# Automatic Gender Identification for Hindi Speech Recognition

D.Shakina Deiv
ABV- Indian Institute of Information Technology & Management
Morena Link Road, Gwalior 474010, India

Gaurav
ABV- Indian Institute of Information Technology & Management
Morena Link Road, Gwalior 474010, India

Mahua Bhattacharya
ABV- Indian Institute of Information Technology & Management
Morena Link Road, Gwalior 474010, India

## ABSTRACT
This paper presents the preliminary work done towards the development of a Gender Recognition System that can be incorporated into the Hindi Automatic Speech Recognition (ASR) System. Gender Recognition (GR) can help in the development of speaker-independent speech recognition systems. This paper presents a general approach to identifying feature vectors that effectively distinguish gender of a speaker from Hindi phoneme utterances. 10 vowels and 5 nasals of the Hindi language were studied for their effectiveness in identifying gender of the speaker. All the 10 vowel Phonemes performed well, while इ, ई, ऊ, ए, ऐ, ओ and औ showed excellent gender distinction performance. All five nasals ङ, ञ, ण, न and म which were tested, showed a recognition accuracy of almost 100%. The Mel Frequency Cepstral Coefficients (MFCC) are widely used in ASR. The choice of MFCC as features in Gender Recognition will avoid additional computation. The effect of the MFCC feature vector dimension on the GR accuracy was studied and the findings presented.

## General Terms
Automatic speech recognition in Hindi

## Keywords
Gender Recognition, Mel-Frequency Cepstral Coefficients, Hindi Phonemes

## 1. INTRODUCTION
To make Information Technology relevant to rural India, voice access to a variety of computer-based services is imperative. The lack of effective Hindi speech recognition system and its local relevance has hindered development of speech applications in our daily life activities like mobile applications, weather forecasting, agriculture, healthcare etc [1]. A major concern in the current research of the speech interface is the robustness of Automatic Speech recognition (ASR) system. Gender recognition is an important step in speech and speaker recognition, as it not only increases the robustness and accuracy of the systems but also can expectedly do so, without much increase in the computational complexity. Automatic gender recognition technique can assist the development of speaker-independent speech recognition systems[2], help 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[3]. Gender classification 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. Therefore, automatic gender detection can be a tool in a content-based multimedia indexing system [4].

The approaches to automatic gender recognition can be classified 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 [5] 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.

Pitch is a very strong source of information for gender identification of adult male and female speakers [6]. 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 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 ($F_0$) and cepstral features, namely LPCCs, MFCCs, PLPs are compared in [7] for robust Automatic Gender Recognition. It concludes that features provide similar performance under clean conditions, but cepstral features yield robust system in noisy condition.

In [8] a vector quantization approach for speaker recognition using MFCC and inverted MFCC for text independent speaker recognition has been proposed and evaluated. Experiments show that the method gives tremendous improvement and it can detect the correct speaker from speech sample of very short duration, even 0.5sec, which is very interesting for the present study also. In [9] 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. Combining MFCC and Pitch leads to enhancement of the Performance of the Gender Recognition as demonstrated in [10].

A glottal excitation feature based Gender Identification System using ergodic HMM in [11] demonstrates the importance of information in the excitation component of speech (pitch) for gender recognition task.

In[12], four approaches for age and gender recognition using telephone speech have been compared; *namely*, a parallel phone recognizer, a system using dynamic Bayesian networks to combine several prosodic features, a system based solely on

linear prediction analysis, and Gaussian mixture models based on MFCCs. It was reported that the parallel phone recognizer is comparable to a human listener but loses performance on short utterances. Several popular methods for gender classification have been investigated in [13] by processing emotionally colored speech.

Using a set of selected *mel*-warped cepstral coefficients it is shown in [14] and [15] that the gender of a speaker can be correctly identified with a performance of about 93% from clean speech. It is interesting to note that the time needed for identification decreases significantly for selected speech segments like vowel phonemes.

Gender Recognition (GR) can help in the development of speaker-independent speech recognition systems. Phoneme based features for effective gender identification have been analyzed in English [14] [15]. In their analysis of phoneme based features for gender identification with neural networks [16] have shown that vowels extracted from a few sentences give valuable gender information. They have used MFCC coefficients.

Hindi belongs to the Indo Aryan family of languages and is written in the Devanagari script. There are 11 vowels and 35 consonants in standard Hindi. Hindi is mostly phonetic in nature and therefore usually has a one to one mapping between the orthography and their pronunciation. They possess a large no of phonemes like retroflex and aspirated stops which are absent in English and other European Languages.

In this paper, a general approach to identifying feature vectors that effectively distinguish gender of a speaker from Hindi phoneme utterances is presented. 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. The application of Mel Frequency Cepstral Coefficients for Automatic gender recognition is analyzed under the following conditions.

- The effect of data selection (selected phonemes of the Hindi Language - ten Hindi vowels अ, आ, इ, ई, उ, ऊ, ए, ऐ, ओ, औ five nasals ङ, ञ, ण, न and म)
- The effect of feature vector dimension on the Automatic Gender Recognition (AGR) accuracy.

## 2. THEORY BEHIND THE GENDER RECOGNITION EXPERIMENTS

The proposed gender recognition system has two modules, feature extraction and feature matching. The system is shown in block diagram as in Fig.1

### 2.1 Feature Extraction

Feature extraction transforms the raw speech signal into a compact but effective representation that is more stable and discriminative than the original signal. Earlier studies on Speech features like pitch ($F_0$), cepstral features, namely Linear Prediction Cepstral Coefficients (LPCC), MFCCs, Perceptual Linear Prediction (PLP) for robust Automatic Gender Recognition have concluded that features provide similar performance under clean conditions, but cepstral features yield robust systems under noisy conditions [5]. Moreover, the feature vector used in gender recognition should preferably be the same as that of speech recognition if possible, as it avoids additional computation. Hence the MFCC vectors are chosen as the speech feature for this study.

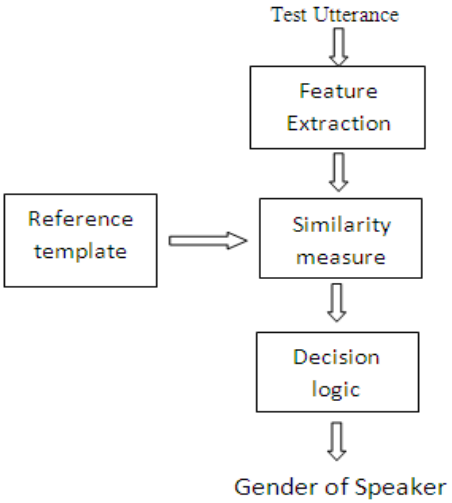


Fig.1 Voice based Gender Recognition System

### 2.2 Mel-Frequency Cepstral Coefficients

Mel-Frequency Cepstral Coefficients (MFCCs) are widely used features for automatic speech recognition systems to transform the speech waveform into a sequence of discrete acoustic vectors. 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 [17]. The MFCC technique makes use of two types of filter, namely, linearly spaced filters and logarithmically spaced filters. The Mel frequency scale has a 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.

#### *2.2.1 Pre–emphasis*

The signal is passed through a filter which emphasizes higher frequencies. This process will increase the energy of the signal at higher frequency. The Pre-emphasis of the speech signal is realized with this simple FIR filter $H(z) = 1 - a z^{-1}$ where a is from interval [0.9,1].

#### *2.2.2 Framing*

The process of segmenting the digitized speech samples into frames with a length within the range of 10 to 30 msec.

#### *2.2.3 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

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

Where N = number of samples in each frame.
Let Y[n] = Output signal
X (n) = input signal
Where N = number of samples in each frame.
The result of windowing the signal is

$$\mathbf{Y(n) = X(n) \times W(n)}$$

### *2.2.4 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 is calculated.

$$\mathbf{Y(\omega) = FFT[h(t) * x(t)] = H(\omega) * X(\omega)}$$

### *2.2.5 Mel Filter Bank Processing*

The most important step in this signal processing is mel-freqency transformation. Compensation for non-linear perception of frequency is implemented by the bank of triangular band filters with the linear distribution of frequencies along with 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.

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

where f is frequency in linear range and $f_{mel}$ the corresponding frequency in nonlinear mel-frequency range.

### *2.2.6 Discrete Cosine Transform*

The next step is to calculate the logarithm of the output of filters. Finally, the log mel spectrum is converted back to time domain. The result is called the Mel Frequency Cepstral Coefficients. The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. The mel spectrum coefficients and their logarithm are real numbers. Hence they can be converted to the time domain using the discrete cosine transform (DCT). The result is the Mel Frequency Cepstral Coefficients.

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

where n is the number of Mel-frequency cepstral coefficients and K is the number of mel-frequency band filters (filterbank channels) in the bank of filters.

## 2.3 Feature Matching

The acoustic feature vectors which will serve as the reference patterns are generated from the training (sample) datasets. The test feature vectors extracted from test data are then to be matched with the reference vectors to identify the gender of the speaker. The distance metric used for feature matching is the Euclidean distance which is simple and effective. The Euclidean distances $D_m$ and $D_f$ , distances of test feature vector from male reference and female reference respectively are calculated as follows.

$$D_m(X, M) = 0.5[(X - M)^T (X - M)]^{1/2}$$

$$D_f(X, F) = 0.5[(X - F)^T (X - F)]^{1/2}$$

where X is the test template , M is the male reference template and F is the female reference template. T denotes transpose.

# 3. EXPERIMENTAL PROCEDURE

## 3.1 Database

1. The ten Hindi vowels अ, आ, इ, ई, उ, ऊ, ए, ऐ, ओ, औ as uttered by 10 males and 10 females were recorded , each vowel being uttered 10 times by every speaker. (10 vowel phonemes x 10 times x 20 speakers = 2000 utterances)
2. The five Hindi nasals ङ, ञ, ण, न and म as uttered by 5 males and 5 females were recorded , each vowel being uttered 5 times by every speaker. (5 nasal phonemes x 5 times x 10 speakers = 250 utterances)
   The above sets of speech data was (in .wav format) used as training data.
3. The 10 Hindi vowels as uttered by 17 males and 17 females were recorded to be used as test data. Hindi nasals ङ, ञ, ण, न, म as uttered by 17 males and 17 females were recorded to be used as test data.

All the utterances were recorded using good quality microphones under office noise condition. The Wave-surfer software was used for recording. All the speakers are natives of the Hindi heartland of India. Their age group is 18 to 30.

## 3.2 Coding the data

We used the HMM Tool Kit (HTK) to parameterize the raw speech waveforms into sequences feature vectors. We use Mel Frequency Cepstral Coefficients (MFCCs), which are derived from FFT-based log spectra. Coding can be performed using the tool HCopy configured to automatically convert its input into MFCC vectors. A configuration file (config) specifies all of the conversion parameters [18]. The target parameters are to be MFCC using *C*0 as the energy component, the frame period is 10msec (HTK uses units of 100ns)

```
# Coding parameters
TARGETKIND = MFCC_0
TARGETRATE = 100000.0
SAVECOMPRESSED = T
SAVEWITHCRC = T
WINDOWSIZE = 250000.0
USEHAMMING = T
PREEMCOEF = 0.97
NUMCHANS = 40
CEPLIFTER = 22
NUMCEPS = 41
```

The target parameters are to be MFCC using *C*0 as the energy component, the frame period is 10msec (HTK uses units of 100ns), the output should be saved in compressed format, and a crc checksum is added. The FFT uses a Hamming window and the signal should have first order pre-emphasis applied using a coefficient of 0.97. The filterbank should have 40 channels and 40 MFCC coefficients will be output.

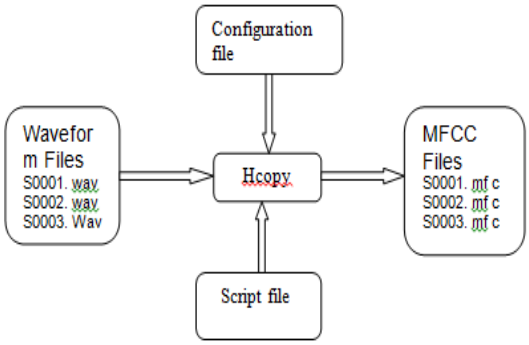


**Fig.2 Extraction of feature vector using HTK**

## 3.3 Reference Template

We have noted in the beginning that the parameters used in gender recognition should be the same as speech recognition if possible, to keep additional computation at the minimum. For this reason, MFCC feature vectors are used. ASR uses feature vectors of length 39. This number, 39, is computed from the length of the parameterized static vector (MFCC 0 = 13) plus the delta coefficients (+13) plus the acceleration coefficients (+13). As it is proved in earlier studies that the difference coefficients do not perform well in gender recognition, the delta and delta-delta coefficients (which are used in ASR) are not considered in this study. A set of 40 Mel-Frequency Cepstral Coefficients (C1- C40) are calculated 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, for example अ separately, and those of female utterances of the phoneme separately. The result is the Mean MFCC vector for male for vowel अ and the Mean MFCC vector for female for vowel अ.

Data averaging should emphasize the speaker's gender information and increase the between-to-within gender variation ratio [6]. 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. Tables 1.a and 1.b summarize the reference template1 for male and female for vowels and nasals respectively. Template1 contains 8 coefficients in the case of vowels [ C2, C4, C5, C6, C7, C8, C9, C12 ]. Template1 For nasals contain10 coefficients, given by [ C2, C4, C5, C6, C7, C8, C9, C10, C11, C12]. Tables 2.a and 2.b summarize the reference template2, for male and female for vowels and nasals respectively. Template 2 contains 18 coefficients for both vowels and nasals

**Table 1.a. Reference Template for female and male (Vowels– Template 1)**

| Mfcc | Vector | C2 | C4 | C5 | C6 | C7 | C8 | C9 | C12 |
|---|---|---|---|---|---|---|---|---|---|
| अ | Female | -3.01732 | -2.22697 | 12.34937 | -1.31137 | 6.118943 | 0.868838 | 9.333874 | 3.687245 |
| | male | -6.61076 | -7.47876 | 5.716305 | -3.15283 | 2.599154 | -1.48249 | 6.934847 | -1.4178 |
| आ | Female | -4.91045 | -3.68455 | 13.72922 | 0.08639 | 3.80616 | 0.13905 | 7.62748 | 2.28018 |
| | male | -7.3424 | -9.66287 | 7.369072 | -1.74082 | 5.472796 | -3.41779 | 5.691005 | -2.82381 |
| इ | Female | 0.364584 | -2.45413 | 8.89341 | -1.04307 | 4.549505 | 0.916575 | 8.755095 | 4.77403 |
| | male | -4.62436 | -5.03479 | 2.471889 | -4.96252 | 0.383785 | -4.14997 | 3.387061 | -0.09768 |
| ई | Female | 0.989559 | -4.96487 | 8.325724 | -1.57415 | 1.547011 | -0.48445 | 7.180009 | 3.874273 |
| | male | -3.29283 | -5.11443 | 1.294441 | -5.27009 | -0.89033 | -6.08726 | 2.463846 | -1.82052 |
| उ | Female | -0.33495 | -2.97757 | 8.164573 | -1.48726 | 7.305703 | 2.259218 | 7.286869 | 4.293372 |
| | male | -4.73775 | -5.79164 | 3.080199 | -5.5864 | 1.682439 | -2.05236 | 2.780175 | 1.134654 |
| ऊ | Female | -0.04538 | -4.98242 | 6.83402 | -2.84342 | 5.929446 | 1.841857 | 5.29011 | 3.37801 |
| | male | -3.18527 | -7.08749 | 2.374152 | -5.70529 | 1.032836 | -2.15448 | 2.060518 | -0.25102 |
| ए | Female | -1.26874 | -4.55126 | 6.23498 | -2.01947 | 6.3867 | 0.180184 | 9.387017 | 4.354564 |
| | male | -5.40633 | -6.22587 | 0.099586 | -7.15058 | 0.801316 | -1.83765 | 2.992681 | 0.310077 |
| ऐ | Female | -2.9789 | -3.51072 | 6.431976 | -1.85723 | 7.316853 | 0.796123 | 7.938667 | 3.736423 |
| | male | -6.556 | -6.89128 | 2.454335 | -7.7334 | 1.874405 | -0.39752 | 4.198624 | -0.74678 |
| ओ | Female | -0.41999 | -3.47033 | 7.944145 | -1.65603 | 7.839662 | 2.121131 | 7.814222 | 3.660802 |
| | male | -4.27628 | -10.1502 | 2.263077 | -4.02615 | 1.745509 | -1.60553 | 3.301879 | 0.874481 |
| औ | Female | -2.26082 | -4.26959 | 9.385092 | -0.50471 | 7.301936 | 2.00286 | 7.360946 | 3.253318 |
| | male | -5.28556 | -10.4674 | 2.824581 | -2.79567 | 2.370375 | -1.99766 | 3.604201 | -1.40094 |

**Table 1.b. Reference Template for female and male (Nasals - Template 1)**

| Mfcc | C2 | C4 | C5 | C6 | C7 | C8 | C9 | C10 | C11 | C12 |
|---|---|---|---|---|---|---|---|---|---|---|
| ङ | 2.94365 | 4.2964 | 12.46139 | 5.26765 | 8.274413 | 6.61156 | 15.2658 | 10.4151 | 11.0971 | 8.0148 |
| | -5.2329 | -7.8077 | 2.08325 | -7.3271 | -0.39163 | -6.29774 | 3.479424 | -1.36817 | 2.194736 | -2.54587 |
| ञ | 2.77729 | 3.39003 | 11.34258 | 5.94154 | 8.335077 | 6.3342 | 13.8347 | 10.2988 | 12.1477 | 8.02219 |
| | -7.3819 | -6.84375 | 1.687444 | -8.79797 | 1.105993 | -6.10932 | 3.289267 | -2.95982 | 0.970673 | -1.93048 |
| ण | 1.82462 | 5.56848 | 12.77806 | 5.99367 | 8.756295 | 6.84717 | 14.0318 | 9.47775 | 12.5577 | 7.73374 |
| | -8.0814 | -7.4109 | 6.181852 | -6.28666 | 2.047826 | -4.7078 | 4.696307 | -2.1045 | 2.438999 | -1.57216 |
| न | 2.89874 | 4.63934 | 12.06502 | 5.17869 | 8.397909 | 6.13445 | 13.5093 | 9.74418 | 11.9592 | 7.67169 |
| | -7.10079 | -7.9324 | 4.540996 | -6.64978 | 1.529433 | -5.1258 | 3.736215 | -2.31374 | 2.462116 | -2.16338 |
| म | 2.88674 | 5.01082 | 13.63256 | 6.45716 | 7.806689 | 4.88959 | 13.1784 | 10.7839 | 12.4048 | 7.08390 |
| | -6.90437 | -8.9153 | 5.129677 | -4.66884 | 1.624701 | -5.7902 | 1.606473 | -3.04279 | 4.440774 | -2.01825 |

**Table 2.a. Reference Template for female and male (Vowels Template 2 )**

| Template 1 + | | C13 | C14 | C15 | C16 | C17 | C18 | C22 | C23 | C24 |
|---|---|---|---|---|---|---|---|---|---|---|
| अ | Female | 7.078848 | 4.387388 | 4.274611 | 4.459731 | 4.833524 | 4.142724 | 0.175642 | -0.17464 | -0.19559 |
| | male | 2.328673 | 0.785536 | 1.301281 | 0.957844 | 1.87722 | -0.87819 | 0.010821 | -0.02578 | -0.02696 |
| आ | Female | 8.511875 | 3.575557 | 4.08985 | 4.251504 | 4.546258 | 4.521629 | 0.302444 | -0.25606 | -0.48561 |
| | male | 2.217316 | 0.494733 | 1.374111 | 1.392261 | 0.748895 | -1.44357 | -0.01293 | 0.009056 | 0.129895 |
| इ | Female | 7.273791 | 3.734413 | 5.668063 | 4.841799 | 5.755881 | 4.391319 | 0.206302 | -0.17005 | -0.20975 |
| | male | 5.150984 | 0.660798 | 1.949596 | 0.378322 | 1.429462 | -0.55297 | 0.028314 | -0.01932 | -0.09033 |
| ई | Female | 7.28744 | 3.243261 | 6.53466 | 6.144399 | 6.42444 | 5.325356 | 0.166334 | -0.15344 | 0.08688 |
| | male | 4.301093 | 0.516831 | 1.409028 | -0.19262 | 1.33204 | -0.49967 | -0.0982 | 0.007074 | -0.02447 |
| उ | Female | 8.226756 | 4.577606 | 5.552359 | 5.205183 | 5.535379 | 4.094204 | 0.128494 | -0.1238 | -0.13079 |
| | male | 5.100364 | -0.22165 | 1.009491 | -0.7336 | 1.306759 | -0.9346 | 0.008909 | -0.04471 | -0.03375 |
| ऊ | Female | 8.033082 | 4.573306 | 6.028065 | 5.99527 | 5.981675 | 4.827958 | 0.058196 | -0.06899 | 0.013771 |
| | male | 5.263575 | 0.919737 | 0.674776 | -0.90614 | 0.917999 | -1.44899 | -0.06602 | 0.00302 | 0.062404 |
| ए | Female | 6.406713 | 2.997172 | 5.155061 | 5.023839 | 6.079622 | 4.178942 | 0.174535 | -0.14677 | -0.1736 |
| | male | 3.006762 | -0.9618 | 0.315245 | -0.3867 | 0.827201 | -1.18775 | -0.01823 | 0.020771 | 0.145847 |
| ऐ | Female | 6.717406 | 2.814474 | 5.108949 | 4.706288 | 5.947813 | 4.357084 | 0.253658 | -0.18001 | -0.10736 |
| | male | 4.496842 | -1.35262 | -0.85873 | -1.87617 | 1.457027 | -0.99309 | -0.03451 | -0.03367 | 0.103064 |
| ओ | Female | 6.90482 | 2.705242 | 4.100156 | 4.301329 | 5.589296 | 4.185041 | 0.25259 | -0.16677 | -0.27696 |
| | male | 4.753529 | -0.90991 | 0.215185 | -0.6765 | 0.305167 | -1.30791 | 0.038058 | -0.01929 | 0.111089 |
| औ | Female | 6.704798 | 2.995721 | 4.444988 | 4.357859 | 5.453042 | 3.612749 | 0.318132 | -0.18424 | -0.33823 |
| | male | 3.50472 | -1.46067 | -0.02768 | -0.59376 | 1.557149 | -1.55945 | -0.03031 | -0.02112 | 0.156502 |

**Table 2.b. Reference Template for female and male (Nasals Template 2 )**

| Template 1 + | | C13 | C14 | C15 | C16 | C17 | C18 | C22 | C23 | C24 |
|---|---|---|---|---|---|---|---|---|---|---|
| ङ | Female | 10.699 | 7.673844 | 9.12353 | 7.4002 | 7.50175 | 5.835663 | 0.395795 | -0.16965 | -0.32255 |
| | male | -0.37913 | -2.39425 | -1.83593 | -1.40936 | 0.496025 | -0.7568 | -0.10423 | 0.004142 | 0.060873 |
| ञ | Female | 11.12132 | 8.022715 | 9.75440 | 7.961456 | 7.785399 | 5.606303 | 0.330703 | -0.14898 | -0.30737 |
| | male | -0.59169 | -2.51066 | -1.65395 | -1.69069 | 0.731818 | -0.89815 | -0.06848 | -0.01173 | -0.02902 |
| ण | Female | 10.3195 | 7.816601 | 9.06098 | 7.590654 | 7.413413 | 5.305706 | 0.401825 | -0.17635 | -0.38986 |
| | male | -1.65261 | -2.07568 | -0.17222 | -1.238 | 0.007525 | -0.83295 | -0.10481 | 0.022599 | 0.115518 |
| न | Female | 10.31345 | 7.911247 | 9.25385 | 7.678782 | 7.513164 | 5.604248 | 0.38615 | -0.15228 | -0.26144 |
| | male | -0.44651 | -2.01583 | -0.54735 | -1.01143 | 0.59927 | -0.8863 | -0.08732 | 0.02209 | 0.158169 |
| म | Female | 10.62705 | 7.983824 | 9.38655 | 7.89045 | 7.433727 | 5.669573 | 0.354713 | -0.16549 | -0.24447 |
| | male | -1.3681 | -2.55767 | -0.43482 | -0.91517 | 0.879118 | -0.91692 | -0.12059 | 0.028867 | 0.177248 |

## 3.4 Performance Evaluation

For each test utterance, the mean vector was calculated by averaging the data over all the frames. This is the test template for each utterance. For evaluating the effectiveness of the feature vector in gender discrimination, each template per utterance per subject is tested against the reference templates of both the genders. The gender of the test speaker is decided using the closest match. The Euclidean distance measure is used as the distance metric for matching.

### *3.4.1 Euclidean Distance measure*

The Euclidean distances $D_m$ (Distance from male reference template) and $D_f$ (Distance from male reference template) were calculated for each test utterance. The matching is done based on the closest distance. Consider a test speaker whose test template is X. If the distance of X from the male reference vector M, ($D_m$) is smaller than ($D_f$) that from female reference vector , then X is closer to M meaning the speaker is a male.

## 4. RESULTS AND DISCUSSION

The results of the experiments were analyzed to find the effect of reducing the feature vector dimension on the accuracy of Gender Recognition.

## 4.1 Effect of Feature Vector Dimension on Gender Recognition Accuracy

We consider 40 coefficients C1 to C40 of the MFCC vector (Template1) in this study as some studies suggest that higher coefficients are very efficient in identifying gender of the speaker. The study of the data led to the choice of effective coefficients and formation of a reference template-2 which has only 17 coefficients.

**Table 3. Experimental results for vowel आ**

| Speaker | Euclidean distance using Template1 | | Result | Euclidean distance using Template 2 | | Result | Euclidean distance using Template 3 | | Result |
|---|---|---|---|---|---|---|---|---|---|
| | $D_m$ | $D_f$ | | $D_m$ | $D_f$ | | $D_m$ | $D_f$ | |
| $\mathbf{T_{M1}}$ | 8.515076 | 8.968088 | M | 10.42375 | 10.95137 | M | 14.18581 | 14.13491 | F |
| $\mathbf{T_{M2}}$ | 10.00855 | 16.73634 | M | 12.37827 | 24.09554 | M | 13.98208 | 26.06662 | M |
| $\mathbf{T_{M3}}$ | 7.516861 | 15.48994 | M | 10.204 | 23.24074 | M | 11.50694 | 25.02997 | M |
| $\mathbf{T_{M4}}$ | 3.605093 | 13.50391 | M | 9.342803 | 23.42708 | M | 10.2663 | 24.74012 | M |
| $\mathbf{T_{M5}}$ | 8.108048 | 14.10092 | M | 11.03528 | 22.31352 | M | 12.80475 | 24.24446 | M |
| $\mathbf{T_{M6}}$ | 6.530588 | 15.2539 | M | 10.54471 | 23.38007 | M | 12.30085 | 25.53016 | M |
| $\mathbf{T_{M7}}$ | 6.230884 | 11.8323 | M | 9.67339 | 19.43026 | M | 11.67101 | 21.74333 | M |
| $\mathbf{T_{F1}}$ | 24.41748 | 16.61677 | F | 30.70315 | 18.93019 | F | 33.78143 | 22.08581 | F |
| $\mathbf{T_{F2}}$ | 28.27455 | 19.62651 | F | 32.28214 | 20.5145 | F | 36.22771 | 24.19149 | F |
| $\mathbf{T_{F3}}$ | 28.35313 | 18.99901 | F | 34.84782 | 21.84275 | F | 39.11009 | 25.93385 | F |
| $\mathbf{T_{F4}}$ | 26.23526 | 16.14464 | F | 31.52047 | 18.48565 | F | 37.31131 | 25.02258 | F |
| $\mathbf{T_{F5}}$ | 26.47049 | 17.67999 | F | 32.65124 | 19.8344 | F | 36.10766 | 23.01636 | F |
| $\mathbf{T_{F6}}$ | 26.17623 | 16.46328 | F | 31.32692 | 18.0766 | F | 35.0953 | 22.1639 | F |
| $\mathbf{T_{F7}}$ | 24.75679 | 15.68334 | F | 29.93977 | 17.32881 | F | 33.6508 | 21.27342 | F |

Hence Template2 is [C2, C4, C5, C6, C7, C8, C9, C12, C13, C14, C15, C16, C17, C18, C22, C23, C24].

As stated earlier, ASR uses only 13 static coefficients, that is up to C12 . To gauge the effect of reducing the dimension of the MFCC feature vector on Gender Recognition, template-3 with selected coefficients up to C12 (dimension is 8) was formed which is [C2,C4,C5,C6,C7,C8,C9,C12].

The following observations are made from the experimental results.

- The female gender recognition is better compared to the male.
- The template-1 has performed as well as template-3 in gender recognition accuracy, if not better.
- The template-2 has performed better than template-3 in gender recognition accuracy.

It is noteworthy that template which contains selected MFC coefficients, Template1 [C2, C4, C5, C6, C7, C8, C9, C12 ] has performed as good as the template containing all the 40 coefficients from C1 to C40. This will mean that we can use selected coefficients from the feature vector extracted for speech recognition without losing out on gender recognition accuracy if the data selection is done carefully. In the case of text dependent gender recognition in Hindi, we can save extra computation on feature extraction for gender recognition modules of ASR. More tests will be conducted with phonemes extracted from continuous Hindi speech.

**Table 4. Effect of Feature Vector Dimension**

| Hindi phonemes in devanagri | Identification rate | | | | | |
|---|---|---|---|---|---|---|
| | Template1 | | Template2 | | Template3 | |
| | M | F | M | F | M | F |
| अ | 94.12 | 100 | 94.12 | 100 | 94.12 | 100 |
| आ | 100 | 100 | 100 | 100 | 94.12 | 100 |
| इ | 100 | 100 | 100 | 100 | 100 | 100 |
| ई | 100 | 100 | 100 | 100 | 100 | 100 |
| उ | 94.12 | 100 | 100 | 100 | 100 | 100 |
| ऊ | 100 | 100 | 100 | 100 | 100 | 100 |
| ए | 100 | 100 | 100 | 100 | 100 | 100 |
| ऐ | 100 | 100 | 100 | 100 | 100 | 100 |
| ओ | 100 | 100 | 100 | 100 | 100 | 100 |
| औ | 100 | 100 | 100 | 100 | 100 | 100 |
| ङ | 94.12 | 100 | 100 | 100 | 94.12 | 100 |
| ञ | 94.12 | 100 | 94.12 | 100 | 94.12 | 100 |
| ण | 100 | 100 | 100 | 100 | 100 | 100 |
| न | 100 | 100 | 100 | 100 | 100 | 100 |
| म | 100 | 100 | 100 | 100 | 100 | 100 |

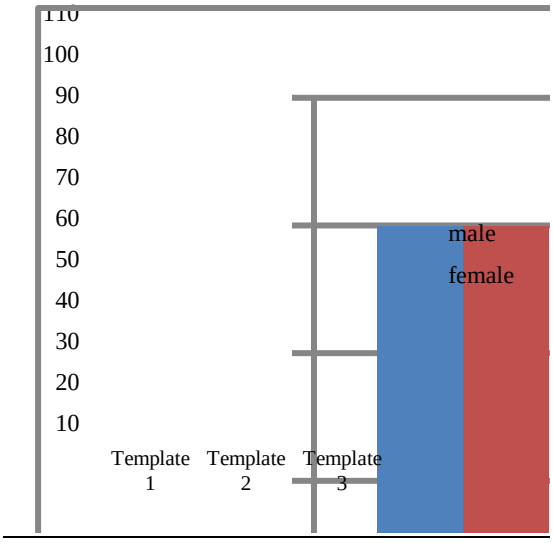


**Fig 3:** Feature vector Dimension Vs Gender Identification accuracy for vowel आ

## 4.2. Effectiveness of Various Phonemes in Gender Recognition

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). Table 5 gives a general overview of the results.

**Table 5.a Effect of Data selection**

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

The experimental results showed that the vowels इ, ई, ऊ,ए, ऐ, ओ and औ have shown very good distinction between the genders using template1. Nasals ण, न and म perform as well as the vowels do as seen from table 5.b.

**Table 5.b Effect of Data selection**

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

This leads to the expectation 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.The customary Greetings in Hindi like ‘Namaskar’ and ‘Pranam’ which have a rich mix of vowels and nasals in them may be tested for this purpose. More tests will be conducted with vowels and nasals extracted from continuous Hindi speech.

## 5. CONCLUSION

As mentioned at the outset, the preliminary study towards materializing a gender recognizer module for Hindi ASR system is presented. Seven of the vowels and five of the nasals that were studied show excellent gender discriminating ability. The work will be extended to formulate a system capable of recognizing the gender of a speaker given the first, very short, single Hindi word utterance, with acceptable accuracy. Further studies will be conducted to validate the effect of feature dimension on Recognition Accuracy.

## 6. REFERENCES

[1] Kumar, K. and Aggarwal, R. K.2011 Hindi Speech Recognition System using HTK. Int. J. of Computing and Business Research, ISSN (Online) : 2229-6166, Volume 2, Issue 2.

[2] Sedaaghi, M. H. 2009 A Comparative Study of Gender and Age Classification in Speech Signals. Iranian Journal of Electrical & Engineering. Vol. 5, No. 1

[3] Childers, D. G., Wu, K. and Hicks, D. M. 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.

[4] Harb, H. and Chen, L. 2006 Gender Identification using a general Audio Classifier. Multimedia Tools and Applications, volume 34, No. 3, 375-395.

[5] Abdulla, W. and Kasabov, N. 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.

[6] Wu, K. and Childers, D.G. 1991 Gender recognition from speech. Part I: Coarse analysis. J.Acoust. Soc. of Am., 90(4):1828–1840.

[7] Pronobis, M. and Doss, M.M. Analysis of F0 and Cepstral Features for Robust Automatic Gender Recognition. https://docs.google.com

[8] Singh, S. and. Rajan, E.G. 2011 Vector Quantization Approach for Speaker Recognition using MFCC and Inverted MFCC. Int. J. of Computer Applications (0975 – 8887). Volume 17, No.1.

[9] Feld,M., Burkhardt, F. and Muller, C. 2010 Automatic Speaker Age and Gender Recognition in the Car for Tailoring Dialog and Mobile Services, INTERSPEECH-2010, 2834-2837

[10] Ting, H., Yingchun, Y. and Zhaohui, W. 2006 Combining MFCC and Pitch to Enhance the Performance of the Gender Recognition Proc. Int. Conf. on Signal processing.

[11] Rajeshwara Rao, R. and Prasad, A. 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.

[12] Metze,F., Ajmera, J., Englert,R., Bub,U., Burkhardt, F., Stegmann,J., Muller, C. , Huber,R., .Andrassy, B.,. Bauer, J. G and Little, B. 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

[13] Sedaaghi, M. H 2008 Gender Classification in Emotional Speech. Speech Recognition, Technologies and Applications, pp. 550, www.intechweb.org

[14] Milan Sigmund 2008 Gender Distinction using Short Segments of Speech Signal. Int. J. of Computer Science and Network Security, Vol.8, No.10.

[15] Milan Sigmund 2008 Automatic Speaker Recognition by Speech Signal. Frontiers in Robotics, Automation and Control.

[16] Gurgen FS, Fan T and Vonwiller J.1994 On the Analysis of Phoneme based features for Gender Identification with Neural Networks. SST 1994. Australian Speech Science and Technology Association Inc.

[17] Hasan, M.R., Jamil, M, Rabbani, M.G. and Rahman, M. S.2004 Speaker Identification using Mel frequency Cepstral Coefficients. Proc.3rd Int. Con. on Electrical & Computer Engineering, Dhaka, Bangladesh.

[18] The HTK Book for HTK Version 3.4, 2009 Cambridge University Engineering Department.

.