exactly the same result as evaluating the DFT definition directly; the only difference is that an FFT is much faster. Let x0, ...,xN-1 be complex numbers. Evaluating this definition directly requires O(N2) operations: there are N outputs Xk, and each output requires a sum of N terms. An FFT is any method to compute the same results in O(N log N) operations. In our proposed algorithm, a well-known split radix FFT (RS-FFT) algorithm was used. This is a divide and conquer algorithm that recursively breaks down a DFT of any composite size N = N1N2 into many smaller DFTs of sizes N1 and N2,along with O(N) multiplications by complex roots of unity traditionally called twiddle factors, When we perform FFT on a frame, we assume that the signal within a frame is periodic, and Continuous. If this is not the case, we can still perform FFT but the in continuity at the frame's first and last points is likely to introduce undesirable effects in the frequency response which is known as spectral leakage.

4) ***Windowing:*** In windowing, each data frame has to be multiplied with a window function in order to keep the continuity of the first and the last points in the frame. If the signal in a frame is denoted by s(n), n= 0,…N-1, then the signal after windowing iss(n)*w(n), where w(n) is the window function. In our research, we used different types of window functions, such as Hamming, Hanning, Rectangular, Bohman, Triangle, Welch, Kaiser and Blackman windows.

Welch window

$$welch(i) = 1 - \left[\frac{i - N/2}{N/2}\right]^2$$



3 Commonly used as a window for power spectral estimation.

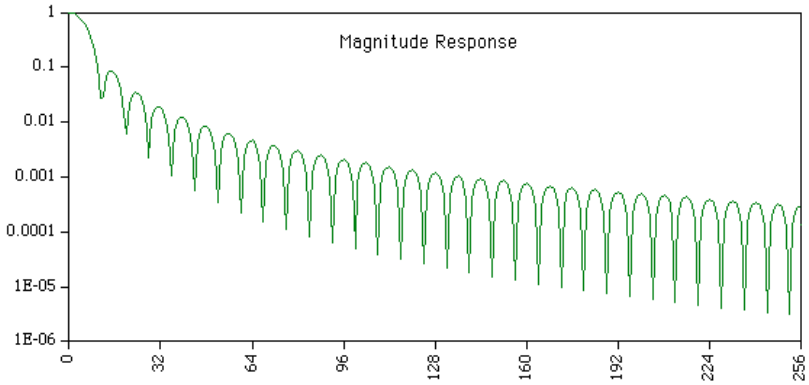


Fig 4: window for power spectral estimation.

B) The Language Model

The Hamming window which is used for the purpose is defined by the equation

W(n).54-0.46cos[2 n/(N-1)]

where, N = number of samples in

frame Let $Y(n)$ = Output signal

and $X(n)$ = input signal Y(n)=X(n)W(n)

The Fast Fourier transform (FFT) is used to convert each frame of N samples from time domain into fre- quency domain. Thus the components of the magnitude spectrum of the analyzed signal are calculated.

$$Y(w)=FFT[h(r)*x(r)]=H(w)X(w)$$

The most important step in this signal processing is Mel-frequency transformation. Compensation for non- linear perception of frequency is implemented by the bank of triangular band filters with the linear distribution of frequencies along the so called Mel-frequency range. Mel-frequency range is described by the following equation.

$$\log_{10}\left(1+\frac{f}{100}\right)Hz$$

$$fmel=2595$$

Where *f* is frequency in linear range and *fme t*he corresponding frequency in nonlinear Mel-frequency range.The *apriori* probability of a word sequence based on syntax, semantics and pragmatics of the language to be recognized. It can be realized by the Language Model which contains a list of words and their probability of occurrence in a given sequence, and is independent of the acoustic signal. The probability of a word sequence is given below.

$$p(w1,w2,w3,w4,\ldots\ldots..wn)=p(w)$$

probability of nth word is

$$p\left(w_1^n\right)=p\left(w_1\right)p\left(w_2\left|w_1\right.\right)p\left(w_2\left|w_1^2\right.\right)\cdots p\left(w_n\left|w_1^{n-1}\right.\right)$$

$$p\left(w_1^n\right)=\prod_{k=1}^{n}p\left(w_k\left|w_1^{k-1}\right.\right)$$

Language Model or Grammar essentially defines constraints on what the Speech Recognition Engine expect as input can.

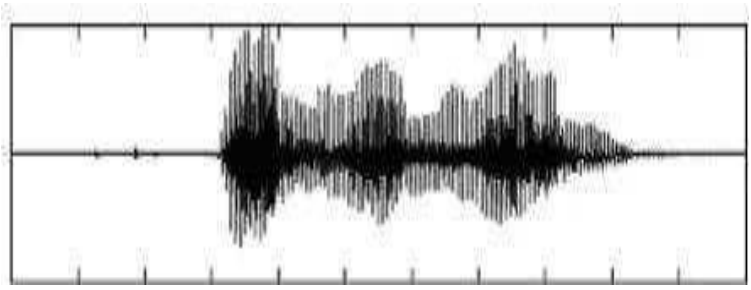


Fig 5: speech input information

The number of feature vectors that

constitute the observation sequence is of variable size. These kinds of sequences are best modeled by Hidden Markov Models.

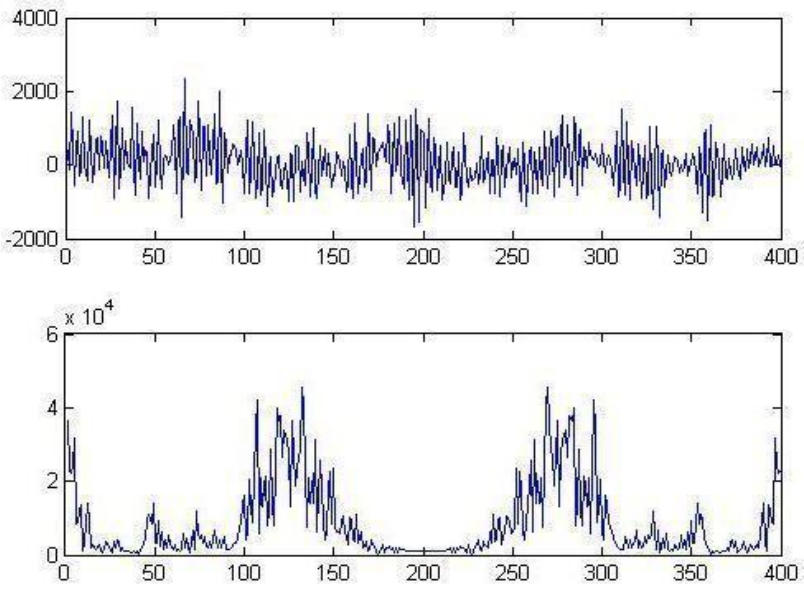


Fig 6: Mel frequency cepstral coefficient speech system.

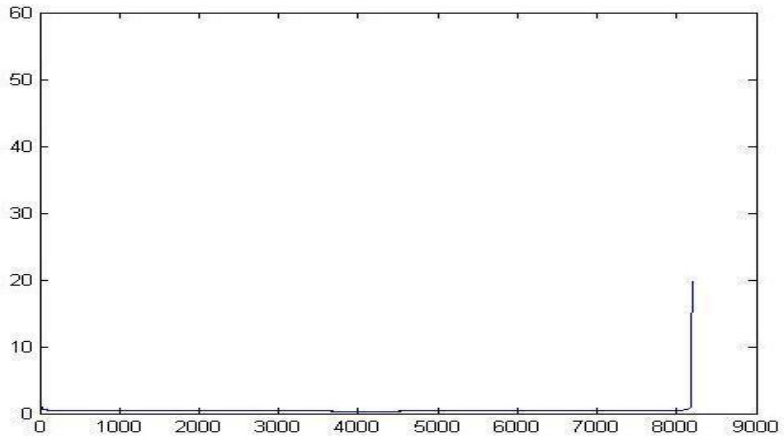


Fig 7: Noise free speech frequency system.

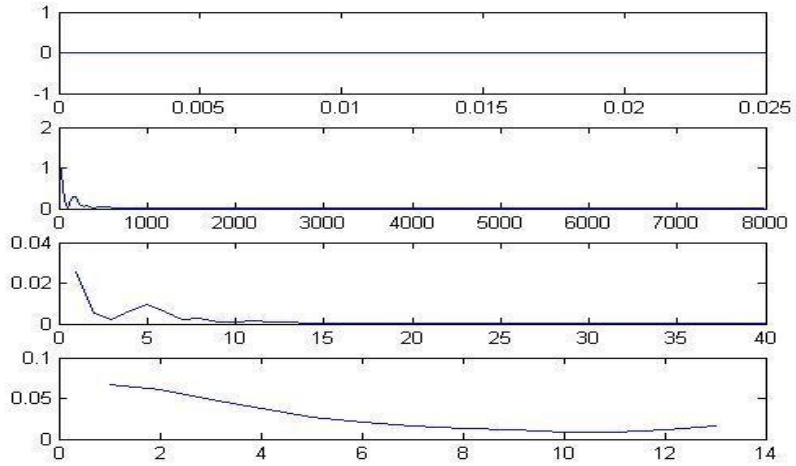


Fig 8: Example of an image with acceptable resolution.

CONCLUSION

The study was carried out for the various emotional speech databases developed, the study of systems using Prosodic features for emotion recognition and lastly the research conducted for Indian language. It was observed during the study the majority of the work in being carried out for English language. The first attempt of work for the English language was done in 1996. During the study it was observed that 16 databases are available for English language. Similarly the work for development of emotional speech database and recognition/analysis of the databases has also been done for German, Italian, Chinese, Japanese, French, Swedish and other languages. The work being conducted for Indian languages is limited to Hindi and Telugu languages. No attempts were made for the development of emotional speech database for any other Indian languages. The ignorance for the other Indian languages motivated for carrying the research for emotion recognition from Marathi speech.

ACKNOWLEDGMENT

The authors would like to thank the Authorities for providing the infrastructure to carry out the research

REFERENCES

1. Pratik K. Kurzekar, Ratnadeep R. Deshmukh, Vishal B. Waghmare, Pukhraj P. Shrishrimal," Continuous Speech Recognition System: A Review", Asian Journal of Computer Science and Information Technology, 2014; 4(6): 62–66.
2. Chao Huang, Eric Chang, Tao Chen "Accent Issues in Large Vocabulary Continuous Speech Recognition (LVCSR)", Microsoft Research China, MSR-TR, 2001; 69: 1-27.
3. Santosh K. Gaikwad, Bharti Gawli, Pravin Yannawar, "A Review of Speech Recognition Technique", International Journal of Computer Applications, November 2010; 10(3): 0975–8887.
4. M. A. Anusuya, S. K. Katti, "Speech Recognition by Machine: A Review", International Journal of Computer Science and Information Security (IJCSIS), Vol. 6, No. 3, pp. 181-205, 2009.
5. X. D. Huang, "A Study on Speaker - Adaptive Speech Recognition", Proc. DARPA Workshop on Speech and Natural Language, pp. 278-283, February 1991.
6. Pukhraj Shrishrimal, R. R. Deshmukh, Vishal Waghmare, (2012, July) "Indian Language Speech Database: A Review". International Journal of Computer Application (IJCA) Vol 47, No.5 pp.17-21

7. Yu Zhou, Yanqing Sun, Lin Yang, Yonghong Yan, "Applying articulatory features to speech emotion recognition", 2009 International Conference on Research Challenges in Computer Science, 978-0-7695-3927-0/09, IEEE 2009
8. Ganesh Janvale, Vishal Waghmare, Vijay Kale and Ajit Ghodke, "Recognition of Marathi Isolated Spoken words Using Interpolation and DTW techniques", ICT and critical infrastructure: proceeding of the 48th Annual of Computer Society of India Vol I. Advances in Intelligent system 3-319-03107-1_3, Print ISBN 978-3-319-031066 Online ISBN 978-3-319-03107-1, January 2014.
9. Vishal B Waghmare, Ratnadeep R Deshmukh, Pukhraj P Shrishrimal (2012, July) "A Comparative Study of the Various Emotional Speech Databases". International Journal on Computer Science and Engineering, Vol 4, issue 6, pp. 1236-40
10. Klaus R. Scherer, "What are emotions? And how can they be measured?" (2005) Trends and developments: research on emotions, Social Science Information Vol 44 – no 4, pp. 695–729
11. Vishal B Waghmare, Ratnadeep R. Deshmukh (2014, February) "Development of Artificial Marathi Emotional Speech Database" in proceeding of 101st Indian Science Congress, Jammu, India, 2014 .
12. Gong Chenghui, Zhao Heming, Zou Wei, Wang Yanlei, Wang Min, "Preliminary Study on Emotions of Chinese Whispered Speech" International Forum on Computer Science-Technology and Applications, 978-0-7695-3930-0/09, IEEE 2009 pp. 429 – 433
13. Neethu Mohandas, Janardhanan P. S. Nair, Govindaru V., "Domain Specific Sentence Level Mood Extraction from Malayalam Text" 2012 International Conference on Advances in Computing and Communications IEEE 2012 pp 78-81.
14. D. C. Ambrus, "Collecting and recording of an emotional speech database". Tech.rep. Faculty of Electrical Engineering, Institute of Electronics, Univ. of Maribor.(2000)
15. M. Alpert, E. R. Pouget R. R. Silva "Reflections of depression in acoustic measures of the patient's speech", Journal of Affective Disorders, 66, 59–69,(2001)