A Novel Text - Independent Voice based Automatic Gender Recognition System

D. Shakina Deiv Mahua Bhattacharya G.K. Sharma

ABV-Indian Institute of Information Technology and Management, Gwalior, India

[*1] dshakinadeiv@gmail.com, [2] mahuabhatta@gmail.com, [3] gksharma@iiitm.ac.in

Abstract

Voice based gender and age classification can be helpful in a number of Information Technology based applications with speech interfaces. The recognition has to be independent of the text of the input speech if the application is online. In this work three different feature sets were tried for text independent gender recognition. The first set is Mel-Frequency Cepstral Coefficients (MFCC) C_1 to C_{24}. A feature relevance study was undertaken using F-ratio based analysis and the first feature set was transformed accordingly to get the novel second set which is the F-ratio based dimension reduced MFCC. The third set is the weighted version of the second one. All the three sets have performed well, especially the second and third show excellent recognition performance. An effort is made to reduce computational complexity by feature dimension reduction and optimizing the number of Gaussian components of the GMM based gender classifier. This work can be extended for age based speaker classification too. A review of recent works and the previous work of the authors on text – dependent gender recognition are briefly presented for context.

Keywords*: Voice based gender classification; Mel-Frequency Cepstral Coefficients; Text – dependent Gender Classification; Template Matching; Text – independent Gender Classification; F-ratio; Dimension Reduction; Gaussian Mixture Model.*

1. Introduction

Speech signal conveys reliable information pertinent to the identity of the speaker like gender, age, social and regional origin, health and emotional state in addition to the message (Benzeguiba et al., 2006). This fact has led to the formulation of a lot of voice based applications. In this work, a novel and robust method to identify the gender of a speaker from voice is presented.

Automatic recognition of gender and age of a speaker is used to enhance the performance of Speaker-Independent Automatic Speech Recognition (SIASR) Systems. It can also be used to combine the unsupervised speaker tracking and automatic speaker adaptation techniques for achieving better human–computer interfaces. In (Das et al., 2013), the effect of aging on Bengali phoneme recognition with a large vocabulary corpus of aging Bengali speech is analyzed and the

results used to improve speaker adaptation. The speech controlled system presented in (Herbig et al., 2012) is suitable for in-car applications like speech controlled navigation, hands-free telephony or infotainment systems, on embedded devices.

AGR can also help to identify acoustic features important for synthesizing male and female voices and provide guidelines for identifying acoustic features related to dialect, accent, age, health and other speaker idiosyncratic characteristics (Childers et al., 1987). AGR is also used to improve the speaker clustering task which is useful in speaker recognition. In content based multimedia indexing, gender of the speaker is a cue used in the annotation (Harb and Chen, 2006). Patterns of variations of voice source features like physiological and anatomical changes of vocal tract with aging have been reported by (Vipperla et al., 2010).

The approaches to automatic gender recognition can be divided into three broad classes. The first approach uses gender-dependent features such as pitch. The fundamental frequency F_0, with typical values of 110 Hz for male speech and 200 Hz for female speech, is an important factor in the identification of gender from voice. In (Abdulla and Kasabov, 2001) average pitch frequency was used as a gender separation criterion and the System achieved 100% accuracy in gender discrimination, with TIMIT (Texas Instruments and Massachusetts Institute of Technology) continuous speech corpus and Otago isolated words speech corpus. (Zourmand et al., 2013) have presented Gender Classification using Fundamental and Formant Frequencies of Malay Vowels.

Pitch is a very strong source of information for gender identification of speakers however a good estimate of the pitch period can only be obtained from voiced portions of a clean non-noisy signal. It is often very weak or missing in telephone speech due to the band-limiting effect of the telephone channel. The method proposed by (Meena et al., 2013) uses fuzzy logic and neural network to identify the gender of the speaker using the features, Short Time Energy, Zero Crossing Rate and Energy Entropy.

The second is Pattern Recognition which uses cepstral features such as Mel-Frequency Cepstral Coefficients (MFCCs) to discern the gender of a speaker from a spoken utterance. The performance of pitch and cepstral features, namely LPCCs, MFCCs, PLPs are compared in (Pronobis and Doss, 2009) for robust Automatic Gender Recognition (AGR). It was found that all the features provide almost similar performance under clean conditions. But under noisy conditions, cepstral features are more robust than others and yield better results. In (Feld et al., 2010) automatic recognition of age and gender of a speaker is studied under car noise for the purpose of applicability in mobile services.

The third approach uses a combination of knowledge based features and statistical features for improved performance. (Sharma and Garg, 2013) have implemented gender classification using MFCC features in combination with DTW to recognize speaker. A glottal excitation feature based Gender Identification System using ergodic HMM in (Rao and Prasad,

2011) demonstrates the importance of information in the excitation component of speech (pitch) for gender recognition task. In (Metze et al., 2007), four approaches for age and gender recognition using telephone speech have been compared; *namely*, a parallel phone recognizer, a system using dynamic Bayesian networks, a system based solely on linear prediction analysis, and Gaussian mixture models based on MFCCs. Several popular methods for gender classification have been investigated by (Sedaaghi, 2008) in emotionally colored speech.

It is widely reported that text selection improves voice based gender and speaker identification. Vowels of English have been reported to perform well in speaker identification for selected MFCC coefficients as feature vectors. It is shown in (Sigmund, 2008) that the gender of a speaker can be correctly identified using a set of selected MFCC with an accuracy of about 93% from clean speech. It is interesting to note that the time duration of speech needed for identification decreases significantly for selected speech segments like English vowels.

Vowels and nasals are found to be effective in gender identification. They are relatively easy to identify in speech signal and their spectra contain features that reliably distinguish genders. Nasals are of particular interest because the nasal cavities of different speakers are distinctive and are not easily modified (except for nasal congestion). Phoneme based features have been analyzed in English for effective gender identification (Sigmund, 2008). The authors (Deiv et al., 2011) have studied the effect of text selection in Hindi. There are 11 vowels and a phonemes like retroflex and aspirated stops which are absent in English and other European Languages. The study analyzed the long and short vowels and 5 nasals of Hindi for their gender distinguishing ability.

A Text Independent Speaker Identification technique using Integrated Independent Component Analysis with Generalized Gaussian Mixture Model is presented by (Ramaligeswararao et al., 2011). The goal of this work too is to design and implement a gender recognition system that is text independent but to be used in online applications. Section 2 explains the extraction of MFCC Features. Section 3 briefly presents the Hindi Text Dependent AGR system (Deiv et al., 2011) that was developed prior to this study and the relevant conclusions we draw from it. Section 4 explains the Text Independent AGR. In section 5, the experimental results are discussed.

2. Feature Extraction

As mentioned earlier pitch is a very strong indicator of the gender of the speaker. However it is difficult to extract accurate pitch due to the non-stationary and quasi-periodicity of speech signal, as well as the interaction between the glottal excitation and the vocal tract. Speech frames are not always periodic and pitch can be determined only from only the voiced frames. Since our interest is in capturing global features which correspond to source excitation, the low frequency or pitch components are to be emphasized. To fulfill this requirement it is felt that MFCCs are most suitable as they emphasize low frequency and de-emphasize high frequencies

(Rao and Prasad, 2011). Mel Frequency Cepsral Coefficients on the other hand have proved to be very robust in the case of gender, speaker and also speech recognition. However, unlike in speech, the difference coefficients have not proved very efficient in previous studies. So the features are MFCC_0. The first 40 coefficients are studied in Text Dependent AGR and based on the inferences, only the first 24 coefficients are taken up in Text Independent AGR.

2.1. Mel-Frequency Cepstral Coefficients

MFCCs give a measure of the energy within overlapping frequency bins of a spectrum with a warped (Mel) frequency scale. Since speech can be considered to be short term stationary, MFCC feature vectors are calculated for each frame of detected speech. MFCC speech parameterization is designed to maintain the characteristic of human sound perception, as they are based on the known variation of the human ear's critical bandwidths with frequency. The MFCC technique makes use of two types of filter, namely, linearly spaced filters and logarithmically spaced filters. The Mel frequency scale has linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz.

In the sound processing, the Mel-frequency cepstrum is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. The procedure by which the Mel-frequency cepstral coefficients are obtained consists of the following steps.

2.1.1 Pre–emphasis The speech signal is passed through a filter which emphasizes higher frequencies i.e. will increase the energy of the signal at higher frequencies. The Pre-emphasis of the speech signal is realized with this simple FIR filter $H(z) = 1 - az^{-1}$ where a falls in the interval [0.9, 1].

2.1.2 Framing The process of segmenting the digitized speech samples into frames with a length within the range of 10 to 40 msec.

2.1.3 Hamming windowing The segment of waveform used to determine each parameter vector is usually referred to as a window. The Hamming window which is used for the purpose is defined by the equation

$$W(n) = 0.54 - 0.46cos\left[\frac{2\pi n}{N-1}\right];\ 0 \le n \le N-1 \qquad \textbf{(1)}$$

Where, N = number of samples in each frame.

Let Y[n] = Output signal and X (n) = input signal

$$Y(n) = X(n)\, x\, W(n) \qquad \textbf{(2)}$$

Here N = number of samples in each frame. The result of windowing the signal is

$$Y(\omega) = FFT\,[h(t) * x(t)] = H(\omega) * X(\omega) \quad (3)$$

2.1.4 Fast Fourier Transform

Next, the Fast Fourier transform (FFT) is used to convert each frame of N samples from time domain into frequency domain. Thus the components of the magnitude spectrum of the analyzed signal are calculated.

2.1.5 Mel Filter Bank Processing

Compensation for non-linear perception of frequency is implemented by the bank of triangular band filters with the linear distribution of frequencies along the so called Mel-frequency range. Linear deployment of filters to Mel-frequency axis results in a non-linear distribution for the standard frequency axis in hertz. Definition of the Mel-frequency range is described by the following equation.

$$f_{mel} = 2595\, log_{10}\left(1 + \frac{f}{700}\right) Hz \quad (4)$$

Here f is frequency in linear range and f_{Mel} the corresponding frequency in nonlinear Mel-frequency range.

2.1.6 Discrete Cosine Transform

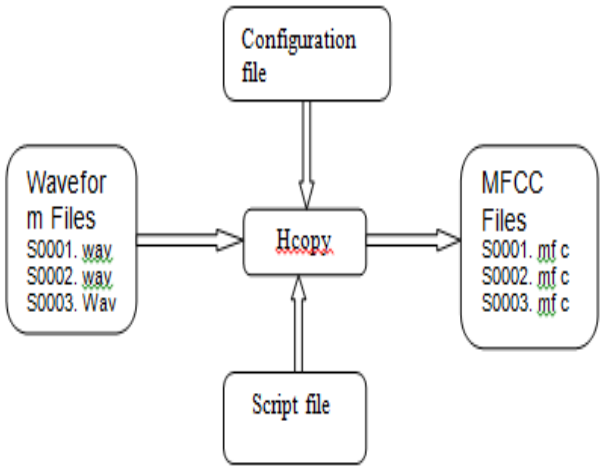
The next step is to calculate the logarithm of the outputs of filters, which affects the dynamics of the signal.

$$c_n = \sqrt{\frac{2}{K}} \sum_{j=1}^{N} (log_{mj}) cos\left(\frac{\pi n}{K}(j - 0.5)\right) \quad (5)$$

Finally, the log Mel spectrum is converted back to time domain. The Mel-spectrum coefficients and their logarithms are real numbers. Hence they can be converted to the time domain using the discrete cosine transform (DCT). The resultant is the Mel Frequency Cepstral Coefficients.

2.2. Coding the data

The speech data were collected and coded into MFCC feature vectors in the following manner. The HMM Tool Kit (HTK) was used for parameterization. Coding was performed using the tool HCopy configured to automatically convert its input into MFCC vectors. A configuration file (config) specifies all of the conversion parameters (Young et al., 2009). The frame period is 10ms. The feature extraction process is shown in figure 1. Both the text dependent and text independent methods use MFCC coefficients. However the databases used for both are different. MFCC_0 with 40 coefficients apart from C_0 was extracted for text dependent AGR and MFCC_0 with 24 coefficients apart from C_0 for text independent AGR. The screen shot of an MFCC feature vector containing 24 coefficients apart from C_0, extracted thus is shown in table 1.




```
# Coding parameters
TARGETKIND = MFCC_0_Z
TARGETRATE = 100000.0
SAVECOMPRESSED = T
SAVEWITHCRC = T
WINDOWSIZE = 250000.0
USEHAMMING = T
PREEMCOEF = 0.97
NUMCHANS = 26
CEPLIFTER = 22
NUMCEPS = 24
```

Figure 1 Extraction of feature vector using HTK

Table 1 Screen shot of extracted MFCC

```
preeti@ubuntu:/media/preeti/Elements/gender/gdw_gmodels$ cd ..
preeti@ubuntu:/media/preeti/Elements/gender$ cd fr_gmodels
preeti@ubuntu:/media/preeti/Elements/gender/fr_gmodels$ HList -h -o -i 13 -s 1 -e 10 test/mfcc_gfemale/aakansha323.mfc
------------------------- Source: test/mfcc_gfemale/aakansha323.mfc -------------------------
  Sample Bytes:  50       Sample Kind:   MFCC_C_K_Z_0
  Num Comps:     25       Sample Period: 10000.0 us
  Num Samples:   200      File Format:   HTK
------------------------------------ Observation Structure ------------------------------------
x:      MFCC_1  MFCC_2  MFCC_3  MFCC_4  MFCC_5  MFCC_6  MFCC_7  MFCC_8  MFCC_9 MFCC_10 MFCC_11 MFCC_12 MFCC_13
      MFCC_14 MFCC_15 MFCC_16 MFCC_17 MFCC_18 MFCC_19 MFCC_20 MFCC_21 MFCC_22 MFCC_23 MFCC_24      C0
------------------------------------------ Samples: 1->10 ------------------------------------------
1:     10.569   1.278   0.431   4.016   8.521   4.258   4.797   5.464   5.613  13.729  12.879   8.507   5.749
        1.411  -0.613   4.764   3.707   2.282   0.681   0.370  -0.287  -0.164   0.082   0.189 -11.007
2:     -9.803   1.523   1.282   2.015   2.354  -1.354  -1.806   4.210   6.223  10.177  12.746   6.790  12.523
        4.402   1.100   3.311   5.538   4.041   0.863  -0.448  -0.637  -0.151   0.115   0.217 -10.553
3:     -8.389   4.063   1.433  -0.886   6.632   3.553   0.658   4.024  11.538   9.536   5.983   3.404   6.929
        6.592  -0.828   5.027   3.273   2.370   2.137   0.187  -0.143  -0.090   0.218   1.100 -10.165
4:     10.018   1.566   0.562   5.822   9.344   9.700   5.894   0.219   4.924   8.456   7.173   4.392   4.447
        5.493   1.456   2.707   5.651   2.175   0.822  -0.380  -1.024  -0.160   0.264   0.571 -10.696
5:     10.677  -0.530   0.712   6.200  10.423   8.100   6.445   6.888   6.514   3.029  10.189   0.854   2.340
        2.656   3.793   2.097   2.232   2.843   1.003  -0.548   0.027  -0.140   0.264   0.091 -10.372
6:     -2.952   4.933   3.031  -5.009  -4.791  -0.839  -2.856  -9.659   1.783   2.614  -0.232  -0.677   5.351
        7.775   8.369   4.350   3.907  -0.852  -1.979  -0.288   0.343   0.019  -0.082   0.227  -4.270
7:      0.157  -1.229  11.467 -16.695  -5.579  -1.880   2.477  -9.372   0.639   9.690  -0.757   5.394   0.602
        5.054   1.343   1.866   3.614  -0.801  -1.387  -0.305   0.791  -0.235   0.166  -0.221   8.785
8:     -1.088  -3.511   9.081 -21.563 -12.523  -6.775  -3.475 -10.731   1.113  10.879  -1.886   2.178  -1.512
        1.697  -5.160   3.008   1.322   1.298   1.609   2.771   1.378  -0.003  -0.103   0.412  10.188
9:      0.543  -1.422  10.264 -20.262 -18.129  -6.392  -6.135  -8.661   1.400   8.246  -0.725   1.036  -2.486
       -1.123  -6.312   5.477   0.348   1.671   2.395   3.050   1.423   0.137  -0.106   1.107   8.808
10:    0.766   0.535  14.470 -17.264 -13.767  -2.977  -2.117  -4.429   7.256  11.655   3.299   6.038   0.996
        1.657  -2.460  -4.050   1.888   3.059   4.457   3.690   1.277   0.258  -0.180   0.808  10.789
------------------------------------------- END -------------------------------------------
preeti@ubuntu:/media/preeti/Elements/gender/fr_gmodels$
```

3. Text Dependent Gender Recognition (TD-AGR)

The methodology is simple in text dependent voice based gender recognition. Template matching experiments were conducted for chosen Hindi vowel and nasal utterances by comparing the test utterance template of gender sensitive MFCC vector with that of the reference one. Figure 2 shows the block diagram of TD-AGR.

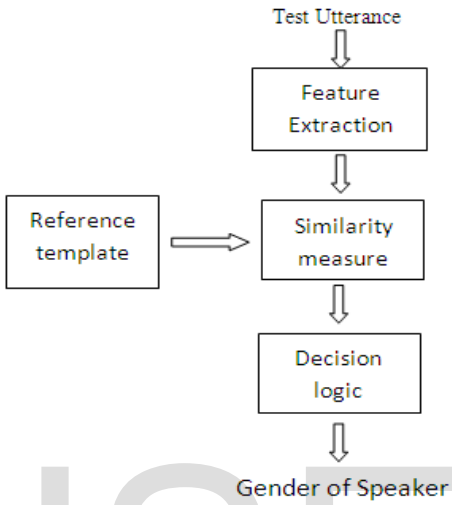


Figure 2 Voice based Text Dependent Gender Recognition System

3.1.Database

The 10 Hindi vowels (10 letters shown in table 2.a) as uttered by 10 males and 10 females were recorded. Each vowel was uttered 10 times by every speaker. The total number of vowel utterances used in training is 2000.

The 5 Hindi nasals (five letters shown in table 2.b) as uttered by 5 males and 5 females were recorded. Each vowel was uttered 10 times by every speaker. The total number of nasal utterances used in training is 250.

The said 10 Hindi vowels as uttered by 17 males and 17 females were recorded to be used as test data. Similarly the five Hindi nasals as uttered by 17 males and 17 females were recorded to be used as test data

All of the above said data were recorded using good quality microphone under office noise condition. The Wave-surfer software was used for recording. All the speakers are natives of the Hindi heart-land of India, educated and in the age group of 18 to 50.

3.2.Experimental Procedure

A set of 40 Mel-Frequency Cepstral Coefficients (C1- C40) were extracted for each utterance, 10 vowels and 5nasals. The mean MFCC vector is calculated by averaging the corresponding MFC coefficients of all male utterances of a particular phoneme, and those of female utterances of the phoneme separately. The result is the Mean MFCC vector for male and the Mean MFCC vector for female for a particular phoneme. Data averaging should emphasize the gender information of the speaker and increase the between-to-within gender variation ratio (Sigmund et al., 2008). The individual features, C1 to C40 (the 40 MFCCs) are analyzed separately for the between-to- within gender variation ratio by studying the mean and standard deviation respectively of each gender group for a particular phoneme. Thus a reference template per gender per utterance is created with those coefficients, which had a low variation within gender and high variation between the genders.

For each test utterance, the mean vector was calculated by averaging the data over all the frames. This is the test template. Each test template is tested against the reference templates of both the genders. The Euclidean distances Dm (Distance from male reference template) and Df (Distance from male reference template) were calculated for each test utterance. The gender of the test speaker is decided using the closest match.

3.3.Effect of Feature Vector Dimension on Gender Recognition Accuracy

Performance of three different combinations of MFCC coefficients, that is, three different MFCC templates for gender recognition performance were compared. The three different reference templates contain different Feature Vector lengths of 12, 24 and 40. The method was template matching and conducted for text dependent Gender Recognition. The template of length 24, having the first 24 MFC coefficients was found to give the best gender recognition results.

3.4. Effect of Data selection on Gender Recognition Vowels and nasals

Effect of Data selection on Gender Recognition Vowels and nasals are relatively easy to identify in speech signal and their spectra contain features that reliably distinguish between genders. Nasals are of particular interest because the nasal cavities of different speakers are distinctive and are not easily modified. Seven of the vowels and three of the nasals of Hindi language show excellent gender discriminating ability as shown in tables 2.a and 2.b.

Table 2.a. Effect of Data selection - Performance of Vowels

अ	आ	इ	ई	उ	ऊ	ए	ऐ	ओ	औ
97	97	100	100	97	100	100	100	100	100

Table 2.b. Effect of Data selection - Performance of Vowels

phoneme	ङ	ञ	ण	न	म
Identification rate in %	97	97	100	100	100

This leads to the conclusion on that the recognition of the gender from the short speech segments in Hindi, with a good mix of vowels and nasals said above should show decent accuracy.

4. Text Independent Gender Recognition (TI-AGR)

Most of the speech based Information Technology (IT) applications are online in nature. Text-dependent speaker classification is out of question in such situations. Moreover, the AGR must identify the gender of the speaker from a reasonably short utterance. Hence after studying the present Text-Independent speaker classifiers and speaker recognizers, an effort is made to design and implement a TI AGR that is suitable for online speech applications.

Most of the speaker classifiers proposed recently use the artificial neural network (ANN), Gaussian mixture model (GMM), hidden Markov model (HMM) or support vector machine (SVM) for classification. A GMM based classifier is used here.

It is shown in (Jian-wei et al., 2009) work that MFCC components from c4 to c16 perform best English voice database while MFCC components from c4 to c18 are suitable for Chinese voice database in speaker recognition task. This work is done on a Hindi database with varying continuous speech Hindi utterances collected by the authors.

The computational complexity and the time required for identifying the gender of a speaker is a function of feature vector dimensionality and the number of mixture components of GMM classifier. In this work, we focus on improving the performance of gender classification and decreasing its time complexity by reducing the dimensionality of the MFCC feature vector through effective feature selection methods and also by optimizing the number of Gaussian mixtures. With this goal, Text-Independent AGR was researched and the findings presented here.

4.1. Database

The database used in this experiment was collected by the authors for the purpose of different experiments on Automatic speech and speaker recognition. The database used for training the text-independent automatic gender recognizer consists of the Hindi speech data collected from 30 speakers. The training database used in this study consists of 1836 sentence utterances in Hindi spoken by 18 males and 16 females. The text corpus includes 43 distinct Hindi sentences.

The test database consists of two sets. The first set consists of 35 speakers, each of the test-speaker speaking a set of utterances (about 20 Hindi sentences), 20 of them are female and 15 male. This is for the utterance based gender test used for checking the text independence of the gender recognizer.

The second set for test consists of 69 speakers each uttering a Hindi sentence. Each utterance lasts for about one to two seconds. All of the utterances were recorded under office noise conditions using good quality microphones. All the speakers are of the age group 18 to 50. The Wave-surfer software was used for recording.

4.2. MFCC Feature Relevance Analysis for Gender Discrimination

The objective of feature relevance analysis is to identify those MFCC coefficients that are gender sensitive. F-Ratio is widely used as the figure of merit for feature selection in speaker recognition applications. It selects the features which maximize the scatter between classes.

4.2.1. F-Ratio based Dimension Reduction

The feature selection can be done in a number of ways. The F-Ratio has been widely used as the figure of merit for feature selection in speaker recognition applications. It is defined as the ratio of the between-class variance and the within-class variance. In the context of features selection for pattern classification, this ratio selects the features which maximize the scatter between classes. F-Ratio is given by

$$J_{fisher} = \frac{between-\ class\ variation}{within-\ class\ variation} \quad (6)$$

$$between-\ class\ variation = \sum_i n_i(\mu_i - \mu)^2/(K-1) \quad (7)$$

$$within-\ class\ variation = \sum_{ij}(\mu_{ij} - \mu_i)^2/(N-K) \quad (8)$$

$n_i = number\ of\ observations$

$\mu_i = sample\ mean\ in\ the\ i^{th}\ group$

$\mu = overall\ mean\ of\ the\ data$

$\mu_{ij} = j^{th}\ observation\ of\ the\ i^{th} group\ out\ of\ the\ K\ groups$

$N = overall\ sample\ size$

The F-distribution has K-1 and N-K degrees of freedom under null hypothesis.

This work aims to choose MFCC coefficients that are effective in gender discrimination between the coefficients using F-Ratio and correlation distance. As we could see, a higher F-Ratio in this

case means better gender discrimination. Figure 3 shows F-ratio values of the different MFCC coefficients.

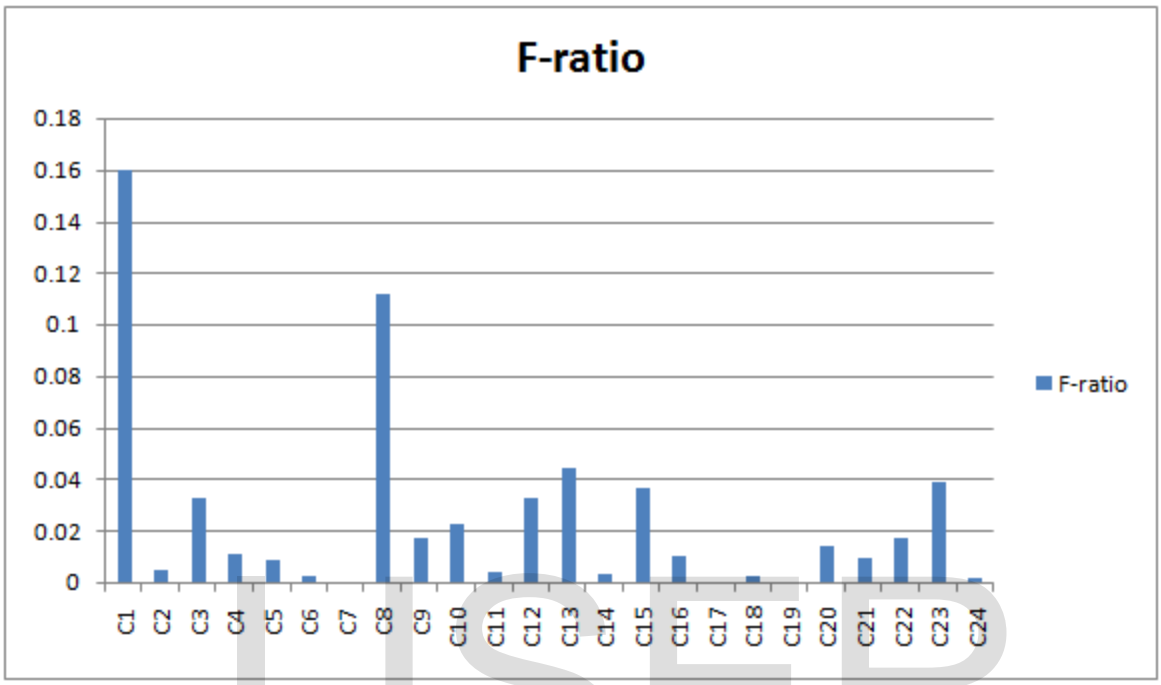


Figure 3 F-ratio values of different MFCC coefficients

The higher the value of F-ratio, the better is the gender discriminating ability of the coefficient.

4.2.2. Dimension Reduced Gender Discriminative MFCC (GD- MFCC)

The Dimension Reduced Gender Discriminative MFCC (GD- MFCC) was generated from MFCC features MFCC_0_Z (C_0, C_1, C_2, ……, C_{24}) as follows. An input feature-transform was deduced from the F-ratio based analysis to eliminate the coefficients that are not gender discriminative. The transformation matrix was such formed as to retain only the coefficients [C_1, C_3, C_4, C_5, C_8, C_9, C_{10}, C_{12}, C_{13}, C_{15}, C_{16}, C_{20}, C_{21}, C_{22}, C_{23}]. Using HTK to implement the input transform, the extracted feature vector was transformed to reduce the dimension to15. This Dimension Reduced MFCC feature is hereafter referred to as the Gender Discriminative MFCC.

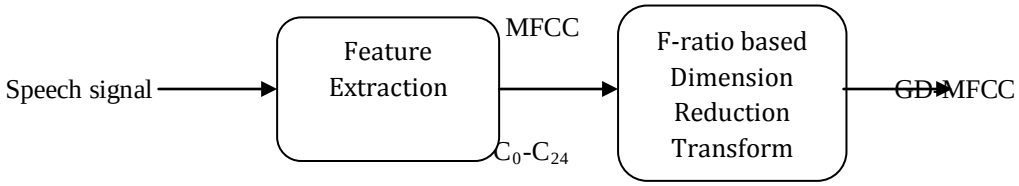


Figure 4 Extraction process of GD_MFCC

GD- MFCC = [C_1, C_3, C_4, C_5, C_8, C_9, C_{10}, C_{12}, C_{13}, C_{15}, C_{16}, C_{20}, C_{21}, C_{22}, C_{23}]

The dimension reduced GD-MFCC was used to train separate GMM models for both genders.

4.2.3. Gender Discriminative Weighted MFCC (GDW- MFCC)

Gender Discriminative Weighted MFCC (GDW- MFCC) was generated from MFCC features MFCC_0_Z (C_0, C_1, C_2, ……, C_{24}) in a manner similar to GD- MFCC except that the feature transformation matrix is different here. The input feature-transform is a weighted version of the previous one that was deduced from the F-ratio based analysis. In addition to eliminating the coefficients that are not gender discriminative, each selected MFCC coefficient is assigned a weight according to the degree of gender discrimination ability shown by it as evident in the F-ratio based analysis. HTK was used to implement the input transform. The MFCC feature vector was transformed to obtain the GDW-MFCC.

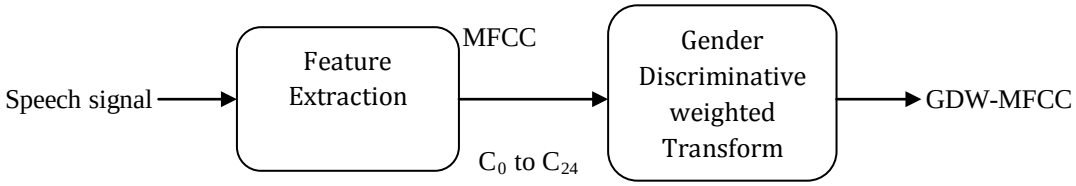
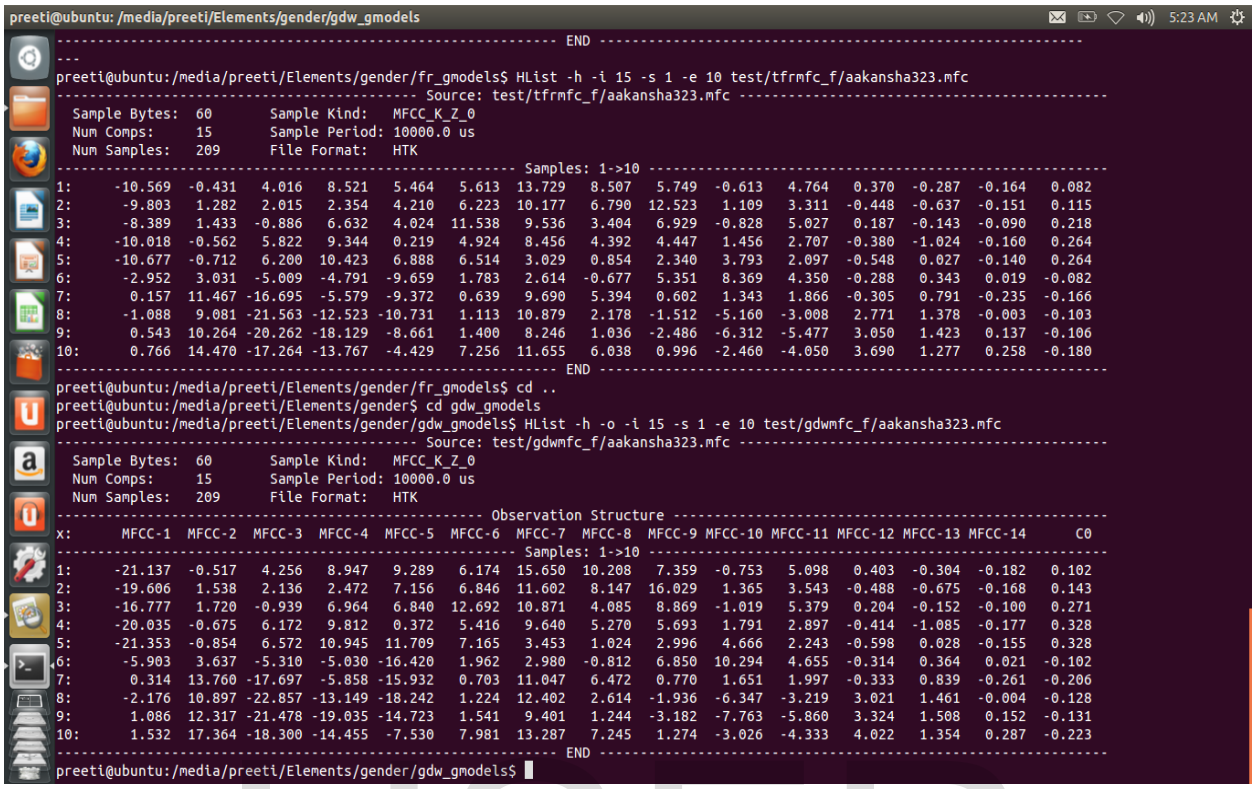


Figure 5 Extraction process of GDW_MFCC

The extracted GDW-MFCC features were also used to train separate GMM models for genders. A screen shot of the extracted feature sets of GD-MFCC and GDW-MFCC are shown in table 3.

Table 3 Screen shot of GD_MFCC and GDW_MFCC extracted using HTK



4.3. The GMM Classifier

A Gaussian mixture model (GMM) is the weighted sum of a number of Gaussian probability density functions (pdfs) where the weights are determined by the distribution. The GMM is chosen for the classification problem for the following reasons. GMM can compactly represent a classification problem since the information is embedded in the Gaussian parameters, namely the mean and the covariance matrix. GMMs are also fast in training and classification. Here GMMS are used for gender classification by training them to represent the distribution of feature vectors of gender specific speech. Then, during classification, a decision is taken for each test utterance by computing the maximum likelihood.

A GMM which is a combination of M Gaussian laws is given by the equation

$$p(x_t|\lambda_s) = \sum_{i=1}^{M} p_i b_i(x(t))$$

M is the number of component densities, x_t is the observed data i.e. the sequence of feature vectors and $x(t)$ a feature vector of dimension D. p_i are the mixture weights for i = 1, .., M and bi(s) is the Gaussian probability distribution function (PDF) associated with the i^{th} mixture component and is given by

$$b_i(x_t) = \frac{1}{2\pi^{D/2}|\sum_i^s|^{1/2}} e^{(-\frac{1}{2})(x-\mu_i)^T \sum_i^{s-1}(x-\mu_i)}$$

Here $\mu_i = mean\ vector$ and $\sum_i^s = covariance\ matrix$ of the i[th] mixture component.

The mixture weights are such that, $\sum_{i=1}^{M} p_i = 1$.

Each class, here gender, is collectively represented by a GMM given by $\lambda_s = \{\mu_i, \sum_i, p_i\}$ where $\mu_i, \sum_i, p_i$ represent the mean, covariance and the mixture weight of the i[th] mixture component.

The parameters of a GMM model can be estimated using maximum likelihood (ML) estimation. The main objective of the ML estimation is to derive the optimum model parameters that can maximize the likelihood of GMM. As direct maximization using ML estimation is not possible, a special case of ML estimation known as Expectation- Maximization (EM) algorithm () is used to extract the model parameters.

The EM algorithm begins with an initial model λ and tends to estimate a new model λ' such that

$$p(x_t|\lambda') \geq p(x_t|\lambda)$$

This is an iterative process where the new model is considered to be the initial model in the next iteration and the entire process is repeated until a certain convergence threshold is obtained.

4.3.1. Implementation of the classifier

The implementation of the classifier is done in the HTK environment as follows. The Gaussian Mixture Model was implemented as a single emitting state HMM. Initially, each state was represented by a single Gaussian and then by using the mixture splitting technique available in HTK-3.4, the number of mixtures increased and the model re-estimated each time using the HERest tool of the HTK. The performance of the model was tested for different number of mixtures and the optimum number of mixture components was estimated experimentally. HVite tool of HTK was used for recognizing the gender of the speaker.

Figures 6 and 7 show the screen-shots of the results of the classifier using GD-MFCC feature and GDW-MFCC feature respectively.

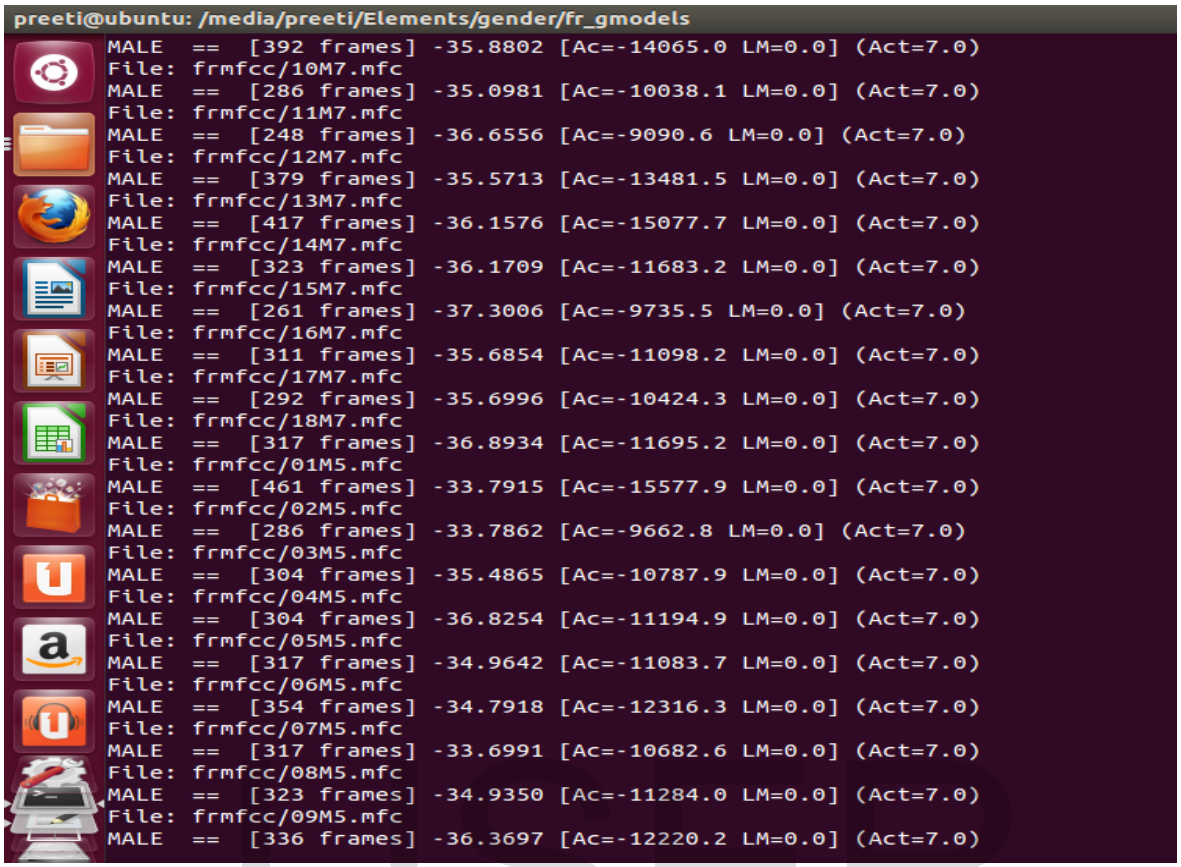


Figure 6 Screen shot of the result given by the TI-AGR using GD_MFCC for a male test speaker

5. Experimental Results and Discussion

The two parameters that can be varied to reduce computational complexity and time complexity while trying to improve gender identification rate are

- Dimension of the Feature Vector
- Order-of-GMM

The gender of the speaker was determined by GMM classification using three sets of feature vectors, namely MFCCs C_0 to C_{24} as features, GD-MFCC features and GDW-MFCC.

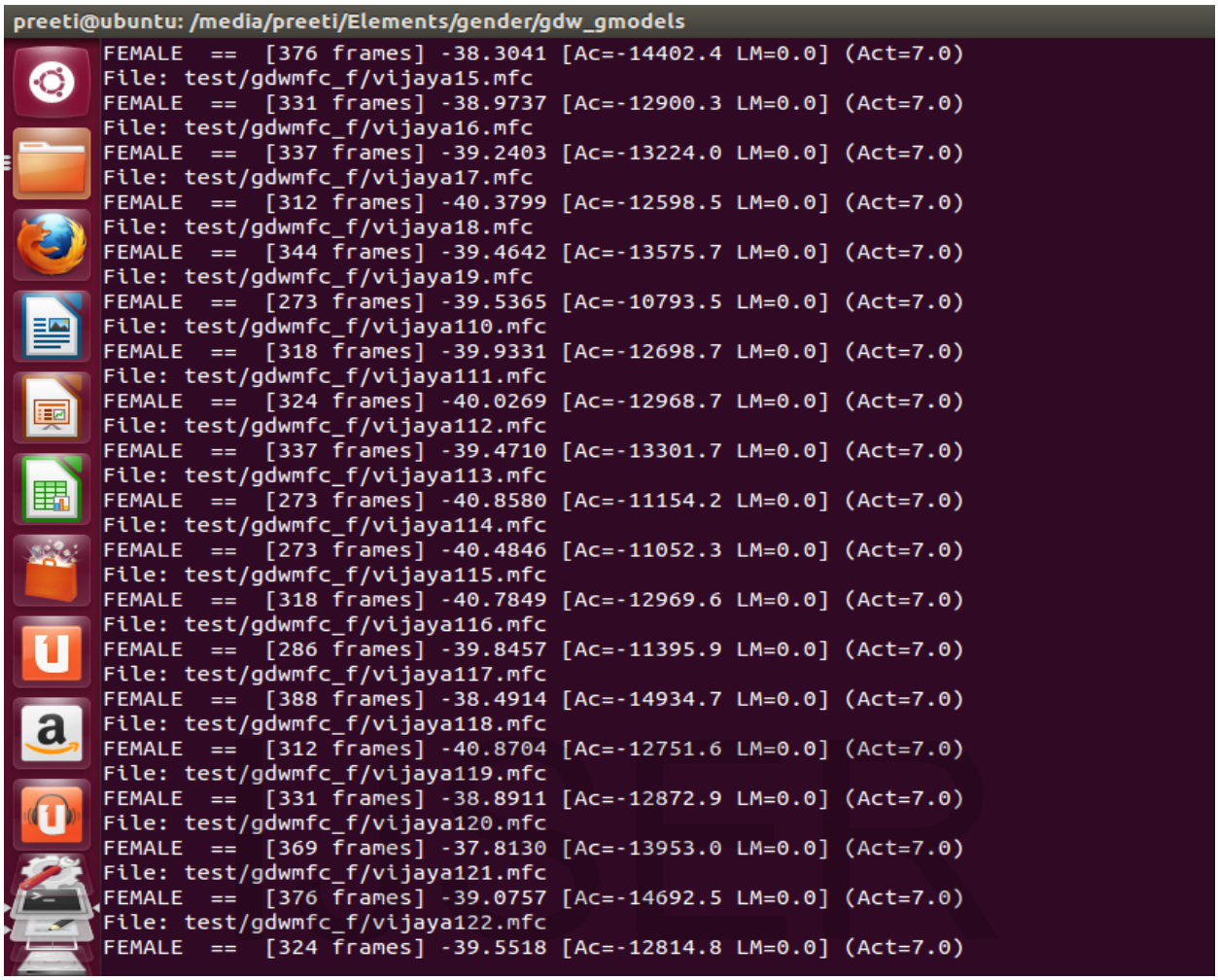


Figure 7 Screen shot of the result given by TI-AGR using GDW_MFCC for a female test speaker

The performance of the classifiers was evaluated on the following three counts and a comparative study is presented.

- Gender Recognition Accuracy
- Computational Complexity
- Time Complexity

Two methods of evaluation were used. In the first method, each of the test-speaker was asked to speak a set of utterances (about 20 Hindi sentences) and for each utterance the gender of the speaker was tested using the GMM classifier. The results were noted. This procedure was repeated for 35 different speakers, 15 males and 20 females for three feature sets, in each case for

varying the number of GMM mixtures stepwise up to 8. Tables 4, 5 and 6 show the performance of the GMM based gender classifier for all the three feature sets.

Table 4 Performance of the TI-AGR using MFCC-25 features in the utterance based test

Speakers	No. of test utterances	No. of correct gender recognitions				
		GMM Mix=1	GMM Mix=3	GMM Mix=4	GMM Mix=6	GMM Mix=8
Female	373	357	366	371	373	373
Male	297	271	289	289	282	278
Total	670	625	659	661	655	651

Beyond 6 Gaussian mixtures, it was noted that the performance of the TI-AGR deteriorated or saturated. Overall the best performance is obtained when the number of Gaussian mixtures is 6.

Table 5 Performance of the TI-AGR using GD-MFCC features in the utterance based test

Speakers	No. of test utterances	No. of correct gender recognitions				
		GMM Mix=1	GMM Mix=2	GMM Mix=4	GMM Mix=6	GMM Mix=8
Female	373	363	371	372	373	373
Male	297	286	273	286	291	284
Total	670	649	644	658	664	657

Table 6 Performance of the TI-AGR using GDW-MFCC features in the utterance based test

Speakers	No. of test utterances	No. of correct gender recognitions				
		GMM Mix=1	GMM Mix=2	GMM Mix=4	GMM Mix=6	GMM Mix=8
Female	373	364	371	372	373	373
Male	297	279	272	285	289	281
Total	670	643	643	657	662	654

In the second method 69 speakers were tested for gender, fixing the number of Gaussian mixtures of the GMM classifier to be at 6. The number of Gaussian mixtures is fixed at 6 based on the experimental results from the first method of evaluation as it performs optimum for all three features mentioned above. The experiment is repeated for all the three features and the results were recorded. Table 7 shows the comparative Performance of the TI-AGR for the three feature sets.

Table 7 Comparative Performance of the TI-AGR for the three feature sets

Speakers	No of speakers tested	MFCC-25		GD-MFCC		GDW-MFCC	
		No. of corr. GR	% Rec. Accuracy	No. of corr. GR	% Rec. Accuracy	No. of corr. GR	% Rec. Accuracy
Male	33	30	90.9	31	93.9	31	93.9
Female	36	34	94.4	35	97.2	35	97.2
All	69	64	92.7	66	95.6	65	95.6

5.1. Effect of MFCC Dimensionality on Gender Recognition

The gender of the speaker was determined by GMM classification using the three sets of features namely

- MFCCs C_0 to C_{24} as features (Feature Dimension 25)
- GD-MFCC features (Feature Dimension 15)
- GDW-MFCC (Feature Dimension 15)

The Gender Recognition Accuracy is defined as the ratio of the number of speakers whose gender is identified correctly to the total number of speakers tested. It was observed that by discarding a number of coefficients that are not gender discriminative the performance of gender classification increases significantly. Figure 8 shows the Gender Recognition Performance of the Text Independent AGR while using the three different feature-sets namely, MFCC_25, GD-MFCC and GDW-MFCC.

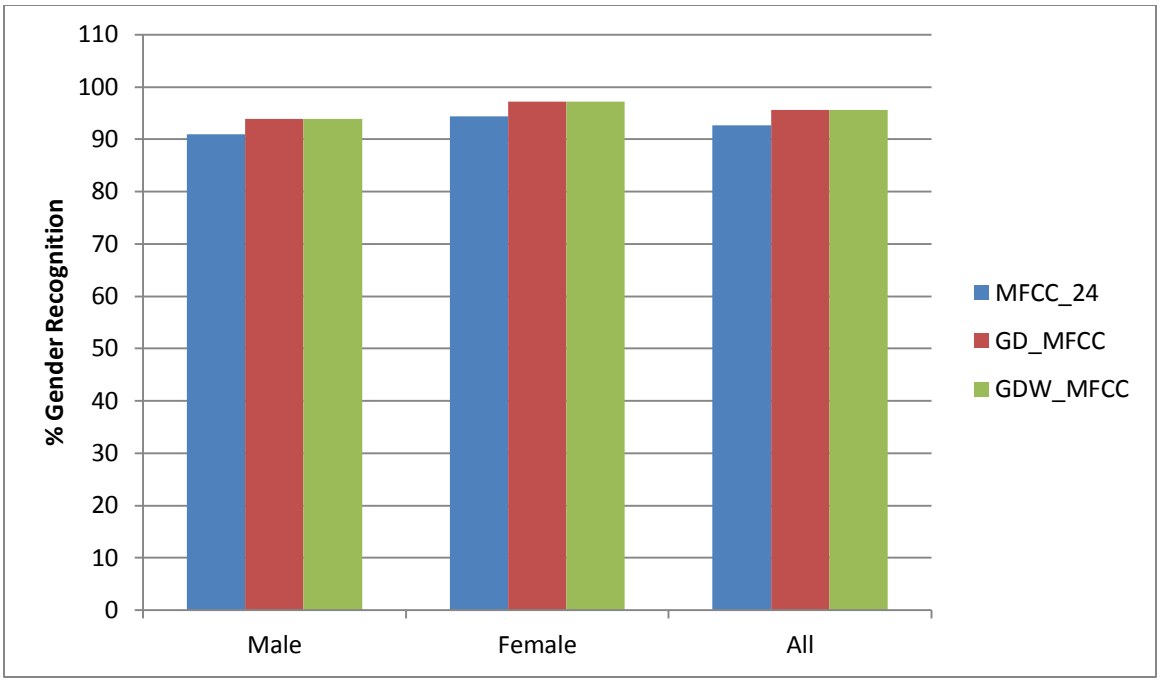


Figure 8 Comparative Gender Recognition Performance of the three TI_AGRs

The dimension reduction has lead to an absolute increase of about 2.9 percent in recognition accuracy. The complexity is reduced by reducing the feature dimension from 25 to 15, a reduction of dimension by 10.

5.2. Effect of Number of mixture components on Gender Recognition Accuracy

The order of GMM is another parameter that can be varied to achieve good trade-off between gender recognition performance and computational and time complexity. Figure 9 shows the Performance of the GMM based Gender Classifier using MFCC_25 features.

The recognition performance for female speakers increases as the number of Gaussian mixture components increases up to 8 whereas the performance for male speakers decreases after the order of GMM goes above 4. Overall the performance is best when the number of Gaussian mixture components is 4 and beyond that the gender recognition performance declines. Hence only 4 Gaussian mixtures are required to obtain maximum gender recognition accuracy when the 24 dimension MFCC vector is used as features. That shows a considerable reduction in complexity when the performance of the text independent AGR goes up to about 95 percent.

The order of GMM classifier was varied for the other two features GD-MFCC and GDW-MFCC, from 1, 2, 4 up to 6 and their gender recognition performance were noted. Figures 10 and 11 show Recognition Performance of the GMM based Gender Classifier when using Dimension Reduced GD-MFCC features and the weighted GDW-MFCC for varying number of Gaussian mixtures.

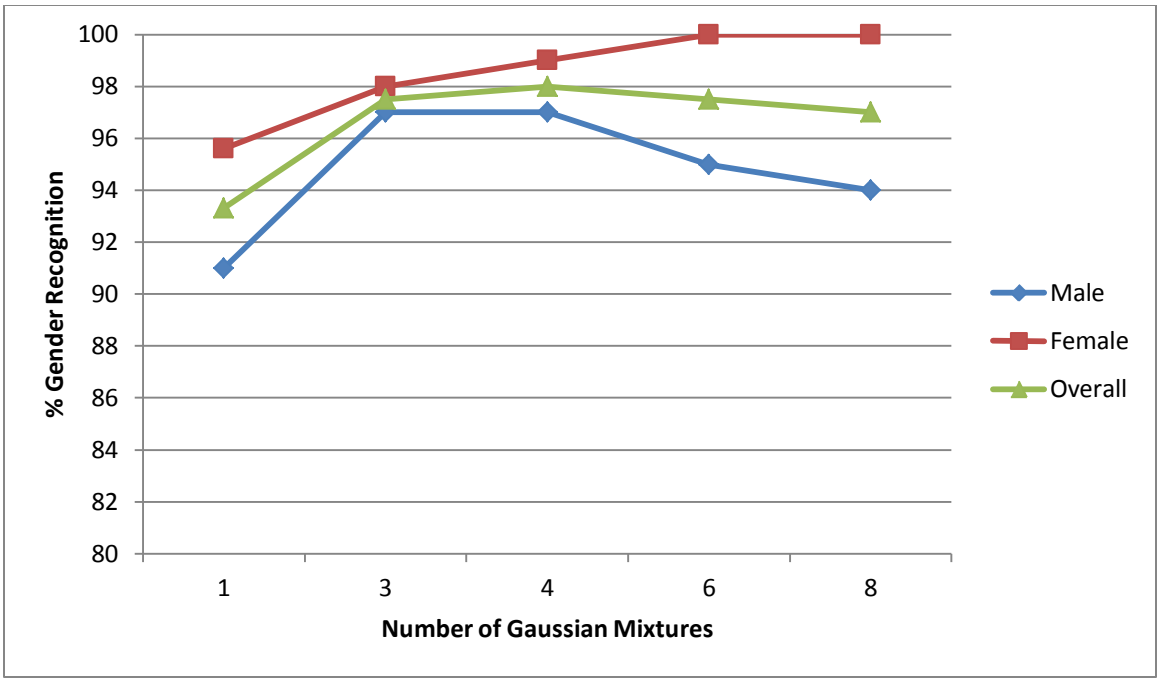


Figure 9 Recognition Accuracy Vs. No. of Gaussian mixtures for the TI-AGR (MFCC-25)

As we see in the graphs shown by figures 10 and 11, the optimum order of GMM is 6 for both the GD-MFCC and GDW-MFCC feature sets. It is noteworthy that both the features are of the same feature vector dimension too. Both are 15 dimensional MFCC features. In both cases, female recognition is better than that of the male as usual. As the order of GMM is increased the gap between the female and male recognition rates reduces as we can observe.

Again we see the Gender Recognition Accuracy increasing and complexity decreasing as the feature dimension reduces by 10 and the number of Gaussian components required to model gender going up marginally. That is a significant improvement over previous AGRs.

Figure 12 shows the comparative performance of the two novel features proposed in this study. It was found that GD-MFCC performed better than GDW-MFCC.

The different weights given to each dimension of the feature vector has not added any benefit to dimension reduction but has deteriorated the gender recognition performance. The dimension reduction has led to the best recognition performance.

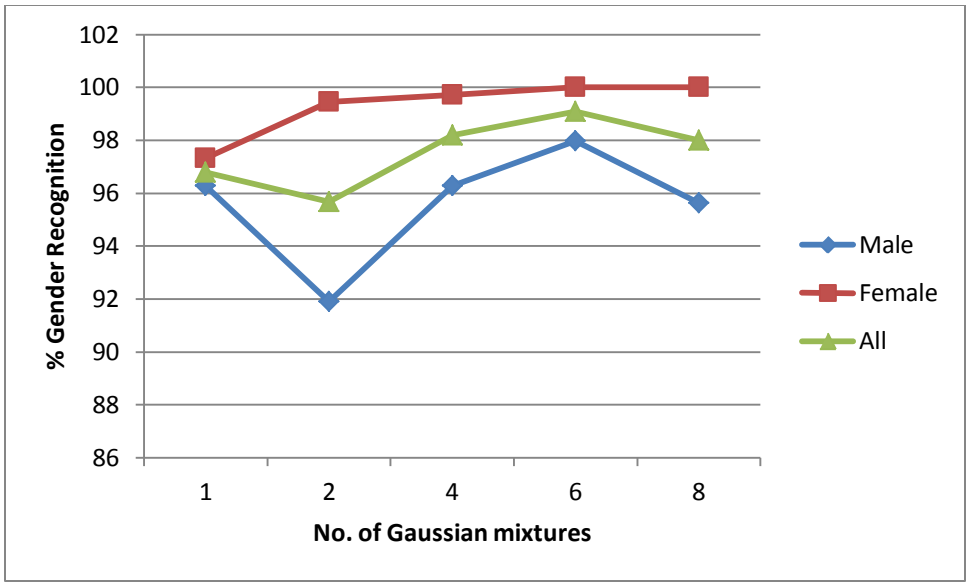


Figure 10 Recognition Performance of the TI-AGR using GD-MFCC

Reducing the dimension of MFCC feature vector by 10 has increased the recognition accuracy by 2.9% absolute. GDW- MFCC however does not improve upon GD-MFCC. This means allotting weights the coefficients has not helped to improve the gender recognition accuracy.

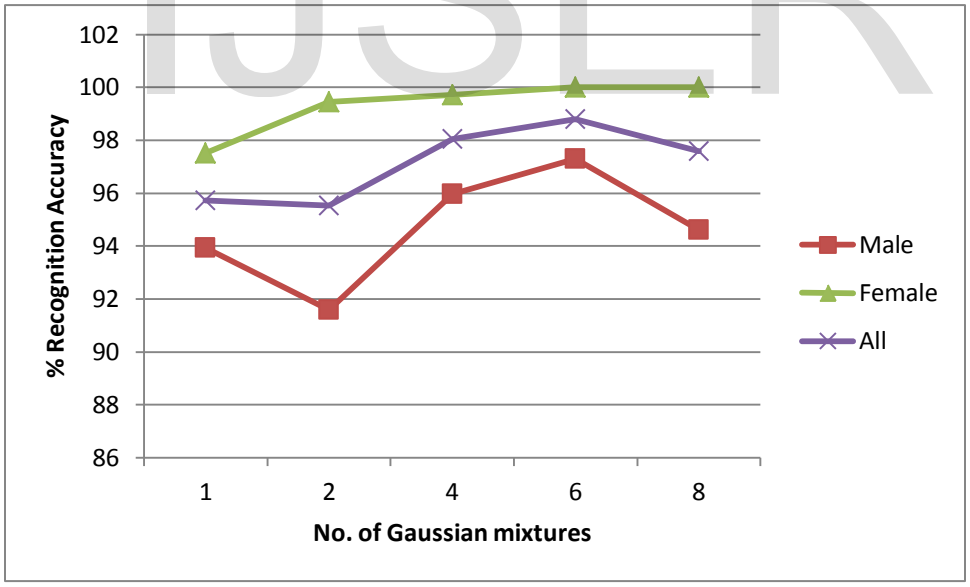


Figure 11 Recognition Performance of the GMM based Gender Classifier using GDW-MFCC features

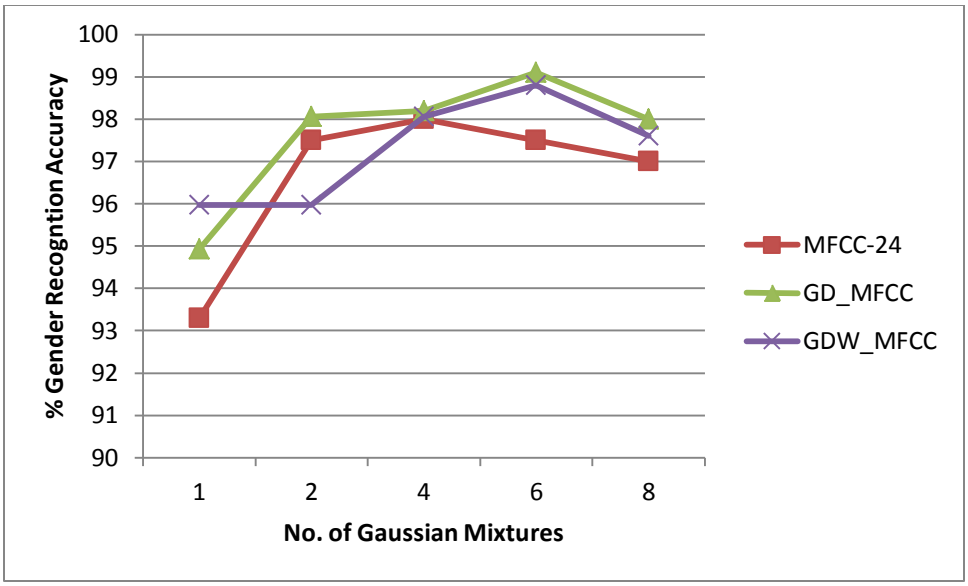


Figure 12 Comparative Gender Recognition Performance of the two novel features.

The results show that both the GD-MFCC and GDW-MFCC feature sets are able to deliver a recognition accuracy of more than 95 %. There is no statistically significant difference in the performance of these two methods though both methods perform excellently. GD-MFCC performs slightly better than GDW-MFCC when evaluated based on the number of utterances for which the gender of the speaker is correctly detected.

However the performance was same when evaluated on the basis of the number speakers whose gender is correctly detected. The GD-MFCC and GDW-MFCC feature sets improve upon MFCC-25 by an absolute 2.9 %. It must also be noted that all the three feature sets perform significantly better for female speakers than male speakers.

The increase in the recognition accuracy when the number of Gaussian mixtures was increased from 1 to 6 was found to be statistically significant at $p < 0.05$.

From the above discussion, the following is clear.

- The decrease in Feature Vector Dimension from 24 to 15 has increased the Recognition Accuracy from by 2.9%
- The recognition performance is the maximum, for the MFCC feature with dimension 25 when the number of Gaussian mixtures is 4, while it is the maximum, for the GD-MFCC feature with dimension 15 when the number of Gaussian mixtures is 6.
- The computational complexity of the proposed novel, text-independent Gender Recognition indeed decreases while the Recognition Accuracy increases considerably
- This Gender Recognizer has been tested and found to work real time and independent of the text of the speech.

6. Conclusion

Using statistical experimental methods to select important components of a feature vector requires a large amount of calculation and processing time. The F-ratio was used as the discriminative measure for gender recognition ability of different coefficients of MFCC. An effort was made to reduce computational complexity and improve recognition accuracy when the speech input is text independent. A significant improvement in recognition accuracy is reported. Dimension reduction based on the F-ratio analysis has yielded good results.

The Performance of the system for different Mixture components shows that the optimal mixture components are 8 for speech signals sampled at 16 kHz. The recognition performance depends on the training speech length selected for training to capture the speaker-specific excitation information. Larger the training length, the better is the performance.

While the text dependent method has resulted in more than 95 percentage recognition accuracy for speech duration of about 0.5 s (vowel and nasal utterance). The text independent AGR also performs excellently delivering a gender recognition accuracy of more than 95 % for an utterance of speech duration of 1 to 2s.

Reference

1. M. Benzeguiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, C. Wellekens (2006). *Automatic Speech Recognition and Intrinsic Speech Variation.* ICASSP- 2006.

2. Biswajit Das, Sandipan Mandal, Pabitra Mitra, Anupam Basu (2013). *Aging speech recognition with speaker adaptation techniques: Study on medium vocabulary continuous Bengali speech.* Pattern Recognition Letters 34 (2013) 335–343. www.elsevier.com/locate/patrec

3. Tobias Herbig, Franz Gerl, Wolfgang Minker (2012) *Self-learning speaker identification for enhanced speech recognition.* Computer Speech and Language 26 (2012) 210–227. www.sciencedirect.com

4. D. G. Childers, K. Wu, and D. M. Hicks (1987). *Factors in voice quality: acoustic features related to gender.* In Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, volume 1, pages 293–296.

5. H. Harb and L. Chen (2006). *Gender Identification using a general Audio Classifier. Multimedia Tools and Applications,* volume 34, No. 3, 375-395.

6. Vipperla, R., Renals, S., Frankel, J., 2010. *Ageing Voices: The Effect of Changes in Voice Parameters on ASR Performance.* EURASIP Journal on Audio, Speech, and Music Process. 2010, 5, p. 10.

7. W. Abdulla and N. Kasabov (2001) . *Improving speech recognition performance through gender separation.* In Proc. Int. Conf. Artificial Neural Networks and Expert Systems (ANNES), pages 218–222, Dunedin, New Zealand.

8. Alireza Zourmand, Hua-Nong Ting, and Seyed Mostafa Mirhassani (2013). *Gender Classification in Children Based on Speech Characteristics: Using Fundamental and Formant Frequencies of Malay Vowels*. Journal of Voice, Vol. 27, No. 2, 2013

9. Kunjithapatham Meena, Kulumani Subramaniam, and Muthusamy Gomathy (2013). *Gender Classification in Speech Recognition using Fuzzy Logic and Neural Network.* The International Arab Journal of Information Technology, Vol. 10, No. 5, September 2013

10. Marianna Pronobis and Mathew Magimai Doss. *Analysis of F0 and cepstral features for robust automatic gender recognition*. Idiap Research Report Idiap-RR-30-2009, 2009. https://docs.google.com

11. M. Feld, F. Burkhardt and C. Muller(2010). *Automatic Speaker Age and Gender Recognition in the Car for Tailoring Dialog and Mobile Services.* INTERSPEECH- 2010, 2834-2837

12. Vijender Sharma and Rakesh Garg, (2013). *Gender and Speaker Recognition Using MFCC and DTW. International Journal of Advanced Research in Computer Science and Software Engineering.* Volume 3, Issue 8, August 2013

13. R. Rajeshwara Rao and A. Prasad (2011). *Glottal Excitation Feature based Gender Identification System using Ergodic HMM.* Int. J. of Computer Applications (0975 – 8887). Volume 17, No.3, pages 0975 – 8887.

14. F. Metze, J. Ajmera, R. Englert, U. Bub, F. Burkhardt, J. Stegmann, C. Muller, R. Huber, B.Andrassy, J. G. Bauer, and B. Little (2007). *Comparison of four approaches to age and gender recognition for telephone applications.* In Proc. 2007 IEEE Int. Conf. Acoustics, Speech and Signal Processing, volume 4, pages 1089–1092. Honolulu

15. M. H. Sedaaghi (2008). *Gender Classification in Emotional Speech. Speech Recognition, Technologies and Applications.* pp. 550, www.intechweb.org

16. Milan Sigmund (2008). *Gender Distinction using Short Segments of Speech Signal*. Int. J. of Computer Science and Network Security, Vol.8, No.10.

17. D.Shakina Deiv, Gaurav and Mahua Bhattacharya (2011). Automatic Gender Identification for Hindi Speech Recognition. *International Journal of Computer Applications* 31(5):1-8. Published by Foundation of Computer Science, New York, USA.

18. N M Ramaligeswararao, V Sailaja and K. Srinivasa Rao (2011). *Text Independent Speaker Identification using Integrated Independent Component Analysis with Generalized Gaussian Mixture Model.* (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 12, 2011

19. S.Young, G.Everman, T.Hain, D.Kershaw, G.Moore, J.Odell, D.Ollason, D.Povey, V.Valtchev and P.Woodland (2009). *The HTK Book (for HTK Version 3.4, 2009)*. Cambridge University Engineering Department.

20. ZHU Jian-wei, SUN Shui-fa, DAN Zhi-ping and LEI Bang-jun.(2009). *MFCC Extraction Based on F-Ratio and correlated distance criterion in speaker Recognition*. International Conference on Multimedia Information Networking and Security.