Volume 3, Issue 8, August 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Gender and Speaker Recognition Using MFCC and DTW

Vijender Sharma [1]
ECE Department, KITM Kurukeshtra
India

Rakesh Garg [2]
ECE Department, KITMKurukshetra
India

Abstract-Gender is an important and most diffrentiative characteristic of a speech. Gender information can also be used to improve the performance of speech and speaker recognition systems. Automatic gender classification is a technique that aims to determine the sex of the speaker through speech signal analysis. However with the increase in biometric security application, practical application of gender identification increased the many fold .The need of gender identification from speech arises several situation such as sorting telephonic call. In this paper we implemented the gender classification method and gender dependant feature such as pitch, formant frequency, jitter etc. We have used MFCC in combination with DTW to recognize speaker.

Keywords: MFCC; DTW; Pitch; Formants; Cepstral coefficients.

I. INTRODUCTION

Nowadays more and more attention has been paid on speaker recognition field. Speaker recognition, which involves two applications: speaker identification and speaker verification, is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves. This technique makes it possible to use the speaker's voice to verify their identity and control access to services such as voice dialing, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers.Figure 1 shows the basic structure of speaker recognition system [1].

Speaker recognition systems generally consist of three major units. The input to the first stage or the front end processing system is the speech signal. Here the speech is digitized and subsequently the feature extraction takes place. There are no exclusive features that convey the speakers identity in the speech signal, however it is known from the source filter theory of speech production that the speech spectrum shape encodes in it the information about speakers vocal tract shape via formants and glottal source via pitch harmonics. Therefore some form or the other of the spectral based features is used in most of the speaker recognition systems. The final process in the front end processing stage is some form of channel compensation [11].

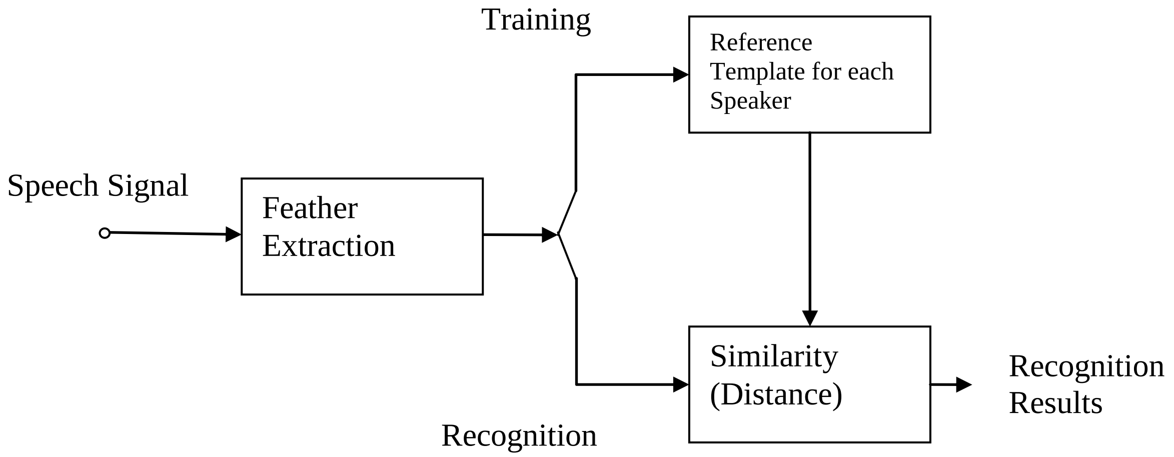

Figure 1: Basic structure of Speaker Recognition system

The process of speaker recognition consists of the training phase and the recognition phase. In the training phase, the features of a speakers.speech signal are stored as reference features. The feature vectors of speech are used to create a speakers model. The numbers of reference templates that are required for efficient speaker recognition depend upon the kind of features or techniques that the system uses for recognizing the speaker. In the recognition phase, features similar to the ones that are used in the reference template are extracted from an input utterance of the speaker whose identity is required to be determined [2].

The recognition decision depends upon the computed distance between the reference template and the template devised from the input utterance.In speaker identification, the distance between an input utterance and all of the available reference templates is computed. The template of the registered user, whose distance with the input utterance template is the smallest, is finally selected as the speaker of the input utterance.

In case of speaker verification the distance is computed only between the input utterance and the reference template of the claimed speaker. If the distance is smaller than the predetermined threshold, the speaker is accepted other the speaker is rejected as an imposter.

II. GENDER RECOGNITION

Gender Recognition (GR) can help in the development of speaker-independent speech recognition systems. The approaches to automatic gender recognition can be done on the basis of following parameters [10].

- Pitch (Fundamental frequency)
- Jitter
- Shimmer
- Formant Frequencies

A. Pitch (Fundamental frequency):

The pitch has aroused the periodicity through vocal cords vibration when madding voiced sound, pitch frequency is a very important parameter using to describe the characteristic of voice excitation source. The variation range of pitch frequency is generally from 50 Hz to 500 Hz, the cycle of the male voice is 50 Hz - 300 Hz, and the female is 100 Hz - 500 Hz [15]. Although each person's different vocal structure lead to different fundamental frequency, because of the pitch frequency's scope is a little small, the gap between different people is little, and the most important is pitch frequency is affected by a lot of factors, such as emotion, tone, it is very difficult to achieve accurate fundamental frequency. Thus, the recognition rate is very low using the fundamental frequency for speaker recognition now. But male fundamental frequency is generally lower than the female; it is a good argument as classification [5].

B. Jitter:

Fundamental frequency is determined physiologically by the number of cycles that the vocal folds do in a second. Jitter refers to the variability of F0, and it is affected mainly because of the lack of control of vocal fold vibration .On the other hand, vocal intensity is related to sub glottis pressure of the air column, which, in turn, depends on other factors such as amplitude of vibration and tension of vocal folds [29]. The novel component in this thesis is the analysis of jitter and shimmer features in order to test their usefulness in speaker verification. These features have been extracted by using the PRAAT voice analysis software. PRAAT reports different kinds of measurements for both jitter and shimmer features, which are listed below.

- **Jitter (absolute):** Jitter is the cycle to cycle variation of the pitch period, i.e., the average of the absolute distance between consecutive periods. It is measured in μ sec.

$$\mathrm{Jitter(absolute)} = \frac{1}{N-1}\sum_{i=1}^{N-1} |T_i - T_{i+1}|$$

  Where Ti is the extracted F0 period length and N is the number of extracted F0 pitch periods. Absolute jitter values, for instance, are found larger in males as compared to females [4].

- **Jitter (relative):** It is the average absolute difference between consecutive periods, divided by the average period. It is expressed as a percentage:

$$\mathrm{Jitter(relative)} = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} |T_i - T_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} T_i}$$

- **Jitter (rap):** It is defined as the Relative Average Perturbation, the average absolute difference between a period and the average of it and its two neighbours, divided by the average period.

- **Jitter (ppq5):** It is the five-point Period Perturbation Quotient, computed as the average absolute difference between a period and the average of it and its four closest neighbours, divided by the average period.

C. Shimmer:

Shimmer is affected mainly because of the reduction in this tension and mass lesions in the vocal folds .Both jitter and shimmer features have been largely used to detect voice pathologies. More recently, they have also been used to determine the classification of human speaking styles and the age and gender of the speakers [3]

- **Shimmer (dB):** It is the variability of the peak-to-peak amplitude in decibels. It is the ratio of amplitudes of consecutive periods. It is expressed as:

$$\mathrm{Shimmer(dB)} = \frac{1}{N-1}\sum_{i=1}^{N-1} |20 \log\left(\frac{A_{i+1}}{A_i}\right)|$$

  Where Ai is the peak-to-peak amplitude in the period and N is the number of extracted fundamental frequency periods. Local shimmer (db) values are found larger in female as compared to males.

- **Shimmer (relative):** It is defined as the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude, expressed as a percentage:

$$\mathrm{Shimmer(relative)} = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i}$$

- **Shimmer (apq3):** It is the three-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of its neighbours, divided by the average amplitude.

- **Shimmer (apq5):** It is defined as the five-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of it and its four closest neighbours, divided by the average amplitude [3].

D. Formant frequencies:

The gender based differences in human speech are partially due to physiological differences such as vocal fold thickness or vocal tract length and partially due to differences in speaking style. The female speakers normally have higher formant frequencies as well as higher fundamental frequency (F0). But for male speakers, the F0 is lower, because of the qualities like aggressiveness, body size, self-assurance, and assertiveness. Voice recognition works based on the premise that a person voice exhibits characteristics are unique to different speaker. The signal during training and testing session can be greatly different due to many factors such as people voice change with time, health condition (e.g. the speaker has a cold), speaking rate and also acoustical noise and variation recording environment via microphone [6].

III. FEATURE EXTRACTION TECHNIQUE

The general methodology of audio classification involves extracting discriminatory features from the audio data and feeding them to a pattern classifier. Different approaches and various kinds of audio features were proposed with varying success rates. The features can be extracted either directly from the time domain signal or from a transformation domain depending upon the choice of the signal analysis approach. Some of the audio features that have been successfully used for audio classification include Mel-frequency cepstral coefficients (MFCC), Linear predictive coding (LPC), Local discriminant bases (LDB) [7].

E. Mel-frequency cepstral coefficients(MFCC) and DTW

The most prevalent and dominant method used to extract spectral features is calculating Mel- Frequency Cepstral Coefficients (MFCC). MFCCs are one of the most popular feature extraction techniques used in speech recognition based on frequency domain using the Mel scale that is based on the human ear scale. MFCCs being considered as frequency domain features are much more accurate than time domain features. MFCCs extraction involves a frame-based analysis of a speech signal where the speech signal is broken down into a sequence of frames.Each frame undergoes a sinusoidal transform (Fast Fourier Transform) in order to obtain certain parameters that then undergo Mel-scale perceptual weighting and de-correlation. The result is a sequence of feature vectors describing useful logarithmically compressed amplitude and simplified frequency information. The Mel-Frequency Cepstral Coefficients (MFCC) represents the real cepstral of a windowed short-time signal derived from the Fast Fourier Transform (FFT) of that signal. The difference from the real cepstral is that a nonlinear frequency scale is used, which approximates the behavior of the auditory system .MFCC is an audio feature extraction technique that extracts parameters from the speech similar to ones that are used by humans for hearing speech, while at the same time, de-emphasizes all other information . As MFCCs take into consideration the characteristics of the human auditory system, they are commonly used in the automatic speech recognition systems [8].

First performing a standard Fourier analysis and then converting the power-spectrum to a mel-frequency spectrum obtain MFCC features. Therefore, MFCC will be obtained by taking the logarithm of that spectrum and by computing its Discrete Cosine transform.The main steps required for the MFCC computations, are clearly shown in Figure 2. The main steps include the followings: pre emphasis, framing, windowing using hamming window, performing Fast Fourier Transform (FFT), applying the Mel-scale filter bank in order to find the spectrum as it might be perceived by the human auditory system, performing the Logarithm, and finally taking the Discrete Cosine Transform (DCT) of the logarithm of the magnitude spectrum to obtain Mel Frequency Cepstral Coefficients.

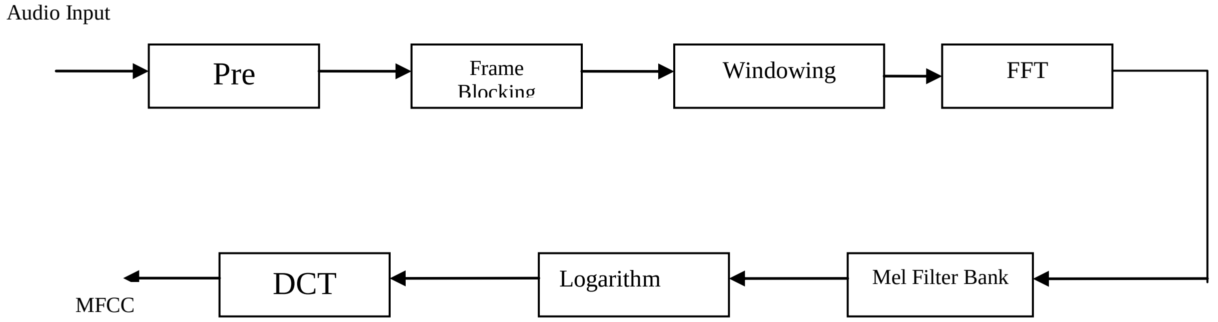

Figure 2: Block Diagram of the Computation Steps of MFCC

A template-based pattern comparison approach, namely dynamic time warping (DTW) algorithm is used in the system. Dynamic time warping is an algorithm used for measuring optimal match between two sequences which may vary in time or speed. The sequences are warped non-linearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension. It was developed for isolated word recognition application and was adapted by Furui for text-dependent speaker verification. For speaker verification task, this approach compares the acoustic features derived from the speech signal collected during enrollment, with the acoustic features derived from the speech signal collected during verification. The result of this comparison is a dissimilarity measure. It has been used in many text-dependent speaker verification applications [9].

IV. SIMULATION SET UP AND RESULTS

In this work gender recognition is done depending upon four acoustic features. Here we have speech samples of males and females .We have calculated four parameters like pitch, jitter, shimmer and formants and depending upon the results we have concluded the gender recognition. Figure 3,4,5,6 shows the speech signal for males (M1, M2, M3, and M4). Speech signal is text "This is Yamunanagar" spoken by all males.

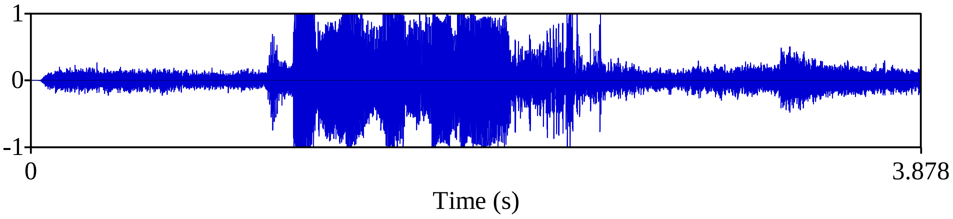

Figure 3: Male 1 Speech Signal "This is Yamunanagar"

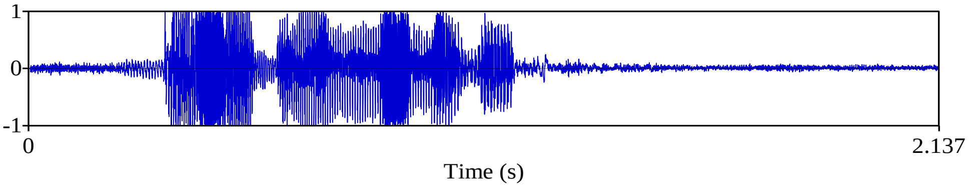

Figure 4: Male 2 Speech Signal "This is Yamunanagar"

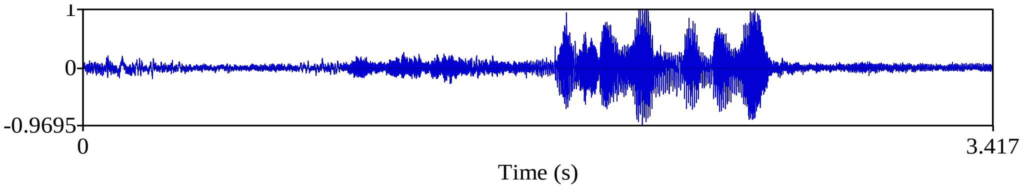

Figure 5: Male 3 Speech Signal "This is Yamunanagar"

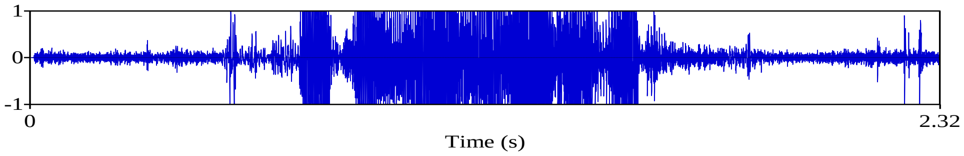

Figure 6: Male 4 Speech Signal "This is Yamunanagar"

Figure 7, 8,9,10 shows the speech signal for females (F1, F2, F3, and F4). Speech signal is text "This is Yamunanagar" spoken by all females.

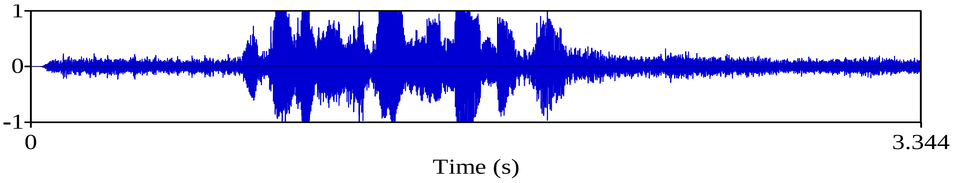

Figure 7: Female 1 Speech Signal "This is Yamunanagar"

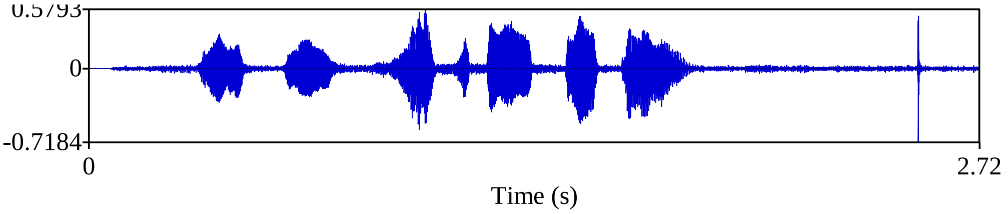

Figure 8: Female 2 Speech Signal "This is Yamunanagar"

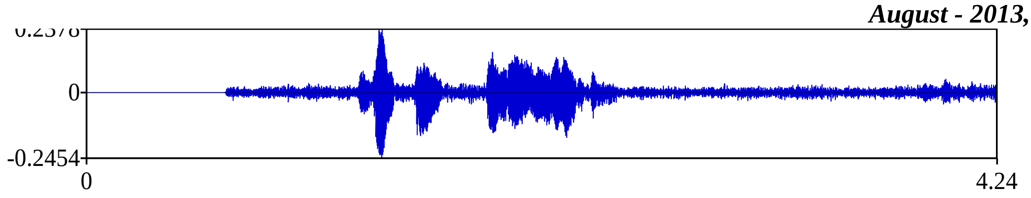

Figure 9: Female 3 Speech Signal "This is Yamunanagar"

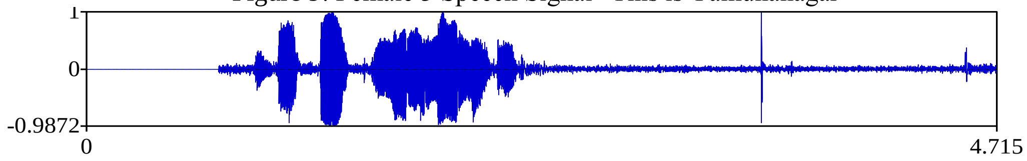

Figure 10: Female 4 Speech Signal "This is Yamunanagar"

Table 1 shows the calculation of mean pitch for 8 speakers 4 male and 4 female. Figure 11 show the bar diagram for male and female and it is concluded that females have higher pitch than males.

TABLE 1
MALE AND FEMALE MEAN PITCH

Pitch Male (M)		Pitch Female	
M1	**Mean pitch: 194.470 Hz**	**F1**	**Mean pitch: 258.455 Hz**
M2	**Mean pitch: 170.985 Hz**	**F2**	**Mean pitch: 285.351 Hz**
M3	**Mean pitch: 176.590 Hz**	**F3**	**Mean pitch: 248.609 Hz**
M4	**Mean pitch 166.701 Hz**	**F4**	**Mean pitch: 250.489 Hz**

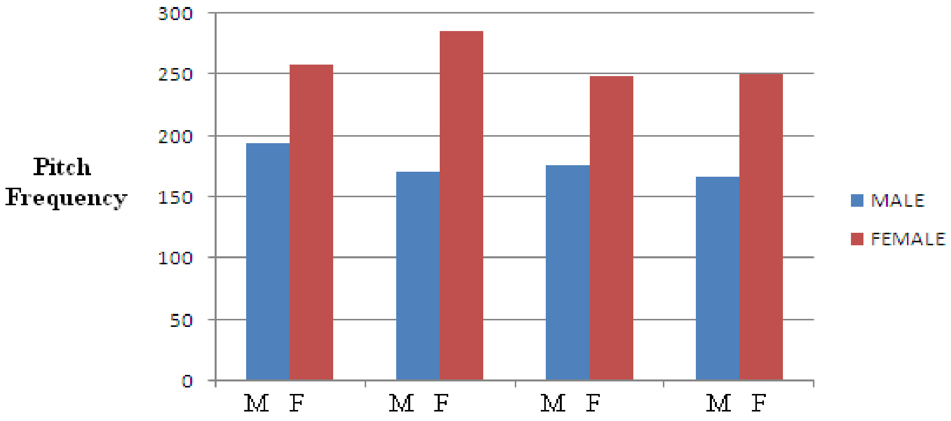

Figure 11: Bar Diagram of Male and Female Mean Pitch

Table 2 shows the calculation of absolute jitter for 8 speakers 4 male and 4 female. Figure 12 show the bar diagram for male and female and it is concluded that males have higher value of absolute jitter than females.

TABLE 2
MALE AND FEMALE JITTER

Jitter male		Jitter female	
M1	**Jitter (absolute): 93.270E-6 seconds**	**F1**	**Jitter (absolute):68.144E-6seconds**
M2	**Jitter (absolute):136.586E-6 seconds**	**F2**	**Jitter (absolute):93.174E-6 seconds**
M3	**Jitter (absolute): 212.329E-6 seconds**	**F3**	**Jitter (absolute):132.5E-6 seconds**
M4	**Jitter (absolute): 202.170E-6 seconds**	**F4**	**Jitter (absolute):111.25E-6 seconds**

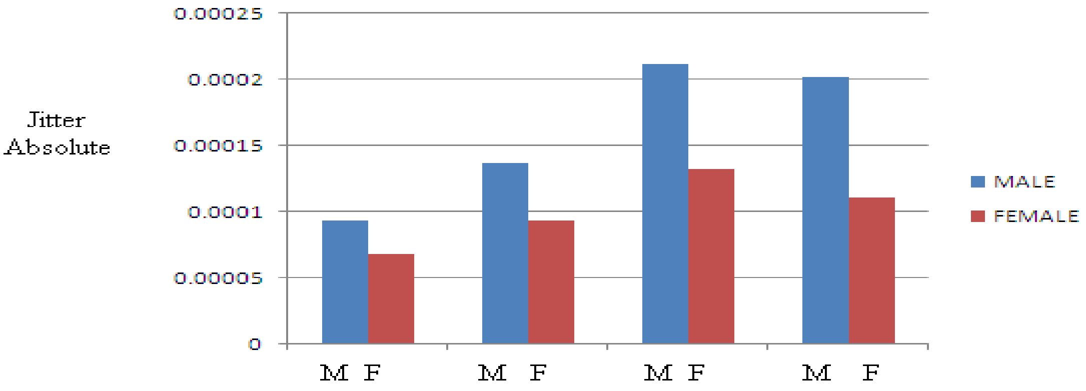

Figure 12: Bar Diagram of Male and Female Jitter

Table 3 shows the calculation of shimmer for 8 speakers 4 male and 4 female. Figure 13 show the bar diagram for male and female and it is concluded that females have higher value of absolute shimmer than females.

TABLE 3
MALE AND FEMALE SHIMMER

Shimmer male		Shimmer female	
M1	**Shimmer (local, dB): 0.830 dB**	**F1**	**Shimmer (local, dB): 1.333 dB**
M2	**Shimmer (local, dB): 0.994 dB**	**F2**	**Shimmer (local, dB): 1.235 dB**
M3	**Shimmer (local, dB): 1.323 dB**	**F3**	**Shimmer (local, dB): 1.517 dB**
M4	**Shimmer (local, dB): 1.151 dB**	**F4**	**Shimmer (local, dB): 1.308 dB**

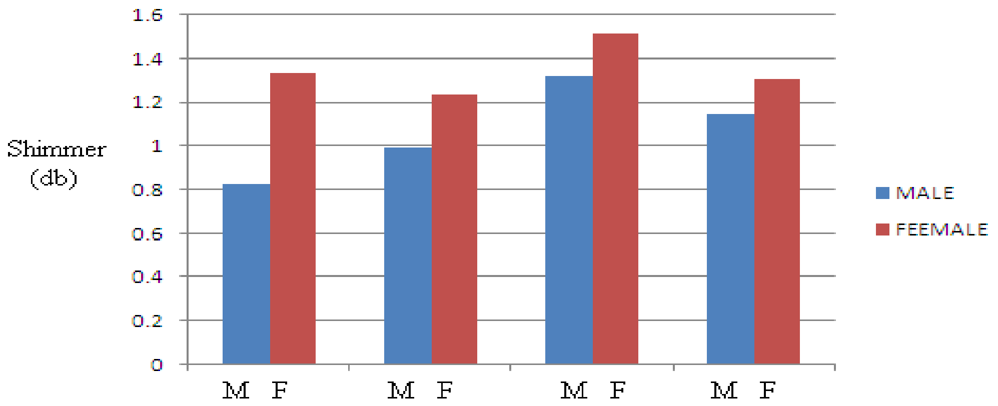

Figure 13: Bar Diagram of Male and Female Shimmer

Table 4 shows the calculation of formant frequencies for 8 speakers 4 male and 4 female. Figure 4.12 show the bar diagram for male and female and it is concluded that females have higher formant frequencies than males.

TABLE 4
MALE AND FEMALE FORMANT F1

Format F1 Male		Formant F1 Female	
M1	**643.4801668223015 Hz (mean F1)**	**F1**	**724.9030792853079 Hz (mean F1)**
M2	**467.1728671841172 Hz (mean F1)**	**F2**	**661.0252047703079 Hz (mean F1)**
M3	**582.8508812013699 Hz (mean F1)**	**F3**	**656.0315133700476 Hz (mean F1)**
M4	**663.684092752845 Hz (mean F1)**	**F4**	**805.573411557635 Hz (mean F1)**

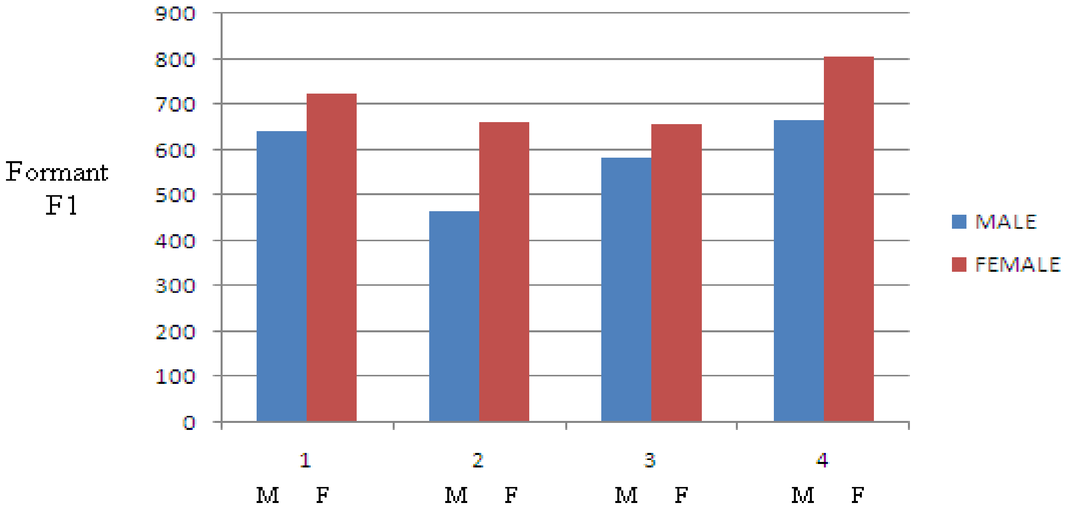

Figure 14: Bar Diagram of Male and Female Formant F1

III. Conclusion

In this paper gender recognition is done depending upon four acoustic features. Pitch and formant are both the most important parameters of the speech signal. In theory, because of the differences of vocal tract structure and sound track, everyone should have their own different characteristics of pitch and formant. Speech signal changes in complex, sound channel and noise have an effect on the signal, and extracting methods are imperfect, so pitch or formant is not an effective parameter in speaker recognition recently, they can only play a supporting role. MFCC is effective for speaker identification, because it combines sensing features of the human ear with producing mechanism of voice.

References

[1]. D.A. Reynolds, R.C. Rose, "*Robust text-independent speaker identification using Gaussian Mixture speaker models*", IEEE Trans. on Speech and Audio Processing, vol.3, no. 1, pp. 72-83, 1995.

[2]. R. A. Cole and colleagues, "*Survey of the State of the Art in Human Language Technology*", National Science Foundation European Commission, 1996.

[3]. Joseph P. Campbell, "*Speaker Recognition: A tutorial*", Proc. IEEE, vol. 85, pp. 1437-1462, September 1997.

[4]. D.A. Reynolds, L.P. Heck, "*Automatic Speaker Recognition*", AAAS 2000 Meeting, Humans, Computers and Speech Symposium, 19 Feb 2000.

[5]. D. Chasan et al, "*Speech reconstruction from MFCCs and pitch*", Proc. ICASSP, 2002.

[6]. L. P. Cordella, P. Foggia, C. Sansone, M. Vento, "*A Real-Time Text-Independent Speaker Identification System*", Proceedings of the ICIAP, pp. 632, 2003.

[7]. Chunsheng Fang, From "*Dynamic time warping (DTW) to Hidden Markov Model (HMM)*", University of Cincinnati,2009.

[8]. Wei-Qiang Zhang, Dengzhou Yang and Jia Liu, Xiuguo Bao, Perturbation "*Analysis of Mel-Frequency Cepstrum Coefficients*", 2010 IEEE.

[9]. Vibha Tiwari-"*MFCC and its applications in speaker recognition*", International Journal on Emerging Technologies, Received 5 Nov., 2009, Accepted 10 Feb., 2010.

[10]. D.Shakina Deiv,Gaurav and Mahua Bhatta Charaya, "*Automatic Gender Identification for Hindi Speech Recognition*", International Journal of Computer Applications (0975 – 8887) Volume 31– No.5, October 2011.

[11]. Nilu Singh, R A Khan and Raj Shree. Article: "*MFCC and Prosodic Feature Extraction Techniques: A Comparative Study*". International Journal of Computer Applications 54(1):9-13, September 2012.