5. SPEECH FEATURE EXTRACTION AND VECTOR QUANTIZATION

5.1 Overview

Speech feature extraction is one of the fundamental steps in any speaker recognition system. This Chapter describes how the speech feature extraction and vector quantization steps can be carried out in a speaker recognition system.

5.2 Speech Feature Extraction

There are several methods of speech feature extraction. Some commonly used methods are Linear Predictive Coding (LPC), Linear Predictive Cepstral Coefficient (LPCC) and Mel Frequency Cepstral Coefficients (MFCC).

5.2.1 Linear Predictive Coding (LPC)

Linear predictive coding (LPC) is one of the earliest standardized coders. LPC has been proven to be efficient for the representation of speech signal in mathematical form. LPC is a useful tool for feature extraction as the vocal tract can be accurately modelled and analysed. Studies have shown that the current speech sample is highly correlated to the previous sample and the immediately preceding samples [30]. LPC coefficients are generated by the linear combination of the past speech samples using the autocorrelation or the auto variance method and minimizing the sum of squared difference between predicted and actual speech sample.

$$\tilde{x}(n) = a_1 x(n-1) + a_2 x(n-2) + \cdots a_M x(n-M) = \sum_{i=1}^{M} a_i x(n-i)$$

$\tilde{x}(n)$ is the predicted $x(n)$ based on the summation of past samples. a_i is the linear prediction coefficients. M is the number of coefficients and n is the sample.

The error between the actual sample and the prediction can then be expressed by

$$\mathcal{E}(n) = x(n) - \tilde{x}(n)$$

$$\mathcal{E}(n) = x(n) - \sum_{i=1}^{M} a_i x(n-i)$$

$$x(n) = \sum_{i=1}^{M} a_i x(n-i) + \mathcal{E}(n)$$

The speech sample can then be accurately reconstructed by using the LP coefficients a_i and the residual error $\mathcal{E}(n)$. $\mathcal{E}(n)$ can be represented by the following in z domain.

$$A(z) = 1 - \sum_{i=1}^{M} a_i z^{-i}$$

The figure below shows the analysis filter

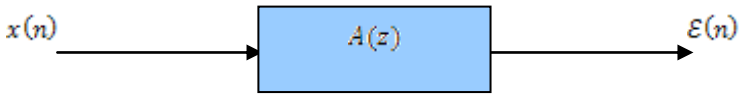

Figure 5.1 Speech Analysis Filter

[31]

The transfer function H(z) can be expressed as an all pole function , where G represents the gain of the system.

$$H(z) = \frac{G}{1 - \sum_{i=1}^{M} a_i z^{-i}}$$

The figure below shows the speech synthesis filter

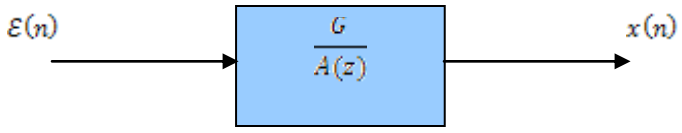

Figure 5.2 Speech Synthesis Filter

[31]

Schroeder [32] mentioned that the LPC model can adequately model most speech sound by passing an excitation pulse through time-varying all-pole filter using LP coefficients. S. Kwong [33] considers LPC as a method that provides a good estimate of the vocal tract spectral envelope. Gupta [34] mentioned that LPC is important in speech analysis because of the accuracy and speed with which it can be derived. The feature vectors are calculated by LPC over each frame. The coefficients used to represent the frame typically ranges from 10 to 20 depending on the speech sample, application and number of poles in the model. However, LPC also have disadvantages. Firstly, LPC approximates speech linearly at all frequencies that is inconsistent with the hearing perception of humans. Secondly, LPC is very susceptible to noise from the background which may cause errors in the speaker modeling.

5.2.2 Linear Predictive Cepstral Coefficients

Linear predictive cepstral coefficients (LPCC) combine the benefits of LPC and cepstral analysis and also improve the accuracy of the features obtained for speaker recognition. LPCC is equivalent to the smooth envelop of the log of the speech that allows for the extraction of speaker specific features. The block diagram of the LPCC is shown in the figure below [31].

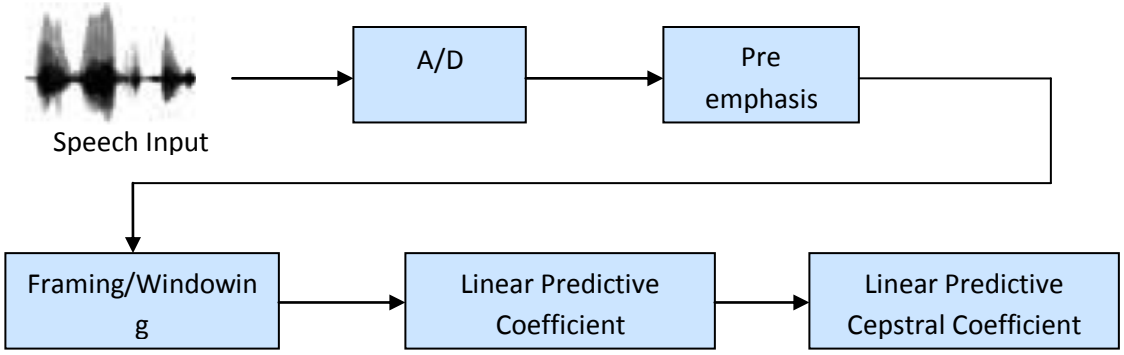

Figure 5.3 Block diagram of Linear Predictive Cepstral Coefficient

LPC is transformed into cepstral coefficients using the following recursive formula

$$c_1 = a_1 \qquad 1 < n \leq p$$

$$c_n = a_n + \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) a_k c_{n-k}$$

where c_i and a_i are the *i*th-order cepstrum coefficient and linear predictor coefficient, respectively. Atal [35] did a study on various parameters for the LPC and found the cepstrum to be the most effective parametric for recognition for speakers. Eddie Wong [36] mentioned that LPCC is more robust and reliable than LPC. However, LPCC also performs poorly under noisy environment.

5.2.3 Mel-Frequency Cepstral Coefficients (MFCC)

Mel-frequency Cepstral coefficient is one of the most prevalent and popular method used in the field of voice feature extraction. The difference between the MFC and cepstral analysis is that the MFC maps frequency components using a Mel scale modeled based on the human ear perception of sound instead of a linear scale [7]. The Mel-frequency cepstrum represents the short-term power spectrum of a sound using a linear cosine transform of the log power spectrum of a Mel scale [31]. The formula for the Mel scale is

$$M = 2595 \log_{10}\left(\frac{f}{700} + 1\right)$$

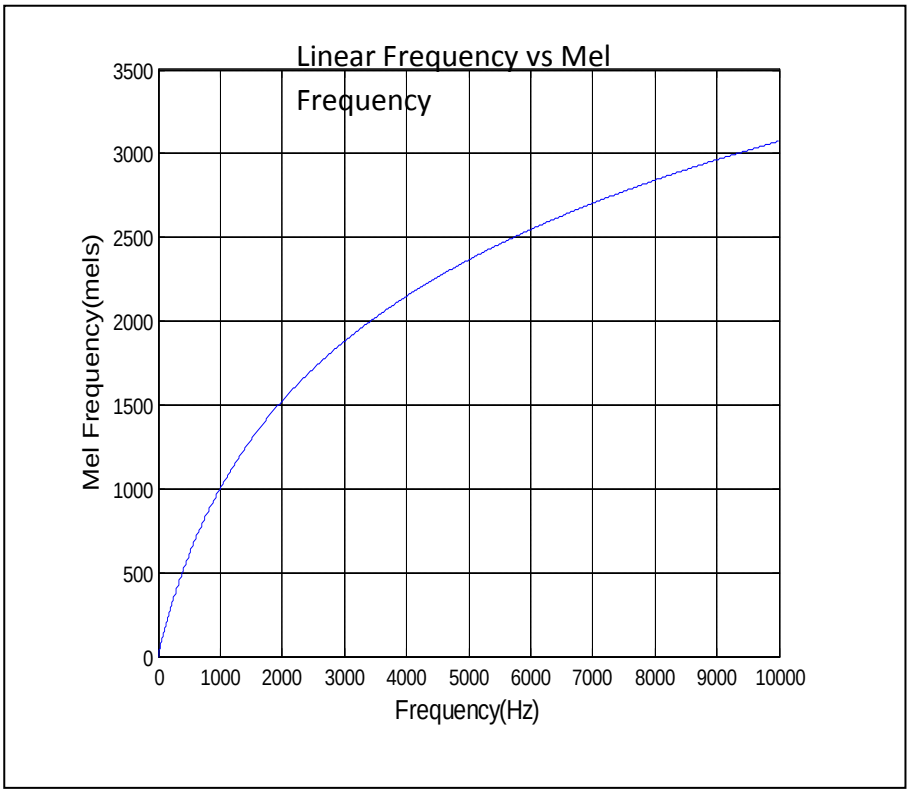

Figure 5.4 Mel Scale plot

Vergin [37] mentioned that MFCC as frequency domain parameters are much more consistent and accurate than time domain features. Vergin [37] listed the steps leading to extraction of MFCCs: Fast Fourier Transform, filtering and cosine transform of the log energy vector. According to Vergin [38], MFCCs can be obtained by the mapping of an acoustic frequency to a perceptual frequency scale called the Mel scale. MFCCs are computed by taking the windowed frame of the speech signal, putting it through a Fast Fourier Transform (FFT) to obtain certain parameters and finally undergoing Mel-scale warping to retrieve feature vectors that represents useful logarithmically compressed amplitude and simplified frequency information [39]. Seddik [40] mentioned that MFCC are computed by applying discrete cosine transform to the log of the Mel-filter bank. The results are features that describe the spectral shape of the signal. Rashidul [7] describe the main steps for extraction of MFCC, shown on figure. The main steps are as follow: pre-emphasis, framing, windowing, perform Fourier fast transform FFT), Mel frequency warping, filter bank, logarithm, discrete Cosine transform (DCT).

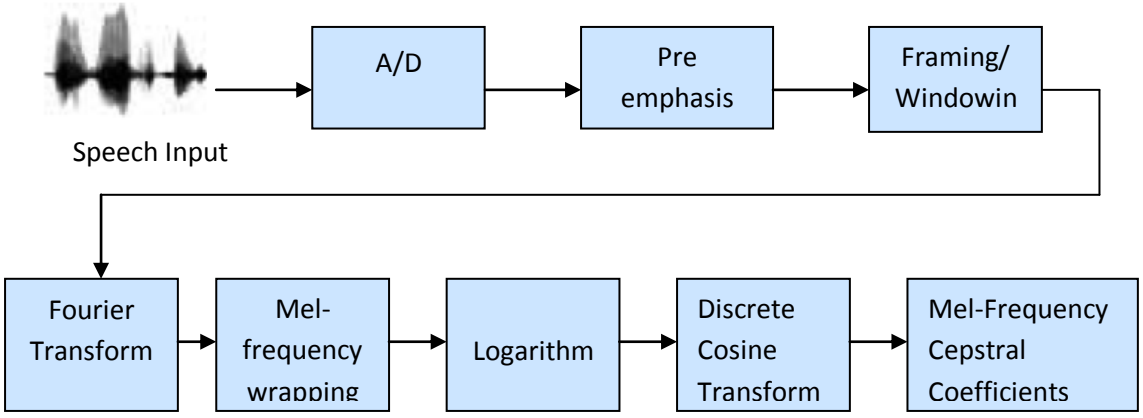

Figure 5.5 Block diagram of Mel-Frequency Cepstral
[31].

The main advantage of MFCC is the robustness towards noise and spectral estimation errors under various conditions [41]. A. Reynolds did a study on the comparison of different features and found that the MFCC provides better performance than other features [42].

5.2.3.1 Sampling

In sampling the data, the digitized audio signal is considered as formed of a set of discrete values on regular intervals, and thus one has to ensure that the sample rate is high enough so that there are sufficient points to characterize the waveform. Sampling should be at least twice the frequency of the waveform as indicated by Nyquist's theorem in (e.g. frequency of 4 kHz should be sampled at 8kHz). Common sampling rates are 8000, 11025, 22050 and 44000. Usually, 10 kHz and above are used.

5.2.3.2 Framing and windowing

Speech is a dynamic and non-stationary process, as the amplitude of the speech waveform varies with time due to variations in the vocal tract and articulators. However, speech analysis usually presumes that the statistical properties of the non-stationary speech process change relatively slowly over time. Although this assumption is not strictly valid, it makes it possible to process short-time speech frames, ranging typically from 10 ms to 40 ms, as a stationary process. Generally speaking, the use of short frame duration and overlapping frames is chosen to capture the rapid dynamics of the spectrum. Speech parameters are extracted on a frame-by-frame basis and the amount of overlap determines how quickly parameters can change from frame to frame. As shown in the Figure 5.8, the speech signal is slowly varying over time and it is called quasi- stationery. The framing process is shown in Figure 5.7.

The speech signal is slowly varying over time (quasi-stationary). When the signal is examined over a short period of time (5-100msec), the signal is fairly stationary. Therefore speech signals are often analyzed in short time segments, which are referred to as short-time spectral analysis. This practically means that the signal is blocked in frames of typically 20-30 msec. Adjacent frames typically overlap each other with 30-50%, this is done in order not to lose any information due to the windowing.

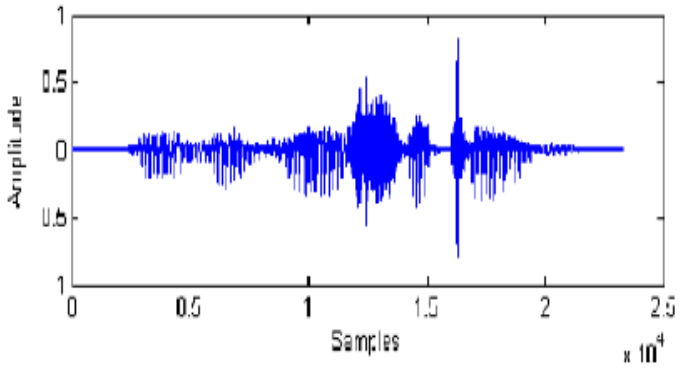

Figure 5.6 Speech signal varying over time (quasi- stationary). [1]

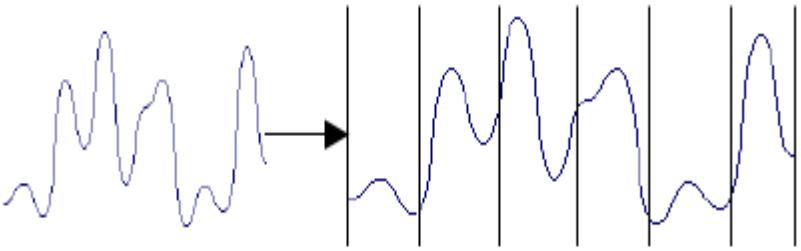

Figure 5.7 Framing the signal. [25]

After the signal has been framed, each frame is multiplied with a window function $w(n)$ with length N, where N is the length of the frame. Typically the Hamming window is used:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \leq n \leq N-1$$

The windowing is done to avoid problems due to truncation of the signal. Windowing means multiplication of a speech signal $s(n)$ by a window $w(n)$ to weight or favor samples by the shape and duration of the window. Coupled with overlapping short-term frames, successive windowing is equal to applying a sliding window to the long-term speech signal. The simplest window has a rectangular shape, weighting all samples of speech signal equally. In fact, not windowing segmented short duration frames at all is equivalent to applying a rectangular window.

Window duration determines the amount of averaging used in power or energy calculation. Window duration and frame duration can be adjusted as a pair. For instance, a frame duration of 20 ms can be coupled with a window duration of 30 ms. An alternative is to choose the window duration equal to the frame duration for simplicity.

5.2.3.3 Hamming window

Hamming window is also called the raised cosine window. The equation and plot for the Hamming window shown below (Figure 5.8). In a window function there is a zero valued outside of some chosen interval. For example, a function that is stable inside the interval and zero elsewhere is called a rectangular window. When signal or any other function is multiplied by a window function, the product is also zero valued outside the interval. Window function has some other applications such as spectral analysis, filter design, and audio data compression such as Vorbis.

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

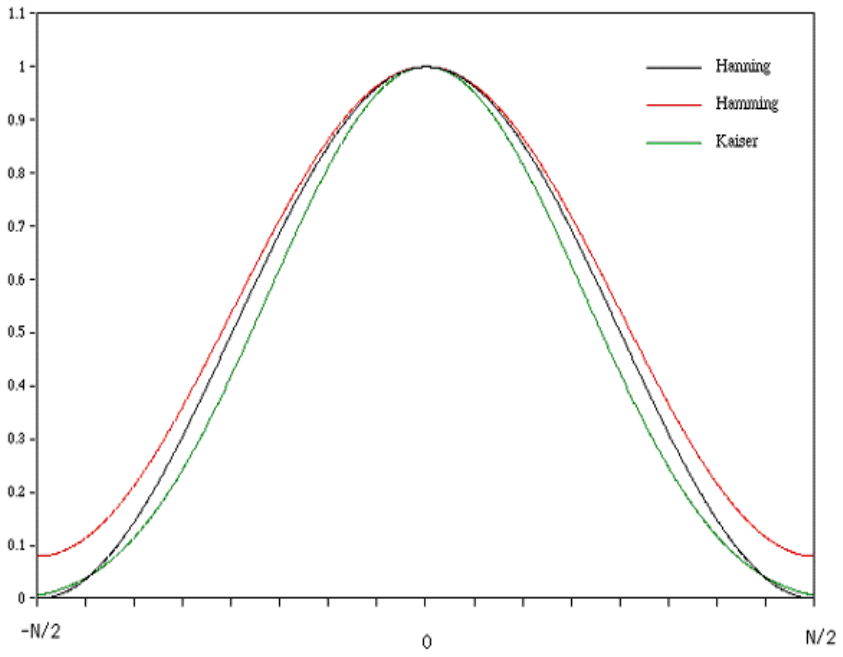

Figure 5.8 Hamming window. [1]

5.2.3.4 Fast Fourier Transform (FFT)

The Discrete Fourier Transform (DFT) or Fast Fourier Transform (FFT) is performed to find frequency components of a signal buried in a noisy time domain. As well, the original signal needs to be Fourier transformed to pass through a set of band-pass filters for the Mel frequency-warping process. Standard Fourier Transform is not used because the audio signal is not known over all time. DFT is therefore a much more usable frequency transformation and is essentially a Fourier representation of a sequence of samples of limited length.

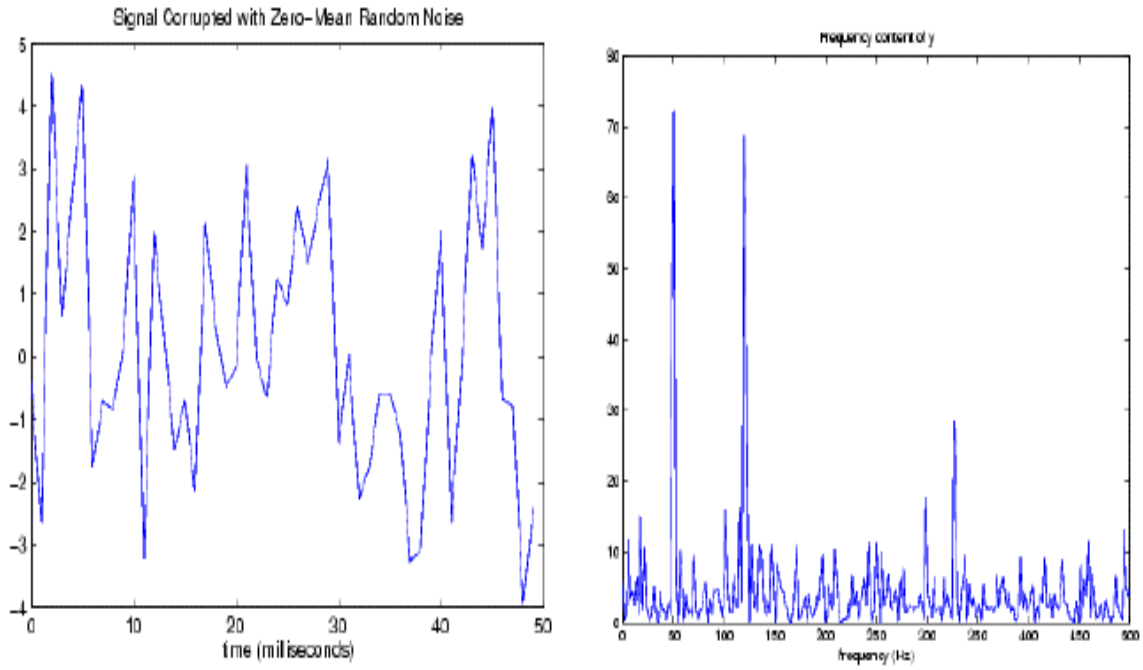

Figure 5.9 Time Domain Signal and its Equivalent Frequency Representation. [25].

5.2.3.5 Mel frequency warping

Mel Frequency Warping smoothes the spectrum and emphasizes perceptually meaningful frequencies. The Fourier Transformed signal is passed through a set of band-pass filters in order to simplify the spectrum without significant loss of data. This is achieved by collecting a number of spectral components into a number of frequency bins. The spectrum is simplified

because using a filterbank separates the spectrum into channels. Filters are spaced uniformly on a Mel Scale and logarithmically on a frequency scale, thus this implies that lower frequency channels are linearly spaced while higher frequency channels are logarithmically spaced.

This is ideal since human perception of audio frequency does not follow a linear scale. Therefore for each tone with an actual frequency f (Hz), a subjective pitch is measured on a scale called the 'mel' scale. The mel-frequency scale is a linearly spaced below 1 kHz and logarithmically spaced for frequencies above that. The pitch of 1 kHz tone is used as a reference point is defined as 1000 Mel's. This is also 40dB above the perceptual hearing threshold. The Mel scale can easily be converted from the frequency scale using the equation:

$$mel(f) = 2595" * log_{10}(1 + {}^{f}/_{700})$$

The subjective spectrum is stimulated using a filter bank spaced uniformly on the mel scale. Spacing of the filterbank is determined by a constant mel frequency interval. The modified spectrum S(ω) thus consists of the output power of these filters when S(ω) is the input. Since this filter bank is applied in the frequency domain, it can be regarded as taking points of the filter windows on the spectrum, where each filter can be viewed as a histogram bin in the frequency domain (Figure 5.10). For smaller frames it is best to use triangular or even rectangular filters because the resolution is too low for the lower frequencies

Each filter in the bank is multiplied by the spectrum so that only one single value of magnitude per filter is returned. This can be achieved through simple matrix operations. This reflects the sum of amplitudes in a particular filter band and thus reduces the precision to the level of human ear. Figure 5.11 shows the results. The x-axis represents the index of a filter and so follows the mel-scale.

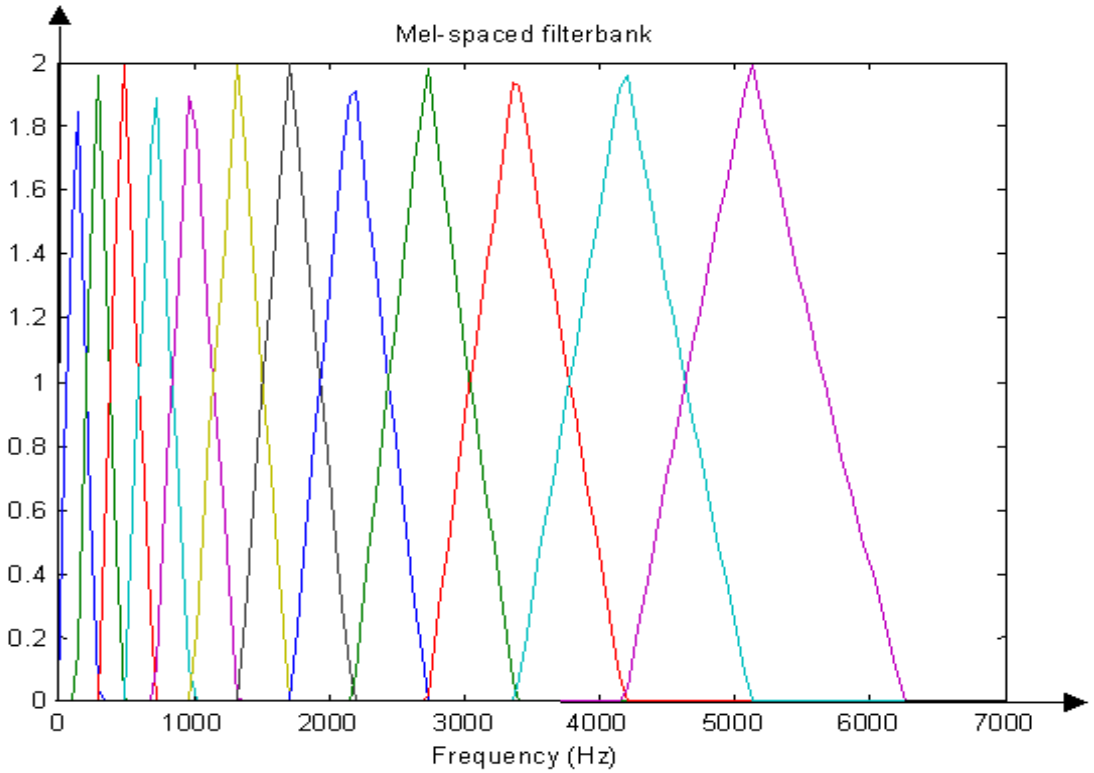

Figure 5.10 Mel Spaced FilterBank. [25]

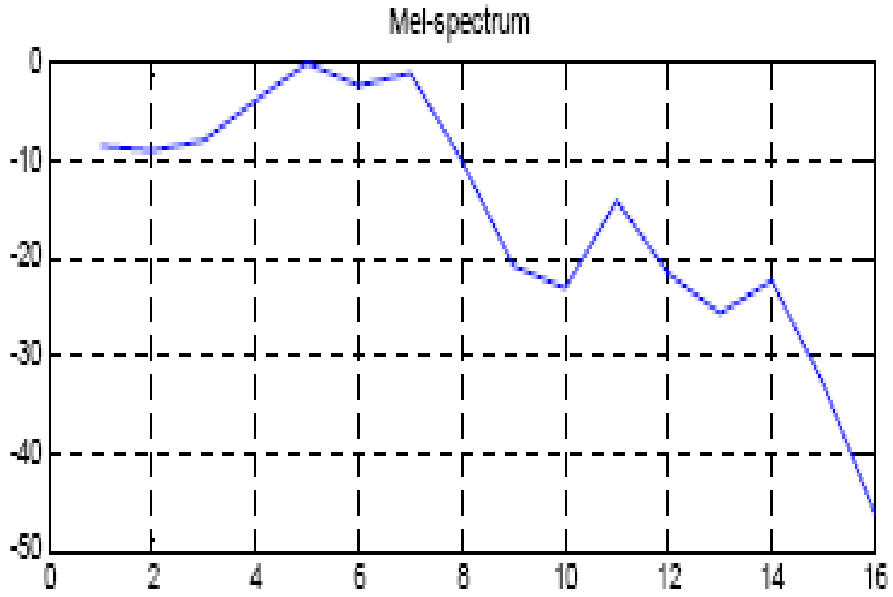

Figure 5.11 Mel Spectrum. [25]

The process of Mel Frequency Warping may be computed in three steps as shown below:

- The area under each filter is constant and sometimes scaled to 1. Let M = desired number of filter banks.
- Distribute these uniformly across the Mel frequency space
- Convert to Hz to get ωi's on linear scale. The relationship between mel and frequency is given by m = ln(1 + f/700) * 1000 / ln(1+1000/700)

5.2.3.6 Discrete Cosine Transform

The final stage performs the Discrete Cosine Transform to decorrelate the mel logarithmic magnitude spectrum to the mel frequency cepstral coefficients MFCC. The cepstrum is the inverse Fourier transform of the frequency spectrum of a signal in logarithmic amplitudes. It displays the ripples and "waveform" of spectral representation in terms of "quefrencies", the unit of which is a second. A common practical procedure is to replace the inverse Fourier with cosine transformation (DCT) since the log- power is real and symmetric so the inverse Fourier with cosine transformation reduces to a Discrete Cosine Transform. In addition, the DCT has the ability to produce more highly correlated feature and cepstral coefficients are more compact since they are sorted in variance order. Figure 5.14 shows Mel spectral vectors of highly correlated components decorrelated into 13 Mel Frequency Cepstral Coefficients.

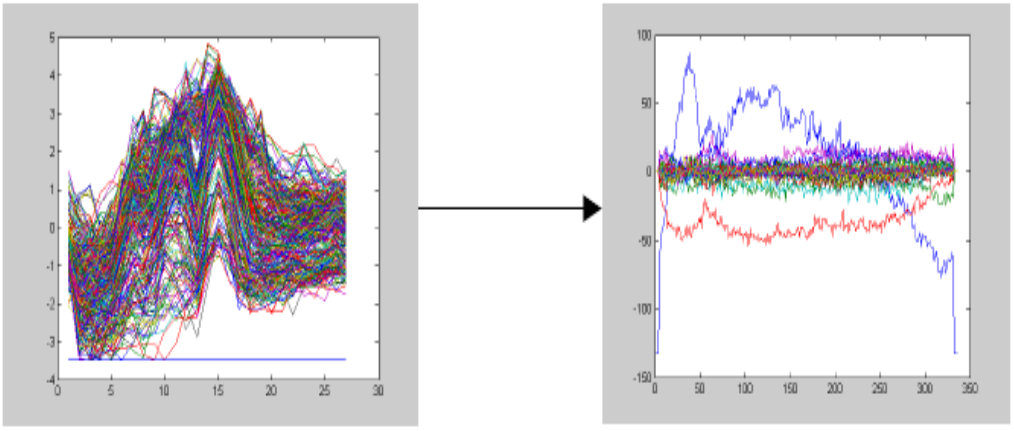

Figure 5.12 Highly Correlated Mel-Spectral Vectors Decorrelated into 13 MFCCs. [25]

The discrete form for a signal x(n) is defined in Equation 5.5 as

$$y(k) = w(k)\sum_{n=1}^{N} x(n)cos\frac{\pi(2n-1)(k-1)}{2N}, \quad k = 1, \ldots\ldots, N \tag{5.5}$$

Where

$$w(k) = \begin{cases} \sqrt{1/N}, & k = 1 \\ \sqrt{2/N}, & 2 \le k \le N \end{cases}$$

By performing DCT, the Mel Cepstrum is obtained and is shown in Figure 5.15. It can be seen that the 0^{th} coefficient C_0 has been excluded. This is because it represents the mean value of the input signal and carries little information. Beth Logan indicated that the zeroth cepstral coefficient contains only magnitude information.

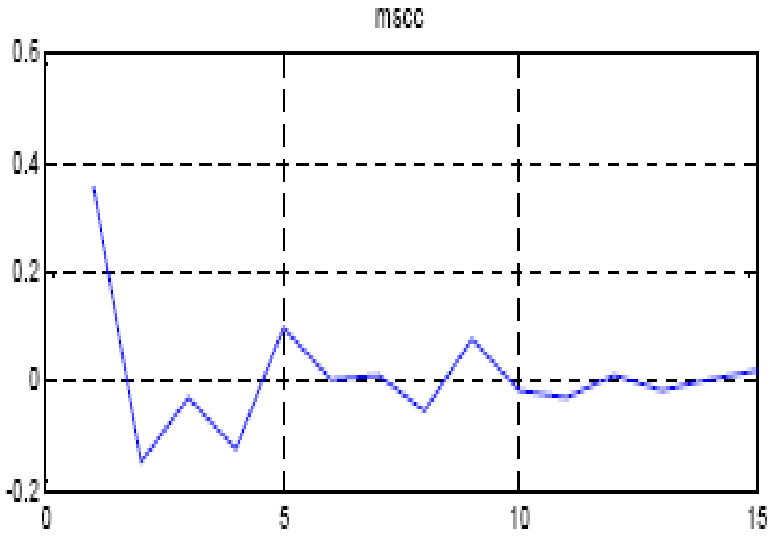

Figure 5.13 Mel Cepstrum. [25]

Observing Figure 5.13, we see that the coefficient amplitudes reduce at the higher frequencies.

5.2.3.7 Cepstrum

In this final step, we convert the log mel spectrum back to time. The result is called the mel frequency cepstrum coefficients (MFCC). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the mel spectrum coefficients (and so their logarithm) are real numbers, we can convert them to the time domain using the Discrete Cosine Transform (DCT). Therefore if we denote those mel power spectrum coefficients that are the result of the last step are $\tilde{S}_0, k = 0,2,\ldots,K-1$, we can calculate the MFCC's, $\tilde{c}_n$, as

$$\tilde{c}_n = \sum_{k=1}^{K} (\log \tilde{S}_k) \cos\left[n\left(k - \frac{1}{2} \right) \frac{\pi}{K} \right], \qquad n = 0,1,\ldots,K\text{-}1 \tag{5.6}$$

Note that we exclude the first component, $\tilde{c}_0$, from the DCT since it represents the mean value of the input signal, which carried little speaker specific information.

5.3 Cepstral Analysis

Among all popular speech parameters, the most functional and efficient ones extract spectral information (in the frequency domain) from speech, because a more concise and easier analysis of speech can be performed spectrally rather than temporarily (in the time domain). Although speech signals demonstrate a range of inter-speaker variations for the same utterance in the time domain, this utterance still exhibits consistency in the frequency domain, to some extent. For this reason, spectral analysis is preferred over temporal analysis to discriminate between phonemes and extract speaker-independent features from speech signal.[17]

Cepstral analysis is a special case of homomorphic signal processing [2]. A homomorphic system is defined as a nonlinear system whose output is a linear superposition of the input signals under a nonlinear transformation. Cepstral analysis has become popular in speech

recognition since its discovery in the late 1960s, due to the powerful yet simple engineering model of human speech-production behind it. According to this linear acoustic model, a speech signal is produced by filtering an excitation waveform through the vocal tract filter as depicted in Figure 5.14.

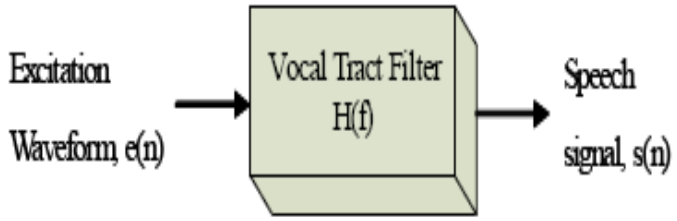

Figure 5.14 Linear Acoustic Model of Human Speech-Production [29].

In this model, the speech signal is expressed as the convolution of an excitation signal $e(n)$ with the vocal tract response $h(n)$. The excitation sequence is either a quasi periodic vocal cord pulse in the case of producing voiced speech or just random noise at the vocal tract constriction, which generates unvoiced speech. Homomorphic signal processing offers a fairly simple method, known as *cepstral deconvolution,* to decouple the vocal tract response from the excitation response, thereby enabling it to model the vocal tract characteristics better. The decomposition of a speech signal $s(n)$ into the excitation sequence $e(n)$ and the vocal tract function $h(n)$ can be described as follows:

$$s(n) = e(n) \otimes h(n) \tag{5.1}$$

Where the operator, "$\otimes$ " represents the convolution operation. Recall that the convolution operation in time corresponds to a multiplication in the frequency domain. Thus, equation becomes:

$$S(f) = E(f) \cdot H(f). \tag{5.2}$$

Note that the complex speech spectrum $S(f)$ is composed of a quickly varying part, excitation spectrum $E(f)$ (which corresponds to high frequency components) and a slowly-varying part, vocal tract response $H(f)$ (which corresponds to low frequency components). Considering that the speech signal is real-valued, the logarithm of equation (5.2) on both sides leads to:

$$\log(|S(f)|) = \log(|E(f).H(f)|) = \log(|E(f)|) + \log(|H(f)|). \qquad (5.3)$$

Now that the signal components in Eq. (5.3) are linearly combined, a linear filter (also known as liftering *operation* in speech engineering terminology) can be applied to remove the noise-like, quickly-varying excitation part from the speech spectrum. Then, the inverse Fourier transform is applied to the remaining component to compute the *real cepstrum*. In short, under a cepstral transformation, the non-linear convolution of two signals $e(n) \otimes h(n)$ becomes equivalent to the linear sum of the cepstral representations of the signals $C_e(S) + C_h(S)$. As a result, the real cepstrum is the inverse Fourier transform of the logarithm of the power spectrum of a speech signal:

$$C_s(n) = \frac{1}{N}\sum_{k=0}^{N-1} \log|S(k)| e^{j2\pi kn/N} \quad \text{, for n = 0,1,…….,N-1.} \qquad (5.4)$$

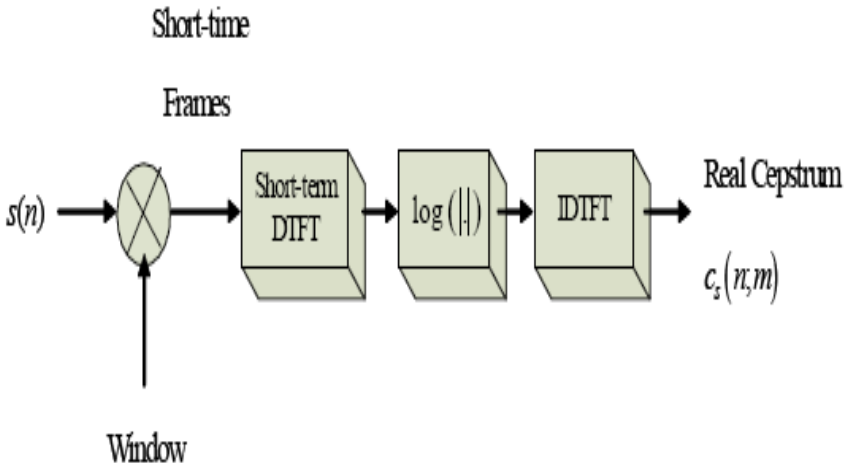

Figure 5.15 A block diagram representation of the short-term real cepstrum Computation [17]

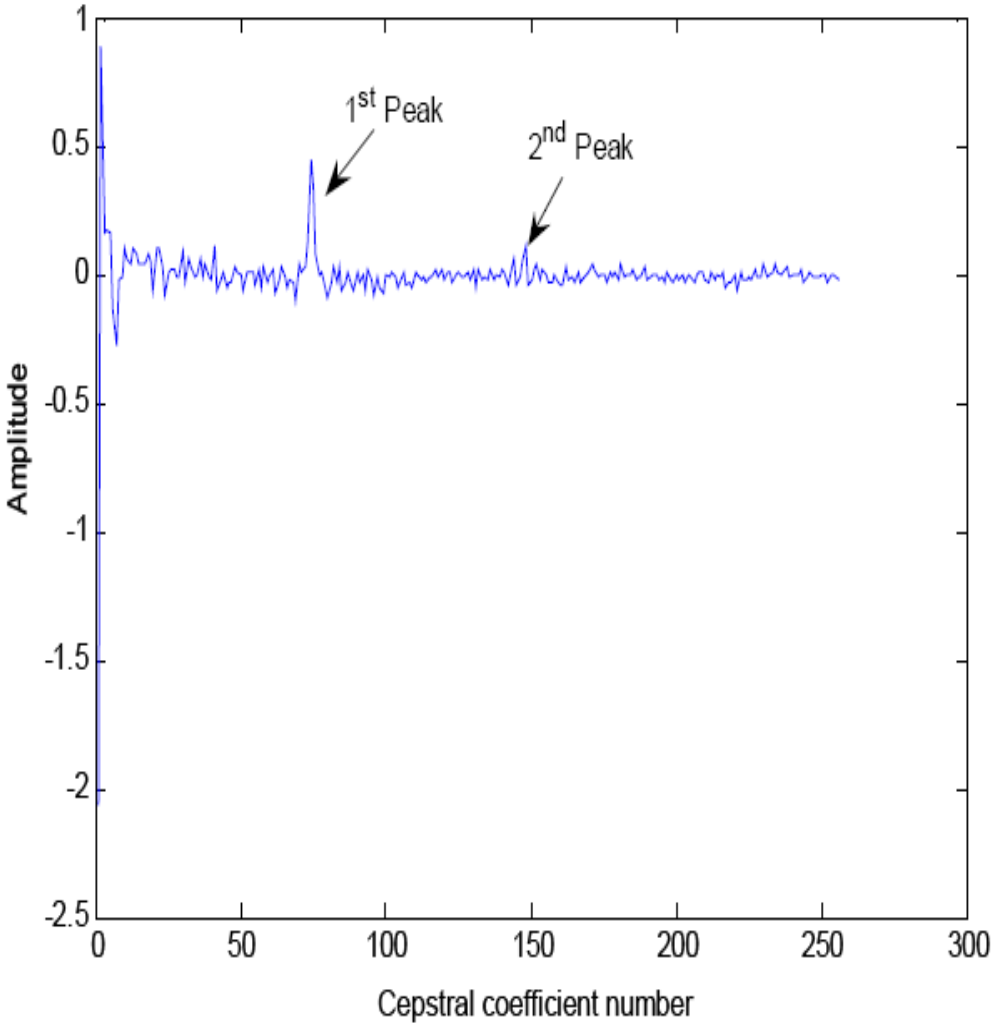

Figure 5.16 The real cepstrum computed for the voiced phoneme, /ae/ in the word "pan." [24]

Figure 5.16 illustrates the fact that the dominant contents of the cepstra are located near the origin, and that a small number of cepstrum coefficients can be used to provide enough spectral information about the phoneme /ae/. In addition, pitch period information may be extracted by the spacing between successive cepstral peak locations (in this case, the pitch period is about 75 samples or equivalently 9.375 ms). Figure 5.17 plots the first 20 coefficients of the real cepstrum computed above, which shows that the first 8-10 coefficients carry most of the spectral information about the voiced phoneme /ae/.

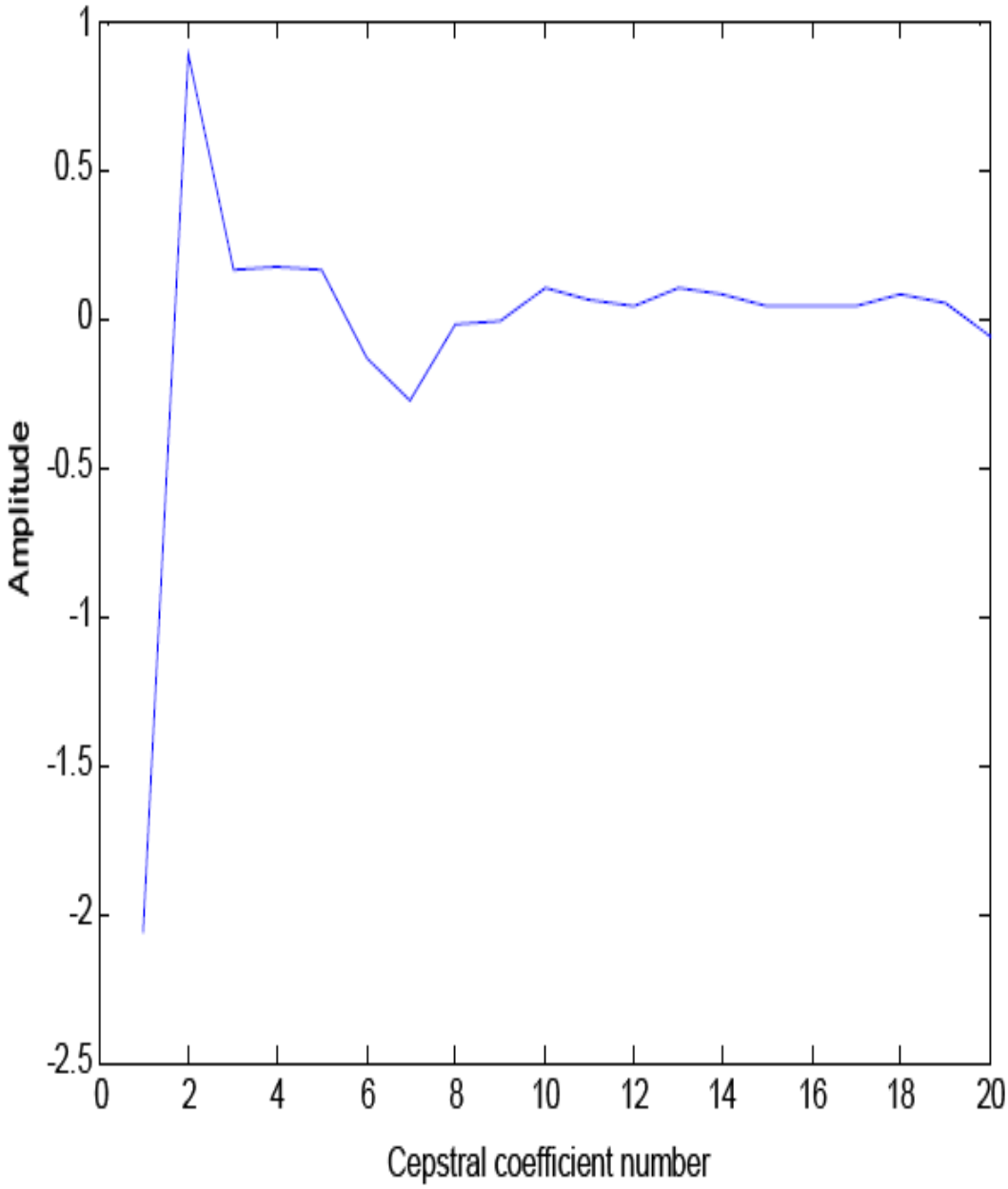

Figure 5.17 The First 20 Coefficients of the Real Cepstrum for the Phoneme /ae/. [24]

5.4 Summary of Feature Extraction Techniques

A summary of the feature extraction technique is compiled in table 5.1 that compares the techniques mentioned in this report in terms of filtering, relevant variables, inputs and corresponding outputs

Table 5.1 Comparison of features extraction in terms of filtering techniques [31]

Process	Technique	Type of Filter	Relevant variables/Data structure	Output
Feature Extraction	Linear Predictive Coding (LPC)	All Pole Filter	Statistical Features Linear Predictive Coefficients	Linear Predictive Coefficients (LPC)
	Linear Predictive Cepstral Coefficients	All Pole Filter	Statistical Features Linear Predictive Cepstral Coefficients	Linear Predictive Cepstral Coefficients (LPCC)
	Mel-Frequency Cepstral Coefficient (MFCC)	Mel-Filter Bank	Statistical Features Mel-Frequency Cepstral Coefficients	Mel-Frequency Cepstral Coefficients (MFCC)

Table 5.2 Comparison of criteria of feature extraction techniques [31]

Criteria	**LPC**	**LPCC**	**MFCC**
Main Task	Features extracted by analysing past speech samples.	Features extracted by combining LPC with spectral analysis	Features extracted based on frequency domain using Mel-scale that represents human hearing
Speaker Dependence	High Speaker dependent	High Speaker dependent	Moderate Speaker dependent
Robustness	Poor	Poor	Good
Motivation Representation	Speech production motivated representation	Speech production motivated representation	Perceptually motivated representation
Filter Bank	All-Pole Filters	All-Pole Filters	Triangular Mel Filters
Typical Applications	Speech compression	Speaker and speech recognition	Speaker and speech recognition

5.5 Summary

In this Chapter the speech feature extraction and vector quantization techniques have been described, and the steps in generating the MFCC coefficients (Framing, Windowing, Fast Fourier Transform, Mel Frequency Wrapping, Discrete Cosine Transform) have been outline.