University of South Florida
Scholar Commons

Graduate Theses and Dissertations | Graduate School

10-30-2014

Age and Gender Recognition for Speech Applications based on Support Vector Machines

Hasan Erokyar
University of South Florida, herokyar@hotmail.com

Follow this and additional works at: https://scholarcommons.usf.edu/etd

 Part of the Electrical and Computer Engineering Commons

Scholar Commons Citation
Erokyar, Hasan, "Age and Gender Recognition for Speech Applications based on Support Vector Machines" (2014). *Graduate Theses and Dissertations.*
https://scholarcommons.usf.edu/etd/5356

This Thesis is brought to you for free and open access by the Graduate School at Scholar Commons. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact scholarcommons@usf.edu.

www.manaraa.com

Age and Gender Recognition for Speech Applications based on Support Vector Machines

by

Hasan Erokyar

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Electrical Engineering
Department of Electrical Engineering
College of Engineering
University of South Florida

Major Professor: Ravi Sankar, Ph.D.
Wilfrido Moreno, Ph.D.
Ismail Uysal, Ph.D.

Date of Approval:
October 30, 2014

Keywords: Speech Processing, Pre-Processing, MFCC, SDC, Pitch, SVMs

Copyright © 2014, Hasan Erokyar

www.manaraa.com

DEDICATION

Dedicated to my parents for their love and support throughout my life.

www.manaraa.com

ACKNOWLEDGMENTS

I would like to acknowledge the guidance and support of my advisor, Dr. Ravi Sankar, whose comments and explanations have taught me a lot about speech and research in general. I would also like to thank Dr. Moreno and Dr. Ismail for their advices and for being on my committee. Also I would like to thank Tete Tevi for his valuable comments. Finally, I'd like to thank my friends and family for their encouragement and support.

www.manaraa.com

TABLE OF CONTENTS

www.manaraa.com

www.manaraa.com

www.manaraa.com

LIST OF TABLES

www.manaraa.com

LIST OF FIGURES

www.manaraa.com

www.manaraa.com

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

www.manaraa.com

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% on ELSDSR [32] speech corpus.

www.manaraa.com

CHAPTER 1

INTRODUCTION

1.1 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. It can also improve the intelligibility of systems and can be helpful in speaker recognition and surveillance systems.

www.manaraa.com

Age and gender recognition is studied to some extent and some of the examples can be seen in the literature [1], [2], [4], and [9]. But there are not many articles using current machine learning algorithms for age and gender recognition application. Some researchers tried different speech features and different classifiers in age and gender recognition which can be found in Chapter 2 under overview of the past history. Researchers have recently used SDC instead of MFCC as a speech feature and so in this research we wanted to explore the viability of this feature for age and gender recognition system. ELSDSR [32] read speech corpus is used in this work which is a good database for age and gender recognition. In this work, we used one of the most popular machine learning algorithms that it Support Vector Machines (SVM) and tried to predict the age and gender of speaker. The final recognition accuracy of the system is 64.20% and out of 4 classes, 3 classes have higher recognition values.

Age and gender recognition is the process of extraction the age and the gender information from the uttered speech. This technique enables speech recognition systems to personalize the ads according to the person's age and gender and also it can have some uses in criminal cases since most of the proofs are of telephone speeches. Also it can be used for processing waiting queue music for different genders and age groups. Not all people appreciate the same type music. Older people might like slow music whereas younger people might like rock or metal music. Another usage of this system can be to try to understand age and gender distribution of a population in an experimental study which gives more details about the experiment.

Age and gender recognition system consist of two parts. The first part is called pre-processing and future extraction. The second part is called classification. In the first part, speech is pre-processed using Digital Signal Processing (DSP) techniques and then some of useful features such as MFCC and pitch information are extracted from the speech. Then these speech features

www.manaraa.com

are fed into machine learning classifier and the system is trained using these speech features vectors. In the classification part, a decision is made comparing these feature vectors and finding the best match using SVM.

Also in the first part, some pre-processing techniques such as Voice Activity Detector and spectral subtraction were applied to the speech signals in order to remove the channel noise, background noise and also remove the silence. These methods improved the classification accuracy of the system. For the second part, SVM with Radial Basis Function (RBF) kernel was trained using different parameters to see the performance differences of the system.

In order to do age and gender recognition, we tried the most popular machine learning algorithm which is SVM, and also combined different speech features together. Two speech features were selected mainly which are pitch and MFCC. Pitch is selected because it is the main discriminator between age and genders. MFCC is selected because it was better than the speech feature SDC. Gender recognition is widely studied in the literature. In this work, we not only studied gender recognition but also incorporated age recognition into the system so the system will make better prediction about the speaker. 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.

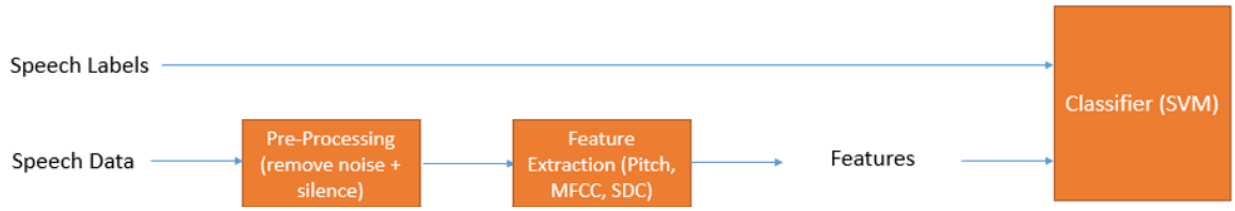

Figure 1. Age and Gender Recognition System Training

www.manaraa.com

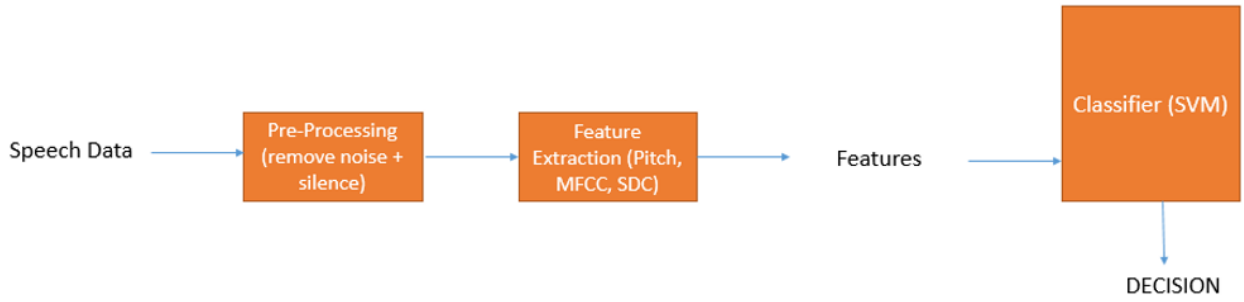

Figure 2. Age and Gender Recognition System Testing

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.

www.manaraa.com

1.2 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.

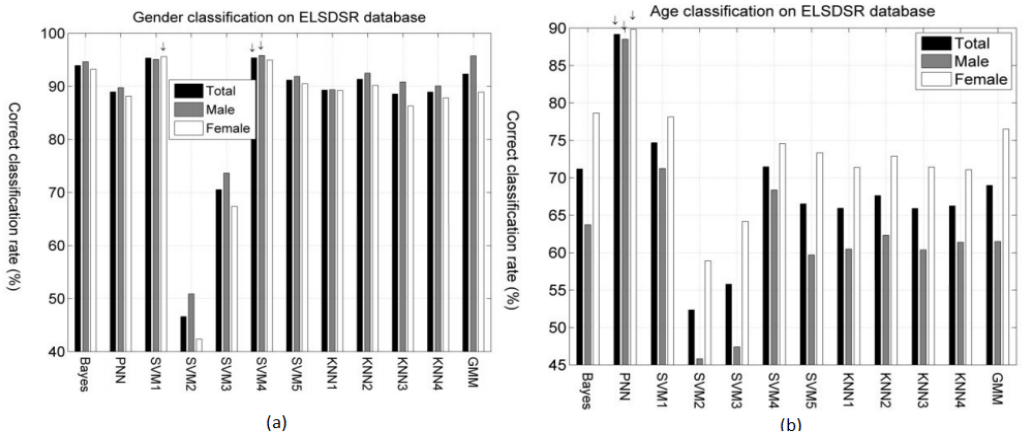

Figure 3. Current Age and Gender Recognition Performance with Different Classifiers on ELSDSR [9]

www.manaraa.com

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 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.

1.3 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.

www.manaraa.com

1.4 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. How does the system work under noisy data? Is it better to train the system under noisy data or clean+noisy data?

The rest of this thesis is organized as follows. In the next sections, each part of the age and gender recognition system is described in details. In chapter 2, some digital signal methods for pre-processing and some useful speech processing topics are discussed. In chapter 3, pre-processing methods such as Voice Activity Detector and spectral subtraction are introduced. In chapter 4, the future extraction parameters and the future extraction methods are explained for acoustic based and pitch based models. In chapter 5, classification part of the age and gender system and decision making are described. In chapter 6, system design and implementation of the

www.manaraa.com

system is mentioned and also various test results are given. Finally, chapter 7 ends with the conclusion and future research.

www.manaraa.com

CHAPTER 2

PAST RESEARCH AND FUSION MODEL

2.1 Overview of Past Research

In [1], Ming Li and his colleagues used seven different methods on the acoustic and prosodic level and also combined all these proposed models to improve age and gender recognition systems accuracy. The proposed methods were GMM, GMM-Mean-SVM, SVM-UWPP-SVM, GMM-UWPP-Sparse representation, GMM-MLLR-SVM and SVM-Prosody. The best results were gotten when fusing all the proposed system together. Fusing four different models trained using short term and long term acoustic and prosodic features, the age and gender classification system was investigated by H. Meinedo on four different datasets [2]. The best results were achieved by linear logistic regression classifier. R. Nisimura and A. Lee investigated a speech guidance system which is also capable of discriminating between adults and children [3]. Both acoustic and linguistic features were used as speech feature vectors and instead of using a GMM based method, an SVM based method was developed and it achieved 92.4 % discrimination accuracy. In 2007, H. Kim and his colleagues used an age and gender classification system for a home robot service. They used MFCC as their speech feature and GMM, MLP and ANN as their classifiers [4]. For gender recognition the accuracy was 94.9% using MFFC as feature vector and GMM as a classifier and for age recognition the recognition accuracy was 94.6%. When they used Jitter and Shimmer as feature vector and ANN as the classifier, the gender recognition accuracy dropped a little bit to 81.09% whereas age recognition accuracy increased a little bit to 96.57%. Wei Li and his colleagues worked on a slightly different problem [5]. They proposed

www.manaraa.com

four different models to train GMM for language and gender recognition task. As speech features they used acoustic features of speech. In [6], an age, gender and accent recognition system was proposed by Phuoc Nguyen and his colleagues. Using vector quantization, GMM and SVM as their classifier, all these classifiers were fused together and the system was tested on Australian speech database which has 108 speakers and 200 utterances for each speaker. Their classification accuracies were good ranging from 97.96% to 98.68%. In 2011, Gil Dobry and his colleagues proposed a new speech dimension reduction method which is named WPPCA [7]. They tested this method on two different tasks. The first task was age group classification and the second task was precise age estimation. They used an SVM with RBF kernel. The performance they observed was better with this dimension reduction method and also due to the decreased speech feature vector size, the training of SVM was much faster and it was less prone to over-fitting. M. H. Bahari and his colleague developed an age and gender recognition system based on WSNMF and GRNN [8]. They did the tests on a Dutch speech database and gender recognition rate was 96% and MAE of age estimation was 7.48 years. In 2009, M. H. Sedaaghi used several classifiers for age and gender recognition systems [9]. These classifiers include probabilistic NN, SVM, KNN and GMM. He used two different databases for this task. The first database he used is DES database and the other one is ELSDSR [32] database. SVM and PNN performed best in terms of gender recognition and age recognition, respectively. In 2010, Tobias Bocklet and his colleagues also developed an age and gender recognition system based on multiple systems and their combination [10]. They used spectral features, prosodic features and glottal features in their work. Their best system used GMM-UBM as classification of age and gender. The classification accuracy for this system was 42.4%. Susanne Schotz studied acoustic speech features such as speech rate, sound pressure (SPL), fundamental frequency (F0) and the other ones and ageing of speech production mechanism in age

www.manaraa.com

recognition [11]. In 2009, Maria Wolters and her colleagues proposed and age recognition system which used acoustic features and lexical features as well [12]. Use of GMM/SVM – super vector system (Gaussian Mixture Model combined together with Support Vector Machine) successfully investigated by Michael Feld and his colleagues in automatic speaker age and gender recognition in the car [13]. Florian Metze and his colleagues investigated four different approaches for age and gender recognition for telephone applications and also made a comparison between humans and their system on the same data set [14]. These approaches were a parallel phone recognizer, a dynamic Bayesian network combining several prosodic features, linear prediction analysis approach and lastly GMM based on MFCC for separation of age and gender. They have discovered that the first approach, parallel phone recognizer was as good as humans in terms of recognition. But on short utterances the accuracy went down. Christian Muller and his colleagues successfully investigated and age and gender recognition system especially for elderly users to respond to their special needs [15]. They used ScanSoft and M3I corpus in their work. Jitter / Shimmer and speech rate was selected as speech features. They had chosen four categories for their classification which are elderly male, non-elderly male, elderly female and non-elderly female. They used different classifiers such as ANN, kNN, NB, SVM and two more for the task. In 2012, Myung-Won Lee and his colleague developed and age and gender age group recognition system for human-robot interaction [16]. They used MFCC and LPCC as speech features and SVM and Decision Tree (DT) as classifiers. In [17], a gender based age recognition system was developed by Frank Wittig and Christian Muller which combined several classifiers in a dynamic Bayesian network. Tobias Bocklet and his colleagues successfully investigated an age recognition system for preschool and primary school age children. They trained a GMM super-vector for every child and trained the system with SVM or SV regression [18]. M. H. Bahari and his colleague did another study on

www.manaraa.com

speaker age estimation using HMM and WSNMF [19]. They also used Least Squares Support Vector Repressor (LS-SVR) as a classifier.

There are a number of reasons that show that automatic age and gender recognition is not an easy task. One of the first reasons is that each person's speech characteristic is unique so that makes classification a hard task. Also another challenge is the noise factor. Noise can be anything other than the speaker's voice. These problems are described in more detail below:

- Each speaker of the language is different. The difference comes from the vocal anatomy of speaker. One male and one female's speech characteristics can be very similar in terms of gender and also people from different age groups can have similar speech characteristics in terms of age classification. So in order to get good recognition results, the system must be trained on lots of data in order for the system to be accurate.
- The biggest problem is the noise factor. The noise can interfere with the actual speech and this can lead to wrong classification. Noise can be anything like, crowd of people noise, street noise, suburban train noise, car noise, restaurant noise or similar kind. So in order to have a reliable age and gender recognition system, some pre-processing techniques need to be applied to raw speech data and this noise needs to be eliminated.

In this thesis, the focus is mainly on two speech features: pitch and MFCC. This thesis can be differentiated from other works in the way that SDC was also used as a speech feature. The reason for using SDC as a speech feature was to see its performance on ELSDSR dataset. Also a hybrid approach of pitch and MFCC was investigated. The dataset used in this work was relatively small due to the time and resource limitations. The algorithms used in this work were less complicated compared to the other works in the literature.

www.manaraa.com

2.2 Age and Gender Recognition Fusion Model

Figure 4 shows a flow chart of our proposed fused age and gender recognition system for speech applications. As it can be seen from the figure, our system consists of two main components. In the first component, age and gender recognition is done based on pitch information of speaker. In the second component, age and gender recognition is done using MFCC which performed better than SDC on ELSDSR [32] corpus. Scores from both pitch based model and MFCC based model are fed into a score matrix and the score matrix generates a final recognition score.

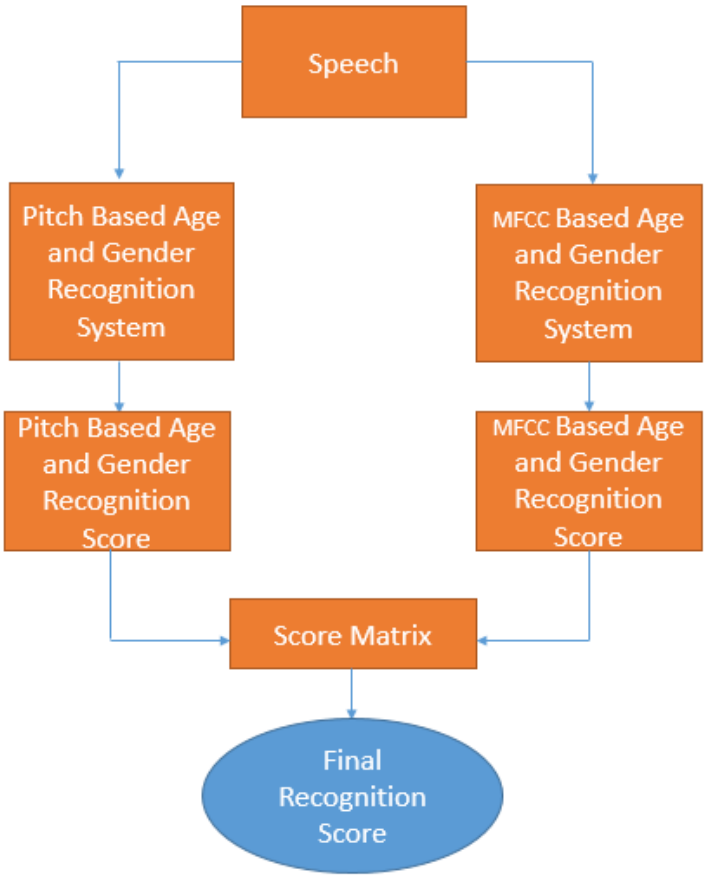

Figure 4. Fused Age and Gender Recognition System Flow Chart

www.manaraa.com

CHAPTER 3

SPEECH BACKGROUND AND PRE-PROCESSING

3.1 Speech Processing

3.1.1 Background

Speech is a natural way of communication between human beings. Speech is produced by the vocal folds vibrations, movement of articulators and also with breathing of air from the lungs. Sounds created using lips, tongue and mouth positions are based on some rules.

Speech signals created by humans are analog in nature. So in order for computers to process speech information, first, the speech must be converted from analog to digital signal. Also, speech signals can be represented in time domain or in frequency domain. While representing speech signals in time domain, on the x-axis there is time and on the y-axis there is amplitude. That does not tell much about the frequency content of the speech. So a better representation would be the frequency domain representation of speech. In this domain, on the x-axis there is frequency and on the y-axis there is magnitude in dB.

Also, speech signals in time domain can be classified into three categories as voiced, unvoiced and silent speech as shown in Figure 5. Voiced sounds are periodic in nature and have higher energy than unvoiced sounds which are aperiodic and noise-like. The silence is when there is no speech and may have energy level related to the background noise.

www.manaraa.com

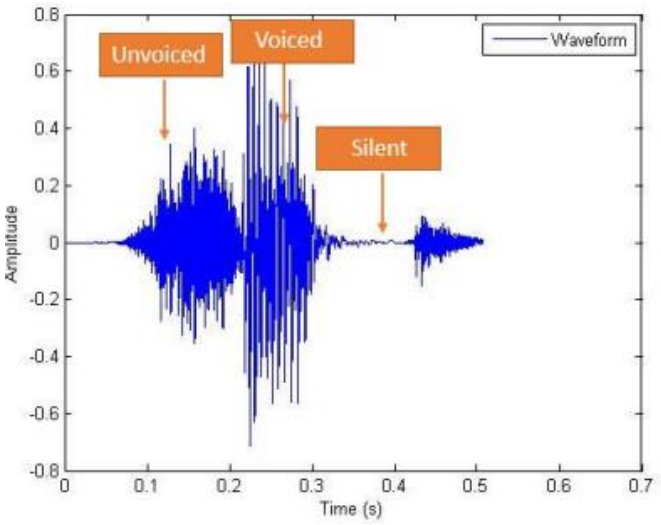

Figure 5. Time Domain Classification of the Speech Utterance 'Sit'

In terms of digital modelling of speech production, there are two factors that come into this equation. The first one is the excitation source which can be periodic or aperiodic. When the vocal fold vibration is modulated by the vocal tract filter it generates a periodic voiced sounds and similarly turbulent air flow from the lungs is modulated by the vocal tract it generates aperiodic noise-like unvoiced sounds. These can described by the discrete-time system (or filter) as shown in Figure 6 where the speech output y(t) is produced by the convolution of the input excitation, x(t) with the vocal tract filter impulse response, h(t) given in equation (1).

$$y(t) = x(t) * h(t) \tag{1}$$

Figure 6. Speech Production Model

www.manaraa.com

To develop an age and gender recognition system for speech applications, one first must look at how speech is formed and the structure of speech. Also the questions about how male and female speech characteristics are different and how the speech characteristics of specific age groups are different will be addressed.

Humans can produce speech starting from 50 Hz and usually this range is between 50 Hz to 3400 Hz. But most of the time, the majority of the energy is between 300 Hz and 3000 Hz. Human ear, on the other hand, can hear frequencies between 20 Hz to 20 kHz. Frequencies above 20 kHz is considered as ultrasonic sounds and human ear is not capable of hearing those frequencies. Also, frequencies below 20 Hz is called infrasonic and humans don't hear it.

In time domain of speech signals, it can be seen that voiced sounds have a repeating periodic pattern. Each of these identifiable patterns is called a cycle. The duration of a cycle is called the pitch period and the fundamental frequency (F_0) or pitch frequency is the inverse of pitch period (T_0). Fundamental frequency shows how high or high low, a person's voice sounds. It is the frequency of his or her vocal cord vibration. Adult males typically have a fundamental frequency between 85 Hz to 155 Hz. Adult females, on the other hand have higher fundamental frequencies. An adult female fundamental frequency is in the range of 165 Hz to 255 Hz. Infants have much higher fundamental frequencies when they speak. It is generally between 250 Hz to 650 Hz. A ten year old boy or girl has a fundamental frequency of 400 Hz. When a person speaks, his/her fundamental frequency changes because of the structure of the language such as intonation and rhythm. So it is not easy to say that there is just one fundamental frequency of a person. However, when the person speaks in a natural voice, it is considered as his/her fundamental frequency.

www.manaraa.com

3.2 Speech Signal Characteristics

When humans speak, the speech signal produced by the vocal tract is an analogue signal. As mentioned earlier, the data on the computers are stored digitally. So when working and processing speech data on computer, first thing, the speech data needs to be converted into digital signal. Also the speech data needs to be sampled at a high rate in order not to lose important information of speech. According to Nyquist–Shannon sampling theorem, the digital sampling frequency must be at least two times bigger than the highest frequency in the analog signal.

3.2.1 Speech Framing and Windowing

Speech signal is a time-varying signal. It is stationary and changes over time. So in order for speech to be processed, it must be divided into non-stationary frames. The general size of speech frames varies between 10ms to 40ms where speech is said to be not changing. Figure 7 shows one frame of the vowel /I/ from the speech utterance 'Sit'.

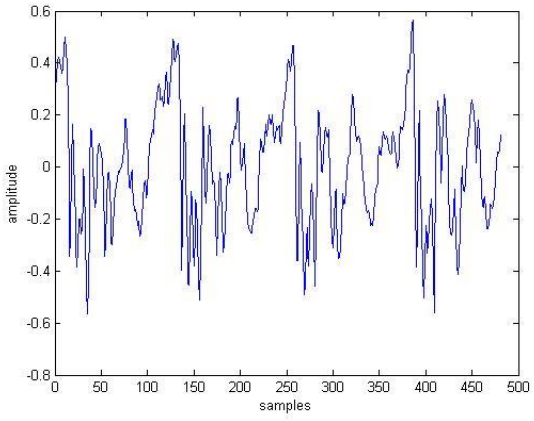

Figure 7. One Frame (30ms) of /I/ Sound from the Speech Utterance 'Sit'

Once the speech signal is cut into frames, the next step is in many cases are windowing. Basically, the speech frame is multiplied by a window function. The most basic windowing function is the rectangular window. When this windowing is applied to the frame, none of the values of the frame changes. That does not make a good job in many cases because there are

www.manaraa.com

discontinuities on the edges of speech signals. It makes it harder to analyze those signals. Rectangular window function can be defined as:

$$\begin{aligned} w[n] &= 1, \quad (0 \le n \le N-1) \\ &= 0, \quad elsewhere \end{aligned} \tag{2}$$

So to reduce this edge effect, instead of using a basic rectangular window, some smoother windows such as Hamming Window are preferred. Getting closer to the edges, the value of the window comes closer to 0, and in the middle the window value is 1. When analyzing long speech signals, the speech windows come one after another and if smoother window functions such as Hamming is used, it is required to overlap the windows to provide signal continuity. Hamming window function can be defined as below:

$$\begin{aligned} w[n] &= 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad (0 \le n \le N-1) \\ &= 0, \quad elsewhere \end{aligned} \tag{3}$$

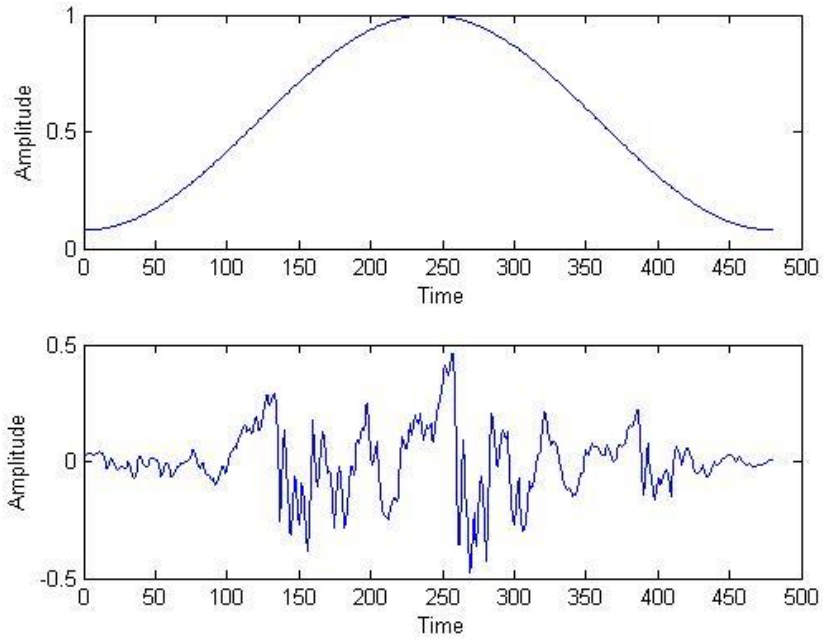

Figure 8. On top: Hamming Window, On Bottom: Hamming Windowed Speech Frame of /I/ from the Speech Utterance 'Sit'

www.manaraa.com

As it can be seen from Figure 8, the middle part of the signal is kept whereas the signals on the edges are tapered. So in order to keep signal continuity, windowed speech frames must be overlapped.

All of the window functions can be accessed using MATLAB's signal processing toolbox.

3.2.2 Fourier Transform of DT Signal and Discrete Fourier Transform (DFT)

All the waveforms no matter what you observe in the universe can be represented as sum of sinusoids of different frequencies. That signal can be a speech signal or electromagnetic signal or the price of your favorite stock versus time or any other kind of signal. Fourier transform is extremely powerful and it converts the signal from time domain into the frequency domain. Any signal can be de-constructed into its sinusoidal components and that helps in understanding the universe and makes life a lot easier for engineers and scientists. Fourier transform is defined as below:

$$X(e^{j\omega}) = \sum_{n=0}^{+\infty} x(n)e^{-j\omega n} \tag{4}$$

For any discrete-time periodic signal x[n] with a period of N, the Discrete Fourier Transform (DFT) can be seen as:

$$X[k] = \sum_{n=0}^{N-1} x[n]e^{-j2\pi(kn)/N} \tag{5}$$

where, k=0, 1, 2, …, N-1.

Figure 9 shows the Discrete Fourier Transformation of the speech utterance 'Sit'.

www.manaraa.com

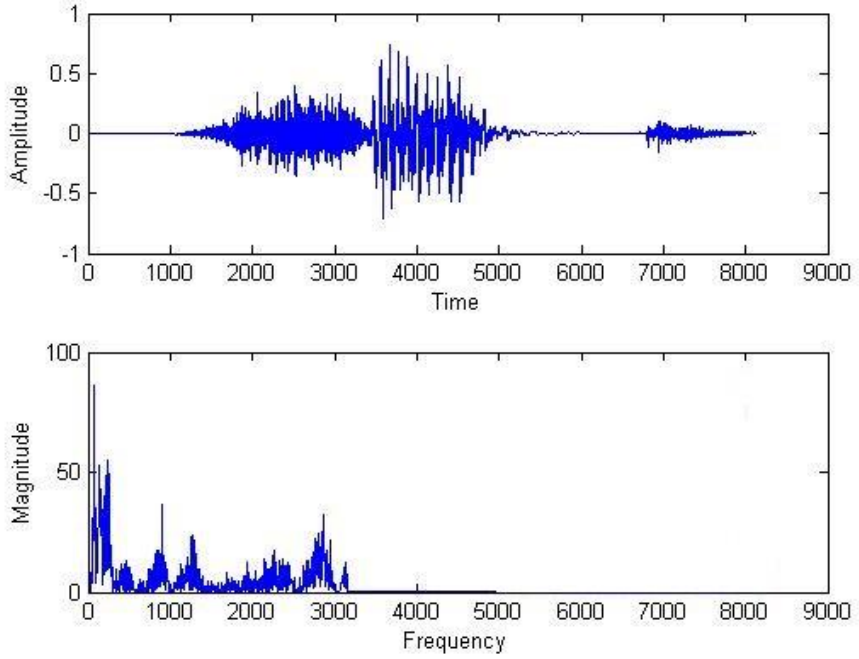

Figure 9. On Top: Time Domain Representation of 'Sit', On Bottom: DFT Applied to the Speech Signal

3.2.3 Discrete Cosine Transform

Discrete Cosine Transform (DCT), is a signal transformation which is similar to Discrete Fourier Transform. It also transforms the signal from time domain into frequency domain in terms of sum of cosine function with different frequencies. But it has some advantages over DFT. It is more efficient then DFT. It can approximate the speech signal with fewer coefficients then DFT. It is used in compression of data for ex: speech data, image data etc. The idea is that, most of the speech frequencies occur in low frequencies so DCT keeps the low frequencies of the speech which make up most of the speech and eliminates high frequencies. There are many different variations of DCT. The most commonly used DCT can be calculated as below:

$$X_k = \sum_{n=0}^{N-1} x_n \cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right], \quad k = 0, \ldots, N-1 \tag{6}$$

where X_k is the kth coefficient of DCT.

www.manaraa.com

Figure 10 represents the Discrete Cosine Transformation of the speech utterance ‘Sit’.

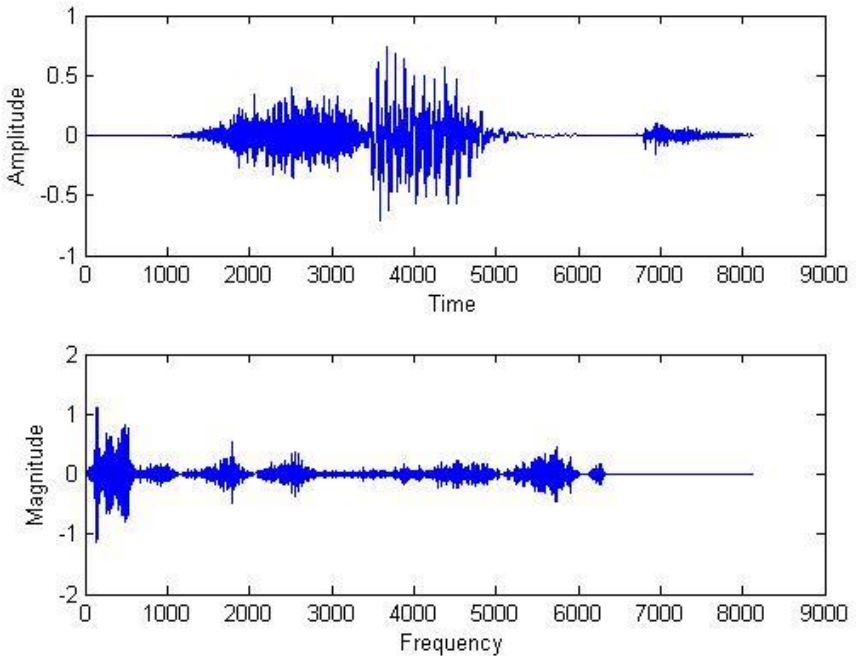

Figure 10. On Top: Time Domain Representation of ‘Sit’, On Bottom: DCT Applied to the Speech Signal

3.2.4 Digital Filters

Digital filters can be used in DSP in two different ways. The first use of digital filters is in terms of separation of signals that have been combined, and the second use of digital filters is in removing some noise and restoration of signals. There are also analog filters but digital filters are much more efficient than analog filters and this is one of the reasons why DSP has become so popular in recent years. In DSP there are three main type of digital filters which are called: high pass filters, low pass filters and band pass filters. A high pass filter filters out the frequencies below the cut-off frequency and passes the frequencies above the cut-off frequency. A low pass filter filters out the frequencies above the cut-off frequency and passes the frequencies below the

www.manaraa.com

cut-off frequency. A band pass filter is a combination of high pass filter and a low pass filter. It only allows a certain range of frequencies to pass through.

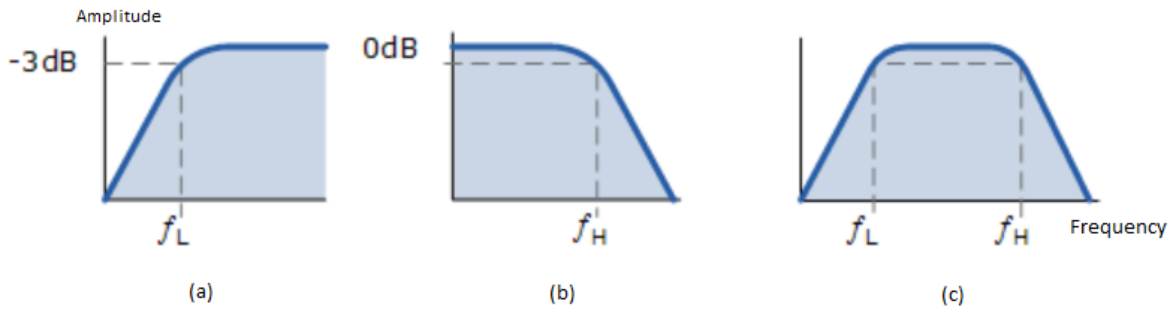

Figure 11. Time Domain Representation of (a): High Pass Filter, (b): Low Pass Filter, (c): Band Pass Filter

3.2.5 Sampling Theorem and Quantization

Aliasing occurs when the discretely sampled signal is not able to reconstruct itself from discrete time to continuous time because it loses some information and cannot capture the changes in the signal. If Nyquist sampling theorem doesn't apply to the continuous signal that will be discretized, aliasing occurs. Nyquist theorem states that in order to reconstruct the signal from its discrete time form, the signal must be sampled at least two times the highest frequency in the analog signal. To state it mathematically:

$$fs \geq 2f_c \tag{7}$$

where *fs* is the sampling frequency and *fc* is the maximum frequency contained in the original signal.

Quantization in signal processing is converting the continuous range of values into the smaller set of discrete values. There happens some errors in this conversion because of the difference between sampled signal and quantized signal and that is called the quantization error. If the quantization is done with low levels, the error gets bigger. For example, CDs are sampled at

www.manaraa.com

44.1 KHz and with a quantization level of 16, which means that the amplitude of the continuous time signal can be quantized into 2^16 different values.

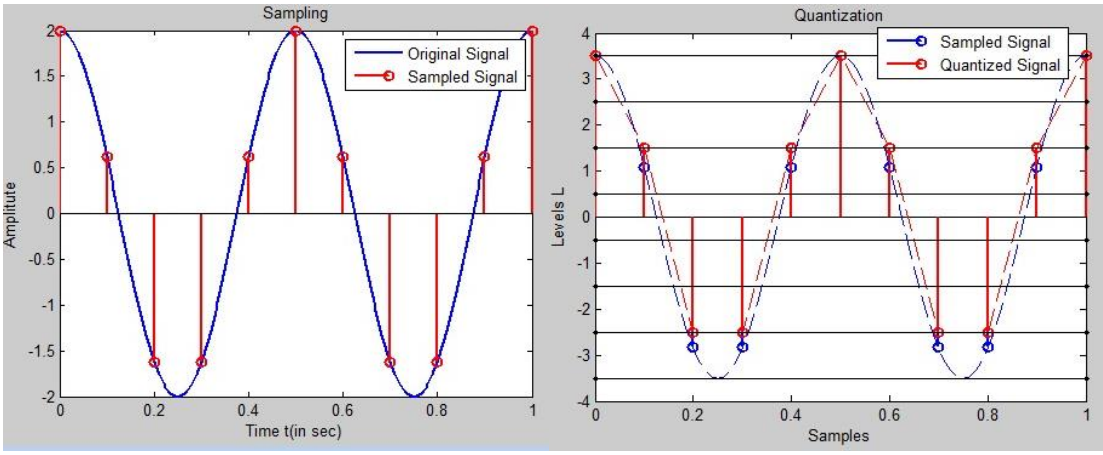

Figure 12. Left: Original Signal and the Sampled Signal, Right: Sampled Signal and Then Quantized Signal

3.3 Pre-Processing

Speech pre-processing is done using some algorithms to improve the quality of the speech. What quality means here is that, it can represent its clarity, intelligibility and pleasantness. Speech pre-processing techniques are so important in speech recognition. Without using some good speech pre-processing techniques, speech recognition cannot be done in the real world. Because in the real world, there are all sorts of noise as described before. There might be street noise, there might be human noise in the background, and there might be some instrument noise. So removing these background noises are so important. Also for many other purposes, speech pre-processing algorithms are being used nowadays. Just to give some examples of its use, it is used for restoring historical recordings, for wiretapping criminal investigations and also to improve the sound quality of hearing devices for older people in Europe. For all of this reasons, there will also be used some speech pre-processing methods in this project as described before.

www.manaraa.com

3.3.1 Signal to Noise Ratio

Signal to noise ratio is proportion of the real signal to the noise produced by the recording device. Each recording device has some kind of noise in it. To measure signal to noise ratio, there are two steps in the process. The first process is, without applying any input signal to the input, the output needs to be measured. In the second part, when a speech signal is applied to the input, the output of the device is recorded. This value is divided by the value in the first part. In many cases, the noise of the device can be eliminated. But if the input sound file is very weak, in this case even small noise from the device can cause distortion. The performance of the signal can be understood by looking at SNR. The larger signal to noise ratio is, the better the quality of the speech is. Also signal to noise ratio is measured in dB which makes it easier to deal with big numbers. Signal to noise ratio (SNR) can be measured using this formula below:

$$SNR = \frac{P_{signal}}{P_{noise}} \tag{8}$$

where, *Psignal* is the average power of the signal and *Pnoise* is the average power or the noise.

This can also be better expressed in logarithmic terms using dB:

$$SNR_{dB} = 10\ log_{10}(\frac{P_{signal}}{P_{noise}}) \tag{9}$$

Figure 13 shows an example of a sine wave with noise added to it where Pn/Ps = 0.1. MATLAB's voicebox toolbox has some useful speech processing functions and using this toolbox, SNR can be found using the function name 'snrseg'.

www.manaraa.com

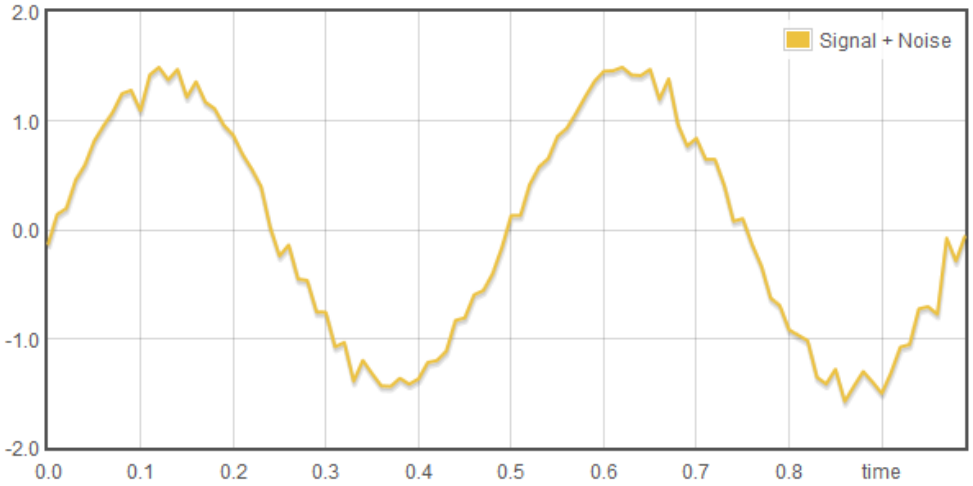

Figure 13. Example of Sinusoidal with Signal to Noise Ratio Pn/Ps = 0.1

3.3.2 Spectral Noise Subtraction

Spectral Noise Subtraction is one of the most widely used speech noise removal methods until now. It assumes that speech contains real speech data and some background noise in it. This technique work in the frequency domain and it assumes that the noisy input speech can be expressed in terms of speech spectrum and the background noise spectrum. Figure 14 shows the diagram of spectral noise subtraction method. The noise spectrum is estimated from the speech samples where only noise is present and then it is subtracted from the original noisy speech spectrum to get the clean speech spectrum.

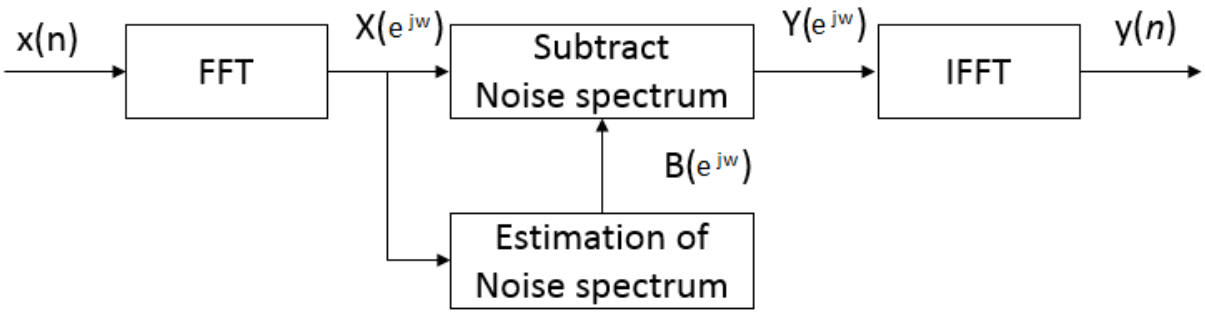

Figure 14. Spectral Subtraction Method

www.manaraa.com

where x(n) is the discrete time noisy speech sequence, which is x(n) = s(n) + b(n) where s(n) is the clean signal and b(n) is the background noise. $X(e^{j\omega})$ is the FFT of x (n). $B(e^{j\omega})$ is the estimation of noise spectrum. So $Y(e^{j\omega}) = X(e^{j\omega}) - B(e^{j\omega})$ and finally y(n) is the discrete time clean speech sequence.

The idea of spectral subtraction is to find the clean speech spectrum $S(n)$ by subtracting the mean of the noise spectrum N[n] from the input spectrum X[n]:

$$|\hat{S}[n]|^b = |X[n]|^b - \overline{|N[n]|^b} \tag{10}$$

where b can take on two different values. If b=1, it is for magnitude spectral subtraction and if b=2 it is for power spectral subtraction. Sometimes, the magnitude of clean speech spectrum might be negative. In this case, the result is zero. So the equation changes to:

$$|S[n]|^b = \begin{cases} |X[n]|^b - \overline{|N[n]|^b} & if\ |X[n]|^b - \overline{|N[n]|^b} > 0 \\ 0 & otherwise \end{cases} \tag{11}$$

Finally estimation of clean speech signal can be found by applying Inverse Fourier Transform.

MATLAB's voicebox toolbox has a function called 'specsub' which can perform spectral subtraction.

3.3.3 Cepstral Mean Normalization

Speech recorded with different microphones in the same environment might have different characteristics. That is normal because each different type of microphone has different type of characteristics and transfer function. Even if the same type of microphones are used in a recording environment, still speech recorded from different microphones can have different sound characteristics. For example, if the microphones are at different distances from each other or sometimes from the acoustics of the room, their characteristics will be different. Cepstral Mean

www.manaraa.com

Normalization (CMN) is a pre-processing method that tries to solve this problem. It finds the cepstral mean vectors of the speech across the utterance from each frame. In the end, it subtracts this from the cepstral coefficients of that utterance. The formulation of CMN can be seen as below:

$$y[n] = x[n] - \frac{1}{N}\sum_{n=1}^{N} x[n] \tag{12}$$

where, x[n] and y[n] are the time-varying cepstral vectors before and after the filtering, and N is the total number of frames.

3.3.4 RASTA Filtering

Relative Spectra (RASTA) filtering is another pre-processing method that tries to remove the channel noise in the speech. As can be seen in spectrogram, the actual speech is much more fluctuated then the noise in the channel. So in order to eliminate the speech signal from this channel noise, a band filter can be used in the frequency domain. Using this band pass filter slow variations in the spectrogram or constant magnitudes in the speech file can be removed. The transfer function of RASTA filter can be written as below [33]:

$$H(z) = 0.1z^4.\frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}} \tag{13}$$

3.3.5 Voice Activity Detector

Voice Activity Detector (VAD) is one of the most widely used pre-processing methods in speech applications [21]. It basically tries to determine which speech frames contains speech and which speech frames doesn't. When a person speaks, he/she cannot speak without stopping so there will be some pauses and maybe some hesitations etc. in his/her speech. So in order to extract some useful speech features from the speech, those non-speech sounds or silences should be removed from the speech samples. Some other applications of VAD can be seen in audio

www.manaraa.com

conferencing, echo cancellation, speech recognition, speech encoding and hands-free telephony [20]. Generally right after speech is digitized, VAD is applied to the speech signal.

Assuming that speech is corrupted by additive noise, there can be written two hypotheses for each frame of the speech:

H0: where speech is absent in the frame X = N

H1: where speech is present in the frame X = N + S

where S, N, X are the L dimensional DFT coefficient vectors of speech, noise and noisy speech and their *k*th element is S_k, N_k and X_k.

DFT coefficients of the each process is independent Gaussian random variables so a Gaussian statistical model can be used. So the probability density functions for H0 and H1 can be found by:

$$p(X|H_0) = \prod_{k=0}^{L-1} \frac{1}{\pi \lambda_N(k)} \exp\{-\frac{|X_k|^2}{\lambda_N(k)}\} \tag{14}$$

$$p(X|H_1) = \prod_{k=0}^{L-1} \frac{1}{\pi[\lambda_N(k) + \lambda_S(k)]} \exp\{-\frac{|X_k|^2}{\lambda_N(k) + \lambda_S(k)}\} \tag{15}$$

where, $\lambda_N(k)$ and $\lambda_S(k)$ are the variances of N_k and S_k. kth frequency band likelihood ratio is:

$$\Lambda_k \triangleq \frac{p(X_k|H_1)}{p(X_k|H_0)} = \frac{1}{1 + \xi_k} \exp\{\frac{\gamma_k \xi_k}{1 + \xi_k}\} \tag{16}$$

where $\xi_k \triangleq \frac{\lambda_S(k)}{\lambda_N(k)}$ and $\gamma_k \triangleq \frac{|X_k|^2}{\lambda_N(k)}$,

The decision rule for individual frequency bands can be seen as:

$$\log \Lambda = \frac{1}{L} \sum_{k=0}^{L-1} \log \Lambda_k \gtrless_{H_0}^{H_1} \eta \tag{17}$$

www.manaraa.com

λ_N can be estimated using noise statistic estimation procedure so only ξ_k needs to be estimated. ML based decision rule can be estimated for ξ_k:

$$\xi_k{}^{(ML)} = \gamma_k - 1 \tag{18}$$

Finally the decision rule can be found as:

$$\log \Lambda^{(ML)} = \frac{1}{L} \sum_{k=0}^{L-1} \{ \gamma_k - \log \gamma_k - 1 \} \gtrless_{H0}^{H1} \eta \tag{19}$$

The left hand of the equation cannot be less than zero so this likelihood ratio is biased to H1. In the paper [21], Sohn proposes another method to reduce the bias which is better in terms of SNR estimation and reduces the fluctuation of estimated likelihood in noised frames.

www.manaraa.com

CHAPTER 4

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.

4.1 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 15 shows a model for this age and gender recognition system:

Figure 15. Model of Age and Gender Recognition System

www.manaraa.com

4.1.1 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 16 represents a diagram for MFCC computation.

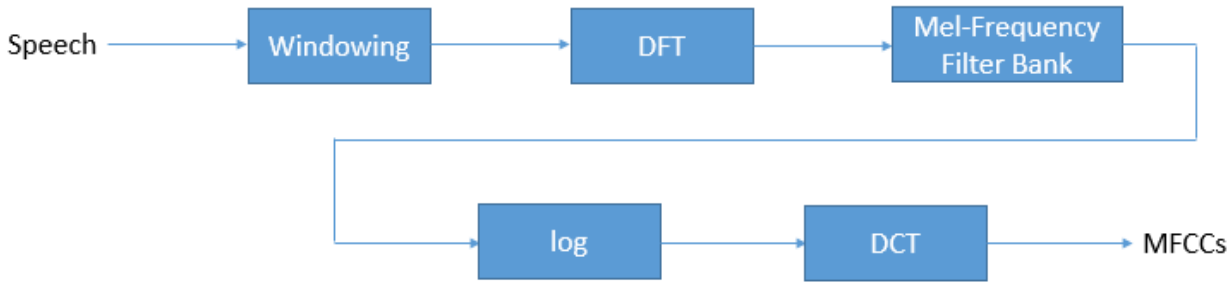

Figure 16. Steps for MFCC Computation

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.

www.manaraa.com

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. The reason for doing that is to decorrelate the overlapped filterbank energies for better classification.

In the second step while converting from frequency to Mel Scale this formula is used:

$$M(f) = 1125\, In\, (1 + \frac{f}{700}) \tag{20}$$

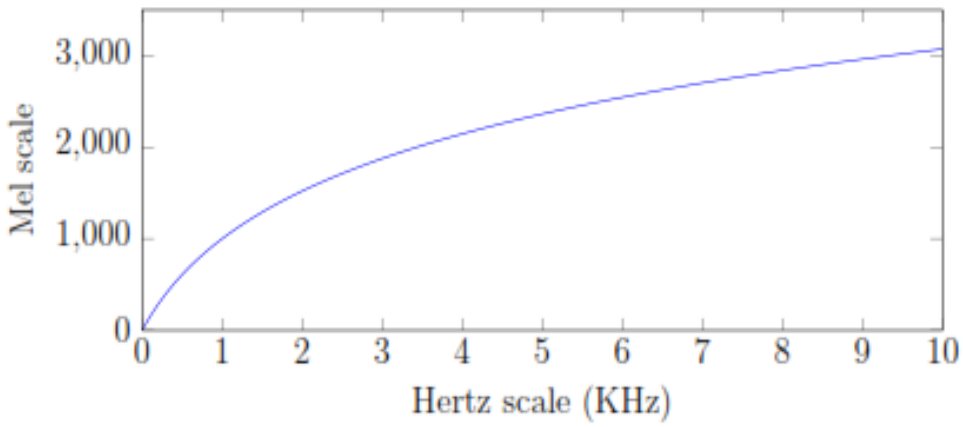

Figure 17. Plot of Mel Frequency Scale

www.manaraa.com

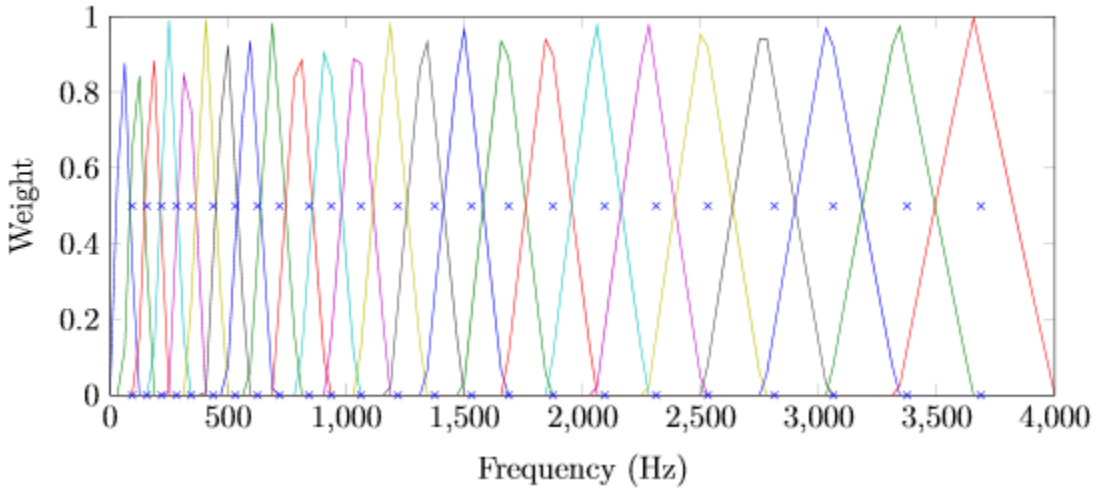

Figure 18. Mel Filter Bank Using 24 Filters

4.1.2 Shifted Delta Cepstral (SDC)

Despite MFCC's success in speech applications, a better set of features derived from the MFCC's have been discovered. It is called Shifted Delta Cepstral. Some of the uses of SDC can be seen in the literature from those papers [22], [23] and [24] where SDC were used in language identification and language recognition tasks. Its advantage over phone-based approaches in this problem is that it is fast, so costs less computationally and performs similar to phone-based approaches which are computationally heavy.

As mentioned before SDC is derived from MFCC. There are four parameters in SDC and these are N, d, P, k, where: N is the number of cepstral coefficients computed at each frame, d is the time advance and delay for the delta computation, k is the number of blocks whose delta coefficients are concatenated to form the final feature vector, and P is the time shift between consecutive blocks. SDC feature vector can be calculated for each frame t as:

www.manaraa.com

$$SDC_{(t)} = \begin{cases} \Delta_c(t,0) \\ \Delta_c(t,1) \\ \dots\dots \\ \Delta_c(t,k-1) \end{cases} \quad (21)$$

where, $\Delta_c(t) = c(t + iP + d) - c(t + iP - d)$

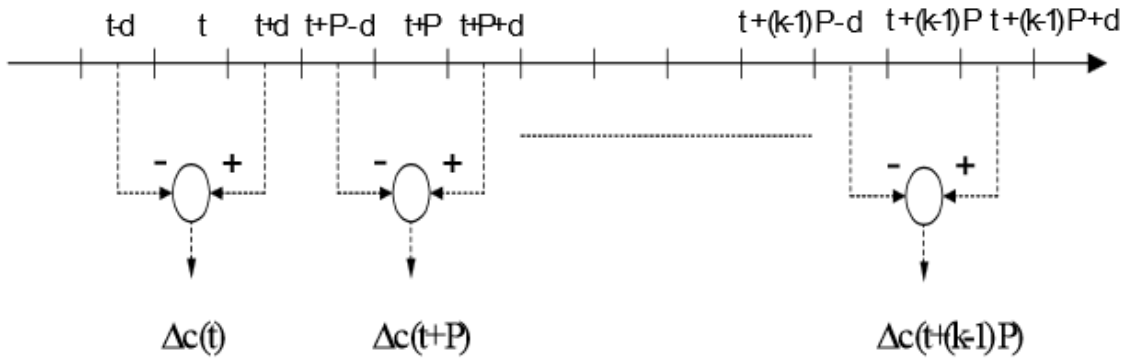

Figure 19. SDC Feature Vector Computation at Frame t with Parameters N,d,P,k [22]

For each SDC feature vector, SDC has kN parameters compared to the conventional cepstral features and this makes this method competitive with the phone-based approaches in age and gender recognition.

4.1.3 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. They recorded 28 young women voices in 1945 and the average fundamental frequency was 229 Hz. After such a long time, they did these recordings again in 1993 and this time the mean of the fundamental frequency measures was dropped down to 181.2 Hz [25]. There was also a similar study done by Hollien and Shipp [26]. They recorded the sounds of 25 men from

www.manaraa.com

each decade and the conclusion of the study is that, fundamental frequency dropped through early and middle adulthood and it rose again later on. Fundamental frequency extraction is still not an easy task just because of the changes of the vocal tract filter. So it is in fact not a constant value.

Sun proposed a pitch determination algorithm in [27], based on subharmonic to harmonic ratio. He did some tests on CSTR's database and his algorithm seems to be superior to any previous algorithms. In this work, his algorithm was used in order to extract pitch information of the speakers in training set and test set.

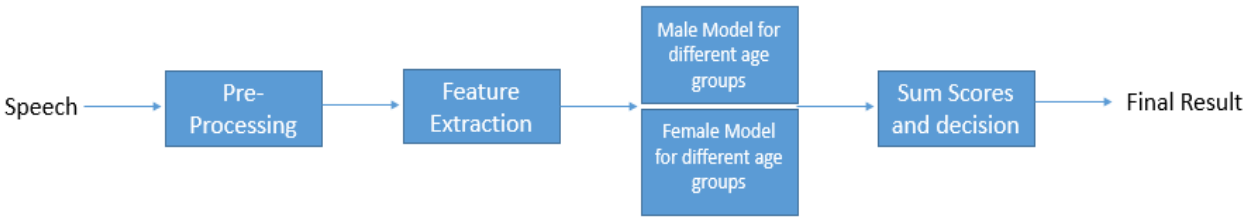

Figure 20. Age and Gender Recognition Model Trained Using MFCC as Features

4.2 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.

4.3 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

www.manaraa.com

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].

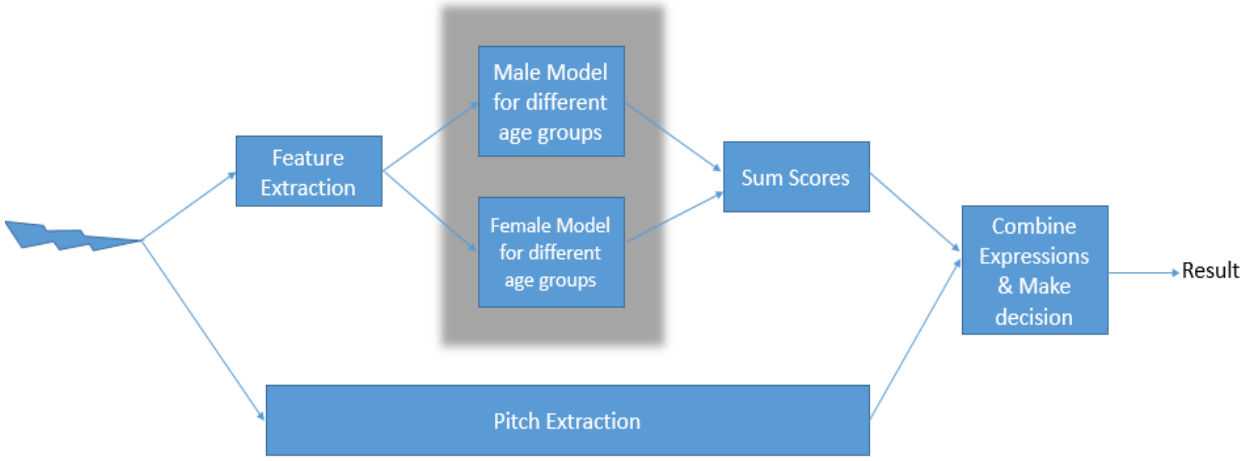

Figure 21. Fused Age and Gender Recognition Model

4.4 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].

www.manaraa.com

CHAPTER 5

CLASSIFICATION FOR AGE AND GENDER RECOGNITION

As mentioned earlier, in order to create an age and gender recognition system, the last step is to use a classifier and according to the result of the classifier make a decision. There are many classifiers that can be used to do the job but here our focus will be particularly on SVM because of its high performance.

5.1 Overview

With the advances of the technology, and the rapid growth of the internet, age and gender recognition became more and more important. There are many different classifiers up to this date but SVM is definitely one of the most used ones in speech applications. It is a supervised machine learning algorithm and can be used for classification as well as regression. It uses some kernel tricks to transform the data and then it tries to find decision boundaries in the data. Depending on which side of the boundary the test sample lays, the test sample is classified. This is true for binary classifications and in this case the data is linearly separable. However in many cases, the data is not linearly separable and so the SVM maps the data into multi-dimensional spaces and tries to find an optimal separating plane. Obviously, building a speech application by hand is time-consuming and can also lead to wrong results. But learning from previous examples as in SVM, is a great way to implement an age and gender recognition system.

There are many benefits to using SVMs. SVM classifier was introduced by V. Vapnik in the 90s [18]. It has a well-founded computational learning theory and the theory behind the classifier is elegant and easy to understand.

www.manaraa.com

Compared to the other text classifiers, SVMs show a substantial increase in terms of performance and reliability and they are also robust.

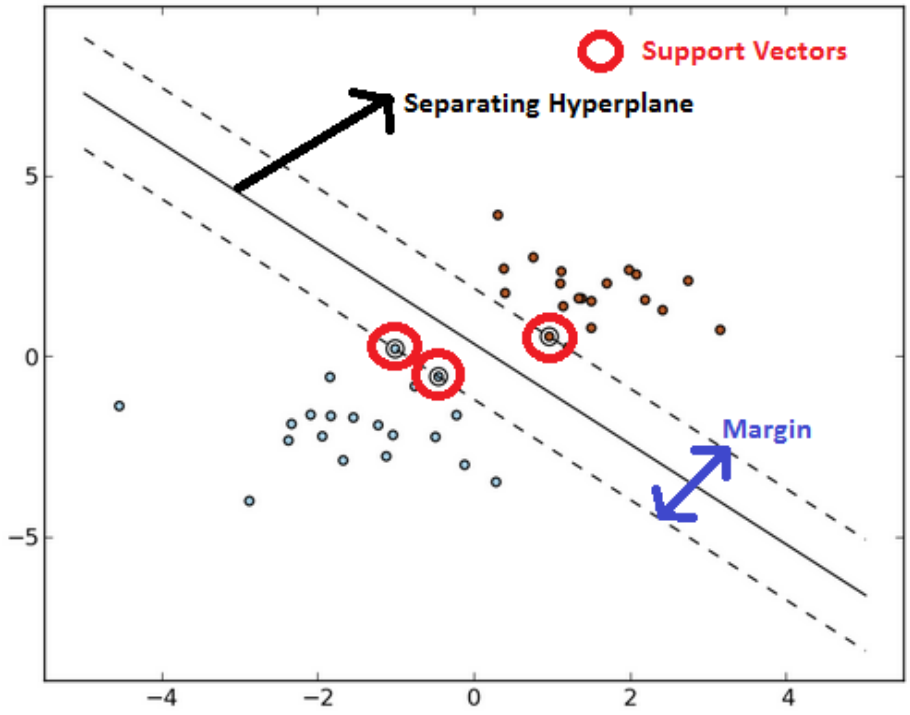

Figure 22. SVM Decision Boundary and Margins and Support Vectors

5.2 Support Vector Machine (SVM)

The optimization objective function for Support Vector Machines can be described as below:

$$min_\theta C \sum_{i=1}^{m}\left[y^{(i)}\, cost_1\left(\theta^T x^{(i)}\right) + \left(1 - y^{(i)}\right) cost_0\left(\theta^T x^{(i)}\right)\right] + \frac{1}{2}\sum_{i=1}^{n}\theta_j^2 \qquad (22)$$

where the idea is to try to minimize this cost function.

And also the hypothesis of SVM is:

$$h_\theta(x) = \begin{cases} 1, & if\ \theta^T x \geq 0 \\ 0, & otherwise \end{cases} \qquad (23)$$

www.manaraa.com

If the parameter vector θ transpose times x is greater or equal than zero, it will be classified as 1 and otherwise it will be classified as zero.

SVM really does a good job in defining the decision boundaries of the training data. Even if there are some outliers in the training set, meaning that there are some positive examples near the negative examples and vice versa, not chosen big regularization parameter C will keep the original decision boundary without the outliers. An illustration of this can be seen in Figure 23.

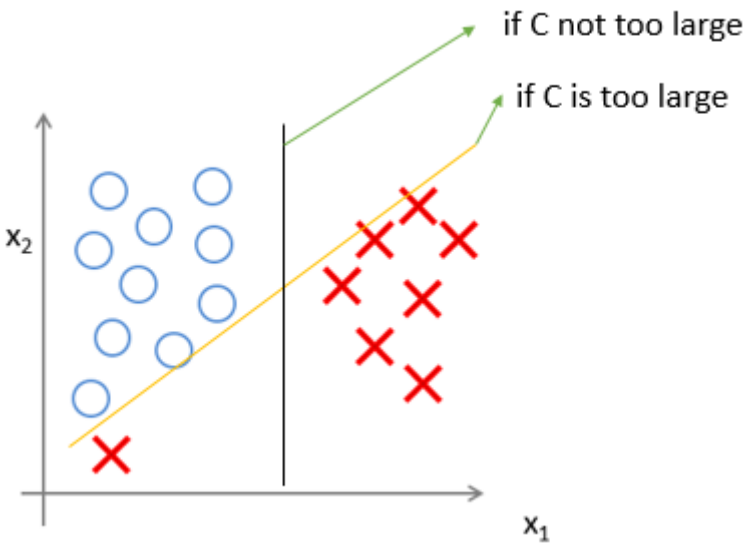

Figure 23. Regularization Parameter (C) in SVM

Also the decision boundary for SVM can be seen as below:

$$min_\theta \frac{1}{2} \sum_{j=1}^{n} \theta_j^2 \tag{24}$$

$$\begin{aligned} \theta^T x^{(i)} &\geq 1 \qquad if\ y^{(i)} = 1 \\ \theta^T x^{(i)} &\leq -1 \qquad if\ y^{(i)} = 0 \end{aligned} \tag{25}$$

SVM has different types of kernel functions and each of these kernel functions give different classification results. For SVM classifier, actually it is a good idea to try different kernels on the training and test data set and see the performance of each other. However the two most popular and most widely used kernels are linear kernel and RBF kernel.

www.manaraa.com

There are kernels for SVM such as Radial Basis Function (RBF) kernel, or linear kernel, Gaussian kernel or polynomial kernel. There is no general rule for selecting a kernel when using SVM for classification problems. Sometimes, RBF might give good results, sometimes a different kernel might give better results. But if number of features n is large, and number of training examples m is small, using a linear kernel might be a good option. If n is small and m is large, using a RBF kernel can be better.

Given some similarity function between the actual training example and a landmark, the training algorithms of a SVM can be written as below:

$$f_i = similarity\ (x, l^{(i)}) \tag{26}$$

$$= \exp(-\frac{||x - \ l^{(i)}||^2}{2\ \sigma^2} \tag{27}$$

$$min_\theta C \sum_{i=1}^{m} [y^{(i)}\ cost_1\ (\theta^T\ f^{(i)}) + (1 - \ y^{(i)})\ cost_0\ (\theta^T\ f^{(i)})] + \frac{1}{2} \sum_{j=1}^{n} \theta_j^2 \tag{28}$$

There are actually two parameters that should be considered when training a SVM. These parameters are C (regularization parameter) and gamma (parameter of RBF). Choosing a large C value leads to lower bias and high variance. Choosing a small C value leads to higher bias and lower variance. In terms of selecting the gamma value, a large gamma value leads to features changing more smoothly and so higher bias and lower variance and a small gamma is the opposite.

5.3 Decision Making

Once the probability matrices are calculated from different classifiers, the system takes all the probabilities as an input and generates an output probability based on the maximum probability. Each classifier is weighted a different coefficient and the fused model works better than the single classifiers.

www.manaraa.com

CHAPTER 6

SYSTEM DESIGN AND IMPLEMENTATION

In the previous chapters, all the elements of an age and gender recognition system have been described. These include pre-signal processing, feature extraction and finally classification. In this chapter, all these methods will be put into practice and a real age and gender recognition system will be implemented.

6.1 Toolboxes

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.

6.1.1 Signal Processing (Voicebox) 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 Voicebox toolbox [28] 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 from [29] was used. And also extracting pitch, the library provided by [30] was used.

6.1.2 Machine Learning (LIBSVM) Toolbox

To do the classification part after feature extraction step, one very popular toolbox called 'LIBSVM' [31] was used. LIBSVM is a library for Support Vector Machines written by Chih-Chung and Chih-Jen Lin.

www.manaraa.com

6.2 System Design

In this section, the system design of an age and gender recognition system will be explained with all its components. Some software design rules were skipped while developing the system due to the fact that this is a research project.

6.2.1 Requirement

The requirement of this project is to build a robust age and gender recognition system which will also give good recognition results under some kind of background noise or silence. So for real world applications and for a variety of speakers, the system will be able to recognize the age and the gender of the speaker. As an initial approach, a gender recognition system was implemented with the whole training and testing data.

6.2.2 First Approach

As a first approach of recognizing the gender of the speaker, the pitch information is used. Pitch is a fundamental difference between males and females. The library provided by [30] was used to extract the pitch information of the training examples. Each training example was windowed at 25 ms. After extracting all the pitch information of all the frames, the mean value of all this windows was taken and it was considered as the pitch of the training example. Human speech normally varies between 100 Hz to 300 Hz so all the frequencies below 100 Hz and above 300 Hz were neglected. Following the pitch information of all the training examples, these features were fed into an SVM for classification.

Here is how this approach was implemented. First, all the training speech examples were loaded into Matlab using the command wavread. Then, using some of the noise reduction techniques like spectral subtraction and voice activity detection was applied to all the input speech

www.manaraa.com

samples to remove background noise and silences. To do these operations Matlab's voicebox toolbox was used and the commands specsub and vadsohn were executed. After these operations, speech pitch extraction algorithm was executed. As mentioned earlier, the frame size of each window was 25ms and the mean pitch value was calculated taking the average of all the windows in one training example. And that mean pitch information was taken as one feature vector in the training set.

The speech dataset used in this work was ELSDSR [32] database. The database is in English language. There are 22 speakers in the dataset. 12 of the speakers are male and 10 of the speakers are female. The age range of the speakers is between 24 and 63. Each speaker has 7 different utterances in the training set. And in the test set, each speaker has 2 utterances. The average speaking time for a speaker in the training set is 83s whereas the average speaking time for a speaker in the test set is nearly 18 seconds.

After extracting pitch information from the speech, these feature vectors were fed into an SVM. When the data was plotted on a 1D scale it was seen that the data was not linearly separable. So in this case, a support vector machine could not be used with a linear kernel. Instead an SVM with RBF kernel was used. The gender recognition result was 97.72%. The gender recognition rate is pretty satisfactory in this test. One reason of this success is that, some pre-processing techniques were applied to speech before feature extraction. And also pitch extraction algorithm used in this work was one of the best algorithms so far which used harmonic-to sub harmonic ratio. This first initial approach did work out well. In the next section, a robust gender plus age recognition system will be described which is more closer to real world applications.

In Figures 24 and 25, the whole data set and classification results can be seen. In the figures, red dots (dark) represent male examples whereas orange dots (light) represent female examples.

www.manaraa.com

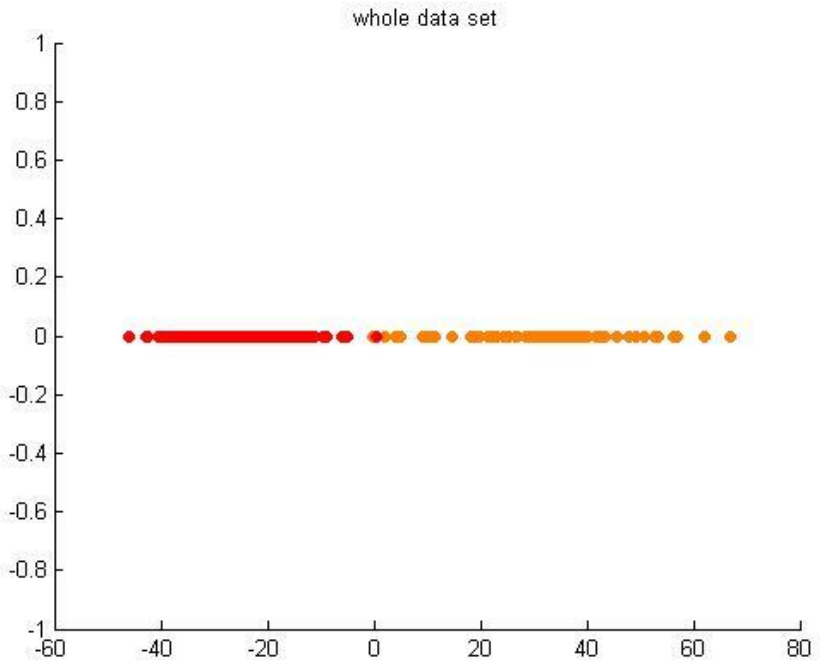

Figure 24. Gender Recognition Whole Data Set (Training + Testing)

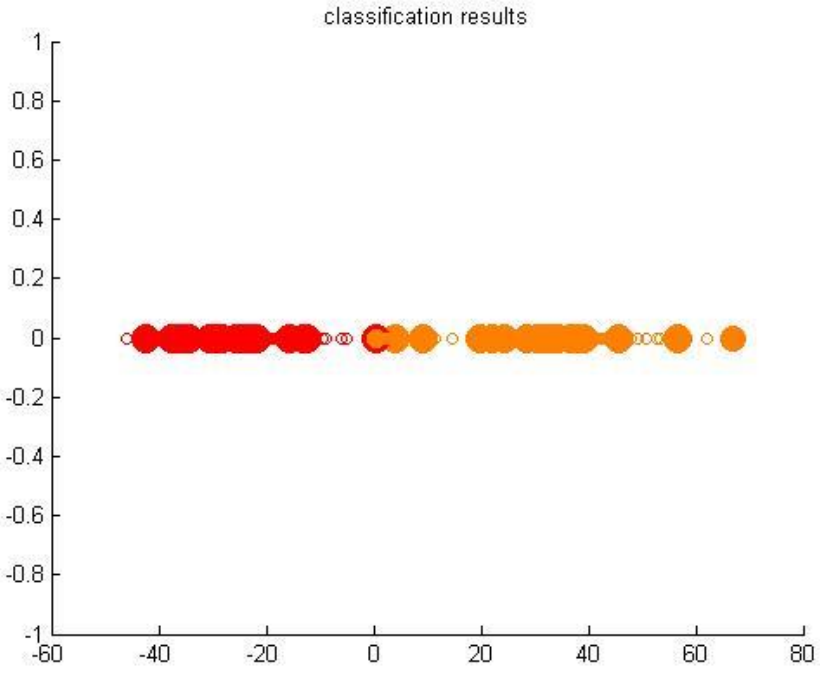

Figure 25. Gender Recognition Classification Results

www.manaraa.com

6.2.3 Algorithm

This approach is very different from the first approach. In this approach, a robust age and gender recognition system will be created. When finding out age and gender of the speaker, pitch information might be useful. But it is a short term feature. So it is not very robust under noisy background sound. So instead of using single feature for this project, a more reliable speech feature will be used. That speech feature is MFCC. MFCC is proven to be the most widely used speech feature in the industry and academia. However, there is a better feature which is derived from MFCC and it is SDC. So and SDC is considered to be a long term speech feature which is better and more robust under noisy conditions. Some of the uses of SDC can be seen in the literature from papers [22], [23] and [24] where SDC was used in language identification and language recognition tasks. So fusing some speech features together is a good idea. And in this project pitch and MFCC features were fused together to improve the accuracy of the age and gender recognition system.

To get better results, some pre-processing techniques were applied to the training data. As the first pre-processing method, voice activity detection was applied to all the training data to remove the silence in them. After voice activity detection, a spectral subtraction method was applied on it to remove all the background noise such as musical noise or white noise.

6.3 SDC Extraction

So two features were 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 pre-processing 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

www.manaraa.com

set to 7-1-3-7. Following this step, the mean value of SDC was normalized to zero and the standard deviation was normalized to one. The block diagram of getting SDC features can be seen in Figure 26.

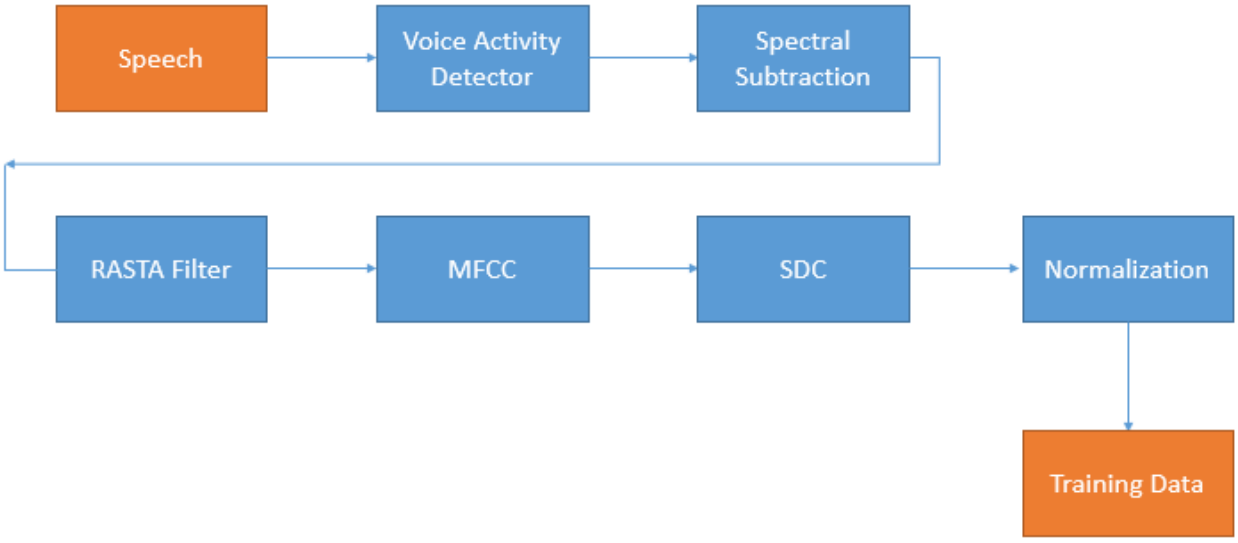

Figure 26. Block Diagram for SDC Feature Extraction Algorithm for the Age and Gender Recognition System

6.4 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 example.

6.5 Model Training

Age and gender recognition is a multi-class classification task. 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

www.manaraa.com

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 parameters 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 steps, pitch extraction algorithm which uses harmonic to sub harmonic ratio was executed on the each training sample. The frame window was set to 25 ms. 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 values below 200 Hz was considered as male speech and the frequencies 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 in Figure 27.

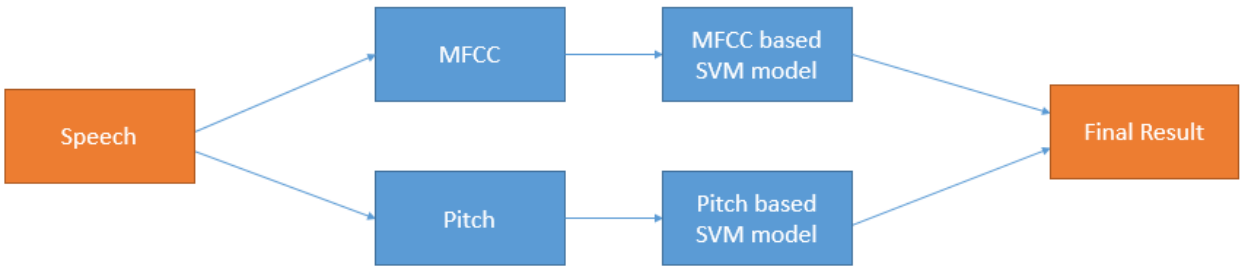

Figure 27. Block Diagram for the Final Fused Age and Gender Recognition System

6.6 Training Dataset

The speech database used in this work is ELSDSR. It was developed by the department of Informatics and Mathematical Modeling (IMM) at Technical University of Denmark [32]. The speakers were from the faculty, Ph.D. students and also master's students. There are 22 speakers

www.manaraa.com

in the database. 12 of the speakers are male and 10 of the speakers are female. The nationalities of the speakers are 20 speakers from Denmark, one speaker from Iceland and one speaker from Canada. There was no rehearsal when creating the database.

The recordings were taken in a chamber in building 321 which is actually a computer lab. They used MARANTZ PMD670 portable solid state recorder for recordings. All the recordings were made in .wav format. The sampling frequency was chosen to be 16 kHz and also bit rate was chosen to be 16 bits.

As mentioned earlier, there are 22 speakers, 12 male and 10 female. Most of these are people are faculty members and Ph. D students. And also there is one international student. The speaker age range is between 24 and 63. It can also be seen from table A.3 in the appendix that female age distribution covers a broader scale whereas male age distribution does not have a large scale. Average female age in the set is 40.6 whereas average male age in the set is 31.3. Speakers in this dataset are not native English speakers and also their dialect is also not the same for everybody. But for our age and gender purposes, this is not a major problem. The average speaking time in the training set is about 83 seconds and the average time in the test set is about 17.6 seconds for all the speakers.

6.7 Feature Selection

Two different experiments were done for getting the first feature vector. MFCC and SDC were compared to see if SDC performed better than MFCC or not. MFCC was applied with 12 coefficients. And SDC was applied with the parameters having 7-1-3-7 values. After these comparison tests, it was seen that MFCC in fact performed better than SDC so instead of using SDC features, MFCC features were picked as the first feature to the system.

www.manaraa.com

6.8 Graphical User Interface (GUI)

A Graphical User Interface was developed for the real time tests. Anyone can record his/her voice by speaking into the microphone and using Matlab's internal recording commands the voice is recorded and then it is plotted in the time domain and in the frequency domain as a spectrogram. As soon as the speech is plotted, pre-signal processing techniques and feature extraction is applied to the speech and then the classification algorithms are run. Depending on the final score, a decision is made and a pop up window is opened with the age and gender information of the speaker.

6.9 Experiments and Results

In this section, all the results and the performances of the tests will be discussed. As mentioned earlier, an age and gender recognition system can be developed in some different ways. Using different speech features and also fusing some speech features, this system can be developed. And all the tests that are done here used a closed-set which means that the training and testing data was from the same source.

The aim of doing different experiments is to compare different speech features in analyzing and age and gender recognition system. As mentioned earlier, different speech features were used to test the system. These features are the pitch frequency, MFCC and finally SDC. And also a better way of fusing pitch and MFCC was also tested. For pitch information, SVM classifier with nonlinear RBF kernel was trained. The classifiers selects a nonlinear threshold to separate between each labels. For the other features MFCC and SDC, again a nonlinear SVM with RBF kernel was trained. SVM was selected as a classifier because of its power and ease of usage. Also for all the tests, the same training test was used so a healthy comparison on the performance could be made.

www.manaraa.com

6.9.1 Pitch Based Models

Pitch based model used only pitch information of the speaker in order to recognize his/her age and gender. First, training data was loaded into Matlab. Then pre-processing algorithms such as VAD and spectral subtraction were used to remove the silence and also background noise from the training. Following that, pitch extraction algorithm which uses harmonic to sub harmonic ratio was executed. The fundamental frequency values vector contained the calculated fundamental frequencies for all the frames. To get the actual pitch of the training example, the mean value was calculated in the end. Then for the classification purpose, a non-linear SVM was trained. In the testing part, the same pre-processing, noise enhancement and extraction techniques was applied on the test data and this pitch value was passed into SVM for the classification.

Various tests were performed in order to better understand the pitch and age and gender recognition relation in humans.

6.9.1.1 One Male and Female Speaker for Each Age Group

For this test, a non-linear SVM with RBF kernel was trained for classification of age and gender. The training set consisted of utterances from just one speaker of each age and gender groups. The total training time was about 6 minutes in total. The model was tested for the test set. The test results can be seen in Table 1.

Table 1: Pitch Based Model Results for One Speaker for each Gender and Age Group

Class 1	40.00%
Class 2	100.00%
Class 3	50.00%
Class 4	10.00%
Overall Class Accuracy	50.00%

www.manaraa.com

The best parameters found using cross correlation were C=4 and gamma= 0.0625. From the table it can be seem that classification percent for class 4 is lower than the other classes. Also overall class accuracy is 50.00% percent which is low. One reason of this is because, the training set was relatively small. Also it is possible that pitch information was overlapped with the other classes. So using only one speaker from each label to recognize age and gender was not a very good idea.

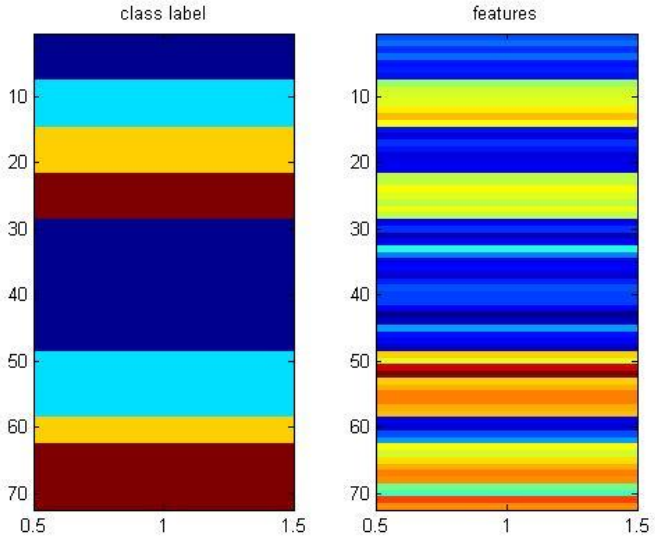

Figure 28. Class Labels and Features for Pitch Based Model for One Speaker for Each Gender and Age Group

In Figure 28, the class labels and also the features can be seen. As mentioned above, the training set consisted of only one speaker for each age category. So in the training set, it can be seen that there are four categories and they are balanced. But when it comes to the test set, the first class has more examples then the other classes as can be seen from the figure. Also from the figure, the features can be seen and in this case it is only pitch information.

www.manaraa.com

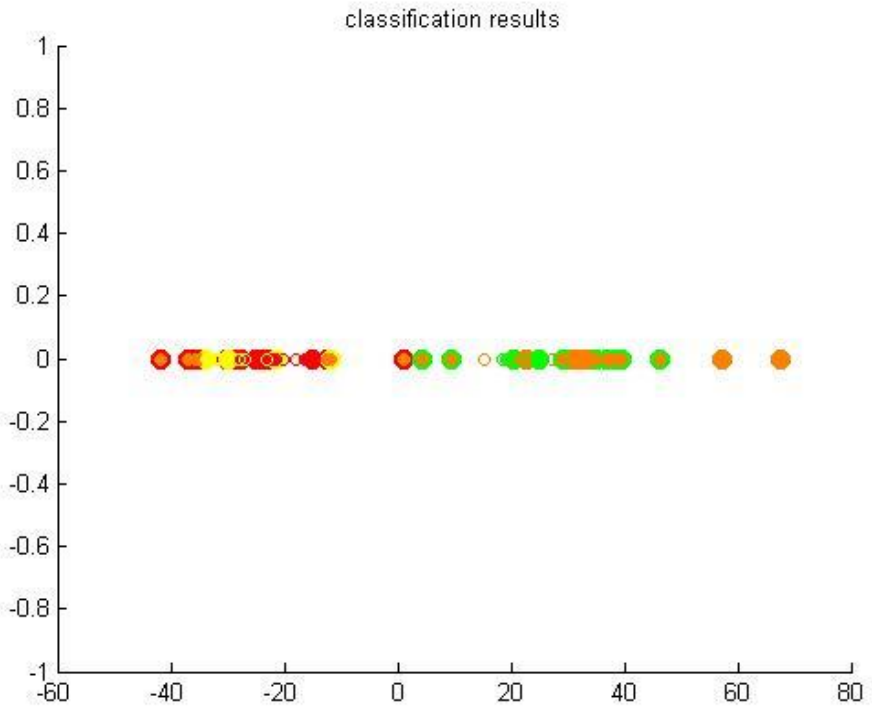

Figure 29. Classification Results for Pitch Based Age and Gender Recognition

Also from Figure 29, the classification results are shown. Unfilled dots represent the samples from the training set. Filled dots represent the data from the test set. And also the different colors represent different classes assigned by SVM classifier. Also edge color represent the true label of the example. The size of the dots represent the confidence level of that example, if it is bigger, the confidence level is higher.

6.9.1.2 Multiple Male Speakers for Each Age Group vs. Multiple Female Speakers for Each Age Group

For this test, a non-linear SVM with RBF kernel was trained for classification of age and gender. The training set consisted of utterances from multiple male speakers for each age group and multiple female speakers for each age group. The total training time was about 30 minutes. And the model was tested on the test set. The test results can be seen in the Table 2.

www.manaraa.com

Table 2: Pitch Based Model Results for Multiple Male and Female Speakers for each Age Group

Class 1	95.00%
Class 2	80.00%
Class 3	25.00%
Class 4	60.00%
Overall Class Accuracy	65.00%

The best parameters found using cross correlation were C=8 and gamma= 0.125. From the Table 2, it can be seen that, in this test actually the overall classification accuracy is better than the previous test. In this first test the classification accuracy for class 1 was 40%. But in this test it went up to 95.00%. And also overall class accuracy went up from 50.00% in the first test up to 65.00%. The reason that the overall classification accuracy increased is because, in this test, the whole training dataset was used to train SVM. So from this example it can be concluded that, the more data someone has, the better the classification accuracies will be. And also it is a good idea to keep in mind that to dataset's formation, some classes were not included in it. This is another issue that in the case of having examples for each classes, the classification accuracies might fell down. So in the next section, different speech features such as MFCC and SDC will come into play and tests will be done using these acoustic speech features.

From the Figure 30, the class labels and feature vectors can be seen for this test. In this test, there are multiple male and female speakers for each age group. But especially it can be seen that for the first class there are more examples in the training set then the other classes. Also on the right, speech features that is pitch in this case can be seen.

www.manaraa.com

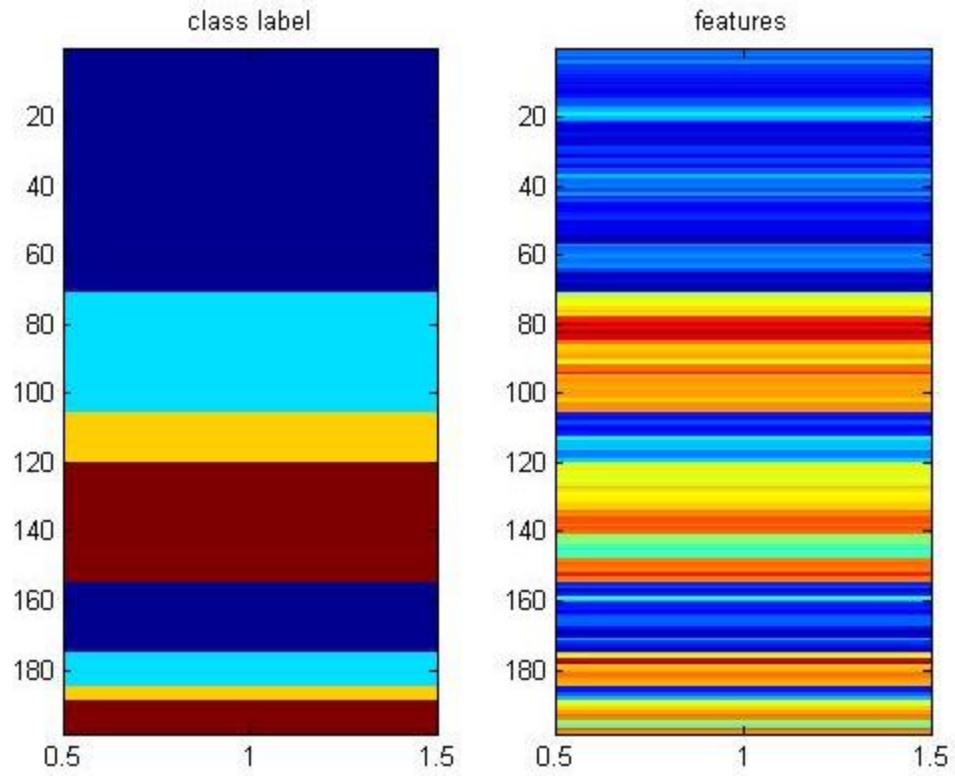

Figure 30. Class Labels and Features for Pitch Based Model for Multiple Speakers for Each Gender and Age Group

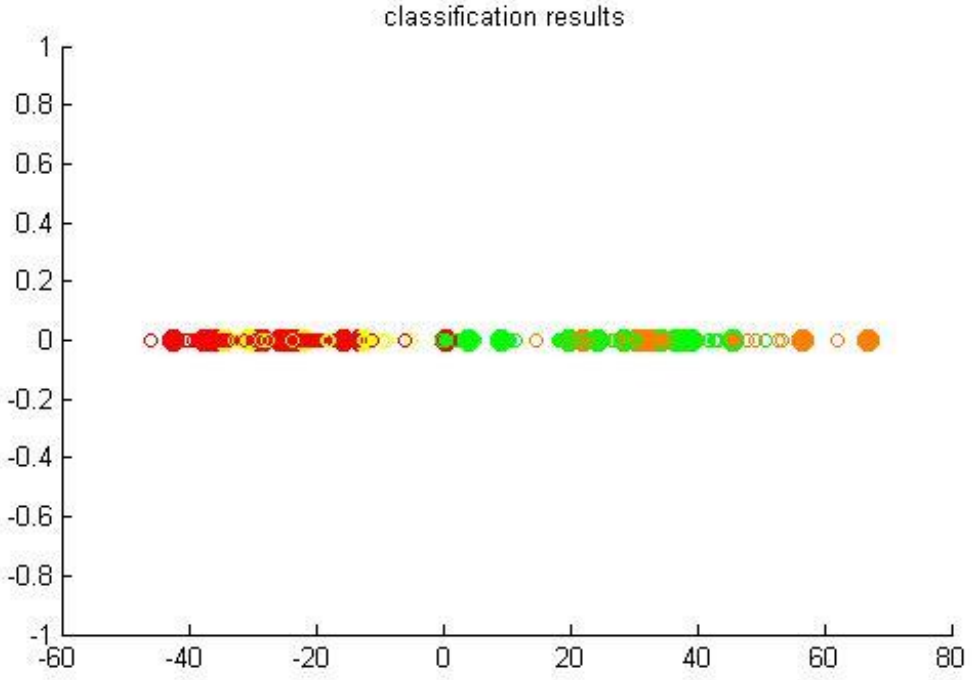

Figure 31. Classification Results for Pitch Based Age and Gender Recognition

www.manaraa.com

Also, from Figure 31, the classification results are shown. Unfilled dots represent the samples from the training set. Filled dots represent the data from the test set. Different colors represent different classes assigned by SVM classifier. Also edge color represent the true label of the example. The size of the dots represent the confidence level of that example, if it is bigger, the confidence level is higher.

6.9.2 Models Based on Acoustic Features

In this section, acoustic speech features such as MFCC and SDC will be used to compare the performance of age and gender recognition system. As a classifier, an SVM with RBF kernel will be trained with different parameters.

6.9.2.1 MFCC Based Model

In this MFCC based model, first, the training data was loaded. Then some pre-processing techniques were applied. Then MFCC was calculated with the 12 coefficients. In order to calculate MFCC, the speech was pre-emphasized and then windowed. Then an FFT was taken. After FFT, a Mel scale filterbank was applied to it and its logarithm was taken. Then finally, DCT was applied to get those coefficients. These features was fed into an SVM with RBF kernel.

A 3-fold cross validation was applied to the training data in order to get the optimum parameters of SVM with RBF kernel. Parameter selection is really important and if the optimum parameters are not selected in the training, the classification accuracy drops significantly. The training dataset is about 30 minutes in duration. There are 22 speakers, 10 male and 12 female. And the average speaking time per speaker is about 83 seconds. Table 3 shows all the test results in this category.

The parameters set manually in this training were C=0.5 and gamma= 0.125. The parameters were not selected using cross validation. From Table 3, it can be seen that classification

www.manaraa.com

percent for class is 3 is 25.86% which is low. But overall class accuracy is 57.90% percent which can be considered as good. One reason of this is because, the parameters used while training SVM were not selected at the end of cross validation. Parameters were selected manually and it shows that parameter selection is really important.

Table 3: Results from MFCC Model Trained Using the One-vs-Rest Multiclass SVM with RBF Kernel with C=0.5 and gamma= 0.125

Class 1	86.91%
Class 2	53.23%
Class 3	25.86%
Class 4	65.61%
Overall Class Accuracy	57.90%

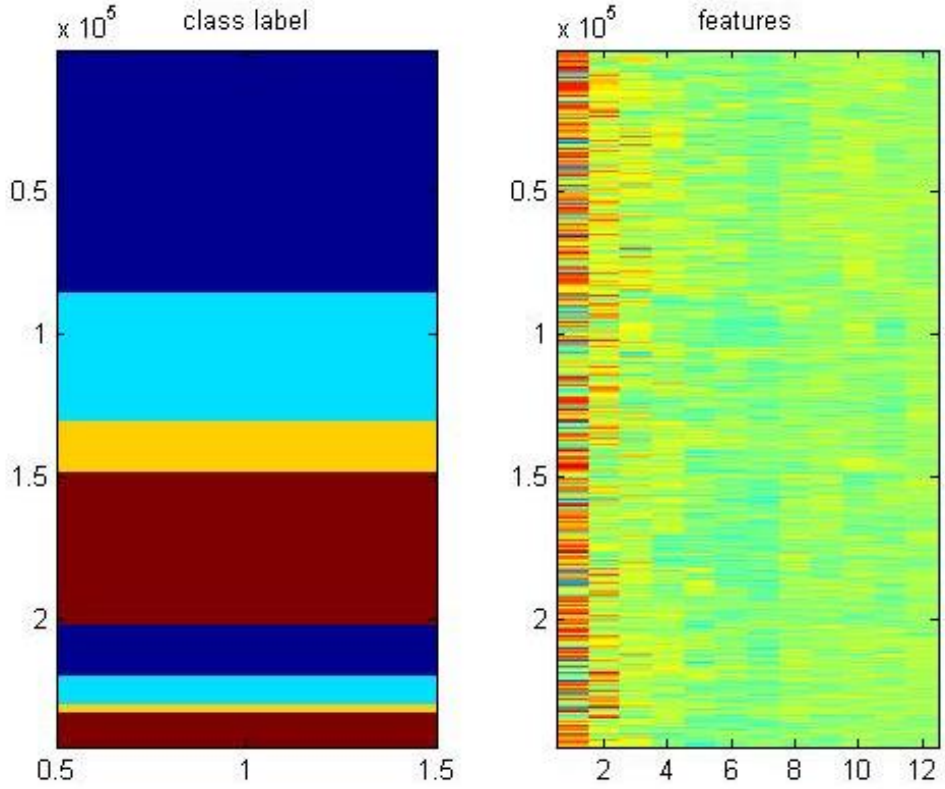

Figure 32. Class Labels and Features for MFCC Based Model Trained Using the One-vs-Rest Multiclass SVM with RBF Kernel

www.manaraa.com

The whole dataset and the classification results were not plotted because it is extremely difficult to visualize multi-dimensional data. In this test, the training and test set made up a matrix of size of 244346x12. In the earlier examples, MDS was used as dimensionality reduction algorithm. But processing a big matrix as in this case takes too much time and also space. Even Matlab was not able to process this matrix and it went out of memory. In this case, also Principal Component Analysis (PCA) was used for dimensionality reduction. But the results from PCA were not satisfactory and the visualization was not good on a 2D plane.

Table 4: Results from MFCC Model Trained Using the One-vs-Rest Multiclass SVM with RBF Kernel with C= 2 and gamma= 0.0625

Class 1	86.89%
Class 2	52.93%
Class 3	25.81%
Class 4	65.41%
Overall Class Accuracy	57.76%

The best values gotten in this test were, C=2 and gamma = 0.0625. The parameters in this test were automatically selected at the end of the cross validation. But not the whole training set were used during cross validation training. Instead, a subset was created out of the training set. Each label contained 2000 examples. The subset had 8000 examples in total each example having 12 features. It turns out that, the classification results shown in Table 4 are similar to the previous test. Probably overall class accuracy could be higher if the whole dataset were used. Even though the results are satisfactory.

www.manaraa.com

6.9.2.2 SDC Based Model

In this SDC based model, first, the training data was loaded into Matlab. Then some pre-processing techniques was applied. Then SDC was calculated with the 84 coefficients. In order to calculate SDC, first MFCC features were calculated as described in the above section. Then finally, SDC coefficients were extracted to get those coefficients. These features were fed into an SVM with RBF Kernel.

A 3-fold cross validation was applied to the training data in order to get the optimum parameters of SVM with RBF kernel. Parameter selection is really important and if the optimum parameters are not selected in the training, the classification accuracy drops significantly. The training dataset is about 30 minutes in duration. There are 22 speakers, 10 male and 12 female. The average speaking time per speaker is about 83 seconds. Table 5 shows all the test results in this category.

Table 5: Results from SDC Model Trained Using the One-vs-Rest Multiclass SVM with RBF Kernel with C=0.5 and gamma= 0.125

Class 1	88.74%
Class 2	31.00%
Class 3	1.72%
Class 4	34.58%
Overall Class Accuracy	39.01%

The parameters set manually in this training were, C=0.5 and gamma = 0.125. The parameters were not selected using cross validation. From Table 5, it can be seen that the classification accuracy for class 3 is very low 1.72%. Classification accuracy for class 1 is high as 88.74%. But overall class accuracy is 39.01% which is low. As mentioned in the MFCC based

www.manaraa.com

model, the reason of this low overall class accuracy is because, the parameters C and gamma were selected manually. In fact, the same values were used with the MFCC based model. From this test, it can be understood that same SVM parameters give different accuracies on different models. In the next test, parameters will be selected at the end of cross validation. But one problem during cross validation was that, it would take so much time days or even weeks to define the parameters C and gamma. So instead, a balanced sub-test was created out of the training set. And cross validation parameters were evaluated based on this subset.

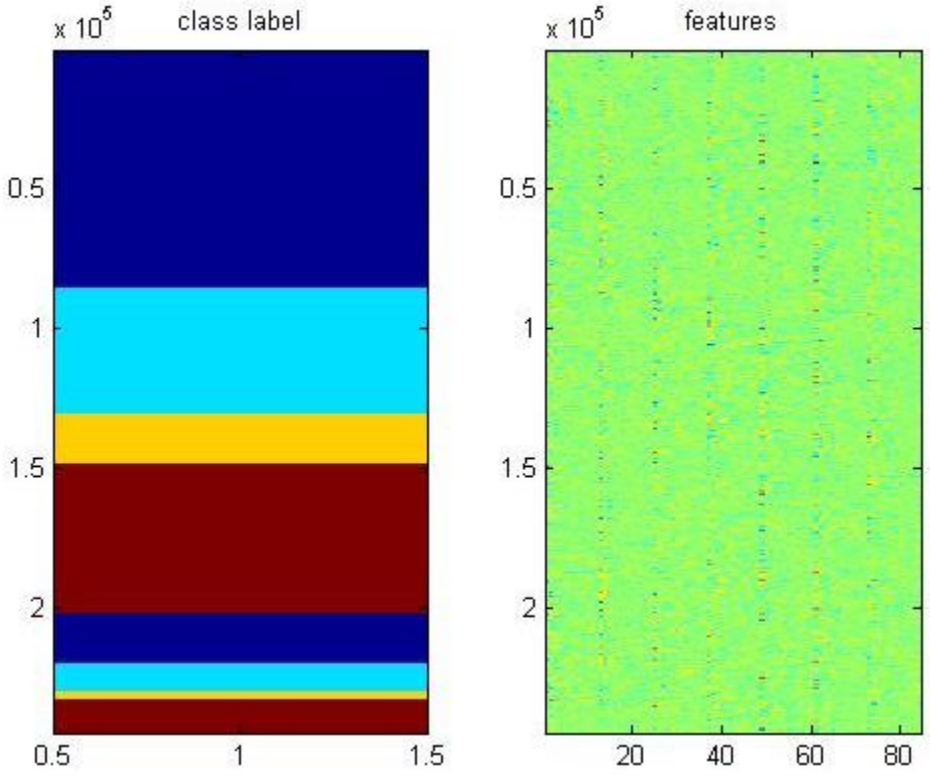

Figure 33. Class Labels and Features for SDC Based Model Trained Using the One-vs-Rest Multiclass SVM with RBF Kernel

As mentioned earlier, the whole dataset and also the classification results were not plotted in this section because of the extreme difficulty of visualization of multi-dimensional data. In this test, the training set and test set made up a matrix of size of 244346x84. In the pitch based model, MDS was used as dimensionality reduction algorithm. But processing a big matrix as in this case

www.manaraa.com

takes too much time and also space. Even Matlab was not able to process this matrix and it went out of memory. In this case also Principal Component Analysis (PCA) was used for dimensionality reduction. But the results from PCA were not satisfactory and the visualization was not good on a 2D plane.

Table 6: Results from SDC Model Trained Using the One-vs-Rest Multiclass SVM with RBF Kernel with C=0.5 and gamma= 0.0625

Class 1	88.74%
Class 2	36.51%
Class 3	1.94%
Class 4	29.52%
Overall Class Accuracy	39.1775%

The training and testing data were same for MFCC based model and for SDC based model. From Table 6, it can be seen that SDC did not perform better than MFCC. This result can be expected since SDC is proven to be better in noisy speech, but since the speech corpus is clean speech, MFCC based model gave better results. Also, parameter change is important but due to the limited resources, cross-validation weren't done on the whole training set.

6.9.2.3 Pitch and MFCC Fused Model

In this section, the last model proposed for age and gender recognition system will be explained. It is a fused model of pitch frequency and MFCC. Training of the model was explained earlier at the beginning of this chapter in model training. Test results can be seen from Table 7.

To fuse two classifiers, first, the probability matrix from pitch based model was oversampled to match the size of the probability matrix from the MFCC based model. Then the probability values of both classifiers were normalized in order to obtain more accurate results.

www.manaraa.com

Two different coefficients were applied to the classifiers. According to the observation, pitch based model were given a coefficient of 0.6 and MFCC based model were given a coefficient of 0.4. Finally, a weighted sum were applied of the probability matrices to obtain the final probabilities. From Table 7, it can be seen that the overall accuracy is just 0.8 percent below the pitch based model. That is not a big loss. The good thing about combining the scores of two classifiers is that, class 3 accuracy jumped up to 74.15% which means that it has more probability to correctly detect that particular age and gender. Nearly for all the tests, class 3 accuracy were low. The reason that class 3 accuracy is low is because the number of training examples were small in size. Also here can be noted that parameter selection is really important in SVM training with RBF kernel.

Table 7: Results from Pitch and MFCC Fused Model

Class 1	91.83%
Class 2	67.67%
Class 3	24.17%
Class 4	73.15%
Overall Class Accuracy	64.20%

Due to the time and resource limitations, tests were not conducted on a different or bigger dataset. But from the fused model it can be seen that, the model works fine.

www.manaraa.com

CHAPTER 7

CONCLUSION

7.1 Summary

The goal of this thesis was age and gender recognition for speech applications. To build that kind of system, all the necessary steps were explained. These steps included pre signal processing techniques, speech feature extraction techniques and finally classification algorithm.

Three different model were proposed for this task. In the first model, only the pitch information was used to try to recognize age and gender of the speaker. In the second model, MFCC and SDC were used as speech features and the classification was done. In the last model, a fused system was proposed which combined pitch and MFCC features together. Various tests were done with different data sizes and also different classification parameters to measure the performance of the system. Due to the time and resource limitations, the tests were done on a relatively small dataset.

According to the test results, it was seen that MFCC performed better as a single speech feature. The reason for this was explained in the earlier sections. And also the best recognition value was achieved with the pitch and MFCC fused model. This fused model gave an accuracy of 64.20% when tested on ELSDSR [32] database and with reasonable C and gamma parameters. And also a GUI implementation was created in Matlab which also gave an option of recoding someone's voice through the microphone and then evaluating his/her age and gender based on the model.

www.manaraa.com

7.2 Future Recommendation

The final pitch and MFCC based fused model gave recognition values up to 64.20% which can be considered as a good value. But in order to increase the performance of the system some additional steps can be implemented. First of all, the training data was small in size. So with the training data size much bigger, the system performance can get better. The system can be trained on super computers so the training time can be reduced. Some noisy dataset can be used and especially SDC based model tests can be observed. Pitch, MFCC and SDC were used as features in this project. So some other speech features can be added into the system which eventually can increase the performance of the system.

www.manaraa.com

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

www.manaraa.com

[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

[17] F. Wittig, and C. Müller, "Implicit Feedback for User-Adaptive Systems by Analyzing the Users' Speech.", 2003

[18] T. Bocklet, A. Maier, and E. Nöth, "Age determination of children in preschool and primary school age with GMM-based supervectors and support vector machines/regression" *Text, Speech and Dialogue*, 2008

[19] M. H. Bahari, and H. V. Hamme, "Speaker age estimation using Hidden Markov Model weight supervectors" *Information Science, Signal Processing and their Applications (ISSPA)*, 2012.

[20] Voice activity detection [online] Available: http://en.wikipedia.org/wiki/Voice_activity_detection#Applications

[21] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection" *Signal Processing Letters*, 1999

[22] T. Carrasquillo, P. A., E. Singer, and M. A. Kohler, "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features"*INTERSPEECH*, 2002

[23] W. M. Campbell, E. Singer, P. A., and T. Carrasquillo, "Language recognition with support vector machines" *ODYSSEY04-The Speaker and Language Recognition Workshop*, 2004

www.manaraa.com

[24] F. Allen, E. Ambikairajah, and J. Epps, "Language identification using warping and the shifted delta cepstrum" *Multimedia Signal Processing*, 2005

[25] A. Russell, L. Penny, and C. Pemberton, "Speaking Fundamental Frequency Changes Over Time in Women A Longitudinal Study" *Journal of Speech, Language, and Hearing Research*, 1995

[26] H. Hollien, and T. Shipp, "Speaking fundamental frequency and chronologic age in males" *Journal of Speech, Language, and Hearing Research*, 1972

[27] X. Sun, "A pitch determination algorithm based on subharmonic-to-harmonic ratio", 2000

[28] M. Brookes, "Voicebox: Speech processing toolbox for matlab", Mar 2011

[29] M. Sahidullahl, "Mfcc to SDC Matlab Library", 2011

[30] X. Sun, "A pitch determination algorithm based on subharmonic-to-harmonic ratio", 2000

[31] C. C. Chang, and C. J. Lin, "LIBSVM: a library for support vector machines." *ACM Transactions on Intelligent Systems and Technology*, 2011

[32] L. Feng, "Speaker Recognition, Informatics and Mathematical Modelling", Technical University of Denmark, 2004

[33] H. Hermansky, and N. Morgan, "RASTA processing of speech." *Speech and Audio Processing*, 1994

[34] S. Davis, and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences" *Acoustics, Speech and Signal Processing*, 1980

www.manaraa.com

APPENDICES

www.manaraa.com

Appendix A: ELSDSR Dataset

Table A.1 ELSDSR Dataset [32]

# of Speakers	22 speakers (12M/10F)
Speaker Age Distribution	24 to 63
Sampling Frequency	16 kHz
# of Sessions/	One
Type of Speech	Read Speech
Microphone	MARANTZ PMD670
Recording Format	.wav
Acoustic Environment	Chamber (Computer Room)
Language	English

Table A.2 Recorder Setup [32]

Input	Setup								
	Auto Mark	Pre Rec	Analog Out	MIN Atten	Repeat	ANC	EDL Play	Level Cont.	S. Skip
MIC (MONO)	OFF	ON	OFF	20dB	OFF	FLAT	OFF	MANUAL	ON 20dB

www.manaraa.com

Table A.3 Information about Speakers [32]

Speaker ID	Age	Nationality
FAML	48	Danish
FDHH	28	Danish
FEAB	58	Danish
FHRO	26	Icelander
FJAZ	25	Canadian
FMEL	38	Danish
FMEV	46	Danish
FSLJ	24	Danish
FTEJ	50	Danish
FUAN	63	Danish
Average	40.6	
MASM	27	Danish
MCBR	26	Danish
MFKC	47	Danish
MKBP	30	Danish
MLKH	47	Danish
MMLP	27	Danish
MMNA	26	Danish
MNHP	28	Danish
MOEW	37	Danish
MPRA	29	Danish
MREM	29	Danish
MTLS	28	Danish
Average	31.3	

Table A.4 Duration of Reading for Training and Test Text [32]

No.	Male		Train(s)	Test(s)	Female		Train(s)	Test(s)
1		MASM	81.2	20.9		FAML	99.1	18.7
2		MCBR	68.4	13.1		FDHH	77.3	12.7
3		MFKC	91.6	15.8		FEAB	92.8	24.0
4		MKBP	69.9	15.8		FHRO	86.6	21.2
5		MLKH	76.8	14.7		FJAZ	79.2	18.0
6		MMLP	79.6	13.3		FMEL	76.3	18.2
7		MMNA	73.1	10.9		FMEV	99.1	24.1
8		MNHP	82.9	20.3		FSLJ	80.2	18.4
9		MOEW	88.0	23.4		FTEJ	102.9	15.8
10		MPRA	86.8	9.3		FUAN	89.5	25.1
11		MREM	79.1	21.8				
12		MTLS	66.2	14.05				

www.manaraa.com