IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 5, Ver. IV (Sep – Oct. 2014), PP 97-104
www.iosrjournals.org

Design and Development of Database and Automatic Speech Recognition System for Travel Purpose in Marathi

Pooja V. Janse[1], Ratnadeep R. Deshmukh[2]
[1,2]*(Department of Computer Science and IT, Dr. B. A. M. University, Aurangabad – 431004, India)*

Abstract: *Past research in mathematics, acoustics, and speech technology have provided many methods for converting data that can be considered as information if interpreted correctly. In order to find some statistically relevant information from data, it is important to have mechanisms for reducing the information of each segment in the audio signal into features. These features should describe each segment in such a characteristic way that other similar segments can be grouped together by comparing their features. Preprocessing of speech signals is considered a crucial step in the development of a robust and efficient speech or speaker recognition system. This paper deals with result obtained by MFCC and LPC feature extraction technique and SVM feature matching technique.*

Keywords: *Speech recognition, Mel Frequency Cepstral Coefficient (MFCC), Linear Predictive Coefficient (LPC), Support Vector Machine (SVM).*

I. Introduction

Speech is the way of Communication between human being. Speech has the capability to be used as an interface for computer system. Human being has long been motivated to develop the computer that can understand and talk like human. Since 1960, computer researcher has trying ways and means to make computer record, interpret and understand human speech.

The computer system which can understand the spoken language are very useful in various domain like education sector, domestic sector, military sector, medical sector, Travel sector, artificial intelligence sector etc. So to perform any type of research, researcher requires some previous data. Generally databases are fundamental for research [1].

The popularly used cepstrum based methods to compare the pattern to find their similarity are the MFCC, LPC and SVM. The MFCC, LPC and SVM features techniques can be implemented using MATLAB. This paper reports the findings of the voice recognition study using the MFCC, LPC and SVM techniques.

The rest of the paper is organized as follows: Need of Development of Speech Database is given in section 2, the methodology of the study in section 3, the implementation of the study in section 4, which is followed by result and discussion in section 5, and finally concluding remarks are given in section 6.

II. Need Of Development Of Speech Database

As little work is done for travel domain in Marathi, it leads to develop ASR system in Marathi. Our motivation to do the Speech Recognition is for trying to develop the speech interface for the system in the Marathi language for travel domain. The research in the speech domain have attained new heights for English, other European languages and for languages spoken in other developed countries.

A lot of work has been completed for isolated word recognition, connected word and continuous speech. The systems developed for English and other European languages have achieved and accuracy of more 85% in some cases they do have achieved accuracy of 95%. However, the work in speech domain for Indian languages is still behind.

A very little work has been carried out for Marathi Language [2]. Hence, we selected to develop the speech database of Isolated Marathi words for Travel purpose. In this work we have tried to capture maximum variation of Marathi language in the Aurangabad district grouped according to their category i.e. Malls, Cinema halls, Markets, Temples, Playgrounds, Station and Airport, Cultural halls, Hotels, Tourist Places and Restaurants.

III. Methodology

We developed the Text corpus grouped according to their category i.e. Malls, Cinema halls, Markets, Temples, Playgrounds, Station and Airport, Cultural halls, Hotels, Tourist Places and Restaurants [3]. Then we selected 100 speakers from different Taluca places from Aurangabad. We recorded speech samples from speaker and then extracted feature for further analysis.

The methodology followed by us for the proposed work is shown in figure.

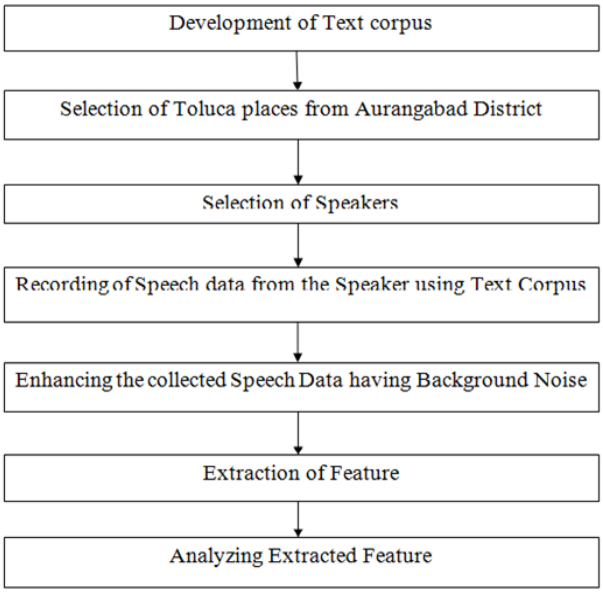

Fig 3.1: Methodology adopted for the proposed work

IV. Implementation

A. Data Collection Procedure

In this stage, the steps followed for developing speech corpora are described. The recording media is chosen first and then the data has been recorded using high quality microphones and laptop using PRAAT for recording speech signal.

1) Speaker Selection

The speech data will be collected from the native speakers of Marathi Language. The selected speakers will be from different regions of Aurangabad District. They would be comfortable with reading and speaking the Marathi Language. The speakers are classified on the basis of gender.

2) Speech Data collection

We used PRAAT software for recording the speech. We used Sennheiser PC360 and Sennheiser PC350 headset for recording the speech samples. The PC360 and PC350 headsets are having noise cancellation facility and the signal to noise ratio (SNR) is less. The steps followed for recording the speech samples was as follows:
Step 1: Selected speakers were asked regarding any problem with reading or speaking the Marathi words.
Step 2: Speakers were given the basic information about the headset used and when to speak the word.
Step 3: The sampling frequency was set to 22050 Hz with Mono sound type.
Step 4: The speaker was asked to read each word and the recorded sample was saved as .wav file.
Step 5: Step 4 was repeated for all 372 utterances that were recorded from the speaker. All the steps were repeated for all the 100 speakers.

3) Data Collection Statistics

The speech data is collected from 100 speakers. Each speaker will be asked to speak 124 words with 3 utterances. 372 utterances of words will be collected from every speaker. Total 37200 utterances of words are recorded. Till date we have collected 37200 utterances from 100 speakers in which 50 male and 50 female speakers.

4) Recording Environment

The speech data will be recorded using high quality microphones like Sennheiser PC 350 and Sennheiser PC 360 with the help of open source PRAAT speech software. The data is recorded in Noisy environment. The purpose of recording in noisy environment is to develop robust ASR System.

The main strength of PRAAT is its graphical user interface. PRAAT also provides the functionality of General analysis (waveform, intensity, spectrogram, pitch, duration) Spectral analysis, pitch analysis, voice analysis, format analysis, intensity analysis, PCA and many facilities.

The signals were greatly different due to many factors such as people voice change with time, health condition (e.g. the speaker has a cold), speaking rate and also acoustical noise and variation recording environment via microphone.
Following tables gives detail information of recording procedure and metadata.

Table: Information about data collection procedure

Process	Description
1) Speaker	50 Female 50 Male
2) Tools	PRAAT, Microphone Sennheiser PC360 and Sennheiser PC350
3) Environment	Noisy
4) Utterance	Three utterance of each word
5) Sampling Frequency, fs	22050 Hz

Table: Metadata about Speech Database

Process	**Description**
Total Number of Words Selected	124
Utterances Recorded	Three utterance of each word
Total Utterance per Speaker	372
Total Speaker	100
• Male Speaker	50
• Female Speaker	50
Total Male Speaker Utterances	18600
Total Female Speaker Utterances	18600
Total Utterances	37200
Total Size of Database	3.33 GB
• Male Database Size	1.64 GB
• Female Database Size	1.69 GB
Software Used for Recording	PRAAT
Tools	Microphone Sennheiser PC360 and Sennheiser PC350
Recording Frequency	22050 Hz

B. Speech Recognition

Speech recognition (SR) is the translation of spoken words into text. It is also known as "automatic speech recognition or ASR", "computer speech recognition", "speech to text or STT". Speech Recognition is an inter-disciplinary research domain. Speech Recognition is the process of converting a speech signal to a sequence of words, by means of an algorithm implemented as a computer program.

Research in speech processing and communication for the most part, was motivated by people desire to build mechanical models to emulate human verbal communication capabilities. Speech is the most natural form of human communication and speech processing has been one of the most exciting areas of the signal processing. Speech recognition technology has made it possible for computer to follow human voice commands and understand human languages. The main goal of speech recognition area is to develop techniques and systems for speech input to machine.

The disciplines that have been applied to one or more speech recognition problems are as follows: signal processing, Physics (i.e. acoustics), Pattern Recognition, Communication and Information theory, Linguistics, Physiology, Computer Science and Psychology. There are various spoken languages in the world. The communication among human being is dominated by spoken language. Hence it is natural to expect speech as an interface between human and machine.

1) Speech Feature Extraction and Analysis

The main objective of the proposed study is development of standard speech database and using that developed database for development of Automatic Speech Recognition System. For developing an Automatic Speech Recognition system we need to extract the feature from the acquired/recorded speech and then apply the recognition algorithm.

The extraction of the best parametric representation of acoustic signals is an important task to produce a better recognition performance. The efficiency of this phase is important for the next phase since it affects its behavior.

Theoretically it is possible to recognize speech directly from the digital waveform of the speech. However, as speech is time varying the idea to perform some form of feature extraction came into existence which is used to reduce the variability of speech signal. In the context of Automatic speech recognition feature extraction is the process of retaining the useful information from the speech signal while the unnecessary and

unwanted information is removed which involves the speech signal analysis. However, while removing the unwanted information from the speech signal some useful information may also lose.

2) Feature Extraction using MFCC and LPC

- ***Mel Frequency Cepstral Coefficient (MFCC)***

The extraction of the best parametric representation of acoustic signals is an important task to produce a better recognition performance. The efficiency of this phase is important for the next phase since it affects its behavior. MFCC is based on human hearing perceptions which cannot perceive frequencies over 1Khz. In other words, in MFCC is based on known variation of the human ear's critical bandwidth with frequency. MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz. A subjective pitch is present on Mel Frequency Scale to capture important characteristic of phonetic in speech [4].

The overall process of the MFCC is shown in figure.

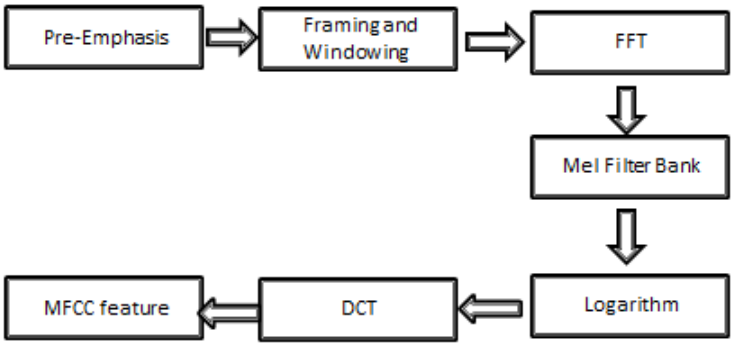

Fig 4.1: MFCC Block Diagram

As shown in Figure, MFCC consists of seven computational steps. Each step has its function and mathematical approaches as discussed briefly in the following:

Step 1: Pre–emphasis

This step processes the passing of signal through a filter which emphasizes higher frequencies. This process will increase the energy of signal at higher frequency.

$$Y[n] = X[n] - 0.95\ X[n-1]$$

Let's consider a = 0.95, which make 95% of any one sample is presumed to originate from previous sample.

Step 2: Framing

The process of segmenting the speech samples obtained from analog to digital conversion (ADC) into a small frame with the length within the range of 20 to 40 msec. The voice signal is divided into frames of N samples. Adjacent frames are being separated by M (M<N). Typical values used are M = 100 and N= 256.

Step 3: Hamming windowing

Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines. The Hamming window equation is given as: If the window is defined as W (n), $0 \leq n \leq N-1$ where

N = number of samples in each frame

Y[n] = Output signal

X (n) = input signal

W (n) = Hamming window, then the result of windowing signal is shown below:

$$Y[n] = X[n] \times W[n]$$

$$W(n) = 0.54 - 0.46 \cos\{2\pi n|N-1\} \qquad 0 \leq n \leq N-1$$

Step 4: Fast Fourier Transform

To convert each frame of N samples from time domain into frequency domain. The Fourier Transform is to convert the convolution of the glottal pulse U[n] and the vocal tract impulse response H[n] in the time domain. This statement supports the equation below:

$$Y(w) = FFT[h(t) \times X(t)] = H(w) \times X(w)$$

Step 5: Mel Filter Bank Processing

The frequencies range in FFT spectrum is very wide and voice signal does not follow the linear scale. The bank of filters according to Mel scale is then performed. After that the following equation is used to compute the Mel for given frequency f in HZ.

F (*Mel*) = [2595 × log 10 [1 + F] 700]

Step 6: Discrete Cosine Transform

This is the process to convert the log Mel spectrum into time domain using Discrete Cosine Transform (DCT). The result of the conversion is called Mel Frequency Cepstrum Coefficient. The set of coefficient is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector.

- ***Linear Prediction Coefficient (LPC)***

LPC (Linear Predictive coding) analyzes the speech signal by estimating the formants, removing speech signal, and estimating the intensity and frequency of the remaining buzz. The process is called inverse filtering, and the remaining signal is called the residue. In LPC system, each expressed as a linear combination of the previous samples. This equation is called a linear called as linear predictive coding [5].

LPC Analysis-

The next processing step is the LPC analysis, which converts each frame of p + 1 autocorrelations into LPC parameter.

Following figure shows block diagram of MFCC+LPC.

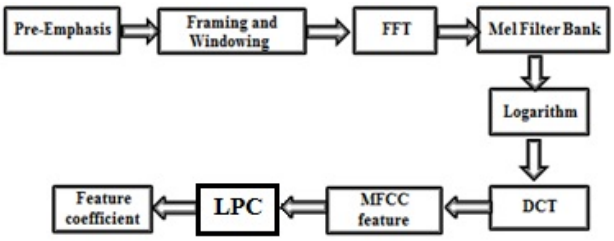

Fig 4.2: Block Diagram of MFCC+LPC

C. ***More Features Extracted:***

- **Pitch:** It is the main feature of an audio file. The perceived pitch of a sound is just the ear's response to frequency, i.e., pitch is just the frequency. Pitch = frequency of sound.
- **Standard Deviation:** Standard deviation shows how much variation or dispersion exists from the average (mean), or expected value. A low standard deviation indicates that the data points tend to be very close to the mean; high standard deviation indicates that the data points are spread out over a large range of values.
- **Energy Intensity:** This feature represents loudness of an audio signal, which is correlated to amplitude of signal.
- **Energy Entropy:** It expresses abrupt changes in the energy level of an audio signal. In order to calculate this feature, frames are further divided into K-sub windows of fixed duration.
- **Short Time Energy:** The amplitude of the speech signal varies appreciably with time. In particular, the amplitude of unvoiced segment is generally much lower than the amplitude of voiced segments. Short Time energy provides a convenient representation that reflects these amplitude variations. The major significance of this is that it provides a basis for distinguishing voiced speech from unvoiced speech.
- **Zero Crossing Rate:** It is the rate of sign-changes along a signal, i.e., the rate at which the signal changes from positive to negative or back. This feature has been used heavily in both speech recognition and music information retrieval, being a key feature to classify percussive sounds.
- **Spectral Centroid:** It is the weighted mean frequency. It indicates where the "center of mass" of the spectrum is. Because the spectral centroid is a good predictor of the "brightness" of a sound, it is widely used in digital audio and music processing as an automatic measure of music timbre.
- **Spectral Roll off:** Spectral Roll off point is defined as the Nth percentile of the power spectral distribution, where N is usually 85% or 95%. This measure is useful in distinguishing voiced speech from unvoiced: unvoiced speech has a high proportion of energy contained in the high-frequency range of the spectrum, where most of the energy for voiced speech and music is contained in lower bands.

$$\sum_{n=1}^{Rt} M_t\ (n) = 0.85 \sum_{n=1}^{N} M_t\ (n)$$

Where Rt is the frequency below which 85% of the magnitude distribution is concentrated.

- **Spectral Flux:** It is a measure of how quickly the power spectrum of a signal is changing, calculated by comparing the power spectrum for one frame against power spectrum for the previous frame. More precisely, it is usually calculated as the Euclidean distance between the two normalized spectra.

These features have been extracted for every uploaded wave file and then database of these features is prepared [6] [7].

D. *Word Recognition using SVM*

Word Recognition is a process where word uttered by the user has to be recognized by the speech recognition system. For recognition purpose we used SVM Algorithm. All the Trained dataset is put into Reference Frames one after the other. Now these Reference Features and Test Features are acts as inputs to the SVM Algorithm program.

SVM is a concept in computer science for a set of related supervised learning methods that analyze data and recognize patterns. Since SVM is a simple and efficient computation of machine learning algorithms, and is widely used for pattern recognition and classification problems, and under the conditions of limited training data, it can have a very good classification performance compared to other classifiers.

Following figure shows the flow of word recognition process.

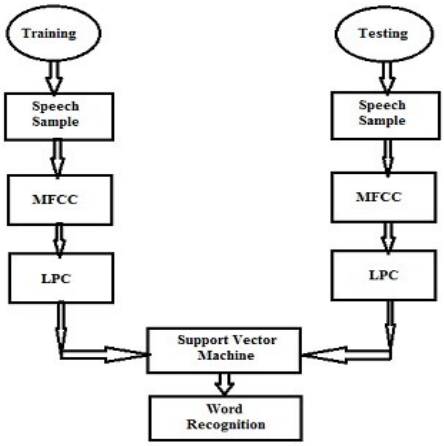

Fig 4.3: Word Recognition System

This is a basic diagram of our ASR system which is basically divided into two parts training side and testing side. From collected database first 2 utterances we have stored as data for training and 3rd utterance we are going to use as a test file. Then we extract the feature using the combination of MFCC and LPC and compare the 3rd utterance with the two utterances which are stored training data. Same procedure we are going to use for all the files which are stored in our database one by one. After extraction of these files SVM is to compare these files and measure its similarity by calculating minimum distance between them.

V. Result And Disscution

The input voice signal is shown in figure 5.1.

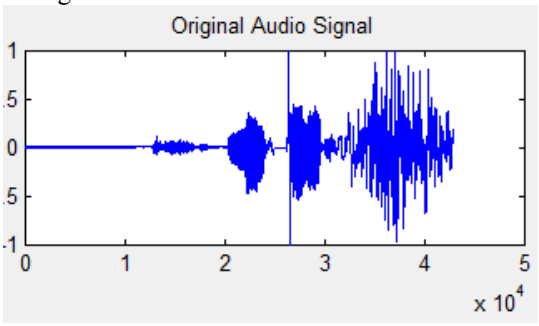

Fig 5.1: Original input signal

Figure 5.1 is used for carrying the voice analysis performance evaluation using MFFC. A MFCC cepstral is a matrix, the problem with this approach is that if constant window spacing is used, the lengths of the input and stored sequences is unlikely to be the same.

Figure 5.2 shows the MFCC output of two different speakers. The matching process needs to compensate for length differences and take account of the non-linear nature of the length differences within the words.

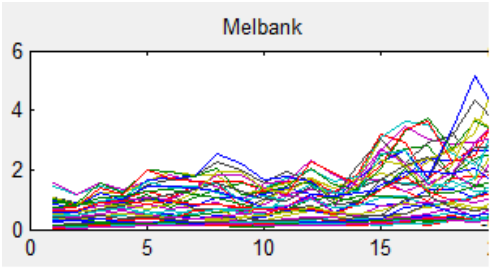

Fig 5.2: Melbank generated of speech signal

After applying MFCC algorithm we apply LPC Autocorrelation analysis so that we can extract better features of speech signal. Following figure shows LPC coefficient.

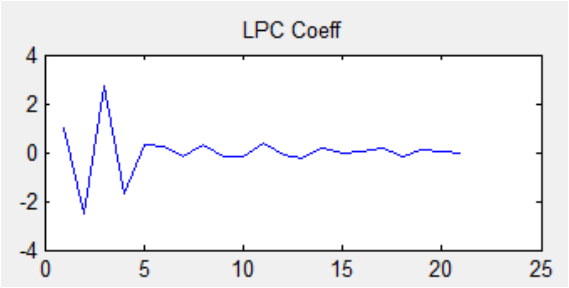

Fig 5.3: LPC Coefficient

The input test voice matched optimally with the training template which was stored in the database. The finding of this study is consistent with the principles of voice recognition where comparison of the template with incoming voice was achieved via a pair wise comparison of the feature vectors using SVM.
After applying SVM we get following Distance Matrix.

Table: Distance Matrix for Cinema Hall

	SATYAM	PVR	FAME TAPDIYA	BIG CINEMMAS	E-SQUARE	APSARA	AMBA	GOLDI	NUPUR	ROKSI	SADIYA
SATYAM	0	0.532	0.3678	0.3638	0.3031	0.4311	0.274	0.319	0.5079	0.38	0.5183
PVR	0.5318	0	0.4529	0.573	0.289	0.8221	0.5854	0.701	0.2032	0.642	0.4293
FAME TAPDIYA	0.3678	0.453	0	0.3	0.3351	0.4917	0.3346	0.444	0.5109	0.353	0.6382
BIG CINEMMAS	0.3135	0.423	0.2803	0	0.2797	0.4923	0.3746	0.412	0.4189	0.316	0.4379
E-SQUARE	0.3031	0.289	0.3351	0.371	0	0.6084	0.3643	0.435	0.2776	0.414	0.3583
APSARA	0.4932	0.952	0.6008	0.5012	0.7082	0.3679	0.4124	0.367	0.9479	0.483	0.9208
AMBA	0.274	0.585	0.3346	0.3215	0.3643	0.4069	0	0.261	0.6244	0.369	0.6502
GOLDI	0.5337	0.96	0.6447	0.5756	0.7697	0.311	0.4879	0	1	0.598	0.98
NUPUR	0.5079	0.203	0.5109	0.585	0.2776	0.8405	0.6244	0.692	0	0.627	0.2983
ROKSI	0.3827	0.34	0.3268	0.4075	0.2733	0.6397	0.5216	0.553	0.28	0	0.4078
SADIYA	0.5183	0.429	0.6382	0.5638	0.3583	0.8006	0.6502	0.634	0.2983	0.595	0

VI. Conclusion

- After doing the literature survey we developed the speech database of isolated word for travel purposes in Marathi language as no such database is available till date.
- After the completion of the database collection for the feature extraction technique we selected Mel Frequency Cepstral Coefficient (MFCC) and Linear Predictive Coding (LPC).
- We have used the mean and standard deviation techniques as well as some extra audio features for the accuracy at the speaker level.
- If we combine some more technique like Hidden Markov Model (HMM), Wavelet transform etc. for speech recognition we can get better accuracy.

Acknowledgements

This work is supported by University Grants Commission. The authors would like to thank the University Authorities for providing the infrastructure to carry out the research.

References

[1] M.A.Anusuya,S.K.Katti," Speech Recognition by Machine: A Review", International Journalof Computer Science and Information Security,Vol. 6, No. 3, 2009,pp. 181-205. R.E. Moore, *Interval analysis* (Englewood Cliffs, NJ: Prentice-Hall, 1966).

[2] Chalapathy Neti, Nitendra Rajput, Ashish Verma, "A Large Vocabulary Continuous Speech Recognition system for Hindi", In Proceedings of the National conference on Communications, Mumbai, 2002, pp. 366-370.

[3] Tejas Godambe and Samudravijaya K., "Speech Data Acquisition for Voice based Agricultural Information Retrieval", presented at the 39th All India DLA Conference, Punjabi University, Patiala, 14-16th June 2011.

[4] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", JOURNAL OF COMPUTING, VOLUME 2, ISSUE 3, MARCH 2010, ISSN 2151-9617.

[5] Leena R Mehta , S.P.Mahajan , Amol S Dabhade Comparative Study Of MFCC And LPC For Marathi Isolated Word Recognition System" International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering Vol. 2, Issue 6, June 2013.

[6] Shruti Aggarwal, Naveen Aggarwal, "Classification of Audio Data using Support Vector Machine", IJCST Vol. 2, Issue 3, September 2011.

[7] Aastha Joshi, "Speech Emotion Recognition Using Combined Features of HMM & SVM Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 8, August 2013.