Methodology for Gender Identification,Classification and Recognition of Human Age

Shivaji J Chaudhari
ME Second Year Student
Dr D Y Patil School of Engineering and technology, Lohegaon, Pune.

Ramesh M Kagalkar
Assistant Professor
Dr D Y Patil School of Engineering and technology, Lohegaon, Pune.

ABSTRACT
The human voice is comprised of sound made by a human being using the vocal cord for talking,singing, laughing, crying and shouting. It is particularly a piece of human sound creation inwhich the vocal cord is the essential sound source, which play an important role in the conversation.The applications of speech or voice processing technology play a crucial role in humancomputer interaction. The system improves gender identification, age group classification, ageand emotion recognition performance. The research work uses new and efficient methods forfeature extraction of speech or voice and classification of standard method on the various audiodatasets. Mel Frequency Cepstral Coefficients feature extraction and selection is performed tofind a more suitable feature set for building speaker models. The proposed system uses GaussianMixture Model is a supervector for system feature selection and feature modelling. SupportVector Machine classification and feature matching technique is used to classify the featurefor different age groups like child, teenage, young, adult and senior to increase the resultantperformance and accuracy. The database is created using the audio files for each age group ofspeaker and for each emotion as an input, performs feature extraction and identifies the gender,classify age group, recognize age and emotion.

Keywords
Mel Frequency Cepstral Coefficient (MFCC), Gaussian Mixture Model (GMM),support vector machine (SVM), Expectation-Maximization (EM), Maximum a Posteriori (MAP), Hidden Markov Models (HMMs), Suprasegmental Hidden Markov Models (SPHMMs), Interactive Voice Response System (IVRs).

1. INTRODUCTION
Human interaction with computers in done in many ways and the interface between humanand the computer is crucial to facilitate this interaction. Maximum desktop applications, internetusing browsers like Firefox, chrome and internet explorer. The computers make use ofthe prevalent Graphical User Interfaces (GUI). Voice User Interfaces (VUI) are used for speechrecognition and synthesizing systems. Human Computer Interaction (HCI) aims to improve theinterface between users and computers by making computers more usable and receptive to usersneed. There are many speaker characteristics that have useful applications. The most popularinclude gender, age, health, language, dialect, accent, sociolect, idiolect, emotionalstate and attentional state. These characteristics have many applications in dialogue systems,speech synthesis, forensic, call routing, speech translation, language learning, assessment systems,speaker recognition, meeting browser, law enforcement, human robot interaction andsmart workspaces. For example, the spoken dialogue system provides services in the domains offinance, travel, scheduling, tutoring. The systems need to gather information from the userautomatically in order to provide timely and relevant services. Most telephones based servicestoday use spoken dialogue systems to either route calls to the appropriate agent or even handle. The complete service with an automatic system. For example, shopping systems can recommendsuitable goods appropriate to the age and gender of the shopper. The speaker specific characteristicsof the signal can be exploited by listeners and technological applications to describeand classify speakers, based on age, gender, accent, language, emotion or health very importantcharacteristic of human speech or voice based interfaces is the dependability of the phonetic,syntactic and lexical properties of the utterance or word spoken by the user. Human voice based gender, age group and precise age estimation are difficult. First, usually there is adifference between the age of a speaker as perceived the identified age and their actualage is estimated age [1], [2], [3].

The law enforcement has been concerned about different biometric featuresto identifythe each human uniqueness. Different biometric features can be used for unique human identification such as fingerprints, facial, hand geometry pattern, signature dynamics and voicepatterns. In some criminal cases, the available evidence is in the form of recorded conversationsor phonetic voice. The speech patterns can include unique and important information tolaw enforcement personnel [4], [6], [14].

1.1 Voice Features
The long speech features extracted from the longer segments of speech signal such as entire sentences,words, syllables are known as supra segmental or prosodic features. They normallyrepresent the speech properties like rhythm, stress, intonation, loudness and duration. Acousticcorrelates of prosodic features are pitch, energy, duration and their derivatives. The emotionspecific information about shapes and sizes of the vocal tract, responsible for producingdifferent sound units and the associated movement of articulators are captured using spectralfeatures. The characteristics of glottal activity, specific to the emotions are estimated usingexcitation source features. [4], [6], [9], [11], [12].

The discourse information on emotion recognition has been combined with acoustic correlatesto improve the overall performance of emotion classification, repetition or correctioninformation was used for the discourse information, also adopted repetition as their discourseinformation.[1], [4], [14], [18], [21].

2. LITERATURE SURVY
In [1] presents a dimension reduction technique which aims to improve greater efficiency and the accuracy of speaker's age group and precise age estimation systems based on the human voice signal. Two different gendersbased age estimation

approaches studied, the first is the age group (senior, adult, and young)classification and the second is an accurate age estimation using regression technique. Thesetwo approaches use the GMM super vectors as features for a classifier model. Age groupclassification assigns an age group to the speaker and age regression estimates the speaker'sprecise age in years.

In paper [5] presents a gender detection is an extremely useful task for an extensive varietyof voice or speech based applications. In the spoken language systems INESC ID, the gender identificationcomponent is initialand the basic component of our voice processing system, where it is utilizedprior to speaker clustering, in order to avoid mixing speakers between male and female gender in thesame cluster. Gender information (male or female) is also used to create gender dependentacoustic module for speech recognition.

In [6] introduce a new gender detection and an age estimation approach is proposed. Todevelop this method, after deciding an acoustic features model for all speakers of the sample database, Gaussianmixture weights are extricated and connected to build a supervector for each speaker. Then,hybrid architecture ofGeneral Regression Neural Network (GRNN) and Weighted Supervised Non Negative Matrix Factorization(WSNMF) are developed using the created supervectors ofthe training data set. The hybrid method is used to detect the gender speaker while testing andto estimate their age. Different biometric features can be used for forensic identification.Choosing a method depends on its use and efficientreliability of a particular application and the available data type.In some crime cases, the available evidenceor proof might be in the form of recorded voice.Speech patterns can include unique and important information for law enforcement personnel.

In [7] mainly focused on enhancing emotion recognition and identification performancebased on a two stages that is combination of gender recognizer andemotion recognizer. The system work is a gender dependent, text independent andspeaker independent emotion recognizer. Both Hidden Markov Model (HMM) and Supra segmentalHidden Markov Model (SPHMM) have used as classifiers in the two stage architecture.This architecture has been evaluated on two different and separate speechdatabases. The two databases are emotional prosody speech andtranscripts database and human voice collected database.

In [8] explores the detection of specific type emotions using discourse informationand language in combination with acoustic signal features of emotion in speech signals. The main focus ison a detecting type of emotions using spoken language data obtained froma call center application. Most previous work in type emotion recognition has used only theacoustic features information contained in the speech. The system contains three sources ofinformation, lexical, acoustic and discourse is used for speaker's emotion recognition.

In [9] develop models for detecting various characteristics of a speaker based on spoken thetext alone. These characteristics or attributes include whether the speaker is speaking nativelanguage, the speakers age and gender, the regional information reported by the speakers. Theresearch explores various lexical features information as well as features inspired by linguistic(a languagerelated) informationand a number of word and dictionary of affect in language. This system suggeststhat when audio or voice data is not available, by exploring effective audio feature sets onlyfrom uttered text and system combinations of multiple classification algorithms, researcherbuild statistical models to detect these attributes of speakers, equivalent to frameworks that canexplore the audio information.

In [10] present speaker characteristicrecognition and identification field has made extensive use of speaker MAP adaptation techniques.The adaptation allows speaker model feature parameters to be estimated using lessspeech data than needed for Maximum Likelihood (ML) training method. The Maximum LikelihoodLinear Regression (MLLR) and Maximum a Posteriori (MAP) techniques have typicallybeen used for speaker model adaptation. Recently, these adaptation techniques have been incorporatedinto the feature extraction stage of the SVM classifier based speaker identification and recognition systems.

In [15] humans, emotional speech recognition contributes much to create harmonious humanto machine interaction, additionally with many potential applications. Three approachesto augment parallel classifier are compared for recognizing emotions from a speech by thespeech database. Classifier applied on prosody, spectral, MFCC and other common features.One is standard classification schemes (one versus one) and two methods are Directed AcyclicGraph (DAG) and Unbalanced Decision Tree (UDT) that can form a binary decision tree classifier.The hierarchical classification technique of feature driven hierarchical SVMs classifiersis designed, it uses different feature parameters to drive each layer and the emotion can be subdividedlayer by layer. Finally, analysis of the classification rate of those three extends binaryclassification, DAG system performs the best for testing database and standard classifier is notfar behind, the UDT is the poorest because of relying on upper layer order processing.

In [16] The extraction and matching process is implemented after the signal preprocessingis performed. The non parametric method for modeling the human voice processing system. The nonlinear sequence alignment called as Dynamic Time Warping (DTW) used asfeatures matching techniques. This paper presents the technique of MFCC feature extractionand wrapping technique to compare the test patterns.

3. PROBLEM STATEMENT

This helps to identify the gender, then classify the speaker agegroup belong to a certain category, then further system will process to recognize exact age andalso system display emotional state of the speaker with his/her profile detail which is stored inthe database. The objective of the system is to extract the feature and compare with databaseto identify the gender and also it classify the certain speaker age group, this two task helps toget increase the system performance and accuracy. On the other side, perform feature selectionfor speaker classification and matching using popular classification techniques, so efficientclassifier classifying speaker characteristics.This system applies techniques like MFCC feature extraction algorithm, GMM modeling technique, SVM classifier and matching technique. The main issue in voice or speech processingresearch to achieve high efficiency and performance of different age group and differentlanguage dependent speaker and to reduce the large size of the dimension of feature matrixusing many techniques.

4. SYSTEM DESIGN

4.1 System Architecture

System Architecture divided into two phases that is

A. Training phase
B. Testing phase

Most of the operations are same in the training phase and testing phase in figure 1.

Training Phase

The training phase used large audio dataset for training the system using the MFFC featureextraction technique applied for extracting the unique feature of audio/voice file and create thefeature vector. GMM super vector representation and dimension reduction for each featuretype, etc. Training phase applied to the large set sample data set for training purpose.

1. Train the system is an over MFCC features, extracted from speech utterances of speechsessions. The speech sessions used to train the system background model should be diversifiedand uniformly distributed over speaker ages and genders.

2. An adaptation of the speaker model is constructed, MAP estimation is used to adapt themodel to represent the model of a specific speaker. The adaptation is done using theMFCC features extracted by the speaker session.

3. Build the supervector with the help of GMM model is represented by one supervector,formed by concatenating all the M component gaussian means.

$V = (u1, u2, \ldots\ldots.u_i)^T$

Where u_i is the mean vector of the i^{th} gaussian. The training super vectors are formedusing the MAP adaptation models. In the baseline system, the super vectors are usedas feature vectors [1], [3], [7].

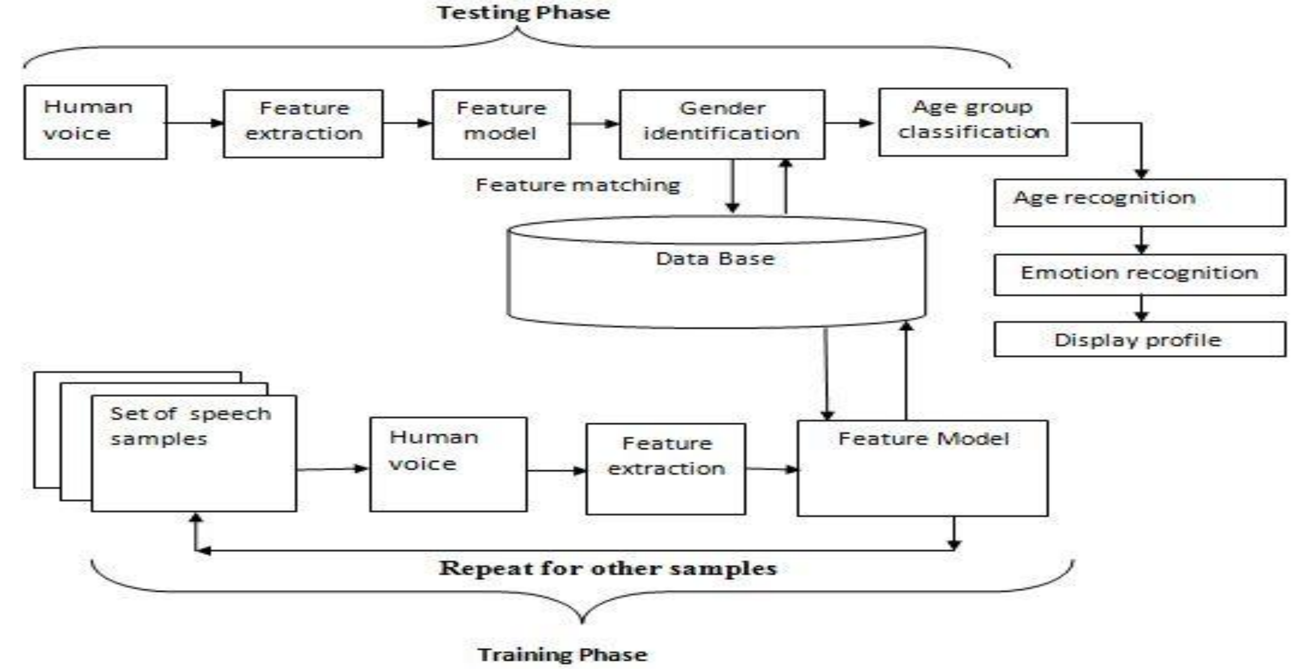

Fig 1: Proposed System Architecture

Testing Phase

In the testing phase, the speech session is processed same as training phase. A GMM modelis trained, a super vector is formed and the dimension reduction projection matrix is appliedon it to create a reduced testing feature vector. SVM classification algorithm and matchingtechnique are applied to classify the result and find the exact result for input voice[1], [5], [6], [7].

4.2 Feature Extraction

The extraction of the best parametric representation of the acoustic signals of the human voiceis an important task to produce a letter recognition performance. The result efficiency of feature extraction phase is important for the next phase like modeling, classification and feature matching since it affects its behavior. Following steps give the detail process of feature extraction of audio file [1], [9],[11], [16].

1. Pre emphasis passing of a signal through a many filter which emphasizes higher frequencies. Itincreases the energy level of the signal at higher frequency.
2. Framing is the process of segmenting the speech or voice samples obtained from Analog to Digital Conversion (ADC) into a predefined small size frame with the length within the specified range of twenty to forty milliseconds. The voice signal is divided into of N sample frames.
3. Hamming window is used as window shape by considering the next block in the feature extraction, processing chain and integrates all the closest frequency lines.
4. Fast Fourier Transform (FFT) convert each frame of N samples from time domain into the frequency domain. To obtain the magnitude, frequency response of each frame performs FFT. The output is a spectrum or periodogram.
5. Mel Filter Bank processes the frequency range in FFT spectrum is very wide and voice signal does not follow the linear scale.
6. Discrete Cosine Transform (DCT) is the process to convert the log mel spectrum into time domain using this process. The result of the conversion is called MFCC. The set of coefficient is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector.
7. Delta delta energy and delta spectrum voice signal the frame changes, such as the slope ofa format at its transitions. Therefore, there is a need to add features related to the changein cepstral features over time.

4.3 Gaussian Mixture Model

A GMM model is a probability density function represented using a weighted sum ofall Gaussian component densities. Modelling technique is commonly used parametric model ofthe probability distribution of features in a proposed system, suchas voice tract related spectral features of signal in a speaker recognition system. The parameters are estimatedfrom training sample voice data using the iterative EM algorithm or MAP estimation from a well trainedprior modelling approach is a well known modelling technique in text independent speakerrecognition systems for frame based features.

$$\lambda \qquad = \sum_{k=0}^{n} \binom{n}{k} x^k a^{n-k}$$

The each component density is a D variate gaussian function of the form, with mean vector$u_i$and covariance matrix$\sum$ i=1. The complete gaussian mixture model is parametrized by thecovariance matrices,mixture weights and mean vectors from all component densities[1], [25].

4.4 Support Vector Machine

SVM is a powerful technique for pattern classification. Classifier map collection inputs into a higherdimensional space and then classify input into separate classes withthe help of hyperplane. A critical aspect is the design of the inner product, that isa kernel function, induced by the high dimensionalmapping and binary classifier classify the training data into two classes and identify the classof testing file[1], [5].

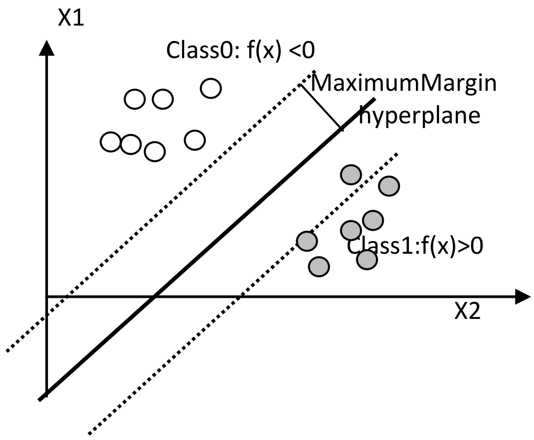

Fig2: Support Vector Machine

$$f(x) = \sum_{k=1}^{n} a_i t_i K(X, X_i) + d$$

Where the t_i are the ideal outputs, i greater than 0. The vectors X_i is support vectors andformedusing the training set by an optimization process. The classifieroutputs are either 1 or -1,depending upon whether the corresponding input data support vector is in class 0 or class 1, is respectively shown in the figure 2. A class decision is based upon whether the value f(x),is above or below a threshold [13], [17] [21].

5. DATA TABLES AND ANALYSIS

The database is the collection of audio files of the speaker from different agegroup.

Table 1: Input dataset classification.

Classified Dataset Name	Age Range (Year	Notation
Child	05-08	C
Teenage	09-17	T
Male Young	18-30	MY
Male Adult	30-60	MA
Male Senior	Greater than 60	MS
Female Young	18-30	FY
Female Adult	30-60	FA
Female Senior	Greater than 60	FS

The free speech is implemented by conversation about topics among two or morepeoples. For example, a conversation between a doctor and a patient, sometimes we collect theconversation between system and human implemented by speech understanding.The formatof speech files is .wav. raw, but wav file preferred because of the quality of sound better ascompared to other type of audio file in table1.

Table 2: Age and gender identification result analysis.

Age group category	Gender identific ation	Age recogni tion
Child	50	75
Teenage	89	85
Male Young	80	80
Male Adult	85	90
Male Senior	90	90
Female Young	95	100
Female Adult	90	100
Female Senior	80	95

The system gives the different and dependent age and gender identification and recognitionresult for testing audio files. The result accuracy and performance based on the quality ofvarious kinds of train data, shown in the table 2.

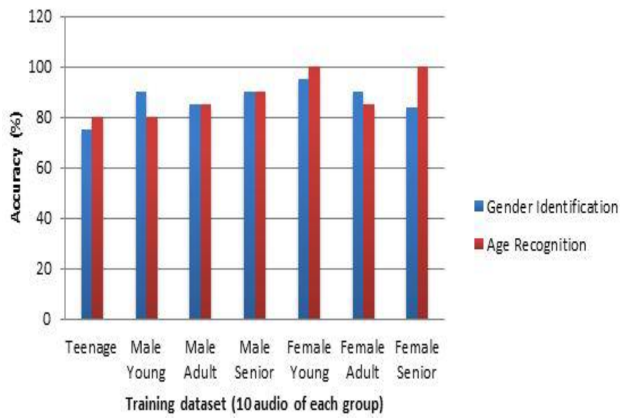

Fig 3: Result of gender identification and age recognition.

The table 2 and figure 3 shows the audio file dataset is created for each age group and for each type of emotion, someaudio files used for system training and remaining sample are given for testing purpose. The resultis percentage is calculated for each group individually, age recognition depends on the genderidentification. The system gives the high average accuracy and performance result is 80.3 %for gender identification and 89.3 % for age recognition.

6. CONCLUSION

Thus the proposed system help to identify, classify and recognize exact speaker age withemotion and displaying profiles of speaker using the trained database. The speaker profile is helpfulin many applications like for advertisement, targeting to particular people, automatically identificationof this feature, age, emotion to provide facility and service to customer in a call center,in some field speaker's voice can be used as the biometric security because each human has a unique voice pattern and unique feature. The the result in the feasible way to increase the accuracy and efficiency of systemoutput.

The future enhancement of the system can be extended to recognize for more complicatednoise sample (.wav file). The health condition of the speaker can also identify separate theindividual speaker classification and age also possible to detect for mix mode gender speaker.

7. ACKNOWLEDGMENTS

The authors would like to thank the researchers as well as publishers for making their resources available and teachers for their guidance. We are thankful to the Ramesh M kagalkar for his valuable guidance and constant guidelines also thank full the computer department staff of DYPSOET, lohegoan, Pune and support. Finally, we would like to extend a heartfelt gratitude to friends and family members

8. REFERENCES

[1] Gil Dobry, Ron M. Hecht, Mireille Avigal and Yaniv Z, SEPTEMBER, 2011. Supervector Dimension Reduction for Efficient Speaker Age Estimation Based on the Acoustic Speech Signal,IEEE transaction V.19, NO. 7.

[2] Hugo Meinedo1 and Isabel Trancoso, 2008Age and Gender Classification using Fusion of Acoustic and Prosodic Features,Spoken Language Systems Lab, INESC-ID Lisboa, Portugal, Instituto Superior Tecnico, Lisboa, Portugal.

[3] Ismail Mohd Adnan Shahin, 2013Gender-dependent emotion recognition based on HMMs and SPHMMs,Int J Speech Technol, Springer 16:133141.

[4] Mohamad Hasan Bahari and Hugo Van h, ITN2008 Speaker Age Estimation and Gender Detection Based on Supervised NonNegative Matrix Factorization, Centre for Processing Speech and Images Belgium.

[5] Shivaji J Chaudhari and Ramesh M Kagalkar, May 2015 Automatic Speaker Age Estimation and Gender Dependent Emotion Recognition, International Journal of Computer Applications(IJCA) (0975 - 8887),Volume 117 No. 17.

[6] Shivaji J. Chaudhari and Ramesh M. Kagalkar, July 2015 A Methodology for Efficient Gender Dependent Speaker Age and Emotion Identification System,International Journal of Advanced Research in Computer and Communication Engineering(IJARCCE) ISSN 2319-5940,Volume 4, Issue 7.

[7] Chul Min Lee and Shrikanth S. Narayanan, 2005 Toward Detecting Emotions in Spoken Dialogs, IEEE transaction 1063-6676.

[8] Tetsuya Takiguchi and Yasuo Ariki, 2006 Robust feature extraction using kernel PCA,Department of Computer and System Engg Kobe University, Japan, ICASSP 1-4244-0469.

[9] Michael Feld, Felix Burkhardt and Christian Muller, 2010 Automatic Speaker Age and Gender Recognition in the Car for Tailoring Dialog and Mobile Services,German Research Center for Artificial Intelligence, INTERSPEECH.

[10] M A. Hossan, Sheeraz Memon and Mark A Gregory, A Novel Approach for MFCC Feature extraction, RMIT university, Melbourne, Australia, IEEE, 2010.

[11] Ruben Solera-Ure, 2008 Real-time Robust Automatic Speech Recognition Using Compact Support Vector Machines,TEC 2008-06382 and TEC 2008-02473.

[12] Wei HAN and Cheong fat CHAN, 2006 An Efficient MFCC Extraction Method in Speech Recognition,Department of Electronic Engineering, The Chinese University of Hong Kong Hong Kong, 7803-9390-06/IEEE ISCAS.

[13] AU Khan and L. P. Bhaiya, 2008 Text Dependent Method for Person Identification through Voice Segment,ISSN- 2277-1956 IJECSE.

[14] Felix Burkhardt, Martin Eckert, Wiebke Johannsen and Joachim Stegmann, 2010A Database of Age and Gender Annotated Telephone Speech, Deutsche Telekom AG Laboratories, Ernst-Reuter-Platz 7, 10587 Berlin, Germany.

[15] Lingli Yu and Kaijun Zhou, March 2014, A Comparative Study on Support Vector Machines classifiers for Emotional Speech Recognition, Immune Computation (IC) Volume2, Number:1.

[16] Rui Martins, Isabel Trancoso, Alberto Abad and Hugo Meinedo, 2009, Detection of Childrens Voices, Intituto Superior Tecnico, Lisboa, Portugal INESC-ID Lisboa, Portugal.

[17] Chao Gao, Guruprasad Saikumar, Amit Srivastava and Premkumar Natarajan, 2011, Open set Speaker Identification in Broadcast News, IEEE 978-1-4577-0539.

[18] Shivaji J Chaudhari and RameshMKagalkar, 2014, A Review of Automatic Speaker Age Classification, Recognition and Identifying Speaker Emotion Using Voice Signal, International Journal of Science and Research (IJSR 2014), ISSN(Online): 2319-7064,Volume 3.

[19] M Ferras, C CLeung, C Barras and Jean Luc Gauvain, 2010, Comparison of Speaker Adaptation Methods as Feature Extraction for SVM-Based Speaker Recognition,IEEE Transaction 1558-7916.

[20] Chao Gao, Guruprasad Saikumar, Amit Srivastava and Premkumar Natarajan, 2011, Open-SetSpeaker Identification in Broadcast News, IEEE 978-1-4577-0539.

[21] ChaoWang, Ruifei Zhu, Hongguang Jia, QunWei, Huhai Jiang, Tianyi Zhang and LinyaoYu, 2013, Design of

Speech Recognition System, IEEE 978-1-4673-2764-0/13.

[22] Manan Vyas, 2013"Gaussian Mixture Model Based Speech Recognition System Using Matlab",Signal and Image Proc An International Journal (SIPIJ) Vol.4, No.4.

9. AUTHORS PROFILE

Shivaji J Chaudhari Research Scholar Dr. D.Y.Patil School of Engineering and Technology, Charoli, B.K.Via Lohegaon, Pune, Maharashtra, India. University of Pune. He received B.E. in Information Technology from SVPM COE Malegaon, Baramati, and Pune University. Currently He completed M.E. in Computer Network from Dr. D. Y. Patil School of Engineering & Technology, Pune, and University of Pune.

Prof Ramesh. M. Kagalkar was born on Jun 1st, 1979 in Karnataka, India and presently working as an Assistant. Professor, Department of Computer Engineering, Dr.D.Y.Patil School Of Engineering and Technology, Charoli, B.K.Via – Lohegaon, Pune, Maharashtra, India. He has 13.5 years of teaching experience at various institutions. He is a Research Scholar in Visveswaraiah Technological University, Belgaum, He had obtained M.Tech (CSE) Degree in 2006 from VTU Belgaum and He received BE (CSE) Degree in 2001 from Gulbarga University, Gulbarga. He is the author of text book Advance Computer Architecture which cover the syllabus of final year computer science and engineering, Visveswaraiah Technological University, Belgaum. One of his research article "A Novel Approach for Privacy Preserving" has been consider as text in LAMBERT Academic Publishing, Germany (Available in online). He is waiting for submission of two research articles for patent right. He has published more than 25 research papers in International Journals and presented few of there in international conferences. His main research interest includes Image processing, Gesture recognition, speech processing, voice to sign language and CBIR. Under his guidance four ME students awarded degree in SPPU, Pune, five students at the edge of completion their ME final dissertation reports and two students started are started new research work and they have publish their research papers on International Journals and International conference.