CHAPTER III
FEATURE EXTRACTION

In the previous chapter, the process of speech production by humans was discussed since the physical traits derived from the speech signal will be used for the speaker recognition task. The LTI model for speech production was described and the convolved nature of speech was emphasized. It was shown that the vocal tract can be modeled as a digital filter for a short speech segment. The voiced and unvoiced segment characteristics were described in detail. The speech signal not only gives the information about the words or message being spoken but also the identity of the speaker can be obtained. The feature extraction module extracts a set of features that gives some speaker-specific information from the speech signal. In this chapter, the various components of front end processing i.e. signal acquisition, pre-processing and feature extraction are discussed in detail. VAD has been used as pre processing step to enhance the performance of speaker recognition system. A hybrid VAD has been proposed and its performance is compared with conventional energy based VAD. Three different features per frame i.e. frame energy, zero crossing rate and fundamental frequency have been used for making VAD decision. The combinations of these three parameters have not been investigated in literature earlier. It is shown that the accuracy of a MFCC-based speaker recognition method improves by adding derivatives. Speech signals from real life situation are investigated to evaluate the effects of text-dependency, length of sample, language, speaking style, emotional state, cheating, microphone and sample quality and noise on recognition rate.

3.1 INTRODUCTION
Speech signal is represented in terms of features for the purpose of speaker identification. This includes physical and learned traits. Learned or high-level features have been recently used in speaker recognition. They are robust and give good performance in noisy environments and channel mismatched cases. On the other hand, physical or low-level spectral features are widely used as they are easy to compute and model. They are much more related to the speech production mechanism
and source-filter modeling. These features are combined into feature vectors to reduce dimensionality and redundancy, while retaining the speaker-specific information.

The most common way to analyze a non-stationary signal is to perform a time-dependent spectral analysis. The signal is divided into a sequence of time frames, during which it may be assumed stationary. The speech signal shows slow variation with time. It is also called a quasi-stationary signal. When it is analyzed for a short span of time i.e. between 20 to 100 ms, the characteristics of the speech signal are found to be reasonably stationary. However, the signal characteristics have a tendency to change over longer periods of time i.e. for 1/5 seconds or even more. Therefore, short-time spectral analysis is used to characterize speech signal. The main criterion for extracting features from an input speech is to strike a balance between the acoustic vector’s dimensionality and its discriminating power. A speaker identification/verification system with small dimensionality will not be effective because it will be less discriminative. On the other hand, performance improves if more training and test vectors are used but this increases complexity.

Some other factors that may affect performance of speaker recognition system include:

Vocal effort—whisper to scream, Speech style—read, extempor or drunk, Speech rate—slow to fast, Speech length (very short vs. very long utterances), Ageing—time between sessions 1 hour – 10+ years, Speaker health, Native language, Number of target speakers, Channel and recording characteristics, Signal to noise ratio, Type of noise—white, babble, music, crosstalk, Room acoustics—reverberation and echoes, Sensor type—high quality microphone, telephone microphone, Distance of sensor from speaker, Bit-error rate in compression and transmission, Single speaker vs. multiple speakers at the same time.

The desired attributes of features are that they should be easy to extract, easy to measure, robust, occur frequently and naturally in speech. Features should be stable and not affected by speaker’s physical and mental state. Nevertheless, no feature has all these attributes. Spectrum based features are the most effective in automatic recognition systems. There are different methods available that gives parametric representation of the speech signal for the speaker recognition task. These include Linear Prediction Coding (LPC), Mel-Frequency Cepstrum Coefficients (MFCC), Cepstral Coefficients using DCT (Markel and Gray, 1976; Deller et al., 2000). MFCCs are the most popular and effective method. It was originally developed for speech recognition but it has
become one of the most common feature extraction techniques in speaker identification. The MFCC mimics human ear as it assumes that the human hearing is an optimal speaker recognizer.

The front end processing unit includes the pre-processing and feature extraction techniques. In pre-processing, raw spoken words utterances are converted into vector form for further signal processing. Components of preprocessing are as follows:

a) pre-emphasis of signal
b) voice activity detection

The aim of pre-processing step is to extract the desired spoken word utterances from the raw spoken word utterances. After applying feature extraction technique on speech sample, feature vectors can be extracted. Feature extraction is a crucial step, as it has strong influence on the recognition rate (Reynolds, 1994; Kajarekar and Hermansky, 2001). The Mel frequency cepstrum can be defined as the short-time power spectrum of a speech signal, which is calculated as the linear cosine transform of the log power spectrum on a non-linear mel scale of frequency. The steps for calculating MFCCs are given below:

- Take the FFT of a windowed signal. Compute its square magnitude. This gives the power spectrum.
- Integrate the power spectrum within the overlapping critical band filter response using triangular overlapping windows called mel filters.
- Compress the spectral amplitudes by taking log.
- Take the IDFT. This gives the cepstral coefficients.
- Perform spectral smoothing. This is done by truncating the cepstrum.

There are different ways of representing the speech signal. The simplest one is the speech waveform which is basically the plot of the sampled points versus time. Figure 3.1 shows a speech signal. Another representation is, so-called, the spectrogram of the signal. Figure 3.2 shows the spectrogram of the input voice signal (Hello). Spectrogram is a visual representation of acoustic signal in frequency domain. It computes the windowed discrete time Fourier transform (DTFT) of a signal using sliding window. Spectrogram is a 2-D plot of frequency against time, where the magnitude at each frequency is represented by the grey scale darkness or by colour in position (t, f) in the display, and the darker regions correspond to higher magnitudes.
3.2 PRE-PROCESSING
Preprocessing of speech signals is considered a crucial step in the development of a robust and efficient speech or speaker recognition system. Its basic diagram is shown in Figure 3.3. First, the continuous speech signal $s(t)$ produced by the speaker and sensed by the microphone has to be converted to the discrete domain. Secondly, the speech signal is segmented into frames. This is done to obtain quasi-stationary units of speech. Finally, a pre-emphasis filter is applied to each frame. Once all this
procedure has been performed, the speech frames are ready to enter the feature extraction subsystem.

### 3.2.1 Pre-Emphasis

The speech signal is a continuous air-pressure signal that can be captured by a microphone. The microphone converts this pressure-signal into a continuous electrical signal. The A/D converter converts this continuous representation into the discrete domain so that it can be processed in the digital domain. To perform this operation, the A/D module takes samples of the original signal with a frequency $F_s$ called *sampling frequency*. According to the *Nyquist* theorem, this sampling frequency has to be at least the double of the input signal’s bandwidth to avoid spectral aliasing. In speech signals this bandwidth can be approximated to 8 kHz since above this limit there is no relevant information. Hence, a sampling frequency of 16 kHz is enough.

Furthermore, most of the information in the speech is located below 4 kHz. This is exploited by the telephone lines which apply a 4 kHz low-pass filtering to the voice signals. Thus, telephone lines reduce significantly the speech signal bandwidth without a significant loss in the quality of speech.

![Diagram of Preprocessor](image)

**Figure 3.3: Preprocessor**

Pre-emphasis is widely used in speech enhancement. Its purpose is to lift the high frequency spectral components in the speech signal. There are various reasons to do this. Firstly, microphones increase the high frequencies roll-off in speech signals. Secondly, the $F3$ and $F4$ formants of speech are much lower than $F1$ and $F2$. 
Therefore, it is necessary to lift these frequency components. Pre-emphasis is done by applying a FIR filter to the speech frame. The filter response is given by Equation 3.1.

\[ H(z) = 1 - a z^{-1} \]  

where ‘a’ typically lies within the range of 1.0 and 0.4 and reflects the degree of pre-emphasis.

Pre-emphasis filtering is traditionally used to compensate for the -6dB/octave spectral slope of the speech signal (Hermansky and Morgan, 1994). Figure 3.4 shows the speech signal with and without pre-emphasis.

![Without pre-emphasis](image1)

![With pre-emphasis (\(\alpha = 0.91\))](image2)

**Figure 3.4: Original Speech Signal without pre-emphasis and with pre-emphasis**

### 3.2.2 Voice Activity Detection

The next step in processing speech data is the reliable separation of speech and non-speech segments. This is achieved by voice activity detector (VAD). VAD is used as a preprocessing step to further improve the accuracy. VAD is a crucial component in applications such as speech transmission, noise reduction, speech recognition and variable rate speech coding. Typical speech conversations are characterized by a speech/non-speech ratio of about 40:60, and thus an effective VAD system improves resource utilization considerably (Kinnunen et al., 2004, Kinnunen 2002).

A typical VAD involves the following three steps:
• Parameterize the speech signal: Time domain or spectral domain based features such as energy, zero crossing rate, spectral shape, cepstral coefficients etc. are extracted from the audio signal.

• Make the initial VAD decision: The decision is made whether a given segment is speech or non-speech. It is typically done framewise. Decision rules, statistical models, or adaptive thresholds are used for this purpose. This might also involve steps like estimating the current SNR or determining the noise type.

• Smooth the VAD decision: Since speech is highly correlated, if the current frame is speech, the next frame is also likely to be speech. Typical VAD algorithms refine the initial VAD decision to avoid rapid transitions from speech to non-speech. When viewed as a pattern recognition problem, VAD has to discriminate two classes, viz. speech (which could be noisy) and non-speech. In clean environments, most VAD algorithms work well, but performance deteriorates considerably in the presence of noise, with many detection errors. Figure 3.5 gives the steps for VAD.

![Figure 3.5: Steps for VAD](image)

The two conventional methods namely traditional method and Statistical method for implementation of VAD are discussed in detail in the subsequent paragraphs.
**Traditional methods for VAD**

Common methods for VAD are based on energy and other time-domain features like zero crossing rate, correlation coefficients, and periodicity measures. Other methods are spectral methods, which include the measurement of spectral distance and formant structure. More complex methods use multiple features to make a VAD decision. However, no specific feature or a combination of features work uniformly well under all conditions. Energy based methods, for instance, do not work well in low SNR conditions. Other methods require SNR estimation, which by itself is a challenging problem in non-stationary environments. In rapidly varying environments, the parameters of the VAD algorithm need to be updated adaptively. Recent methods for VAD require use of long term variability measures. This utilizes long term analysis windows (as opposed to a short term window) for measuring the long term spectral divergence or long term signal variability.

**Statistical methods for VAD**

Other methods for VAD employ statistical models to discriminate between speech and non-speech. One such approach assumes that the statistics of non-speech are stationary over a longer time period than that of speech. To make a VAD decision, the statistics of the current frame are compared with the estimated noise statistics. In another approach, a likelihood ratio test is applied assuming each spectral component of speech and non-speech follow a complex Gaussian distribution. The two competing hypotheses in this case are as follows: given the spectral component $X(k)$ of a frame of noisy speech, then $H_0 : X(k) = N(k)$ and $H_1 : X(k) = N(k) + S(k)$, where $k$ is the index into the discrete Fourier transform coefficient; $N$, $S$ and $X$ represent noise, speech and noisy speech respectively.

Another method, models speech and non-speech as Laplacians (instead of Gaussians). Some statistical methods for VAD make use of the observation that the higher order statistics (HOS) of speech are different from non-speech. The non-speech is assumed to be Gaussian, and the HOS are used to distinguish speech from non-speech. However, performance degrades when speech is corrupted by noise which is not Gaussian or when detecting unvoiced regions in speech (which are Gaussian-like in nature). The method proposed in aims to solve this by combining HOS with low band to full band energy ratios.
The existing algorithms suffer from following drawbacks:

1. In energy based VAD, it is assumed that the energy content in the voiced sample is higher than silence / unvoiced sample. However, it is not same for different speakers and hence it is difficult to set a proper value as threshold.
2. Due to noise spike some frames may have higher energy in the transitional region and may cross the threshold. Energy based algorithm may classify it as voice where as it should have been classified as a silence part.
3. A threshold for the classification of the voiced and unvoiced parts needs to be set in ZCR based VAD techniques. It may not be the same for various speakers. However, it cannot be varied / adjusted for different speakers, resulting in improper classification.
4. In statistical method for VAD, it is assumed that the first few samples of the input speech are due to noise. It may not be always correct. There is likelihood that some portion of the voiced part may have same distribution as that of the noise present in the beginning of the sample. These portions of the voiced part may get wrongly classified as noise.

### 3.3 Proposed Method of Voice Activity Detection

VAD has been used to primarily distinguish speech signal from silence. The challenge is to design a VAD algorithm which can distinguish speech from non-speech in an audio file. As discussed earlier, using single feature for making VAD decision does not give satisfactory performance. A hybrid technique has been proposed. Three different features per frame are used. The first feature is the frame energy. Energy can be easily measured and hence can be used for speech / silence detection. The energy is calculated by summing the square absolute value of each sample in a frame.

In the energy-based VAD, short time energy estimate is given by Equation 3.2 and Equation 3.3.

\[
E_s(m) = \frac{1}{L} \sum_{n=m-L+1}^{m} s^2(n)
\]

\[i.e. \text{ If } E_s > KE_r, \text{ frame is ACTIVE} \]

\[\text{else, frame is INACTIVE} \]

where \(E_s\) is energy of current frame, \(E_r\) is energy of noise frame, \(K\) is the scale factor that allows for adapting the threshold. Figure 3.6 shows the block diagram of the
energy based VAD. Threshold value adjustment helps to track time varying changes in the environment.

![Figure 3.6: Energy based VAD](image)

The second feature used is most dominant frequency component of the frame spectrum (Wildermoth and Paliwal, 2000). This feature is obtained by finding the frequency corresponding to the maximum value of the spectrum magnitude.

The third feature used for VAD is the zero-crossing rate (ZCR). The number of zero crossing rates for noise is more than that of speech. A zero-crossing occurs if there is a sign change between two adjacent samples. The summation of every zero-crossing in a frame, gives the zero-crossing rate. Short time zero crossing rates are given by Equation 3.4.

\[
Z_i(n) = \frac{1}{L} \sum_{n=m-L+1}^{m} \frac{\text{sgn}(s(n)) - \text{sgn}(s(n-1))}{2}
\]  

\[\text{--------- } (3.4)\]

In each frame the values of three features i.e. energy, most dominant frequency and zero crossing rate are obtained. They are compared with predefined threshold values and if any two of them fulfills the criteria, then that frame is marked as speech frame otherwise it is discarded. This increases robustness.

The performance of the proposed VAD depends heavily on the preset values of the threshold for detection of voice activity. The VAD proposed here works well when the energy of the speech signal is higher than the background noise and the background noise is relatively stationary. The amplitude of the speech signal samples are compared with the threshold value which is being decided by analyzing the performance of the system under different noisy environments.
3.4 MEL FREQUENCY CEPSTRAL COEFFICIENTS (MFCC)

The feature extraction module is crucial in any speaker recognition systems. This operation is required for two reasons. Firstly, the speaker recognition system needs to operate with low-dimensionality vectors in order to work in real-time mode. Secondly, the feature extraction block removes unnecessary information which is being carried in the speech frames and emphasizes speaker-dependent aspects of speech. MFCC are the most popular feature for speaker recognition task. They were introduced by Davis and Mermelstein in the 1980’s, and have been state-of-the-art ever since. Prior to the introduction of MFCCs, Linear Prediction Coefficients (LPCs) and Linear Prediction Cepstral Coefficients (LPCCs) were the main feature type for automatic speech recognition (ASR). MFCC’s are based on the known variation of the human ear’s critical bandwidth with frequency. Mel-frequency scale is shown in Figure 3.7. It has linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz (Srinivasa Kumar and Mallikarjuna Rao, 2011; Hasan et al., 2004; Paliwal and Alsteris, 2003).

![Figure 3.7: Plot of Pitch Mel Scale versus Hertz Scale](image)
Figure 3.8 shows the block diagram of a MFCC processor. The various steps for implementing MFCC are explained in detail in subsequent paragraphs.

![Figure 3.8: MFCC Processor](image)

**3.4.1 Frame Blocking**

In this step, the continuous speech signal is divided into frames of N samples. The next frame starts after M samples. The value of M is less than N. The first frame consists of the first N samples. The adjacent frames are separated by M samples and there is a overlapping of N - M samples. Usual values of N and M are 256 and 128 respectively. Hence, a speech signal is processed in frames which are overlapping with each other. From each frame, a feature vector is computed. The length of frame is approximately 20-40 msec, with an overlap of about 30-75 %. Usually frame length is fixed, which is simple to implement. Figure 3.9 gives the frame output of the truncated signal.
3.4.2 Windowing

Windowing reduces the effect of the spectral artifacts (spectral leakage/smearing) that arise from discontinuities at the frame endpoints. The signal discontinuities are minimized by using the window to taper the signal to zero at the beginning and end of each frame. Let $w(n)$ represents the window function, then the result of windowing is $y(n)$ represented by

$$y(n) = x(n) w(n), \ 0 \leq n \leq N - 1$$
Hamming window as shown in Figure 3.10 (a) is usually used, which has the form as given in Equation 3.5.

$$w(n) = 0.54 - 0.46 \cos \left( \frac{2\pi n}{N-1} \right)$$

0 ≤ n ≤ N-1, where N is the frame length. Figure 3.10(b) shows effect of applying Hamming window on a signal. Hamming window is used because its spectrum falls off rather quickly thus allowing better isolation. It has a wider main lobe but its side lobes are much softer whereas in rectangular window main lobe is narrower and side lobes are more prominent. The leakage effect produced by rectangular window is more noticeable. But as the main lobe of the Hamming window is wider than that corresponding to the rectangular window, the rectangular window offer a higher spectral resolution as it spreads a wider range of frequencies. However, its marked leakage effect is decisive for choosing the Hamming window.

### 3.4.3 Short Term Fast Fourier Transform

Each frame of N samples is transformed from the time domain into the frequency domain using Fast Fourier Transform (FFT) [52]. The FFT is used for fast implementation of the Discrete Fourier Transform (DFT). For the set of N samples it is given by Equation 3.6.

$$X_k = \sum_{n=0}^{N-1} e^{-j2\pi kn/N}$$

In general $X_k$’s are complex numbers and their absolute values are taken into account. The resulting sequence $\{X_k\}$ is interpreted as follows: positive frequencies 0 ≤ f < $F_s/2$ correspond to values 0 ≤ n ≤ N / 2 -1, while negative frequencies $-F_s/2 < f < 0$ correspond to N / 2 -1 ≤ n ≤ N +1. $F_s$ represent the sampling frequency. The result of FFT is spectrum or periodogram.

### 3.4.4 Mel-Frequency Warping

Psychophysical studies have shown that human perception of frequency content of speech signals does not obey a linear scale. Human ear is more sensitive to low frequencies. For each tone with an actual frequency f, a subjective pitch is measured
on mel scale. Mel is an abbreviation of the word melody. It is a unit of pitch. It is defined to be equal to one thousandth of the pitch ($\varphi$) of a simple tone with frequency of 1000 Hz and amplitude of 40 dB above the auditory threshold. The mel-frequency is given by Equation 3.7.

\[
\text{Mel} (f) = 2595 \log (1 + \frac{f}{700})
\]

Equation 3.7

The filter bank has a triangular band pass frequency response. This filter bank when applied in the frequency domain means applying the triangular-shape windows to the spectrum. At low frequencies there is linear spacing between filters but at high frequencies they are spaced logarithmically. Figure 3.11 shows mel-spaced frequency bank.

![Figure 3.11: Mel Frequency Filter Bank](image)

The value of frequency in Mel for different values of frequency in Hz is given in Table 3.1.

**Table 3.1: Frequency versus Pitch**

<table>
<thead>
<tr>
<th>Pitch f (Hz)</th>
<th>$\varphi$(Mel)</th>
<th>Pitch f (Hz)</th>
<th>$\varphi$(Mel)</th>
<th>Pitch f (Hz)</th>
<th>$\varphi$(Mel)</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>43</td>
<td>867</td>
<td>928</td>
<td>4109</td>
<td>2314</td>
</tr>
<tr>
<td>161</td>
<td>257</td>
<td>1000</td>
<td>1000</td>
<td>5526</td>
<td>2600</td>
</tr>
<tr>
<td>200</td>
<td>300</td>
<td>2022</td>
<td>1542</td>
<td>6500</td>
<td>2771</td>
</tr>
<tr>
<td>404</td>
<td>514</td>
<td>3000</td>
<td>2000</td>
<td>7743</td>
<td>2914</td>
</tr>
</tbody>
</table>
3.4.5 Log Compression and Discrete Cosine Transform

The log Mel spectrum is converted back to time domain. The result is called the Mel Frequency Cepstral Coefficients (MFCC). A good estimation of the spectral properties of the signal for a given frame is obtained from this type of representation of the speech spectrum. The Mel spectrum coefficients are real numbers. They are converted to time domain using Discrete Cosine Transform (DCT). If the Mel power spectrum coefficients are denoted as $S_k$, $k = 1, 2, ..., K$ and the MFCC's are denoted by $C_n$ then $C_n$ is given by Equation 3.8.

$$C_n = \sum_{k=1}^{K} (\log S_k) \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right], \quad n = 1, 2, ..., K$$

The first component of the DCT is neglected as it gives the mean value of the input signal and it carries little speaker specific information.

The performance of MFCCs can be enhanced by adding the first and second derivative coefficients to the MFCCs. This operation increases the dimension of the feature vectors resulting in a higher demand of training data. They represent dynamic information in the speech production procedure.

The steps used to calculate MFCC are summarized as follows:

Take a speech file. 256 samples are taken from the voiced part of the utterance for calculation of MFCCs. Then, FFT is taken and its squared magnitude is calculated to find the power spectrum. Next step is implementation of the FFT points over triangular filters with 50% overlap. The filters are designed in Mel scale, as the unequally spaced filters on the frequency axis in Hz will be equally spaced on the mel scale. The log spectrum for each FFT is then plotted in the mel scale (Niemi-Laitinen, 2005). A triangular filter of width equal to 300 mel is used. So, the first filter will be from 300-600 mel. The next one will extend from 450-750 mel so as to have 50% overlap between the filters. 32 such filters are designed. This adds to 5100 mel. To integrate the FFT bin values over these filters, the number of FFT bins falling in the filter width is obtained and the magnitude of each is multiplied with the weight corresponding to the height at that position and finally the result for all bins falling in the triangle width are added. One integration value per filter is obtained, so a total of 32 values. The IDFT is taken to get cepstral coefficients. Spectrum smoothing is performed by truncating output of IDFT and 12 cepstral coefficients are retained. These are the required MFCC values.
### 3.4.6 Delta and Delta-Delta Coefficients

Until now, no time evolution information is included in either cepstral coefficients or MFCC. However, dynamic information in speech signal is also different from speaker to speaker. This information is often included in the feature set by cepstral derivatives.

MFCC coefficients are static features. They give information about the speech signal over a fixed period of time. In order to capture information about how these vectors change over time, dynamic coefficients called delta and delta-delta coefficients are added. These dynamic features track the temporal variability in the feature vectors and are found to improve the recognition accuracy (Longworth and Gales, 2007; Furui, 1986; Gray, 1984). The first order derivative of cepstral coefficients is called delta coefficients, and the second order derivative of cepstral coefficients is called delta-delta coefficients.

A delta coefficient gives the speech rate, and delta-delta coefficients give something similar to acceleration of speech. The delta coefficients as given in Equation 3.9 are computed via linear regression over a window of consecutive frames with window length of 3-7 seconds.

\[
\Delta c_i = \frac{\sum_{k=1}^{K} (c_{i+k} - c_{i-k})}{2 \sum_{k=1}^{K} k^2}
\]

\[\text{------------------------ (3.9)}\]

where c and \(\Delta c\) corresponds to the static and dynamic coefficients respectively. K is the number of surrounding frames and \(c_i\) is the feature vector for which the delta coefficients are being calculated. Double delta coefficients are computed in a similar way from the delta coefficients.

Figure 3.12 gives the flowchart of Mel-frequency cepstrum and their delta, delta-delta coefficients. Figure 3.13 shows original speech signal, MFCC signal, delta and delta-delta signal.
Figure 3.12: Mel-Frequency Cepstrum and their delta, delta-delta coefficients
3.5 SIMULATION

It is important to select an appropriate database for the evaluation of the performance of any speaker identification system. In this thesis, TIMIT and NOIZEUS databases have been used for closed set speaker identification. The reason for selecting TIMIT database is that it is widely used, acceptable and publicly accessible. Simulation is also performed using noisy data. This can be generated either by adding noise in the TIMIT database or using already recorded speeches in the noisy environment. NOIZEUS database contains recordings in different places: babble, car, exhibition hall, restaurant, street, airport, railway station and train used to simulate the results in real world noisy environment. The simulation is being done in MATLAB (Rudra Pratap, 2006).
The necessary steps for the implementation of VAD and MFCC are given in subsequent paragraphs.

### 3.5.1 Voice Activity Detection

Since it is natural for people to pause while speaking, many of the speech frames will contain no useful information. The proposed VAD algorithm will be applied to the speech signals to identify the specific frames that include speech segments and specific frames that are silent.

The input and parameters to the VAD must match the input and parameters to the MFCC algorithm. For each speech frame found to be silent, the entire frame is removed. The steps involved are:

**Input: Speech sample, Sample rate**  
**Output: Indication of silence frame**

**Parameters:** window size=20ms, step size=10ms

- **Step I:** Segment speech into frames.
- **Step II:** Find energy of each frame.
- **Step III:** Determine maximum energy.
- **Step IV:** Determine zero crossing rate and dominant frequency of each frame.
- **Step V:** Remove any frames which do not satisfy these conditions:
  - a. less than 30 dB of maximum energy
  - b. less than -55dB overall energy.
  - c. number of zero crossing rates are higher than a predefined threshold.

### 3.5.2 MFCC

Steps for calculating MFCC are:

**Input:** Speech sample, sample rate  
**Output:** Matrix of MFCC by frame

**Parameters:** window size=20ms, step size=10ms, nbins=32, d=12(cepst)

- **Step I:** Compute FFT power spectrum.
- **Step II:** Compute mel frequency m channel filter bank.
- **Step III:** Convert to cepstra via DCT.

The quality of the speaker recognition system depends on the proper feature extracted values. MFCC are the most widely used feature set in text-independent speaker recognition. The derivatives of MFCC i.e. ∆MFCC and ∆∆MFCC based speech
feature extraction technique has been used to enhance the efficiency of the system. In the MFCC feature extraction there is a need to decide on parameters. Default values are taken for many parameters. These features may not necessarily be the best choice. TIMIT database is used.

First the number of MFCC needed to be set. Its default value of 12 is taken. The frame length is set to 20 ms which is within the typical range of 10-30 ms normally used. The frame overlap is set to 50% by default which is also well within the typical values of 25-70%. The number of filters is set to 32 by default.

Initially, pre-emphasis of speech is done. The pre emphasis factor of 0.97 is chosen. It is then segmented into frames. Frame length is **20ms** with a frame shift of **10ms**. Hamming window is applied to each frame. The time domain signal is converted into frequency domain by using FFT. A set of **32** overlapped triangular filters are applied to the spectrum. These filters are uniformly spread over the Mel frequency axis. Discrete cosine transform converts the log energy output of filters to a vector of **12** melcepstral (MFCC) features. The first and second time derivatives of MFCC components are obtained and appended to the vector. This leads to **36** feature vectors per frame. The parameters and their values for MFCC extraction are summarized in Table 3.2.

![Table 3.2: The parameters and their values for MFCC extraction](image)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Window length</td>
<td>20 ms</td>
</tr>
<tr>
<td>Window overlap</td>
<td>50%</td>
</tr>
<tr>
<td>Number of MFCCs</td>
<td>12</td>
</tr>
<tr>
<td>Number of mel bands</td>
<td>36</td>
</tr>
</tbody>
</table>

Screenshot of MATLAB window showing MFCC matrix is given in Figure 3.14.
The results of simulation study are discussed below.

3.6 FEATURE EXTRACTION USING MFCC AND ITS DERIVATIVES

The simulations are performed to study the effect of different parameters on performance of MFCC-based recognition system. The effect of number of filters and type of window on MFCC performance is studied. There are large variations between simulations and real conditions, even in noise-free environments. The performance of MFCC in presence of noise is investigated.

3.6.1 Number of Filters in Filter-banks vs. Identification Rate

To investigate the effect of changing the number of filters in the filter-bank on the identification rate, tests are performed using all the test speakers for different numbers of the filter-banks and the identification rate for each value is calculated. It can be seen that the maximum recognition rate is obtained with 32 numbers of filters. Increasing the number of filters results in increasing the distortion measure (Euclidean distance between the test utterance and the speakers’ models in database). Increasing the number of filter banks implies taking more data from the input speech. It results in
increasing the number of terms in the Euclidean distance, and so the distortion measure will increase. Thus, the number of the filter-banks plays a major role for the purpose of improving the recognition accuracy. The variation of recognition rate with change in number of filters is shown in Table 3.3. The corresponding bar chart is shown in Figure 3.14.

Table 3.3: Variation of Recognition Rate with Change in Number of Filters

<table>
<thead>
<tr>
<th>Number of filters</th>
<th>12</th>
<th>22</th>
<th>32</th>
<th>42</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recognition Rate</td>
<td>65%</td>
<td>80%</td>
<td>95%</td>
<td>90%</td>
</tr>
</tbody>
</table>

![Bar chart showing variation of recognition rate with number of filters](image)

Figure 3.15: Variation of Recognition Rate with Change in Number of Filters

3.6.2 Effect of Variation in Type of Window

Considering 32 filters as a standard number of filters, the window type is changed. Two types of windows are used viz. Hamming Window and Rectangular window. Results show that efficiency is more when hamming window is used. The main lobe width of Hamming window is twice that of the rectangular window. Also, the
magnitude of sidelobes is lower than that of rectangular window. There is an approximate difference of 43 dB between pass band and stop band gains in case of Hamming window. Hamming window generates less oscillation in the side lobes and hence, hamming window is generally preferred. The identification rate for two different types of windows is shown in Table 3.4 and the corresponding bar chart is shown in Figure 3.16.

Table 3.4: Type of Window vs. Identification rate

<table>
<thead>
<tr>
<th>Types of windows using 32 filters</th>
<th>Identification Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hamming window</td>
<td>96%</td>
</tr>
<tr>
<td>Rectangular window</td>
<td>85%</td>
</tr>
</tbody>
</table>

Figure 3.16: Identification Rate for different Types of Window

3.6.3 Effect of Adding Derivatives

A white Gaussian noise of different levels i.e. between 5 dB to 30 dB is added to the recorded test samples to test the robustness of described technique in noisy
environments that are unavoidable in most real applications. The features extraction methods that are used are MFCC, ΔMFCC and ΔΔMFCC.

The entire identification system is implemented in MATLAB environment. The performances of speaker recognition system is evaluated by applying Equation 3.10

\[ I = \frac{C}{N} \times 100 \% \]

\[ \text{---------- (3.10)} \]

where, \( I \) represents the percentage of correctly identified speakers called identification or recognition rate, \( C \) is the number of correctly identified speakers and \( N \) is the total number of speakers that have enrolled for identification test. The results obtained are summarized in Table 3.5 and the bar chart representation of the same is shown in Figure 3.17.

**TABLE 3.5: Effect of Adding Derivatives**

<table>
<thead>
<tr>
<th>SNR(dB)</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>12.23</td>
<td>23.67</td>
<td>43.15</td>
<td>63.43</td>
<td>82.43</td>
<td>91.67</td>
</tr>
<tr>
<td>ΔMFCC</td>
<td>15.93</td>
<td>25.15</td>
<td>45.54</td>
<td>66.35</td>
<td>84.35</td>
<td>92.65</td>
</tr>
<tr>
<td>ΔΔMFCC</td>
<td>16.43</td>
<td>26.83</td>
<td>46.52</td>
<td>68.74</td>
<td>86.74</td>
<td>92.89</td>
</tr>
</tbody>
</table>

![Figure 3.17: Variation of Recognition Rate at different values of SNR](image-url)
The recognition rates using delta-delta coefficients are better than the other two cases. These results indicate that in noisy environments the dynamic variants of MFCC algorithm are better suited to robust conditions.

In the next section, effect of VAD on the improvement of performance of Speaker Recognition System is investigated.

3.7 EFFECT OF VAD ON SPEAKER IDENTIFICATION RATE

To improve the performance of any speaker identification system in presence of noise, it is required to remove silence. The silence / unvoiced part of speech signal is more affected by noise than the voiced parts. Also, the features extracted from these regions are primarily due to noise. Speech or speaker dependent information is not present in these features and hence results in error. The proposed VAD algorithm is tested for different values of signal to noise ratio. To get the performance of the proposed VAD algorithm, the results obtained for (a) MFCC - with proposed VAD (b) MFCC-with energy based VAD and (c) MFCC - without VAD are compared. The effect of VAD step on Speaker recognition rate is tabulated in Table 3.6. The corresponding bar chart is shown in Figure 3.18.

Table 3.6: Speaker Identification Rate with VAD

<table>
<thead>
<tr>
<th>Features</th>
<th>Clean speech</th>
<th>Noisy speech 30 dB SNR</th>
<th>Noisy speech 20 dB SNR</th>
<th>Noisy speech 10dB SNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC without VAD</td>
<td>95.65</td>
<td>93.0</td>
<td>65.25</td>
<td>29.5</td>
</tr>
<tr>
<td>MFCC with proposed VAD</td>
<td>98.0</td>
<td>97.5</td>
<td>70.0</td>
<td>34.5</td>
</tr>
<tr>
<td>MFCC with energy based VAD</td>
<td>96.5</td>
<td>96.0</td>
<td>68.0</td>
<td>32</td>
</tr>
</tbody>
</table>
It is clearly evident from the experimental result that with the decrease in signal to noise ratio, the recognition rate improves on using VAD. Hence, the VAD algorithm is more effective for low SNR.

![Graph showing recognition rate for clean speech and different SNR levels with and without VAD](image)

**Figure 3.18: Speaker Identification Rate for Speech degraded by Additive White Gaussian Noise**

When tested on TIMIT database i.e. in the clean environment the proposed algorithm gives almost same performance as that obtained without VAD. But with lower SNR the improvement is more prominent. When SNR is low, most of the speech parts are eliminated as silence by the traditional method. But the proposed algorithm is able to discriminate between speech and non speech parts in a better way and hence eliminates only the unvoiced part. Speaker identification rate is improved up to 7% with the use of proposed VAD technique for noisy speech of 20 dB SNR using MFCC feature vectors.

The effects of various technical, speaker and data related factors which effect the performance of MFCC based system are investigated in the next section.
3.7 FACTORS AFFECTING MFCC PERFORMANCE

Mel Frequency Cepstral Coefficients (MFCCs) have been the most acceptable low-level features for speaker recognition and speech recognition systems. The Mel-Frequency Cepstrum (MFC) is a representation of the short-term power spectrum of a signal, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. It has been found that the performance of MFCC based system degrades due to the presence of various types of noises, deliberate cheating, channel mismatch conditions (Saastamoinen et al., 2005). Many other factors like microphone or speech quality, length and language of speech strongly affect speaker recognition. Emotional state of the speaker has strong influence on speech rate.

A database has been created by recording speech from 30 individuals (15 Males and 15 Females) from 8 different states of India. The details are shown in Table 3.7. A predefined set of sentence in both English and speaker’s native language is recorded. Two set of recordings is done-one with Microphone-1 and second with Microphone-2. To study the effect of deliberate cheating some speakers are made to record by changing their voice. To study the effect of spontaneous speech vs text reading, speakers are asked to speak spontaneously on any general topic.

<table>
<thead>
<tr>
<th>State</th>
<th>Number of speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delhi</td>
<td>5</td>
</tr>
<tr>
<td>Bihar</td>
<td>5</td>
</tr>
<tr>
<td>Tamilnadu</td>
<td>5</td>
</tr>
<tr>
<td>Maharashtra</td>
<td>5</td>
</tr>
<tr>
<td>Haryana</td>
<td>5</td>
</tr>
<tr>
<td>Punjab</td>
<td>3</td>
</tr>
<tr>
<td>Kashmir</td>
<td>2</td>
</tr>
</tbody>
</table>

The effects of following factors are studied:

1. **Language and Data-related factors**: The effects of text dependency, length of sample, language of speech are investigated. The simulations are performed to confirm whether the text content is important. The influence of the length of the sample is studied. Regarding the language, it is investigated, if the recognition rate is
better if a speaker speaks in his native language or in English language. Also, it is investigated if there is effect of language mismatch between training and testing conditions. The effects of language and data-related factors on recognition rate are listed in Table 3.8.

Table 3.8: Language and Data-related factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>Recognition Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text Dependent</td>
<td>97.6</td>
</tr>
<tr>
<td>Text independent</td>
<td>94.4</td>
</tr>
<tr>
<td>Short sample</td>
<td>93.3</td>
</tr>
<tr>
<td>Long sample</td>
<td>94.8</td>
</tr>
<tr>
<td>Language mismatch</td>
<td>88.4</td>
</tr>
</tbody>
</table>

2. **Speaker-Dependent factors**: The effects of speaker dependent factors like text reading vs. spontaneous speech, deliberate cheating and emotional state of speaker are investigated. It has been studied whether a speaker is better recognized using speech guidance (spontaneous speech vs. read text). Also, the effect of disguise is studied, i.e. speaker does not want to be recognized and hence deliberately changes his voice. System fails when deliberate cheating is done; error rates are as high as 80%. The same results are obtained with spontaneous speech.

Human beings can be easily overwhelmed by various emotions. Hence, the effect of emotional state of a speaker on recognition rate is investigated. Simulations are performed to confirm the effect of the emotional state of the speaker on recognition rate. The system proved to be 95% reliable in case of normal conditions (neutral speech samples) but the recognition rates ranges from 40% to 90%, depending on the emotional state of the speaker. Hence, some improvements are required to be made in the conventional systems when the speech samples are corrupted by different emotions. One alternative is to combine the MFCC features with fundamental frequency. The variation in fundamental frequency in different emotional states is investigated.

Fundamental frequency is widely used in speaker recognition. It is an important prosodic feature. The rate of vibration of the vocal cords during the production of the speech is called pitch. It is speaker variant. The fundamental frequency can vary due
to factors like emotional state, ageing and health of the speaker. Some technical factors such as the transmission channel and the recording device may cause fundamental frequency to vary. It contains speaker specific information and can be used to separate males and females. The fundamental frequency of male, female and children are different. This is due to the anatomic differences. The approximate values of frequency are 150 Hz for males, 250 Hz for females and 300 Hz for children.

In Table 3.9, the average value for the fundamental frequency of the five females and five males are shown which have been recorded in different emotional state. Ten different samples for every speaker are recorded for the same emotional state. The average frequency is obtained using these recordings.

TABLE 3.9: Fundamental frequencies for different emotional states of the speaker

(a) Female Speakers

<table>
<thead>
<tr>
<th>Speaker/Emotion</th>
<th>Anger</th>
<th>Fear</th>
<th>Happy</th>
<th>Boredom</th>
<th>Neutral</th>
<th>Sad</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>240 Hz</td>
<td>240 Hz</td>
<td>238 Hz</td>
<td>170 Hz</td>
<td>167 Hz</td>
<td>135 Hz</td>
</tr>
<tr>
<td>2</td>
<td>240 Hz</td>
<td>223 Hz</td>
<td>240 Hz</td>
<td>163 Hz</td>
<td>163 Hz</td>
<td>158 Hz</td>
</tr>
<tr>
<td>3</td>
<td>256 Hz</td>
<td>220 Hz</td>
<td>236 Hz</td>
<td>176 Hz</td>
<td>184 Hz</td>
<td>150 Hz</td>
</tr>
<tr>
<td>4</td>
<td>230 Hz</td>
<td>218 Hz</td>
<td>236 Hz</td>
<td>190 Hz</td>
<td>188 Hz</td>
<td>186 Hz</td>
</tr>
<tr>
<td>5</td>
<td>238 Hz</td>
<td>200 Hz</td>
<td>223 Hz</td>
<td>186 Hz</td>
<td>196 Hz</td>
<td>158 Hz</td>
</tr>
</tbody>
</table>

(b) Male Speakers

<table>
<thead>
<tr>
<th>Speaker/Emotion</th>
<th>Anger</th>
<th>Fear</th>
<th>Happy</th>
<th>Boredom</th>
<th>Neutral</th>
<th>Sad</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>200 Hz</td>
<td>180 Hz</td>
<td>205 Hz</td>
<td>110 Hz</td>
<td>120 Hz</td>
<td>110 Hz</td>
</tr>
<tr>
<td>2</td>
<td>182 Hz</td>
<td>170 Hz</td>
<td>202 Hz</td>
<td>102 Hz</td>
<td>102 Hz</td>
<td>96 Hz</td>
</tr>
<tr>
<td>3</td>
<td>192 Hz</td>
<td>159 Hz</td>
<td>210 Hz</td>
<td>100 Hz</td>
<td>101 Hz</td>
<td>110 Hz</td>
</tr>
<tr>
<td>4</td>
<td>202 Hz</td>
<td>150 Hz</td>
<td>170 Hz</td>
<td>116 Hz</td>
<td>114 Hz</td>
<td>112 Hz</td>
</tr>
<tr>
<td>5</td>
<td>200 Hz</td>
<td>138 Hz</td>
<td>142 Hz</td>
<td>140 Hz</td>
<td>139 Hz</td>
<td>116 Hz</td>
</tr>
</tbody>
</table>

The results show that the emotional state strongly influences frequency variation. It is clear from Table 3.9, that in case of fear, anger and happy state, the fundamental frequency is higher compared to the other emotional states.
3. **Technical Factors:** The technical factors which affect the performance are mismatched microphones, distance from microphone, sampling rate and additive noise. The effects due to mismatch of microphone during training and testing condition are investigated. Also, the effect of distance from the microphone has been investigated. The effects of sampling rate and additive background noise have been investigated.

Accuracy of speaker recognition decreases when there is acoustic mismatch between the training and testing conditions. This can be due to the change in health condition or attitude of a person, microphone used or transmission channel, or due to the noise. The effects of different noise types while using two different microphones are listed in Table 3.10.

**Table 3.10: Effect of various noise conditions with different microphones**

<table>
<thead>
<tr>
<th>Noise during training</th>
<th>Noise during testing</th>
<th>Recognition Rate for M1</th>
<th>Recognition Rate for M2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>Clean</td>
<td>97.2</td>
<td>95.2</td>
</tr>
<tr>
<td>Babble</td>
<td>Babble</td>
<td>92.2</td>
<td>91.3</td>
</tr>
<tr>
<td>Clean</td>
<td>Car</td>
<td>91.4</td>
<td>84.3</td>
</tr>
<tr>
<td>Clean</td>
<td>Babble</td>
<td>91.1</td>
<td>82.8</td>
</tr>
</tbody>
</table>

Results are better when there is no mismatch, the quality of microphone is also important. The most important inference from these results is that training data should be very similar to the testing data.

Distance from the microphone is another important aspect that affects the performance of the speaker recognition system. Table 3.11 shows the variation of the percentage accuracy for recognition with the distance from the microphone.

**Table 3.11: Variation of Recognition Rate with distance from microphone**

<table>
<thead>
<tr>
<th>Speaker No.</th>
<th>Distance from Microphone</th>
<th>Recognition rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>5cm</td>
<td>95%</td>
</tr>
<tr>
<td>2.</td>
<td>10cm</td>
<td>90%</td>
</tr>
<tr>
<td>3.</td>
<td>50cm</td>
<td>80%</td>
</tr>
</tbody>
</table>
It is quite clear that the recognition accuracy decreases with the increase in the distance from the microphone and this is due to reduced value of the signal to noise ratio. The reduced recognition accuracy due to increase in distance can be overcome by using pre-processing steps in the initial stages of the voice processing so as to improve the overall SNR for better recognition accuracy.

3.7 CONCLUSIONS

Pre-processing of the speech signal plays very important role in any speech processing application. Various processes include noise removal, endpoint detection, pre-emphasis, framing, windowing etc. The removal of silence/unvoiced portion is the fundamental step in speaker identification process. The voiced part of the speech signal carries the information which is more important from the perspective of speaker identification. Therefore, the removal of the redundant information in the unvoiced part in the preprocessing step is very significant in reducing the dimensions of features. This reduces the computational complexity of the subsequent stages and the speaker identification rate also improves. Speaker identification rate is improved up to 7 % due to the implementation of the proposed VAD technique. The proposed hybrid technique is a novel one.

The feature extractors that have been considered are MFCC and their derivatives. The results indicate that in noisy environments the dynamic variants of MFCC algorithm are better suited to robust conditions.

Language and data related, speaker dependent and technical factors that affect performance of MFCC have been investigated. Such work has not been reported in literature earlier. The most important result is that there should not be any mismatch between the training and testing conditions. The most degrading factor is noise. For different types of noise during training and testing conditions, the error rate is above 50 %. The error rate is below 15 % for matched noise conditions. The effect of speech and language factors is less than technical factors but deliberate cheating makes an exception.
In the next chapter, the speaker models which are used to train the speaker data are explained. This includes both the Neural Networks and the Support Vector Machine model. The relative performance analysis of the two models is done using different database and the parameters required to yield satisfactory results are also investigated.