Gender Recognition for Speech Applications based on Support Vector Machines

Shikha Panwar[1], Mr. Hitanshu Saluja[2]
[1]M.Tech. Scholar, [2]Assistant Professor
[12]ECE Department, Ganga Technical Campus
Soldha, Bahadurgarh, Haryana, India

Abstract- Automatic age and gender recognition for speech applications is very important for a number of reasons. One of the reasons is that it can improve human-machine interaction. For example, the advertisements can be specialized based on the age and the gender of the person on the phone. It also can help identify suspects in criminal cases or at least it can minimize the number of suspects. Some other uses of this system can be applied for adaptation of waiting queue music where a different type of music can be played according to the person's age and gender. And also using this age and gender recognition system, the statistics about age and gender information for a specific population can be learned. Machine learning is part of artificial intelligence which aims to learn from data. Machine Learning has a long history. But due to some limitations, for ex., the cost of computation and due to some inefficient algorithms, it was not applied to speech recognition tasks. Only for a decade, researchers started to apply these algorithms to some real world tasks, for ex., speech recognition, computer vision, finance, banking, robotics etc. In this thesis, recognition of age and gender was done using a popular machine learning algorithm and the performance of the system was compared. Also the dataset included real-life examples, so that the system is adaptable to real world applications. To remove the noise and to get the features of speech examples, some digital signal processing techniques were used. Useful speech features that were used in this work were: pitch frequency and cepstral representations. 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. 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% on ELSDSR speech corpus.

I. INTRODUCTION

a. Background

The technology has improved significantly over the last decade. With the improvements in new algorithms, big data technologies and data storage methods, it is continuing to improve more. Parallel to these technological improvements, speech recognition systems have improved significantly. Now we are able to talk to our phones to get directions, to ask for some information or to send a text message. Not many years ago, Microsoft demonstrated a speech recognition system in China which not only has a much improved accuracy but it also translated English into Chinese in the real time with the speaker's accent and speech patterns. Due to these improvements and out of need, bio metric age and gender systems have emerged. Biometrics is a branch of computer science that studies the characteristics and traits of humans for identification and access control or surveillance purposes. There are two characteristics in biometric identifiers, physiological characteristics such as face recognition, DNA, retina or fingerprint and behavioral characteristics such as typing rhythm, voice or gait. In this project, our focus is on voice which is one of the behavioral characteristics in biometric identifiers Age and gender recognition for speech applications has many practical applications and it can be useful in many applications such as human-computer interaction or information retrieval.

This system discussed in this thesis can be applied to many speech recognition systems. Age and Gender recognition system training and testing steps can be seen from Figures 1 and 2.

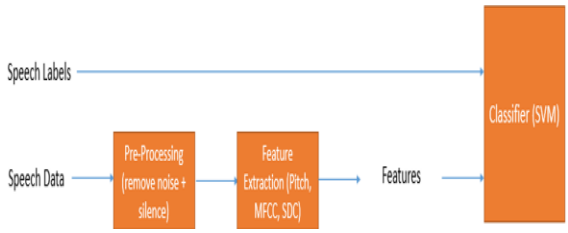

Fig.1: Age and Gender Recognition System Training

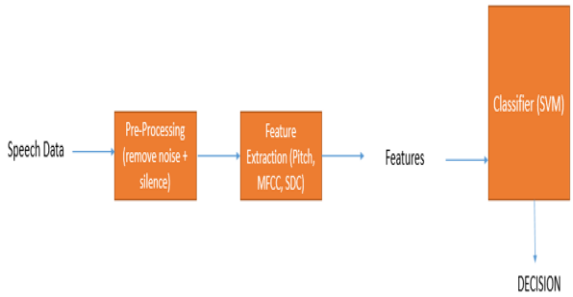

Fig.2: Age and Gender Recognition System Testing

Our goal is this thesis is to create a robust age and gender recognition system for speech applications which also gives good recognition rates under real world conditions. Some of the areas that this system can be used in are explained in more detail below:

- **Phone Ads:** A good use of Age and Gender Recognition System is phone ads. Many big companies play phone ads while a customer waits on the line. So in order to play the same ad for every customer, it can be customized. And the end result is that, the ads become more efficient and possibly the sales increase because the ads are more relevant to the specific gender and age. Criminal cases: This is also a good example to the usage of this system. It turns out that, a lot of times, in criminal cases the evidence is in the form of telephone speech. And by analyzing age and gender of the suspects, number of suspects can be narrowed down.
- **Waiting queue music**: This is also another example to the usage of the system. Waiting queue music on phone lines can be customized according to the age and gender of the caller. This can help increase customer satisfaction of the companies.
- **Statistics of a certain population:** Age and gender recognition system can be handy when researchers or companies collect age and gender information of a certain group of people. That information can help understand the experiment better and make better analysis.

b. The Problem

Age and gender recognition for speech applications has not been researched in the academia widely. This is an open research area. Although, there are some attempts at implementing this system, it does not give better recognition values especially in noisy environments. Since most real world applications require noisy environments, it is particularly important to have a robust age and gender recognition system. And this system should be implemented into speech recognition system. Many researchers tried different speech features and also different classifiers to solve this problem. But still the system has not been perfected. There is no perfect solution to this problem. In order to develop better age and gender recognition systems, researchers need to come with better speech features which capture most of the essential information in the speech. English Language Speech Database for Speech Recognition (ELSDSR) speech corpus was used in this project. It is a dataset made by University of Denmark (DTU) for Speaker Recognition, Age and Gender Recognition purposes. There are 22 speakers in the dataset, 10 of the speakers are female and 12 or the speakers are female. It was recorded in a chamber with 16 kHz sampling frequency and with a bit rate of 16.

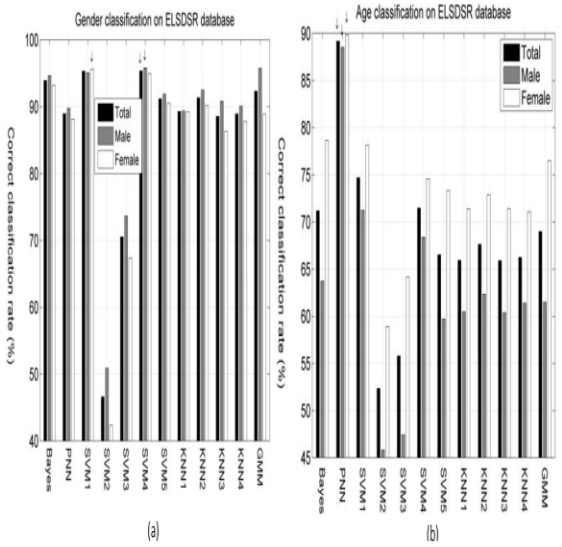

Fig.3: Current Age and Gender Recognition Performance with Different Classifiers on ELSDSR

From Figure 3(a), it can be seen that, gender classification rates are pretty high with many classifiers. The highest scores were achieved with Probabilistic Neural Networks (PNN) and SVM.

Also male recognition rates are very similar to female recognition rates. In Figure 1.3(b), age classification with different classifiers can be seen. In age classification, the rates are lower compared to gender classification which can be expected. The highest classification accuracy was achieved with PNN. The second highest classification accuracy was achieved with SVM.

c. Motivation

The motivation of this thesis is to implement a robust, integrable age and gender recognition system for speech applications which is also robust under noisy conditions for real world applications. To do that, we will use the power of the state of the art machine learning algorithms for pattern recognition. Actually the motivation behind this work can be addressed by answering some of the questions below? Why using age and gender recognition system? When implementing the system, which classifier to use? Is it going to perform well under real world conditions? The advantage of this age and gender recognition system is that, it is going to be easy to successfully implement. With a little effort we will be able to extract age and gender information of the speaker without being invasive. So despite other biometric properties, using the voice of the speaker for these purposes is allowed and it is not as hard as trying to capture his/her iris information to access to some of information. Of course the telephone speech is everywhere. It is very easy to access. It is also a good idea to combine various biometrics when getting age and gender information about the people. But due to the system complexity, cost of implementation and scarcity of data, it is more reasonable to use speech signals in order to recognize age and gender of a speaker since it is one of the easiest form of data that can be accessed.

d. Thesis Goals and Outline

The main goal in this thesis is to improve the age and the gender recognition for speech applications such that, under real world conditions the system will give good recognition values. To do so, state of the art machine learning algorithms will be used for pattern recognition problem. Secondary reasons in this work can be seen below:

- Study the speech features combined with the state of the art machine learning algorithms such as Support Vector Machines (SVM) to create such a recognition system. Decide which machine learning algorithms to use.
- Formulate a text-independent age and gender recognition system
- Choice of robust representation of speech signal for age and gender recognition (future selection), which feature vectors gives the best performance of the system
- Integration of age and gender recognition system into speech recognition system
- Evaluation of age and gender recognition system under real world conditions.

II. LITERATURE SURVEY

Speech recognition has in years has become a practical concept, which is now being implemented in different languages around the world. Speech recognition has been used in real-world human language applications, such as information recovery. Speech in human can be said as the most common means of the communication because the information maintains the basic role in conversation. The conversation or speech that is captured by a microphone or a telephone is converted from acoustic signal to a set of words in speech recognition. A set of word can either be the final result or it can then apply the synthesis to pronounce into sounds, which means speech-to-speech. Its means that, speech recognition can serve as the input to further linguistic processing in order to achieve speech understanding. Speech recognition systems can be characterized by environment, vocabulary, acoustic model, speaking style, speaking mode, language model, perplexity, Signal to Noise Ratio (SNR) and transducer. The literature survey for research was done by referring to the journal papers, conference papers, articles, books, internet and databases. Overall, this chapter describes a review of speech recognition task, speech recognition approaches, current speech recognition system, Tamil speech recognition system as well as different type of methods applied to speech recognition system. Based on the review of the advantages and disadvantages, this thesis discusses the most suitable techniques and methods to develop a speech recognition system.

There are three approaches to speaker independence. The first approach is to use knowledge engineering techniques to find perceptually motivated speech parameters that are relatively invariant between speakers. The justification for this approach is that if an expert spectrogram reader can read spectrograms with high accuracy, it should be possible to find the invariant parameters such as expert employs. Furthermore, if these invariant parameters can be found, then SI recognition is as easy as SD recognition. This idea has been carried out by many researchers (Thompson & Laver 1987, Cole 1986). In particular, Cole achieved high accuracy on very limited task in recognition of English letter (Cole et al. 1983), but has not been successful on more difficult one in phonetic classification (Cole et al. 1986b). The second approach is to use multiple representations for each reference to capture the inter-speaker variability. The most well-known studies were performed byresearchers at Bell Lab on clustering **(Levinson et al. 1979, Rabiner et al. 1979).**Typically, each word in the vocabulary is uttered by many speakers of different sex, age and dialect; these multiples examples are then divided into several clusters, and a prototype is generated from each cluster.

Lee (Lee 2016) introduced a learning algorithm that clusters and generalizes at a sub word level. Like the knowledge engineering approach, the multi representation approach produced good results for limited tasks, but has not be successfully extended to a large vocabulary task. This is because the greater number of divided cluster, the greater the computation involved at the recognition stage. The final approach is using speaker adaptation technique which turns SI models into SD models. Speaker adaptation begins with an

existing set of model parameters, and a small number of adaptation sentences from an adapted speaker. These sentences are used to modify the parameters so that they are adjusted to the adapted speaker. The systems described in this thesis are intended for SD application. This is because training data collected at this study is small and unable to train accurate SI models.

Based on Speech:

In isolated word recognition, the assumption is that the speech to be recognized comprised a single word of phrase and was to be recognized as a complete entity with no explicit knowledge or regard for the phonetic content of the word of phrase. Another implicit assumption is that each spoken utterance has a clearly defined beginning and the ending point that could be found using some type of endpoint detector. This system work well and entirely appropriate in certain application such as "command and control" application, in which user is required to speak the command words at once a time. In continuous speech recognition, the speech to be recognized is a sequences of words uttered in a fluent (continuous) manner. There are reasons why research on isolated word recognition task should be extended to continuous speech recognition. First concern is the speed of speaking. The primary advantage of ASR is speed, since speech is the highest capacity human output channel. Talking is much faster than typing. However, with isolated word speech, this advantage diminishes substantially.

III. RESULT AND DISCUSSION

Even till now book keeping is done to maintain the records of people entering the locker room. This application is developed to automate the locker system. The existing system is weak whereas this system can make the environment full proof secure.

In these days, improvement in techniques for bank locker security is increasing. Banks need to make the whole environment secure and automated. Hence this project plays an important role to maintain the security of the locker room door. This project will be modified according to the requirement of the particular bank .

The remaining area of concern for advancement or modification is to implement the following features to the software:

- Enhance its features and increasing the number of services provided.
- Maintaining log files for unauthorized user.
- Implementing higher level of security.

IV. AGE AND GENDER RECOGNITION SYSTEMS

After some background on how speech is formed and some pre-processing techniques, in this chapter age and gender recognition systems will be described. Also training and testing phases of this kind of systems will be discussed in addition to the features extracted from speech. Also some previous research on age and gender recognition systems and used feature extraction and classification techniques will be reviewed.

a. Acoustic Features

Acoustic features of speech uses some acoustic characteristics (physical characteristics such as loudness, amplitude, frequency etc.) of speech to extract some useful information. In terms of acoustic speech features, one feature dominates all the others. That one is the Mel-Frequency Cepstral Coefficients (MFCC). MFCC has been used in many speech applications from speech recognition to language identification, from gender recognition to age recognition. Some of the uses of MFCC in the literature can be seen from papers [4], [14] or [16]. Figure 4 shows a model for this age and gender recognition system:

Fig.4: Model of Age and Gender Recognition System

b. Mel Frequency Cepstral Coefficients (MFCC)

As mentioned earlier, MFCC is one of the most widely used speech extraction feature in speech applications. Speech is formed by the shape of human's vocal tract and also lips and tongue. So in order to recognize what has been said, the shape of the vocal filter must be modelled. And MFCC tries to model this vocal tract filter in short time power spectrum. MFCC was introduced in 1980s by Davis and Mermelstein [34]. And since then it has been the state-of-the-art speech feature in acoustic domain. Before the invention of MFCC, some other extraction methods were used such as Linear Prediction Coefficients (LPC) and also Linear Prediction Cepstral Coefficients (LPCC). Figure 5 represents a diagram for MFCC computation.

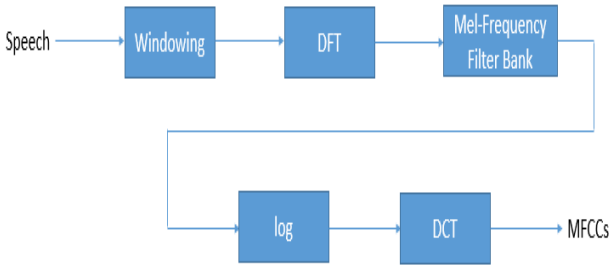

Fig.5: Steps for MFCC Computation

1. The first step in computing MFCC is the windowing. Speech is windowed into 20-40 ms frames. If the speech duration is less, there might now be enough examples for a reliable computation or otherwise if speech duration is more than 40 ms in this case there are much changes in the signal.

2. Second step is where the power spectrum of each frame is calculated. This works as a similar way of the human cochlea and different frequencies cause vibration in different parts of this organ. So in the end, the frequencies found in the signal are calculated.
3. After the second step, the periodogram is found and it contains a lot of information about speech. Actually, if the frequency increases, cochlea cannot make a good job at discerning the frequencies. At low frequencies, filter bank is narrower and at high frequencies filter bank becomes wider. So Mel-Frequency Filter Bank is used to determine the energy levels of frequency regions.
4. Next step, the log is taken of filter bank energies. The reason for this is that, the human ear does not have a linear scale in term of hearing. For example, in order to hear two times louder, the energy must be eight times the first energy. So by doing this, human ear will be modelled better.
5. Final step is to take the DCT of the log filter bank energies.

c. Pitch Extraction Method

One of the features that can be used for age and gender recognition is pitch or fundamental frequency. As mentioned earlier, the male fundamental frequency changes between 85 Hz to 180 Hz, whereas a typical female fundamental frequency is between 165 Hz to 225 Hz. And also these fundamental frequencies change with age. There was as study done by Russell, Penny and Pemberton.

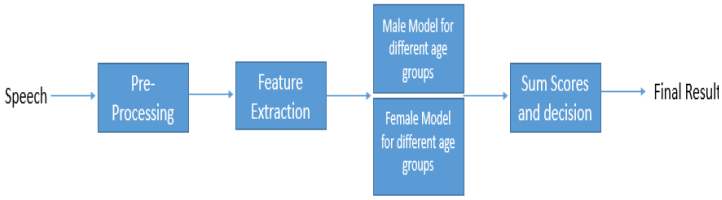

Fig.6: Age and Gender Recognition Model Trained Using MFCC as Features

d. Pitch Based Models

Pitch based model is a model which only takes pitch as a future and it can be used in age and gender identification system. In fact, Microsoft Kinect uses pitch information in order to identify the gender. It basically tries to determine a threshold in the training of the system. It is usually around 200 Hz. And if a new person comes in, if the pitch frequency of him/her is below the threshold, he is a male. Otherwise, she is a female. But this kind of systems work well with clean speech. Also it is not so reliable for determining the age.

e. Models Based on Acoustic Features

Speech applications based on acoustic features of the speech use the acoustic features of the speech such as MFCC or recently introduced SDC. They are of course more reliable compared to pitch based models. After those speech features are extracted from the speech, they are fed into some machine learning classifiers such as Support Vector Machines (SVM) or Gaussian Mixture Model (GMM). SVM has been proven to be a successful classification algorithm for age and gender recognition in [1], [2], [9], and [16].

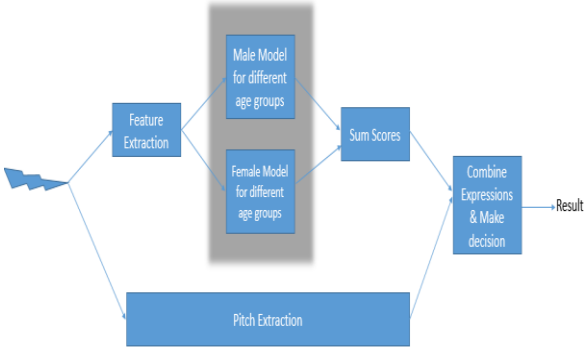

Fig.7: Fused Age and Gender Recognition Model

f. Fused Models

Before 2005, researchers were using only pitch or only MFCC as feature vectors in speech
applications. But after that time, they decided to fuse some features together so the recognition
accuracy will be better than before. Here can be seen some work where researchers fused pitch
and MFCC to make a more robust speech applications. Some of the previous work combining
pitch and MFCC can be seen in [5] and [16].

V. IMPLEMENTATION OF ENHANCED PITCH DETECTION METHOD

a. GRAPHICAL USER INTERFACE

I have made graphical user interface which consist of :FIGURE 8 : ACCESS LEVEL 1 BUTTON

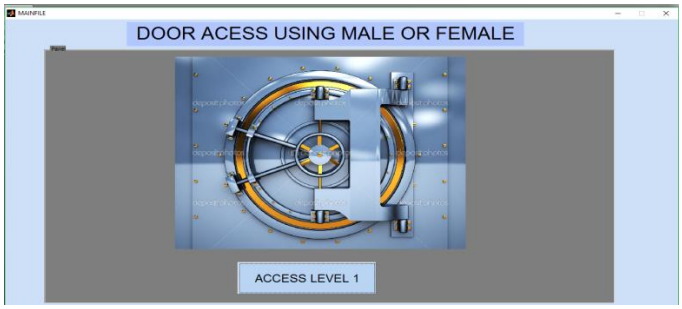

Fig.8: ACCESS LEVEL 1 BUTTON

i. ACCESS LEVEL 1 BUTTON :
It is a button , on clicking this button , the user will enter to the main

screen where following options will be provided :

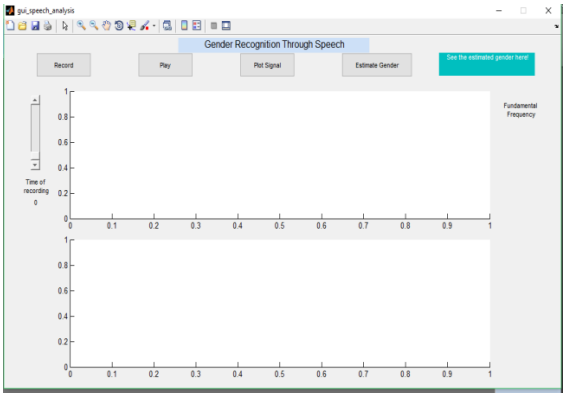

Fig.9: MAIN SCREEN

ii. Time Slider :

Time setting slider for setting the time for recording the voice input .

- Maximum time for recording : 5 sec
- Minimum time for recording : 0 sec

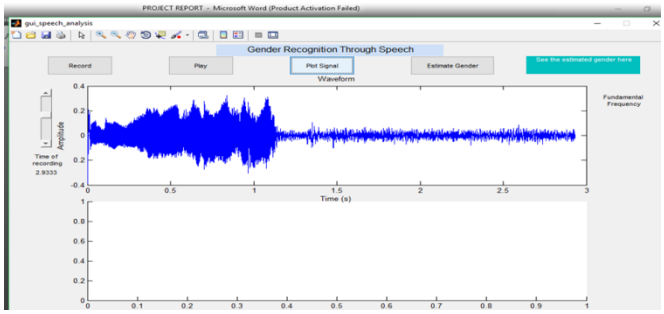

Fig.10: PLOT SIGNAL BUTTON

iii. Estimate GenderButton :

As soon as we click on this button ,it performs four tasks :

- Estimate the gender : whether male or female
- Plot a graph between correlation coefficient and delay
- Open the image linked to voice input
- If the voice input is fast , it signals that and ask for giving the voice input again and that too bit slow .

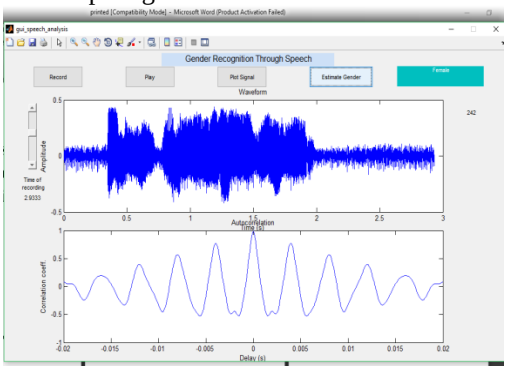

Fig.11 : TASKS PERFORMED BY ESTIMATE GENDER BUTTON

The autocorrelation at 0 delay is computed for appropriate normalization purposes. The autocorrelation function is then searched for its maximum (normalized) value. If the maximum (normalized value) exceeds 0.3, the section is classified as voiced and the location of the maximum is the pitch period. Otherwise, the section is classified as unvoiced. Formula for calculating correlation coefficient :

$$R_n(k) = \sum_{m=-\infty}^{\infty} \left[x(m)w(n-m)x(m+k)w(n-k-m) \right]$$

It basically , generates the result and determine whether

- the input signal is of a male voice or

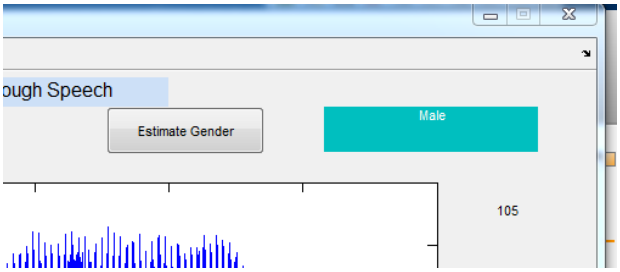

Fig.12: GENDER ESTIMATION OF MALE VOICE

- the input signal is of a female voice .

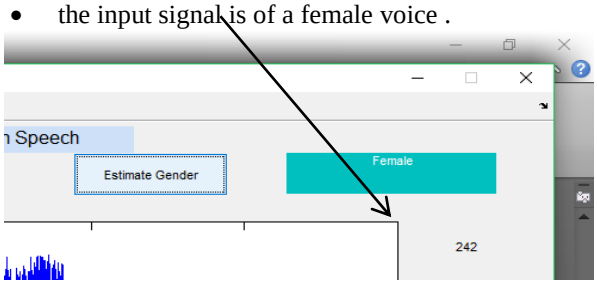

Fig.13 GENDER ESTIMATION OF FEMALE VOICE

The image linked to male voice signal will open having :

- Door Closed : If the input is given in the female locker room

Fig.14: WHEN MALE INPUT HIS VOICE IN FEMALE LOCKER ROOM

- Door Open : If the input is given in the male locker room

Fig.15: WHEN MALE INPUT HIS VOICE IN MALE LOCKER ROOM

The image linked to female voice signal will open having :

- Door Closed : If the input is given in the male locker room

Fig.16: WHEN FEMALE INPUT HER VOICE IN MALE LOCKER ROOM

- Door Open : If the input is given in the female locker room

Fig.17: WHEN FEMALE INPUT HER VOICE IN FEMALE LOCKER ROOM

VI. CONCLUSION AND FUTURE ENHANCEMENT

CONCLUSION

Gender recognition is a task of recognizing the gender from his or her voice. With the current concern of security worldwide speaker identification has received great deal of attention among of the speech researchers. Also a rapidly developing environment of computerization, one of the most important issues in the developing world is speaker identification.

Information about the gender of a person is an important component for effective behavioral analytics. Gender recognition reduces the complexity of automatic speech recognition (ASR) and interactive voice response systems and improves their efficiency.

VII. REFERENCES

[1]. M. Li, K. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer speech and language*, Vol. 27, No. 1, pp. 151-167, Jan. 2013

[2]. H Meinedo and I Trancoso, "Age and Gender Classification Using Fusion of Acoustic and Prosodic Features", *Proc. INTERSPEECH*, pp. 2818-2821, 2010

[3]. R. Nisimura, A. Lee, H. Saruwatari, and K. Shikano, "Public speech-oriented guidance system with adult and child discrimination capability,"*Proc. ICASSP2004*, vol. 1, pp. 433-436, 2004.

[4]. H. Kim, K. Bae, H. Yoon, "Age and gender classification for a home-robot service" Proc. *16th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 122–126, 2007

[5]. W. Li, D. J. Kim, C. H. Kim, and K. S. Hong, "Voice-Based Recognition System for Non-Semantics Information by Language and Gender" *Electronic Commerce and Security (ISECS)*, 2010.

[6]. P. Nguyen, D. Tran, X. Huang, and D. Sharma, "Automatic classification of speaker characteristics" *Communications and Electronics (ICCE)*, 2010.

[7]. G. Dobry, R. M. Hecht, M. Avigal, and Y. Zigel, "Supervector dimension reduction for efficient speaker age estimation based on the acoustic speech signal." *Audio, Speech, and Language Processing*, 2011

[8]. M. H. Bahari, and H. V. Hamme, "Speaker age estimation and gender detection based on supervised non-negative matrix factorization" *Biometric Measurements and Systems for Security and Medical Applications (BIOMS)*, 2011.

[9]. M. H. Sedaaghi, "A comparative study of gender and age classification in speech signals" *Iranian Journal of Electrical & Electronic Engineering*, 2009

[10]. T. Bocklet, G. Stemmer, V. Zeissler, and E. Nöth, "Age and gender recognition based on multiple systems-early vs. late fusion" *INTERSPEECH*, 2010

[11]. S. Schötz, "Acoustic analysis of adult speaker age" *Speaker Classification I*. Springer Berlin Heidelberg, 2007

[12]. M. K. Wolters, V. Ravichander, and R. Steve,"Age recognition for spoken dialogue systems: Do we need it?" *INTERSPEECH*, 2009

[13]. M. Feld, F. Burkhardt, and C. A. Müller, "Automatic speaker age and gender recognition in the car for tailoring dialog and mobile services" *INTERSPEECH*, 2010

[14]. F. Metze, J. Ajmera, R. Englert, and U. Bub, "Comparison of four approaches to age and gender recognition for telephone applications" *Acoustics, Speech and Signal Processing*, 2007

[15]. C. A. Müller, F. Wittig, and J. Baus, "Exploiting speech for recognizing elderly users to respond to their special needs" *INTERSPEECH*, 2003

[16]. M. W. Lee, and K. C. Kwak. "Performance Comparison of Gender and Age Group Recognition for Human-Robot Interaction" *International Journal of Advanced Computer Science & Applications*, 2012