ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 11, November 2018

Age and Gender Identification by Indian Multilingual Speech Sample: A Review

Barkha Shrivastava, Vinay Jain

M. Tech Scholar, Communication System, Shri Shankaracharya Technical Campus (SSGI), Bhilai, CG, India

Associate Professor, Department of Electronics and communication Engineering, Shri Shankaracharya Technical Campus (SSGI), Bhilai, CG, India

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 in which the vocal cord is the essential sound source, which plays an important role in the conversation. The applications of speech or voice processing technology play a crucial role in human computer interaction. The system improves gender identification, age group classification and emotion recognition performance. The performance of the age and gender recognition system depends on the speech features used. As the first speech feature, the fundamental frequency was selected. Fundamental frequency is the main differentiating factor between male and female speakers. Also, fundamental frequency for each age group is different. So in order to build age and gender recognition system, fundamental frequency was used. To get the fundamental frequency of speakers, harmonic to sub harmonic ratio method was used. The speech was divided into frames and fundamental frequency for each frame was calculated. In order to get the fundamental frequency of the speaker, the mean value of all the speech frames were taken. It turns out that; fundamental frequency is not only a good discriminator gender, but also it is a good discriminator of age groups simply because there is a distinction between age groups and the fundamental frequencies. Mel Frequency Cepstral Coefficients (MFCC) is a good feature for speech recognition and so it was selected. Using MFCC, the age and gender recognition accuracies were satisfactory. As an alternative to MFCC, Shifted Delta Cepstral (SDC) was used as a speech feature. SDC is extracted using MFCC and the advantage of SDC is that, it is more robust under noisy data. It captures the essential information in noisy speech better. From the experiments, it was seen that SDC did not give better recognition rates because the dataset did not contain too much noise. Lastly, a combination of pitch and MFCC was used to get even better recognition rates. The final fused system has an overall recognition value of 64.20%.

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).

I. INTRODUCTION

Human interaction with computers in done in many ways and the interface between human and the computer is crucial to facilitate this interaction. Maximum desktop applications, internet using browsers like Firefox, chrome and internet explorer. The computers make use of the prevalent Graphical User Interfaces (GUI). Voice User Interfaces (VUI) is used for speech recognition and synthesizing systems. Human Computer Interaction (HCI) aims to improve the interface between users and computers by making computers more usable and receptive to users need. There are many speaker characteristics that have useful applications. The most popular include gender, age, health, language, dialect, accent, socialist, idiolect, emotional state and attention 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 and smart workspaces. For example, the spoken dialogue system provides services in the domains of finance, travel, scheduling, tutoring. The systems need to gather information from the user automatically in order to provide timely and relevant services.

DOI: 10.15680/IJIRCCE.2018. 0611044

II. LITERATURE SURVEY

In paper [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 genders-based 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. These two approaches use the GMM super vectors as features for a classifier model. Age group classification assigns an age group to the speaker and age regression estimates the speaker's precise age in years.

In paper [2] presents gender detection is an extremely useful task for an extensive variety of voice or speech-based applications. In the spoken language systems INESC ID, the gender identification component is initial and the basic component of our voice processing system, where it is utilized prior to speaker clustering, in order to avoid mixing speakers between male and female gender in the same cluster. Gender information (male or female) is also used to create gender dependent acoustic module for speech recognition.

In paper [3], it introduces new gender detection and an age estimation approach is proposed. To develop this method, after deciding an acoustic features model for all speakers of the sample database, Gaussian mixture weights are extricated and connected to build a super vector for each speaker. Then, hybrid architecture of General Regression Neural Network (GRNN) and Weighted Supervised Non-Negative Matrix Factorization (WSNMF) are developed using the created super vectors of the training data set. The hybrid method is used to detect the gender speaker while testing and to estimate their age. Different biometric features can be used for forensic identification. Choosing a method depends on its use and efficient reliability of a particular application and the available data type. In some crime cases, the available evidence or proof might be in the form of recorded voice. Speech patterns can include unique and important information for law enforcement personnel.

In paper [4], it mainly focused on enhancing emotion recognition and identification performance based on two stages that is combination of gender recognizer and emotion recognizer. The system work is a gender dependent, text independent and speaker independent emotion recognizer. Both Hidden Markov Model (HMM) and Supra segmental Hidden Markov Model (SPHMM) have used as classifiers in the two-stage architecture. This architecture has been evaluated on two different and separate speech databases. The two databases are emotional prosody speech and transcripts database and human voice collected database.

In paper [5], it explores the detection of specific type emotions using discourse information and language in combination with acoustic signal features of emotion in speech signals. The main focus is on a detecting type of emotions using spoken language data obtained from a call centre application. Most previous work in type emotion recognition has used only the acoustic features information contained in the speech. The system contains three sources of information, lexical, acoustic and discourse is used for speaker's emotion recognition.

In paper [6], it develop models for detecting various characteristics of a speaker based on spoken the text alone. These characteristics or attributes include whether the speaker is speaking native language, the speakers age and gender, the regional information reported by the speakers. The research explores various lexical features information as well as features inspired by linguistic (a language related) information and a number of word and dictionary of affect in language. This system suggests that when audio or voice data is not available, by exploring effective audio feature sets only from uttered text and system combinations of multiple classification algorithms, researcher build statistical models to detect these attributes of speakers, equivalent to frameworks that can explore the audio information.

In paper [7], it presents speaker characteristic recognition and identification field has made extensive use of speaker MAP adaptation techniques. The adaptation allows speaker model feature parameters to be estimated using less speech data than needed for Maximum Likelihood (ML) training method. The Maximum Likelihood Linear Regression (MLLR) and Maximum a Posteriori (MAP) techniques have typically been used for speaker model adaptation. Recently, these adaptation techniques have been incorporated into the feature extraction stage of the SVM classifier-based speaker identification and recognition systems.

In paper [8], human's emotional speech recognition contributes much to create harmonious human to machine interaction, additionally with many potential applications. Three approaches to augment parallel classifier are compared for recognizing emotions from a speech by the speech database. Classifier applied on prosody, spectral, MFCC and other common features. One is standard classification schemes (one versus one) and two methods are

DOI: 10.15680/IJIRCCE.2018. 0611044

directed a cyclic Graph (DAG) and Unbalanced Decision Tree (UDT) that can form a binary decision tree classifier. The hierarchical classification technique of feature driven hierarchical SVMs classifiers is designed, it uses different feature parameters to drive each layer and the emotion can be sub divided layer by layer. Finally, analysis of the classification rate of those three extends binary classification, DAG system performs the best for testing database and standard classifier is not far behind, the UDT is the poorest because of relying on upper layer order processing.

In paper [9], the extraction and matching process is implemented after the signal pre-processing is performed. No parametric method for modeling the human voice processing System. The nonlinear sequence alignment called as Dynamic Time Warping (DTW) used as features matching techniques. This paper presents the technique of MFCC feature extraction and wrapping technique to compare the test patterns.

III. METHODOLOGY

FEATURE EXTRACTION

The extraction of the best parametric representation of the acoustic signals of the human voice is 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.

Pre-emphasis is passing of a signal through many filters, which emphasizes higher frequencies. It increases the energy level of the signal at higher frequency.

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.

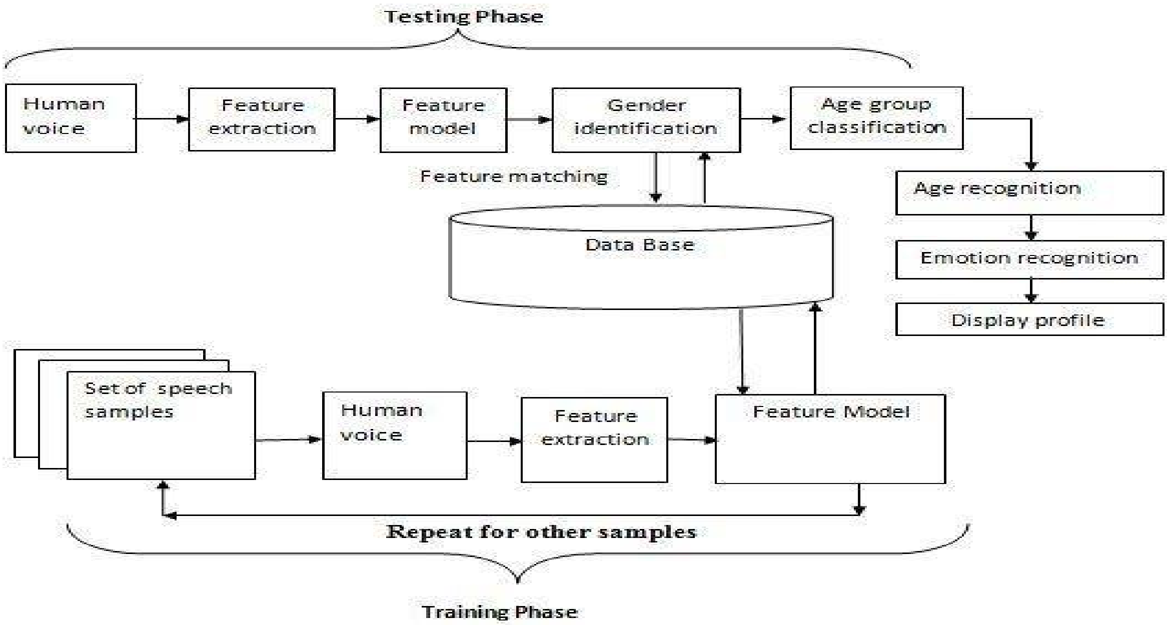

Figure1. Block Diagram of Age and Gender Identification.

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.

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 period diagram.

Mel Filter Bank processes the frequency range in FFT spectrum is very wide and voice signal does not follow the linear scale.

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 coefficients is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector.

Delta energy and delta spectrum voice signal the frame changes, such as the slope of a format at its transitions. Therefore, there is a need to add features related to the change in cepstral features over time.

All these methods will be put into practice and a real age and gender recognition system will be implemented.

This project was developed in Matlab. Some of the Matlab's libraries for signal processing and speech processing which are written by scientists were used to speed up the process and also for their reliability. Below can be seen some short explanations about the toolboxes that have been used.

SIGNAL PROCESSING (VOICE BOX) TOOLBOX

Developing and age and gender recognition system is mainly a signal processing and classification task. To do front end signal processing, and then noise reduction, and then extracting acoustic features such as MFCC, Matlab's Voice box toolbox was used. This toolbox is very powerful and it covers a lot of speech processing tools. For getting SDC features out of MFCC, this library was used.

SDC EXTRACTION

So, two features are used for age and gender recognition in this project. The first feature used was SDC. As mentioned earlier, SDC is derived from MFCC. After applying some preprocessing methods, MFCC was calculated with the 12 coefficients and each frame size being 30ms. Following MFCC calculation, a RASTA filter was applied to the signal to remove the channel noise. After that, SDC was calculated using the library [29] with the N-d-p-k parameters set to 7-1-3-7. Following this step, the mean value of SDC was normalized to zero and the standard deviation was normalized.

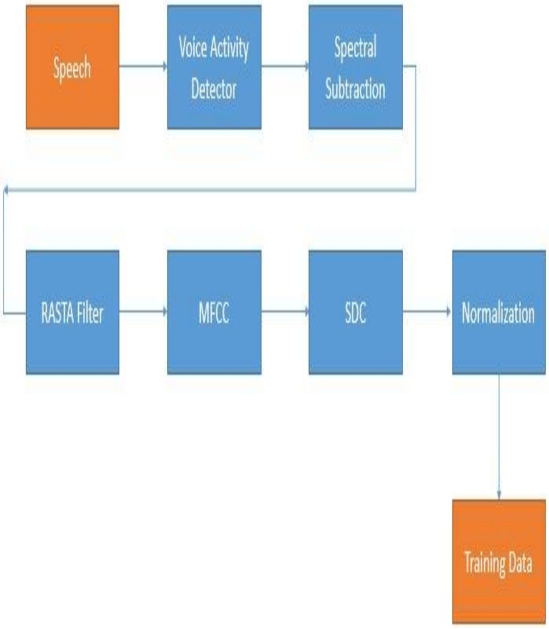

Figure2. Block Diagram for SDC Feature Extraction Algorithm for the Age and Gender Recognition System.

DOI: 10.15680/IJIRCCE.2018. 0611044

PITCH EXTRACTION

For getting the other speech feature pitch, the library provided by [27] was used. The algorithm uses sub harmonic to harmonic ratio to get the pitch information. Also the frequency range was set between 100 Hz and 300 Hz for more reliable pitch estimation. Each frame size was set to 25 ms. after getting the pitch information for all frames in the training set; the mean value for each training set was taken as the pitch of the training. After using the previous pre-processing and SDC feature extraction steps, an SVM was trained for both genders and age groups. Nonlinear RBF kernel was used in this test. The database contained 4 labels. These labels included young adult male, young adult female whose age ranges between 20 and 40 years and also middle age male and middle aged female whose age range between 40 and 65 years old. For SVM training, the algorithm described in this paper was used [31]. The parameter of SVM was first selected manually and second time they were selected with cross validation after training on a balanced sub set.

For pitch calculation, again the same pre-processing techniques were applied to the training set. After that step, pitch extraction algorithm which uses harmonic to sub harmonic ratio was executed on the each training sample. The frame window was set to 25ms. And after getting the pitch value for each frame in all the training examples, the mean value of each speech example was taken as one feature vector in the training set. To make the model simpler, a threshold value was selected to be 200 Hz. The value below 200 Hz was considered as male speech and the frequency above 200 Hz was considered as female speech.

To use pitch and MFCC score together, they were first scaled to the same dimension and later was done a weighted sum to get the final result. The final fused age and gender recognition system can be seen.

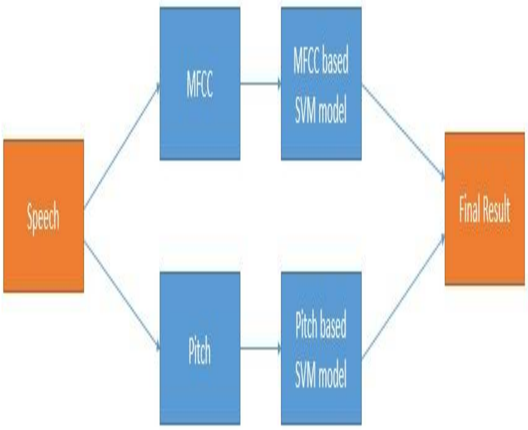

Figure3. Block Diagram for the Final Fused Age and Gender Recognition System

TABLE 1: CRITERIA FOR GENDER AND AGE DETECTION

Age Group	Gender Identity	Age Recognition
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

IV. CONCLUSION

Thus, the proposed system helps to identify, classify and recognize exact speaker age with emotion and displaying profiles of speaker using the trained database. The speaker profile is helpful in many applications like for advertisement, targeting to particular people, automatically identification of 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 result is in the feasible way to increase the accuracy and efficiency of system output.

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

ACKNOWLEDGEMENT

I would like to acknowledge my gratitude to a number of people who have helped me in different ways for the successful completion of my thesis. I take this opportunity to express a deep sense of gratitude towards my guide, Mr.Vinay Jain Associate Professor (ET&T), Faculty of Engineering & Technology, SSTC-SSGI, Bhilai for providing excellent guidance, encouragement and inspiration throughout the project work. Without his invaluable guidance, this work would never have been a successful one. I am thankful to Mr. Chinmay Chandrakar HOD, ET&T Department, and Dr.P. B. Deshmukh, Director, SSTC-SSGI, Faculty of Engineering & Technology, Bhilai for their kind help and cooperation. I would also like to thank Mr. Chandrashekhar Kamargaonkar M.E. In-charge, SSTC-SSGI, Mr. Sharad Mohan Shrivastava, Faculty of Engineering & Technology, Bhilai for his kind support and helpful suggestions. I feel immensely moved in expressing my indebtedness to my parents whose sacrifice, guidance and blessings helped me to complete my work.

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 BaSupervised NonNegative Matrix Factorization, Centre for Processing Speech and Images Belgium.

[5] Shivaji J Chaudhari and Ramesh M Kagalkar, May 2015Automatic Speaker Age Estimation 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 14244-0469.

[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, 78039390-06/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.

DOI: 10.15680/IJIRCCE.2018. 0611044

[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 978-1-45770539.

[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 978-1-45770539.

[21] ChaoWang, Ruifei Zhu, Hongguang Jia, QunWei, Huhai Jiang, Tianyi Zhang and LinyaoYu, 2013, Design of Speech Recognition System, IEEE 978-1-4673-27640/13.

[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.