Chapter 1. Introduction

Speech recognition is the ability of machine to identify "what is spoken by particular speaker".A speech recognition system consists of a microphone, for the person to speak into; speech recognition program; a computer to take and interpret the speech; a good quality sound card for input and/or output; a proper and good pronunciation. Automatic speech recognition systems involve numerous separate components drawn from many different disciplines such as statistical pattern recognition, communication theory, signal processing, combinatorial mathematics, and linguistics.

1.1 Overview of speech recognition:

A speech recognition system consists of five blocks: - Speech Capture Device,DSP module,Pre-processed signal storage,Reference speech pattern,Pattern matching algorithm. A speech capture device generally consist of a microphone and associated analog-to-digital converter which digitally encodes the raw speech waveform module performs endpoint detection to seperate speech from non-speech,converts raw waveform into frequency domain representation and performs further windowing, scaling, filtering and data compression. The goal is to enhance and retain only those components of spectral representation that are useful for recognition purposes. Thereafter preprocessed speech is buffered for recognition algorithm. Stored reference patterns can be matched against the user's speech sample once it has been preprocessed by the DSP module. This information is stored as a set of speech templates or as generative speech models. Pattern matching algorithm must compute a measure of goodness-of-fit between preprocessed signal and all the stored templates.

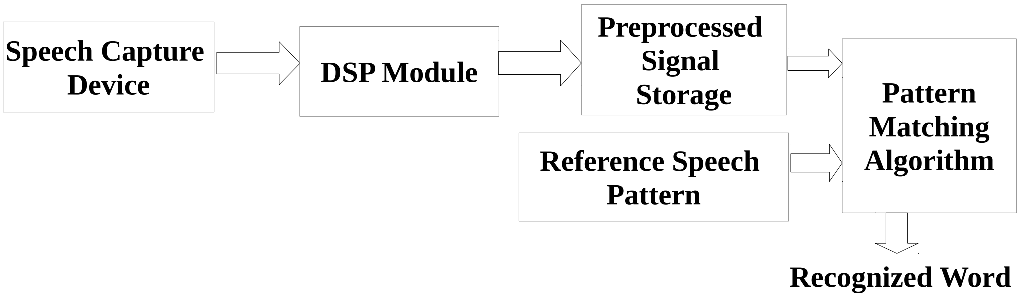


Fig. 1.1 Components of speech recognition system

1.2 Classification of speech recognition systems:

The classification of speech recognizers is used for determining the complexity of speech recognition task accomplished by speech recognizer. Complexity of speech recognizer affects the performance measures like accuracy and speed.

1.2.1 Depending on size of vocabulary:

a) Small vocabulary:

It is clear that smaller the vocabulary size , higher the recognition accuracy and time for getting the result is also small.(speed factor)

b) Large vocabulary:

Opposite is true for large vocabulary.

1.2.2 Depending on particular speaker:

a) Speaker dependent:

Systems that require a user to train the system according to his or her voice.

b) Speaker independent:

Systems that do not require a user to train the system i.e. they are developed to operate for any speaker.

1.2.3 Depending on continuousness of speech that they recognize:

a) Isolated word recognizer:

Accept one word at a time. These recognition systems allow us to speak naturally continuous.

b) Connected word recognizer:

Allow speaker to speak slowly and distinctly each word with a short pause i.e. planned speech.

c) Continuous word recognizer:

Continuous speech recognizers allow users to speak almost naturally, while the computer determines the content.

1.2.4 Spontaneous recognition system

Spontaneous recognition systems allow us to speak spontaneously.

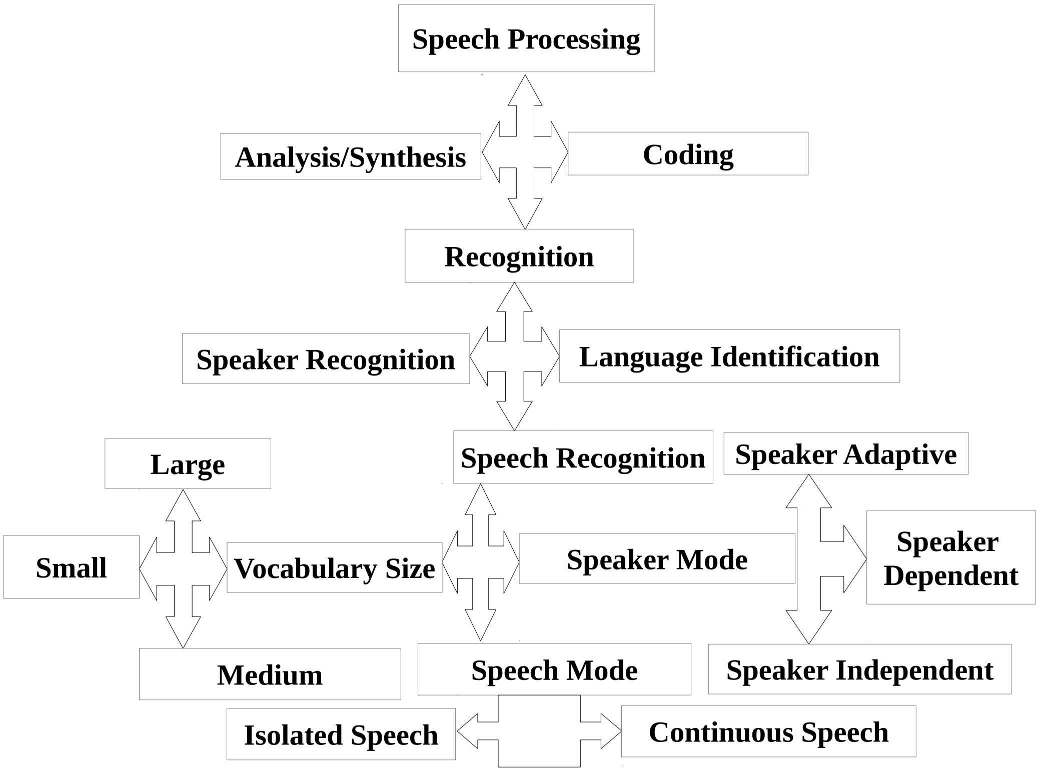


Fig. 1.2 Classification of speech recognition system

1.3 Approaches to speech recognition:

There are various approaches to recognize the speech & given as follows:

1.3.1 Acoustic phonetic approach:

This is earliest approach to speech recognition. In acoustic phonetic approach, input speech signal is first transformed into the spectral form using various transformation techniques such as fourier transform,short time fourier transform,continuous wavelet transform or discrete wavelet transform. After the transformation,some spectral features that represent acoustical properties of signal are extracted. Finally,the speech recognizer tries to find best match from the reference speech pattern[7][8].

1.3.2 Pattern recognition approach:

Patten recognition approach works in two stages:first one is training of speech pattern and second is recognition of pattern by comparison. In training phase,the speech signal is represented by it's feature that forms the test pattern. In comparison phase,according to goodness of fit between the test pattern and reference pattern,the best match is found[7][8].

a) Template based approach:

In template based approach,test pattern T is formed that represents the speech characteristics. After that using some distance formula like in dynamic time warping method,distance between each component of T and each component of reference pattern R is calculated. This distance calculation is done with each refernce pattern in speech database. The minimum distance between test pattern and reference pattern decides the best match. Dynamic time warping (see chapter number-7) and vector quantization[7][8] are the two methods based on template approach.

b) Stochastic approach:

Stochastic approach tries to characterize the statistical properties of signal. Stochastic approach is based on assumption that signal can be characterized as a parametric random process and parameters of stochastic process can be determined in well defined manner. Hidden Markov model is one of the best stochastic approach[9][10].

1.3.3 Artificial intelligence approach (Knowledge based approach) :

The Artificial Intelligence approach is a hybrid of the acoustic phonetic approach and pattern recognition approach. In the AI, an expert system implemented by neural networks is used to classify sounds. The basic idea is to compile and incorporate knowledge from a variety of knowledge sources with the problem at hand[7].

1.3.4 Connectionist approach (Artificial neural networks):

Among the techniques used within this class of methods are use of an expert system (e.g., a neural network) that integrates phonemic, lexical, syntactic, semantic, and even pragmatic knowledge for segmentation and labelling, and uses tools such as artificial NEURAL NETWORKS for learning the relationships among phonetic events. The focus in this approach has been mostly in the representation of knowledge and integration of knowledge sources. In connectionist models, knowledge or constraints are not encoded in individual units, rules, or procedures, but distributed across many simple computing units[7].

1.4 Classifier & support vector machine(SVM):

In practice, the choice of a classifier is a difficult problem and it is often based on which classifier(s) happen to be available, or best known, to the user. One of the powerful tools for pattern recognition that uses a discriminative approach is a SVM. SVMs use linear and non-linear separating hyper-planes for data classification. However,since SVMs can only classify fixed length data vectors, this method cannot be readily applied to task involving variable length data classification.

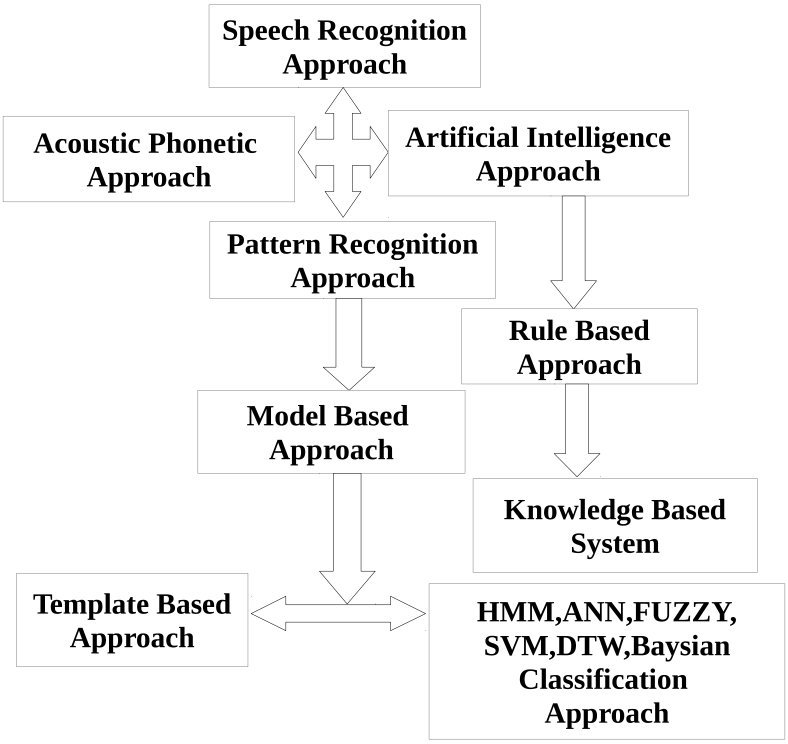

Speech Recognition Approach
Acoustic Phonetic Approach
Artificial Intelligence Approach
Pattern Recognition Approach
Rule Based Approach
Model Based Approach
Knowledge Based System
Template Based Approach
HMM,ANN,FUZZY, SVM,DTW,Baysian Classification Approach

Fig. 1.3 Taxonomy of speech recognition approaches

1.5 Feature Extraction:

In speech recognition, the main goal of the feature extraction step is to compute a parsimonious sequence of feature vectors providing a compact representation of the given input signal. Various methods for speech recognition are broadly classified in following table:

Method	***Property***	***Comments***
Principle Component Analysis	Non-linear feature extraction method, Linear map; fast; eigenvector-based	Traditional,eigenvector method, also known as karhuneu-Loeve expansion; good for Gaussian data.
Linear Discriminant Analysis(LDA)	Non-linear feature extraction method,	Better than PCA for classification;

	Supervised linear map; fast; eigenvector-based	
Independent Component Analysis (ICA)	Non linear feature extraction method, Linear map, iterative non-Gaussian	Blind course separation, used for de-mixing non- Gaussian distributed sources(features)
Linear Predictive coding(LPC)	Static feature extraction method,10 to 16 lower order coefficient,	
Cepstral Analysis	Static feature extraction method, Power spectrum	Used to represent spectral envelope
Mel-frequency scale analysis	Static feature extraction method, Spectral analysis	Spectral analysis is done with a fixed resolution along a subjective frequency scale i.e. Mel-frequency scale.
Filter bank analysis	Filters tuned required frequencies	
Mel-frequency cepstrum (MFFCs)	Power spectrum is computed by performing Fourier Analysis	
Kernel based feature extraction method	Non linear transformations,	Dimensionality reduction leads to better classification and it is used to remove noisy and redundant features, and improvement in classification error
Wavelet	Better time resolution than Fourier Transform	It replaces the fixed bandwidth of Fourier transform with one proportional to frequency which allow better time resolution at high frequencies than Fourier Transform
Dynamic feature extractions i)LPC ii)MFCCs	Acceleration and delta coefficients i.e. II and III order derivatives of normal LPC and MFCCs coefficients	
Spectral subtraction	Robust Feature extraction method	
Cepstral mean subtraction	Robust Feature extraction	
RASTA filtering	For Noisy speech	
Integrated Phoneme subspace method	A transformation based on PCA+LDA+ICA	Higher Accuracy than the existing methods

Table 1.1 Feature extraction methods

1.6 Performance measures:

The performance of speech recognition systems is usually specified in terms of accuracy and speed. Accuracy may be measured in terms of performance accuracy which is usually rated with word error rate (WER), whereas speed is measured with the real time factor. Other measures of accuracy include Single Word Error Rate (SWER) and Command Success Rate (CSR).

1.6.1 Word error rate(WER):

Word error rate is a common metric of the performance of a speech recognition or machine translation system. Word error rate can then be computed as:

$$\textbf{WER} = \frac{S+D+I}{N}$$

where S is the number of substitutions,

D is the number of the deletions,

I is the number of the insertions,

N is the number of words in the reference.

1.6.2 Word recognition rate:

When reporting the performance of a speech recognition system, sometimes word recognition rate (WRR) is used instead:

$$\textbf{WRR} = \textbf{1} - \textbf{WER} = \frac{N-S-D-I}{N} = \frac{H-I}{N}$$

where H is N-(S+D), the number of correctly recognized words.

Chapter 2. Literature Survey (Year Wise)

2.1 Important milestones in speech recognition:

Sr. No.	***Year***	***Description***
1920-1960 [7]		
1	1920	Machine Recognition came into existence
2	1950	ASR by machine
3	1952	A system for isolated digit recognition for a single speaker.
4	1956	Olson & Belar tried to recognize 10 distinct syllables of a single talker
5	1959	Fried & Denes tried to build a phoneme recognizer
1960-1970 [7]		
6	1960	Hardware vowel recognizer
7	1962	Hardware phoneme recognizer
1970-1980 [7]		
8	1970	Isolated word and discrete utterance recognizer,use of dynamic programming and linear predictive coding,large vocabulary speech recognizer by IBM
9	1973	CMU's Heresay I & Heresay II system ,Harphy system able to recognize speech using vocabulary of 1011 words,BBN's HWIM (here what I mean) system
1980-1990 [7]		
10	1980	Key focus on connected word recognizer,A feature-based speech recognition approach by Moshey J. Lasry,use of statistical approach such as HMM.
11	1990	Speech recognizer with significant improvements in statistical approach
1990-2000 [7]		
12	1990	Pattern recognition approach,Instead of using Bayes decision theory,problem is transformed into the optimization problem to reduce empirical recognition error. Bayes Decision theory is rejected due to the fact that distribution function for speech signal could not be chosen accurately
2000-2009 [7]		
13	2000	Use of Variational Bayesian estimation and clustering techniques
14	2005	Large vocabulary continuous speech recognition
15	2007	Analysis of results of spontaneous speech recognition & read speech for large vocabulary,Method for fourier transform phase for feature extraction(Rajesh M. Hegde)
16	2008	Application of conditional random field

Table 2.1 Milestones in speech recognition

2.2 Summary of the technology progress:

In the last 60 years, especially in the last three decades, research in speech recognition has been intensively carried out world wide, spurred on by advances in signal processing algorithms, architectures and hardware. The technological progress in the 60 years can be summarized in the table :

Sr. No.	***Past***	***Present(new)***
1	Template matching	Corpus-based statistical modelling, e.g. HMM and n grams
2	Filter bank/spectral resonance	Cepstral features, Kernel based function, group delay functions
3	Heuristic time normalization	DTW/DP matching
4	Distance-based methods	Likelihood based methods
5	Maximum likelihood approach	Discriminative approach e.g. MCE/GPD and MMI
6	Isolated word recognition	Isolated word recognition
7	Small vocabulary	Large vocabulary
8	Context Independent units	Context dependent units
9	Clean speech recognition	Noisy/telephone speech recognition
10	Single speaker recognition	Speaker-independent/adaptive recognition
11	Monologue recognition	Dialogue/Conversation recognition
12	Read speech recognition	Spontaneous speech recognition
13	Single modality(audio signal only)	Multi-modal (audio/visual) speech recognition
14	Hardware recognizer	Software recognizer
15	Speech signal is assumed as quasi stationary in the traditional approaches. The feature vectors are extracted using FFT and wavelet methods etc.	Data driven approach does not posses this assumption i.e. signal is treated as non-linear and non-stationary. In this features are extracted using Hilbert Haung Transform using IMFs

Table 2.2 Summary of technological progress in last 60 years

Chapter 3. System Requirement Specification

Complete software as well as hardware requirement is essential for success of any system development process. No matter how well coded,the program ,if poorly analysed then specified program will disappoint the user and bring woe to the developer. The system requirement task is process of discovery, refinement, modelling and specification.

System requirement analysis results in specifications of software's/hardware's operational characteristics; indicates software's/hardware's interface with other system elements and establish the constraints that software/hardware must meet. System requirement analysis allows software engineer or analyst or modeller to study the basic requirements that must be necessary to initiate the development process.

In speech recognition process,following software as well as hardware requirements must be taken into consideration:

3.1 Software requirements:

Software	***Version***
MATLAB with DSP toolbox	R2012a

3.2 Hardware requirements:

Expected	***Available***	
200 megahertz Pentium processor	Intel® Core™ i3 CPU M 380 @ 2.53GHz ×4	
A minimum of 64 megabytes of RAM	RAM-3.7 GB	
A minimum of a 16-bit sound card	Intel High Definition Audio Device(Sound Card) 64-bit	
A basic microphone having properties stated in second column	Frequency Response	100Hz-16k Hz
	Sensitivity	-58dB +/- 3dB
	Directivity	Omni-Directional
	S/N Ratio	More Than 60dB
	Impedance	2.2k ohms

Chapter 4. Problem Definition

4.1 Problem statement:

Speech Recognition is the process that allows humans to communicate with computer by speech. So,here problem is determining “what is spoken by particular user”

input speech
Speech Recognizer
recognized speech

Fig. 4.1 Speech recognition

4.2 Steps in speech recognition:

Speech Recognition mainly consist of three step procedure:

1) Speech Acquisition & preprocessing
2) Feature Extraction
3) Classification/Pattern Recognition

4.2.1 Speech acquisition:

Speech acquisition is a process of capturing speech through microphone. It includes converting the acoustical signal to some computer readable code. This process also called as Digital Recording. This part is studied in next chapter (chapter number-5) namely Speech Acquisition in detail. Output of this stage is digital representation of speech.

4.2.2 Feature extraction:

Speech signal is a non stationary signal. Therefore it possess different characteristics which changes person by person. Feature extraction is process of extracting these characteristics of speech. In feature extraction process,we extracts feature like pitch (frequency of fundamental of harmonic series),Timbre (relative heights of peaks in power spectrum compared to fundamental),Loudness(overall heights of peak in power spectrum) and so on. The way used to extract features from speech is studied in detail in upcoming chapter namely Feature Extraction(chapter number-6).Output of this stage is a feature vector or simply features.

4.2.3 Classification & pattern recognition:

Pattern recognition is based on two types of classification:1) supervised classification 2) unsupervised classification. In supervised classification, data is classified into predefined classes and in unsupervised classification the aim is to cluster the data at the input. Speech recognition needs supervised classification since the collected data is classified into phonetic classes or word classes. DTW is used for this task.

Chapter 5. Speech Acquisition & Pre-processing

Speech acquisition includes converting acoustic signal to some computer readable codes. This process can also be referred as digital recording. In order to understand the process of digital recording,some concepts of digital signal processing should be understood. These concepts are the pre-requisites for digital recording.

Speech Acquisition Flow-Chart:

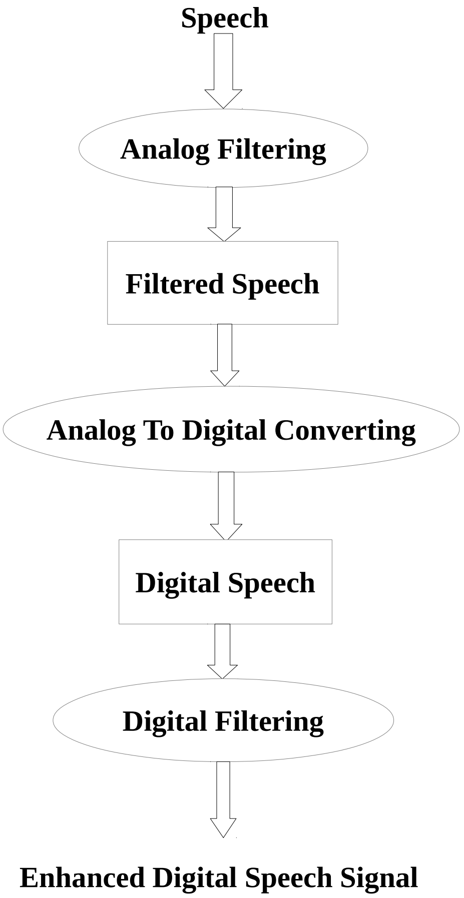


Fig. 5.1 Speech Acquisition by computer for speech recognition

5.1 Pre-requisites for digital recording:

Most signals of practical interest,such as speech,biological signal,seismic signals,radar signals,sonar signals,and various communication signals such as audio and video signals, are analog. To analyze analog signals by digital means,it is necessary to convert them into digital form that is,to convert them to a sequence of numbers having finite precision. This procedure is called as analog-to-digital conversion and the corresponding device is called as analog to digital converter (A/D).A/D conversion is three step procedure:

1) Sampling
2) Quantization
3) Coding

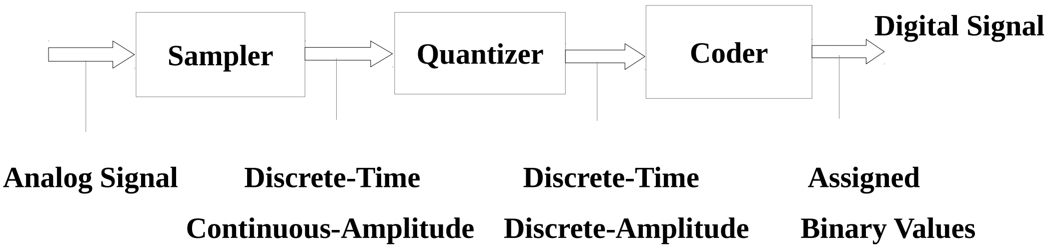


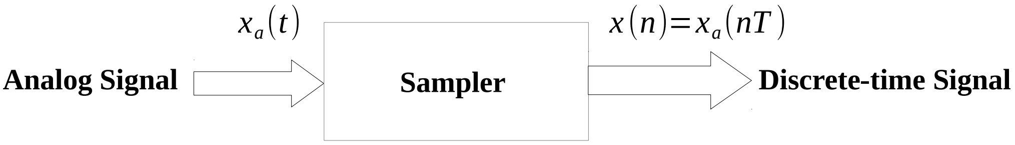
Fig. 5.2 Basic parts of A/D converter

5.1.1 Sampling:

This is the conversion of a continuous-time signal into discrete-time signal obtained by taking “samples” of continuous-time signal at discrete-time instants. There are many ways to sample an analog signal. Periodic or uniform sampling is used generally. This is described by following relation:

$$x(n)=x_a(nT)$$

where x(n) is the discrete-time signal obtained by “taking samples” of analog signal $x_a(t)$ every T seconds. The time interval T between successive samples is called the sampling period or sampling interval and it's reciprocal $\frac{1}{T}$ = F_s is called the sampling rate (samples per second) or the sampling frequency (hertz).



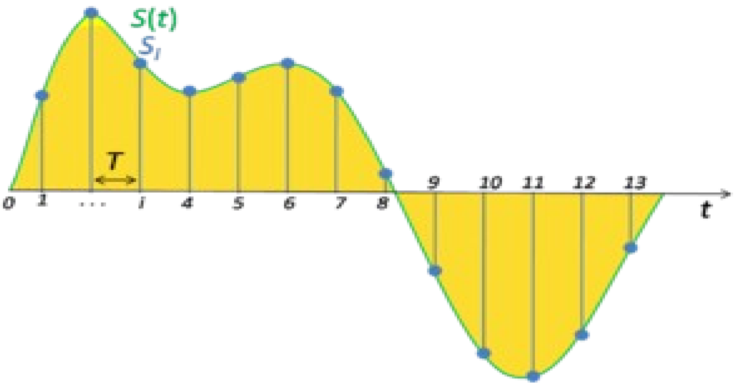


Fig. 5.3 Periodic sampling of an analog signal

5.1.2 Quantization:

This is conversion of a discrete-time continuous-amplitude signal into a discrete-time,discrete-amplitude signal. The value of each signal sample is represented by a value selected from a finite set of possible values. The difference between the unquantized sample $x(n)$ and the quantized output $x_q(n)$ is called the quantization error.

5.1.3 Coding:

In coding process,each discrete value $x_q(n)$ is represented by a b-bit binary sequence.

5.1.4 Sampling Theorem:

In many cases of practical interest ,it is desirable to convert the processed digital signal into analog form. The process of converting a digital signal into an analog signal is known as digital-to-analog (D/A) conversion. All D/A converters "connect the dots" in digital signal by performing some kind of interpolation, whose accuracy depends on quality of D/A converter. For signals having limited frequency content,the sampling theorem is introduced in order to state necessary and sufficient condition while reconstructing analog signal from digital signal. In principle,the analog signal can be reconstructed from the samples,provided that the sampling rate is sufficient high to avoid the problem commonly called aliasing[1][6].

Sampling Theorem Statement:

"If the highest frequency contained in analog signal $x_a(t)$ is $F_{max}=B$ and the signal is sampled at a rate $F_s > 2\mathrm{F}_{max}=2\mathrm{B}$,then $x_a(t)$ can be exactly recovered from it's sample values using the interpolation function:

$$g(t)=\frac{\sin(2*pi*B*t)}{2*pi*B*t}$$

Thus, $x_a(t)$ may be expressed as:

$$x_a(t)=\sum_{n=-\infty}^{\infty} x_a(\frac{n}{F_s})g(t-(\frac{n}{F_s}))$$

where $x_a(\frac{n}{F_s})=x_a(n*T)=x(n)$ are the samples of $x_a(t)$

When the sampling of $x_a(t)$ is performed at the minimum sampling rate $F_s=2\mathrm{B}$,the reconstruction formula becomes:

$$x_a(t)=\sum_{n=-\infty}^{\infty} x_a(\frac{n}{2\mathrm{B}})\sin(\frac{2*pi*B(t-n/2\mathrm{B})}{2*pi*B(t-n/2\mathrm{B})})$$

The sampling rate $F_N=2\mathrm{B}=2\mathrm{F}_{max}$ is called as Nyquist Rate.

5.2 Speech Acquisition:

Sounds are the result of vibrations of objects. For example-the human vocal chords. In general, without the influence of a specific sound vibration, air molecules move around randomly. A vibrating object pushes against the randomly-moving air molecules in the vicinity of the vibrating object, causing them first to crowd together and then to move apart. The alternate crowding together and moving apart of these molecules in turn affects the surrounding air pressure. The air pressure around the vibrating object rises and falls in a regular pattern, and this fluctuation of air pressure, propagated outward, is what we hear as sound.

Microphone used in this project possess properties like dynamic,omni-directional. Dynamic microphone use the principle of electromagnetic induction. It work on the moving- coil principle. They contain a diaphragm that is fixed to a moving coil. The coil is positioned in a static magnetic field generated by a permanent magnet. As stated in first paragraph,sound waves causes air fluctuation. When these fluctuation hit the microphone,they set up vibrations in the diaphragm,which are transferred to the coil. The movement of the coil in the magnetic field induces a signal voltage that is proportional to the incident sound. In this way,sound is captured by microphone and get stored in computer.

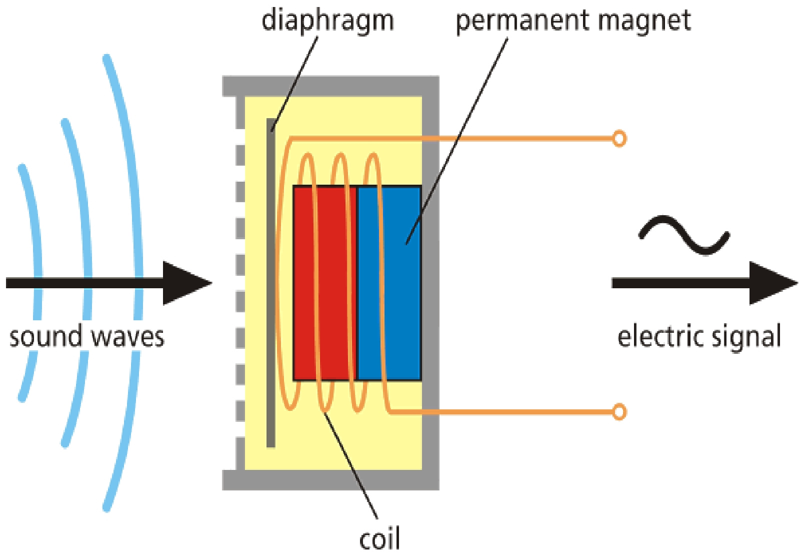


Fig. 5.4 Architecture of dynamic microphone

Our microphone has frequency response 100Hz-16kHz.Therefore,it only pick-ups the frequency between the range of 100Hz to 16 kHz. While picking the frequency in this range,it pick up them along with some redundant frequency components. These frequency components are considered as noise. Some of those frequencies can be filtered. Generally,filters are used to modify the magnitude of signals as a function of frequency. Desirable signals in one range of frequencies (usually called as band , here 100Hz-16kHz components) are passed essentially unchanged ,while unwanted signals (noise) in another band are reduced or attenuated. Generally, analog filtering part is integrated into the microphone. In order to improve the quality of signal,digital filtering is performed which comes under the pre-processing part.

5.3 Pre-processing:

5.3.1 Digital filtering:

Filters are usually classified according to their frequency-domain characteristics as low-pass,high-pass,band-pass,and band-stop or band-elimination filters.

a) Low pass filter:

Ideal magnitude response characteristic of low-pass filter is shown in figure 5.5.It only passes the lower frequency components and rejects higher frequency components.

b) High pass filter:

Ideal magnitude response characteristic of high-pass filter is shown in figure 5.5.It cut-offs the lower frequency components and allows higher frequency components to pass.

c) Band pass filter:

Ideal magnitude response characteristic of band-pass filter is shown in figure 5.5.It passes the frequency components in particular limit. It rejects lower as well as higher frequency components and allows to pass frequency components between lower and higher frequencies.

d) Band stop filter:

Ideal magnitude response characteristic of band-stop filter is shown in figure 5.5.It rejects the frequencies specified in particular limit.

e) All pass filter:

Ideal magnitude response characteristic of all-pass filter is shown in figure. It allows to pass all the frequency components.

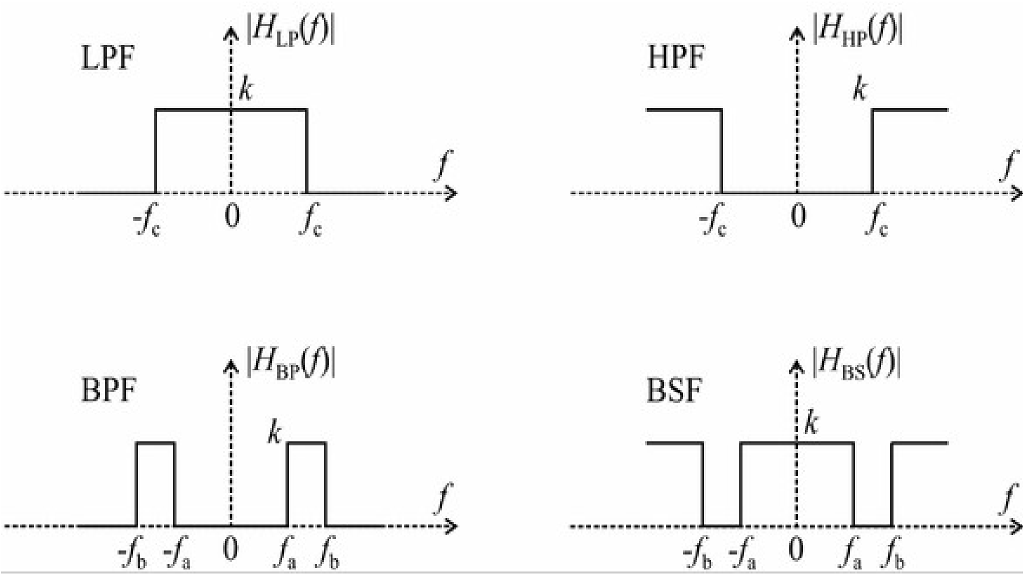


Fig. 5.5 Ideal magnitude response of filters

5.4 Methods for description of filters:

Filters can be described using the following time or frequency domain methods:

5.4.1 Time domain input-output relationship:

A difference equation is used to describe the output of a discrete-time filter in terms of a weighted combination of the input and previous output samples. For example a first-order filter may have the following difference equation -

$$y\ (m) = a\ y\ (m - 1) + x(m)$$

where x(m) is the filter input, y(m) is the filter output and 'a' is the filter coefficient.

5.4.2 Impulse response:

A filter can be described in terms of its response to an impulse input.For example the response of the filter of equation-y (m) = a y (m − 1) + x(m) to a discrete-time impulse input at m= 0 is

$$y(m)=a^m \qquad \text{m= 0, 1, 2, ...}$$

$y(m)=a^m=1,a,(a^2),(a^3),(a^4),......$ for m=0,1,2,3, 4 ... and it is assumed y(-1)= 0. Impulse response is useful because: (i) any signal can be viewed as the sum of a number of shifted and scaled impulses, hence the response a linear filter to a signal is the sum of the responses to all the impulses that constitute the signal, (ii) an impulse input contains all frequencies with equal energy, and hence it excites a filter at all frequencies and (iii) impulse response and frequency response are Fourier transform pairs.

5.4.3 Transfer function, poles and zeros:

The transfer function of a digital filter H(z) is the ratio of the z-transforms of the filter output and input given by -

$$H(z)=\frac{Y(z)}{X(z)} \quad \text{................................(1)}$$

For example the transfer function of the filter of equation- y (m) = a y (m − 1) + x(m) is given by -

$$H(z)=\frac{1}{(1-az^{-1})} \quad \text{.......................................(2)}$$

A useful method of gaining insight into the behaviour of a filter is the pole- zero description of a filter. Poles and zeros are the roots of the denominator and numerator of the transfer function respectively.

5.4.4 Frequency response:

The frequency response of a filter describes how the filter alters the magnitude and phase of the input signal frequencies. The frequency response of a filter can be obtained by taking the Fourier transform of the impulse response of the filter, or by simple substitution of the frequency variable e^{jw} for the z variable $z=e^{jw}$ in the z-transfer function given in equation (1).The frequency response of a filter is a complex variable and can be described in terms of the filter magnitude response and the phase response of the filter.

5.5 Pole-zero plot:

Location of poles and zeros affects the frequency response characteristics of system. We can compute frequency response characteristics from pole-zero plot. Pole-zero plot can be used to design digital filters with desirable frequency response characteristics. Following assumptions are made about the location of poles and zeros while designing the digital filters:

1) All poles should be placed inside the unit circle in order for filter to be stable. However,zeros can be placed anywhere in the z-plane.
2) All complex zeros and poles must occur in complex-conjugate pairs in order for the filter coefficients to be real.

For given pole-zero pattern,the system function H(z) can be expressed as:

$$H(z)=\frac{\sum_{k=0}^{M} b_k z^{-k}}{1+\sum_{k=1}^{N} a_k z^{-k}} \quad \text{....................................(3)}$$

$$\rightarrow H(z) = b_0 \frac{\prod_{k=1}^{M} (1 - z_k z^{-1})}{\prod_{k=1}^{N} (1 - p_k z^{-1})} \quad \text{.........................(4)}$$

where b_0 is gain constant selected to normalize the frequency response at some specified frequency.

M is number of zeros.

N is number of poles.

5.5.1 Pole-zero placement in case of low-pass,high-pass,band-pass filters:

a) In case of low-pass filter:

In case of low-pass digital filter,the poles should be placed near the unit circle at points corresponding to low frequencies (near w=0) and zeros should be placed near or on the unit circle at points corresponding to high frequencies (near w=pi).

b) In case of high-pass filter:

In case of high-pass digital filter,the poles should be placed near or on unit circle at points corresponding to high frequencies (near w= pi) and zeros should be placed near the unit circle at points corresponding to low frequencies (near w=0).

c) In case of band-pass filter:

Band-pass filter should contain one or more pairs of complex conjugate poles near the unit circle.

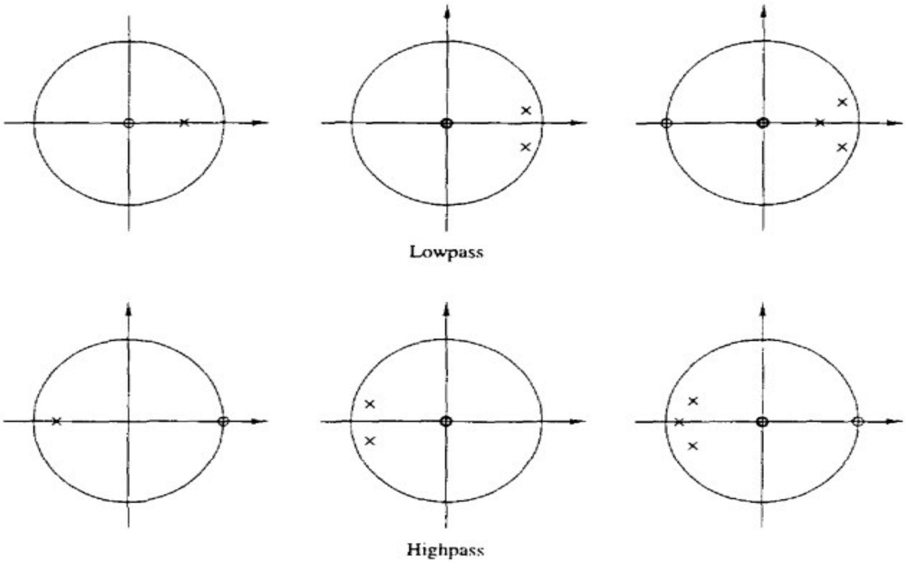


Fig. 5.6 Pole-zero placement for low pass & high pass filters

5.6 Characteristics of practical frequency-selective filters:

Ideal filters are non-causal and hence physically unrealisable for real time signal processing applications. Causality implies that the frequency response characteristics H(w) of filter cannot be zero except at a finite set of points in frequency range. In addition, H(w) cannot have an infinitely sharp cut-off from pass-band to stop-band that is drop from unity to zero abruptly.

Although the frequency response characteristics possessed by ideal filters may be desirable,they are not absolutely necessary in most practical applications. If we relax these conditions,it is possible to realize causal filters that approximate the ideal filters as closely as we desire. In particular, it is not necessary to insist that the magnitude | H(w) | be constant in the entire passband of the filter. A small amount of ripple in the passband is usually tolerable (see following fig. 5.7).Similarly,it is not necessary for the filter response |H(w)| to be zero in the stop-band. A small, non-zero value or small amount of ripple in stop-band is also tolerable. The transition of frequency response from pass-band to stop-band defines the transition band or transition region of filter.

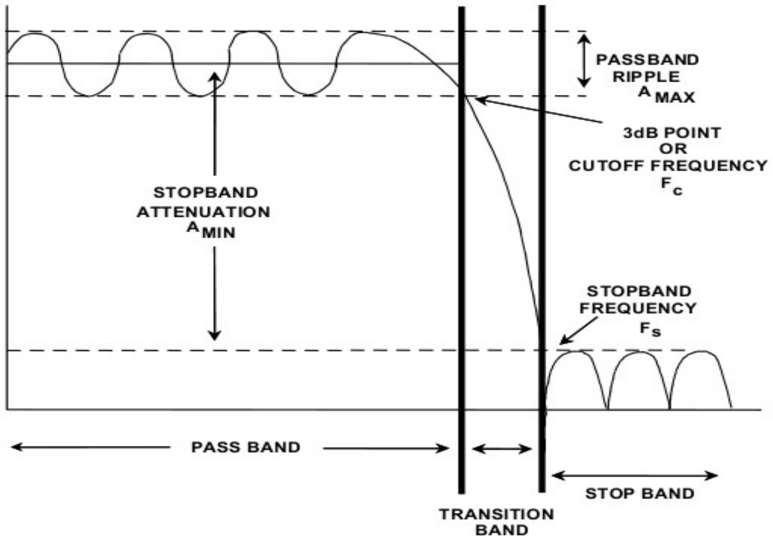


Fig. 5.7 Magnitude characteristics of physically realizable filter

5.7 FIR (Finite Impulse Response) and IIR (Infinite Impulse Response) Filters:

5.7.1 Comparative study of FIR ans IIR filter:

Sr. No.	***FIR***	***IIR***
1	FIR filters have linear phase meaning that no phase distortion of signal occurs. Filter has linear phase iff it's coefficients are symmetrical around centre coefficient i.e. first coefficient is same as last,second coefficient is same as second last and so on.	On other hand,IIR filter always cause some phase distortion (non-linear phase)
2	FIR filters are always stable and have finite length impulse response because there is no feedback in the filter.	IIR filters generally have an infinite length impulse response and may have infinite magnitude output,becomes unstable under some conditions.
3	FIR filters can be designed with a specified amount of quantization noise which can be made as small as necessary	This is not possible in case of IIR filters

4	FIR filters may provide sharper cut-off frequency response but it takes more resources (multipliers and adders) and also the filter order is high as compared to IIR filters. As a result of this,more memory and calculations are required to achieve filter response characteristics.	IIR filters provide much sharper cut-off frequency response as compared to same order FIR filter.
5	FIR tap is simply coefficient. The number of taps is indication of i) the amount of memory required to implement the filter. ii) the number of calculations required to implement the filter. iii) the amount of filtering,the filter can do,in effects more the taps means more stop-band attenuation,less ripple.	
6	MAC (Multiply-Accumulate): MAC is the operation of multiplying a coefficient by corresponding delayed data sample and accumulating the result. FIR filter requires one MAC per tap	
7	Transition Band: The band of frequencies between passband and stop band,Narrower the transition band,more taps are required to implement the filter.	
8	Delay Of FIR filter: The FIR filter which has N taps,the delay is equal to $\frac{N-1}{2F_s}$	

Table 5.1 Comparative study of FIR and IIR filters

5.7.2 FIR Filters:

There are various methods of design of FIR filters. Some of them are enlisted below:

1) Design of linear phase FIR filter using windows.
2) Design of linear phase FIR filter by frequency sampling method.
3) Design of optimum equiripple linear phase FIR filters.
4) Design of FIR differentiator.
5) Design of Hilbert Transformer.

Each one has it's own advantages and disadvantages. Since FIR filters takes more resources (multiplier and adder) to give sharper cut-off frequency response,IIR filters are used. Also the order of these filters while giving the sharper cut-off frequency response is high as compared to IIR filters.

5.7.3 IIR Filters:

Some of IIR filters are enlisted below. Each IIR filter has it's own advantages and dis-advantages (in case of ripple in pass-band and in stop-band).Therefore,their use varies with application to application.

1) Butter-worth Filter.
2) Chebyshev Type-1.
3) Chebyshev Type-2.
4) Elliptic Filter
5) Bessel Filter.

5.7.4 Comparative Study Of IIR Filters:

Butter-worth filter has no ripple in the pass-band or the stop-band, and because of this is sometimes called a maximally flat filter. The Butter-worth filter achieves its flatness at the expense of a relatively wide transition region from pass-band to stop-band.

The Chebyshev Type-1 filter has a smaller transition region than the same- order Butter-worth filter, at the expense of ripples in its pass-band.

The Chebyshev Type-2 filter has ripples in their stop-band,hence also called as inverse Chebyshev Type-1 filter.

Elliptic filters has ripples in both pass-band and stop-band.

Since butter-worth filter has no ripple in pass-band and stop-band,I have used it in Speech Recognition project. Results and conclusions of speech acquisition and preprocessing are shown in chapter number-8.

Chapter 6. Feature Extraction

Speech is time-varying a signal. The information or features contained in the signal is very difficult to analyze. Traditionally,features of speech signal are extracted using Fourier Transform and Short-Time Fourier Transform. But FT and STFT has some drawbacks. FT gives the frequency information of the signal, which means that it tells us how much of each frequency exists in the signal, but it does not tell us when in time these frequency components exist. On other hand,STFT gives time-frequency information but it gives it only for fixed interval of time i.e. STFT gives fixed resolution at all times. These drawbacks are overcomed by use of Wavelet Transform.

6.1 Acoustical features of speech:

6.1.1 Volume:

This feature represents the loudness of the audio signal, which is correlated to the amplitude of the signals. Sometimes it is also referred to as energy or intensity of audio signals. It is identified by identifying the overall heights of peaks in the power spectrum.

6.1.2 Pitch:

This feature represents the vibration rate of audio signals, which can be represented by the fundamental frequency, or equivalently, the reciprocal of the fundamental period of voiced audio signals. It is the fundamental frequency in power spectrum.

6.1.3 Timbre:

This feature represents the meaningful content of audio signals, which is characterized by the waveform within a fundamental period of voice signals. It is identified by identifying relative heights of peaks in the power spectrum. Following figure shows power spectrum:

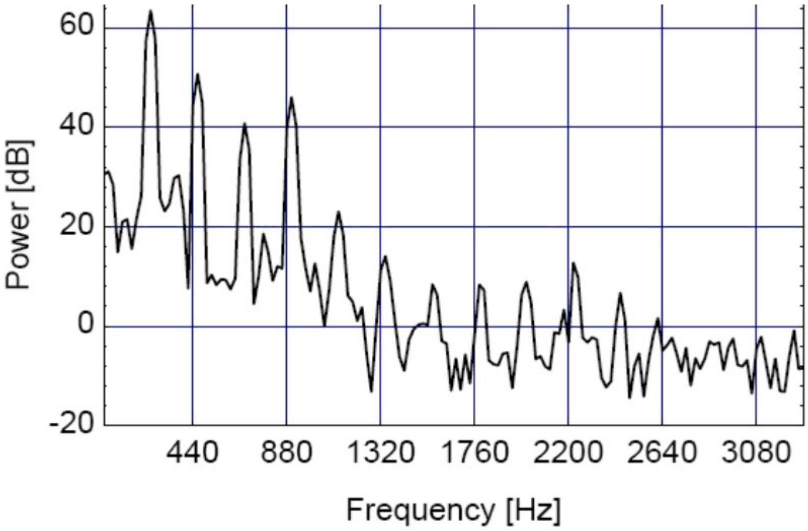


Fig. 6.1 Power spectrum of phone

In above graph, we can observe the fundamental peak (biggest peak in the graph) which is between 0 and 440 Hz,harmonic peaks (peaks between first and last highest peak,avoid the smallest peaks between them).

6.2 Features in time-domain:

6.2.1 Average energy:

The average energy indicates the loudness of the audio signal. We can calculate the average energy by mean-square value.

$$E=\frac{1}{N}\sum_{n=0}^{N-1} x(n)^2$$

where E is the average energy of audio signal x(n) . N is the total number of samples in this audio signal.

6.2.2 Zero-crossing rate:

The zero-crossing rate (ZCR) indicates the frequency of signal amplitude sign change. The average zero-crossing is calculated as follows:

$$ZC=\frac{1}{2\mathrm{N}}\sum_{n=1}^{N} sgn\,x(n)-sgn\,x(n-1)$$

where sgn x(n) is the sign of x(n) .

ZCR has the following characteristics:

1) In general, ZCR of both unvoiced sounds and environmental noise are larger than voiced sounds.

2) It is hard to distinguish unvoiced sounds from environmental noise by using ZCR alone since they have similar ZCR values.

3) ZCR is often used in conjunction with the volume for end-point detection. In particular, ZCR is used for detecting the start and end positions of unvoiced sounds.

6.2.3 Silence ratio:

The silence ratio indicates the proportion of the silence in the audio signal. Silence is defined as an interval where the absolute amplitude values are below a certain threshold. The silence ratio is the ratio between the sum of silent intervals and the total length of the audio signal. In general, two thresholds are defined: one is used to determine if an audio sample is silence and the other is used to determine if those silence samples are silence interval. For two fixed threshold, an audio sample whose value is below the threshold is considered as silent and the number of silent is above the other threshold is considered as silence interval.

6.3 Features in frequency domain:

6.3.1 Spectrum:

Sound spectrum represents the amplitude of the signal varying with frequency. The energy distribution across the frequency range is seen from the spectrum.

6.3.2 Bandwidth:

The bandwidth indicates the frequency range of a sound.The simplest definition of the bandwidth is the frequency difference between the highest frequency and lowest frequency of the non-zero spectrum components.

6.3.3 Harmonic:

In harmonic sound, the spectral components are the multiples of the lowest frequency. The lowest frequency is called fundamental frequency. The method that decides if a sound is harmonic is to check if the frequency of the dominant components is the multiples of the fundamental frequency.

6.3.4 Pitch:

This feature represents the vibration rate of audio signal.

6.3.5 Timbre:

This feature represents the meaningful content of audio signal.

6.4 Techniques to analyse audio signals:

6.4.1 Fourier Transform:

Express any signal/function in terms of a weighted sum of cosine and sine functions.

$$FT\, x(t) = X(w) = \int_{-\infty}^{\infty} x(t)\,\mathrm{e}^{-jwt} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\, w = 2 * pi * F$$

Disadvantages:

a) For some signal, it may not practical to operate in Fourier transform. The reason is that in most of cases, only the interval $t_0 \le t \le t_1$ is interested.

b) The Fourier transform defines the global representation of the frequency content of the signal over a total period of time over which the signal x(t) exists; it does not give access to the signal's spectral variations during this interval of time.

6.4.2 Time frequency analysis (Short time fourier transform):

STFT is obtained by applying FT to successive portion of signal by means of sliding window of finite size.

$$STFT_{(g(w,b))} x(t) = X_g(w,b) = \int_{-\infty}^{\infty} x(t)\, g(t-b)\, e^{(-jwt)}$$

where g(t) is sliding window.

Disadvantage:

Once a particular analyzing window (fixed length of window) has been chosen in a given STFT, the quantities delta t and delta w remains unchanged over the entire analysis procedure.

6.4.3 Continuous wavelet transform:

In order to get good resolution, the analysis of low-frequency signals, which require long time-domain analysis windows and vice-versa to the high frequency signals. The Wavelet transform gives a better trade-off between time and frequency resolutions than the fixed length windows used in the STFT.

Disadvantage:

An important drawback which affects the continuous wavelet transform is its high level of redundancy which results from the facts that both the dilation parameter and the translation parameter are continuous quantities.

6.4.4 Discrete wavelet transform:

CWT is function of two parameters and therefore contains high amount of extra (redundant) information when analyzing a signal. Instead of continuously varying the parameters we analyze the signal with small number of scales with varying number of translations at each scale. This is discrete wavelet transform. Due to this advantages,DWT is used in analysis of audio signal.

6.5 Wavelet transform:

Wavelets means 'small wave'.So,wavelet analysis is about analyzing signal with short duration finite energy functions. They transform the signal under investigation into another representation which presents the signal in more useful form. This transformation is called as 'wavelet transform'.We manipulate wavelet in two ways. The first one is translation. We change the central position of wavelet along the time axis. The second one is scaling.

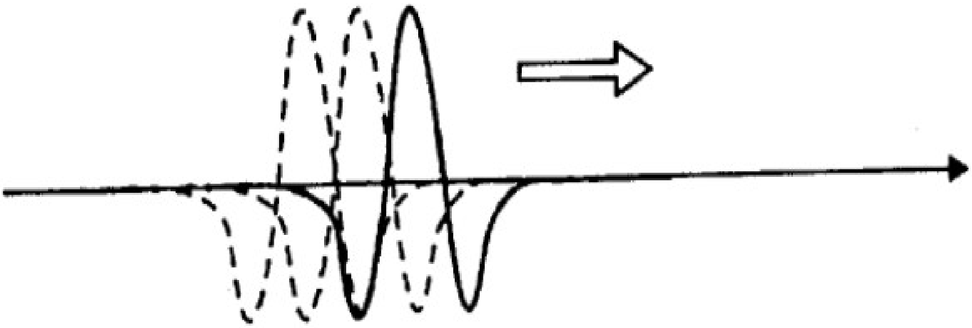


Fig. 6.2 Translation of wavelets

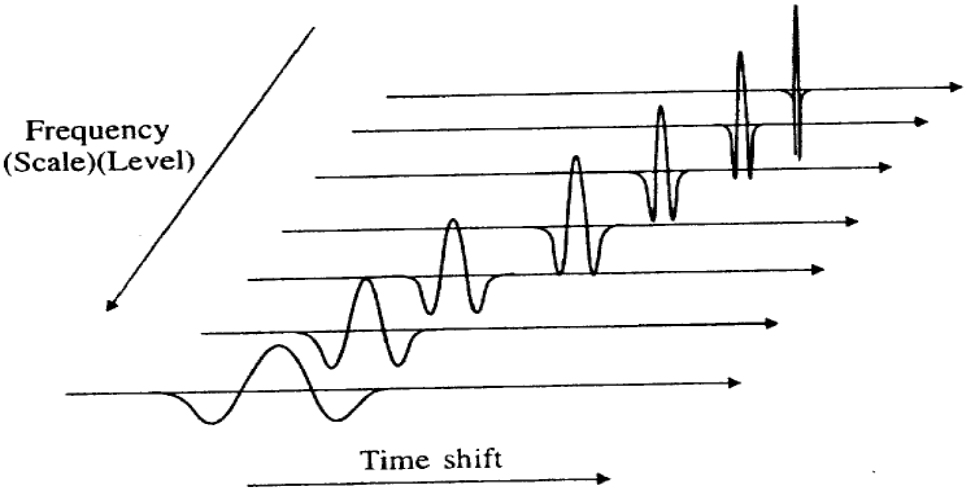


Fig. 6.3 Change in scale of wavelets

without understanding the continuous wavelet transform,we are unable to understand discrete wavelet transform. Let's take the tour of continuous wavelet transform,how it works?

6.6 Continuous wavelet transform(CWT):

The continuous wavelet transform was developed as an alternative approach to the short time fourier transform to overcome the resolution problem. The wavelet analysis is done in a similar way to the STFT analysis, in the sense that the signal is multiplied with a function, similar to the window function in the STFT, and the transform is computed separately for different segments of the time-domain signal. However, there are two main differences between the STFT and the CWT:

1. The Fourier transforms of the windowed signals are not taken, and therefore single peak will be seen corresponding to a sinusoid, i.e. negative frequencies are not computed.
2. The width of the window is changed as the transform is computed for every single spectral component, which is probably the most significant characteristic of the wavelet transform.

The continuous wavelet transform is defined as follows :

$$X_{WT}(\tau, s) = \frac{1}{\sqrt{|s|}} \int_{-\infty}^{\infty} x(t)\psi^* \left(\frac{t-\tau}{s}\right) dt.$$

As seen in the above equation, the transformed signal is a function of two variables, tau and s, the translation and scale parameters, respectively. psi(t) is the transforming function, and it is called the mother wavelet . The term mother wavelet gets its name due to two important properties of the wavelet analysis as explained below:

The term wavelet means a small wave. The smallness refers to the condition that this (window) function is of finite length (compactly supported).The wave refers to the condition that this function is oscillatory. The term mother implies that the functions with different region of support that are used in the transformation process are derived from one main function, or the mother wavelet. In other words, the mother wavelet is a prototype for generating the other window functions. The term translation is used in the same sense as it was used in the STFT; it is related to the location of the window, as the window is shifted through the signal. This term, obviously, corresponds to time information in the transform domain.

The parameter scale in the wavelet analysis is similar to the scale used in maps. As in the case of maps, high scales correspond to a non-detailed global view (of the signal), and low scales correspond to a detailed view. Similarly, in terms of frequency, low frequencies (high scales) correspond to a global information of a signal (that usually spans the entire signal), whereas high frequencies (low scales) correspond to a detailed information of a hidden pattern in the signal (that usually lasts a relatively short time).Cosine signals corresponding to various scales are given as examples in the following Fig. 6.4

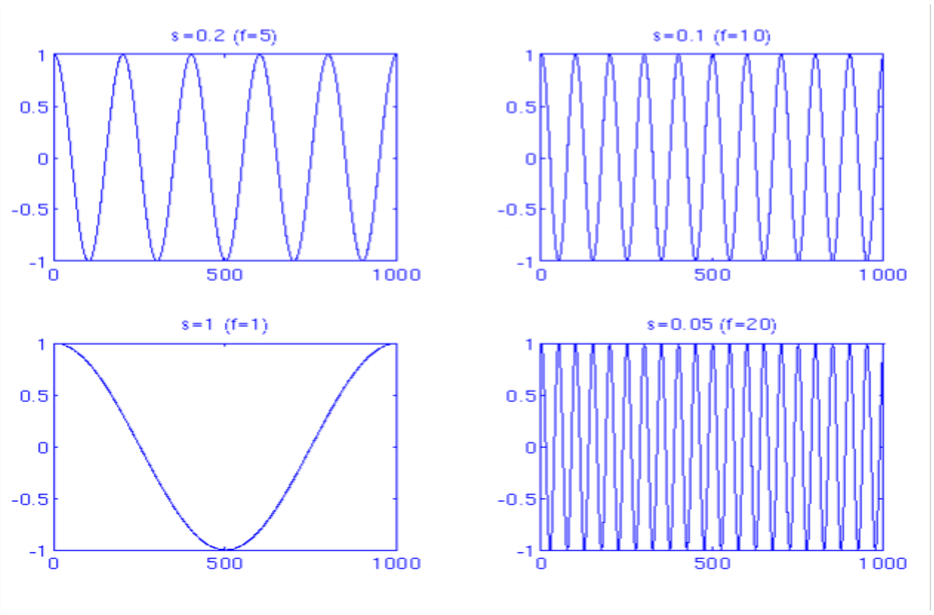


Fig. 6.4 Cosine signal corresponding to various scale

In practical applications, low scales (high frequencies) do not last for the entire duration of the signal, unlike those shown in the figure, but they usually appear from time to time as short bursts, or spikes. High scales (low frequencies) usually last for the entire duration of the signal. Scaling, as a mathematical operation, either dilates or compresses a signal. Larger scales correspond to dilated (or stretched out) signals and small scales correspond to compressed signals. All of the signals given in the figure are derived from the same cosine signal, i.e. they are dilated or compressed versions of the same function. In the above figure, s=0.05 is the smallest scale, and s=1 is the largest scale. In terms of mathematical functions, if f(t) is a given function, f(st) corresponds to a contracted (compressed) version of f(t) if s > 1 and to an expanded (dilated) version of f(t) if s < 1.However, in the definition of the wavelet transform, the scaling term is used in the denominator, and therefore, the opposite of the above statements holds, i.e. scales s > 1 dilates the signals whereas scales s < 1, compresses the signal.

6.7 Computation of CWT:

Let x(t) is the signal to be analyzed. The mother wavelet is chosen to serve as a prototype for all windows in the process. All the windows that are used are the dilated (or compressed) and shifted versions of the mother wavelet. There are a number of functions that are used for this purpose. Once the mother wavelet is chosen, the computation starts with s=1 and the continuous wavelet transform is computed for all values of s, smaller and larger than ``1". However, depending on the signal, a complete transform is usually not necessary. For all practical purposes, the signals are band-limited, and therefore, computation of the transform for a limited interval of scales is usually adequate. For convenience, the procedure will be started from scale s=1 and will continue for the increasing values of s, i.e., the analysis will start from high frequencies and proceed towards low frequencies. This first value of s will correspond to the most compressed wavelet. As the value of s is increased, the wavelet will dilate. The wavelet is placed at the beginning of the signal at the point which corresponds to time=0. The wavelet function at scale ``1" is multiplied by the signal and then integrated over all times. The result of the integration is then multiplied by the constant number 1/sqrt{s}. This multiplication is for energy normalization purposes so that the transformed signal will have the same energy at every scale. The final result is the value of the transformation, i.e., the value of the continuous wavelet transform at time zero and scale s=1. In other words, it is the value that corresponds to the point tau =0 , s=1 in the time-scale plane.

The wavelet at scale s=1 is then shifted towards the right by tau amount to the location t=tau and the above equation is computed to get the transform value at t=tau , s=1 in the time- frequency plane. This procedure is repeated until the wavelet reaches the end of the signal. One row of points on the time-scale plane for the scale s=1 is now completed.

Then, s is increased by a small value. This is a continuous transform, and therefore, both tau and s must be incremented continuously. However, if this transform needs to be computed by a computer, then both parameters are increased by a sufficiently small step size. This corresponds to sampling the time-scale plane.

The above procedure is repeated for every value of s. Every computation for a given value of s fills the corresponding single row of the time-scale plane. When the process is completed for all desired values of s, the CWT of the signal has been calculated. The fig. 6.5 below illustrate the entire process step by step:

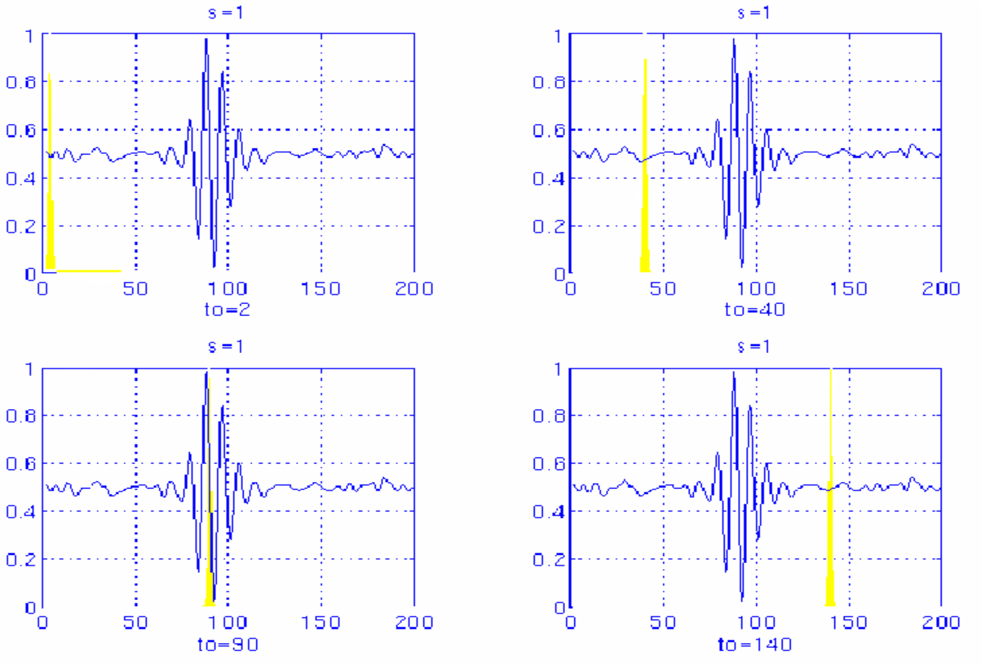


Fig. 6.5 Signal and wavelet function with different values of tau

In fig. 6.5, the signal and the wavelet function are shown for four different values of tau. The scale value is 1, corresponding to the lowest scale, or highest frequency. Four distinct locations of the wavelet function are shown in the figure at to=2, to=40, to=90, and to=140. At every location, it is multiplied by the signal. The product is non-zero only where the signal falls in the region of support of the wavelet, and it is zero elsewhere. By shifting the wavelet in time, the signal is localized in time, and by changing the value of s, the signal is localized in scale (frequency).

If the signal has a spectral component that corresponds to the current value of s (which is 1 in this case), the product of the wavelet with the signal at the location where this spectral component exists gives a relatively large value. If the spectral component that corresponds to the current value of s is not present in the signal, the product value will be relatively small, or zero. The signal in Figure 6.5 has spectral components comparable to the window's width at s=1 around t=100 ms.

The continuous wavelet transform of the signal in Figure 6.5 will yield large values for low scales around time 100 ms, and small values elsewhere. For high scales, on the other hand, the continuous wavelet transform will give large values for almost the entire duration of the signal, since low frequencies exist at all times .

6.8 Time & frequency resolutions:

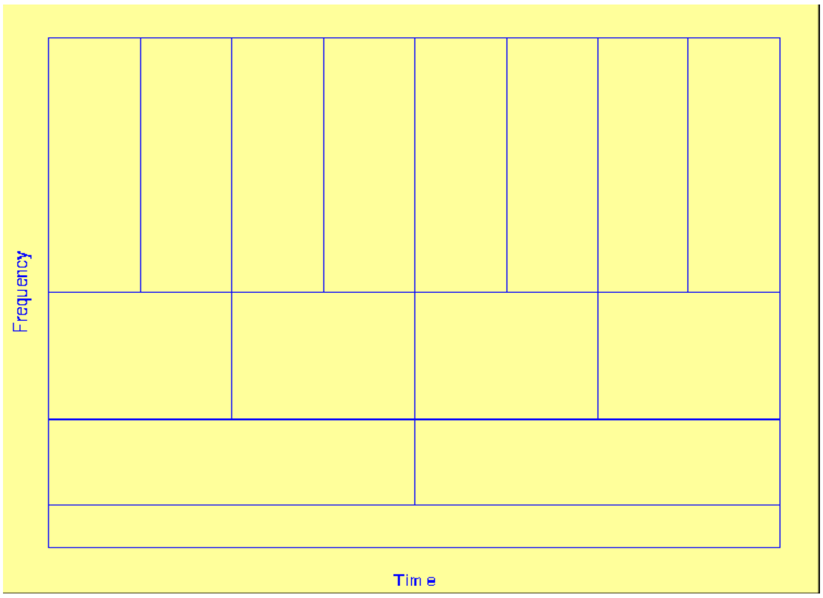


Fig. 6.6 Time & frequency resolution

Fig. 6.6 is used to explain how time and frequency resolutions should be interpreted. Every box in Fig. 6.6 corresponds to a value of the wavelet transform in the time-frequency plane. Boxes have a certain non-zero area, which implies that the value of a particular point in the time-frequency plane cannot be known. All the points in the time-frequency plane that falls into a box are represented by one value of the WT.

In above fig. 6.6 ,although the widths and heights of the boxes change, the area is constant. That is each box represents an equal portion of the time-frequency plane, but giving different proportions to time and frequency. Note that at low frequencies, the height of the boxes are shorter (which corresponds to better frequency resolutions, since there is less ambiguity regarding the value of the exact frequency), but their widths are longer (which correspond to poor time resolution, since there is more ambiguity regarding the value of the exact time). At higher frequencies the width of the boxes decreases, i.e., the time resolution gets better, and the heights of the boxes increase, i.e., the frequency resolution gets poorer.

6.9 Discretization of CWT:

It is apparent that neither the FT, nor the STFT, nor the CWT can be practically computed by using analytical equations, integrals, etc. It is therefore necessary to discretize the transforms. As in the FT and STFT, the most intuitive way of doing this is simply sampling the time-frequency (scale) plane. Again intuitively, sampling the plane with a uniform sampling rate sounds like the most natural choice. However, in the case of WT, the scale change can be used to reduce the sampling rate.

At higher scales (lower frequencies), the sampling rate can be decreased, according to Nyquist's rule. In other words, if the time-scale plane needs to be sampled with a sampling rate of N_1 at scale s_1 , the same plane can be sampled with a sampling rate of N_2 , at scale s_2 , where, $s_1<s_2$ (corresponding to frequencies $f_1>f_2$) and $N_2<N_1$. The actual relationship between N_1 and N_2 is :

$$N_2=\frac{s_1}{s_2}*N_1 \quad \text{or} \quad N_2=\frac{f_2}{f_1}*N_1$$

In other words, at lower frequencies the sampling rate can be decreased which will save a considerable amount of computation time.

During discretization,the scale parameter s is discretized first on a logarithmic grid. The time parameter is then discretized with respect to the scale parameter, i.e., a different sampling rate is used for every scale. In other words, the sampling is done on the dyadic sampling grid shown in Fig.6.7:

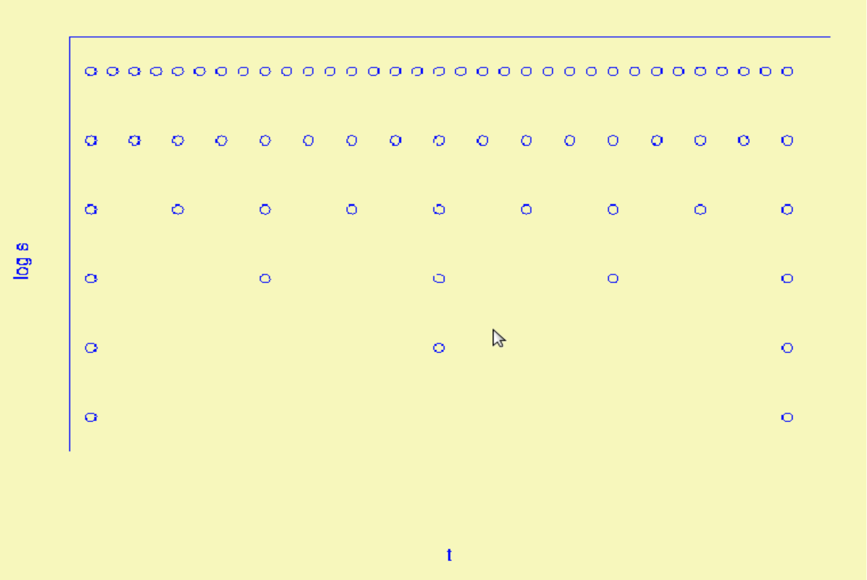


Fig. 6.7 Discretization of scale & time

The CWT assigns a value to the continuum of points on this plane. Therefore, there are an infinite number of CWT coefficients. First consider the discretization of the scale axis. Among that infinite number of points, only a finite number are taken, using a logarithmic rule. The base of the logarithm depends on the user. The most common value is 2 because of its convenience. If 2 is chosen, only the scales 2, 4, 8, 16, 32, 64,...etc. are computed. If the value was 3, the scales 3, 9, 27, 81, 243,...etc. would have been computed. The time axis is then discretized according to the discretization of the scale axis. Since the discrete scale changes by factors of 2, the sampling rate is reduced for the time axis by a factor of 2 at every scale.

The lowest scale (s=2), only 32 points of the time axis are sampled (for the particular case given in Fig 6.7). At the next scale value, s=4, the sampling rate of time axis is reduced by a factor of 2 since the scale is increased by a factor of 2, and therefore, only 16 samples are taken. At the next step, s=8 and 8 samples are taken in time, and so on. Expressing the above discretization procedure in mathematical terms, the scale discretization is $s=s_0^j$ and translation discretization is $tau=k.s_0^j(tau_0)$ where $s_0>1$ and $tau_0>0$. Note, how the translation discretization is dependent on scale discretization with s_0 .The continuous wavelet function :

$$\psi_{\tau,s}(t) := \frac{1}{\sqrt{|s|}}\psi\left(\frac{t-\tau}{s}\right)$$

By inserting value of s and tau,

$$\psi_{j,k}(t) = \frac{1}{\sqrt{s^j}}\psi\left(\frac{t - k\tau s^j}{s^j}\right)$$

6.10 Sub-band coding & multi-resolution analysis:

The continuous wavelet transform was computed by changing the scale of the analysis window, shifting the window in time, multiplying by the signal, and integrating over all times. In the discrete case, filters of different cut-off frequencies are used to analyze the signal at different scales. The signal is passed through a series of high pass filters to analyze the high frequencies, and it is passed through a series of low pass filters to analyze the low frequencies.

The resolution of the signal, which is a measure of the amount of detail information in the signal, is changed by the filtering operations, and the scale is changed by up-sampling and down-sampling (sub-sampling) operations. Sub-sampling a signal corresponds to reducing the sampling rate, or removing some of the samples of the signal. For example, sub-sampling by two refers to dropping every other sample of the signal. Sub-sampling by a factor n reduces the number of samples in the signal n times. Up-sampling a signal corresponds to increasing the sampling rate of a signal by adding new samples to the signal. For example, sampling by two refers to adding a new sample, usually a zero or an interpolated value, between every two samples of the signal. Up-sampling a signal by a factor of n increases the number of samples in the signal by a factor of n.

The procedure starts with passing this signal (sequence) x(n) through a half band digital low-pass filter with impulse response h[n]. Filtering a signal corresponds to the mathematical operation of convolution of the signal with the impulse response of the filter. The convolution operation in discrete time is defined as follows:

$$x(n)*h(n)=\sum_{-\infty}^{\infty} x(k)h(n-k)$$

A half band low-pass filter removes all frequencies that are above half of the highest frequency in the signal. For example, if a signal has a maximum of 1000 Hz component, then half band low-pass filtering removes all the frequencies above 500 Hz. After passing the signal through a half band low-pass filter, half of the samples can be eliminated according to the Nyquist’s rule, since the signal now has a highest frequency of p/2 radians instead of p radians. Simply discarding every other sample will sub-sample the signal by two, and the signal will then have half the number of points. The scale of the signal is now doubled. Note that the low-pass filtering removes the high frequency information, but leaves the scale unchanged. Only the sub-sampling process changes the scale. Resolution, on the other hand, is related to the amount of information in the signal, and therefore, it is affected by the filtering operations. Half band low-pass filtering removes half of the frequencies, which can be interpreted as losing half of the information. Therefore, the resolution is halved after the filtering operation. Note, however, the sub-sampling operation after filtering does not affect the resolution, since removing half of the spectral components from the signal makes half the number of samples redundant anyway. Half the samples can be discarded without any loss of information. In summary, the low-pass filtering halves the resolution, but leaves the scale unchanged. The signal is then sub-sampled by 2 since half of the number of samples are redundant. This doubles the scale. This procedure can mathematically be expressed as :

$$y(n)=\sum_{-\infty}^{\infty} h(k)x(2\mathrm{n}-k)$$

Having said that, we now look how the DWT is actually computed: The DWT analyzes the signal at different frequency bands with different resolutions by decomposing the signal into a coarse approximation and detail information. DWT employs two sets of functions, called scaling functions and wavelet functions, which are associated with low pass and high-pass filters, respectively. The decomposition of the signal into different frequency bands is simply obtained by successive high-pass and low-pass filtering of the time domain signal. The original signal x[n] is first passed through a half-band high-pass filter g[n] and a low-pass filter h[n]. After the filtering, half of the samples can be eliminated according to the Nyquist's rule, since the signal now has a highest frequency of p/2 radians instead of p. The signal can therefore be sub-sampled by 2, simply by discarding every other sample. This constitutes one level of decomposition and can mathematically be expressed as follows:

$$y_{high}(k)=\sum_{n} x(n)g(2\mathrm{k}-n) \quad \& \quad y_{low}(k)=\sum_{n} x(n)h(2\mathrm{k}-n)$$

where $y_{high}(k)$ & $y_{low}(k)$ are the outputs of the high-pass and low-pass filters, respectively, after sub-sampling by 2. This decomposition halves the time resolution since only half the number of samples now characterizes the entire signal. However, this operation doubles the frequency resolution, since the frequency band of the signal now spans only half the previous frequency band, effectively reducing the uncertainty in the frequency by half. The above procedure, which is also known as the subband coding, can be repeated for further decomposition. Fig. 6.7 illustrates this procedure.

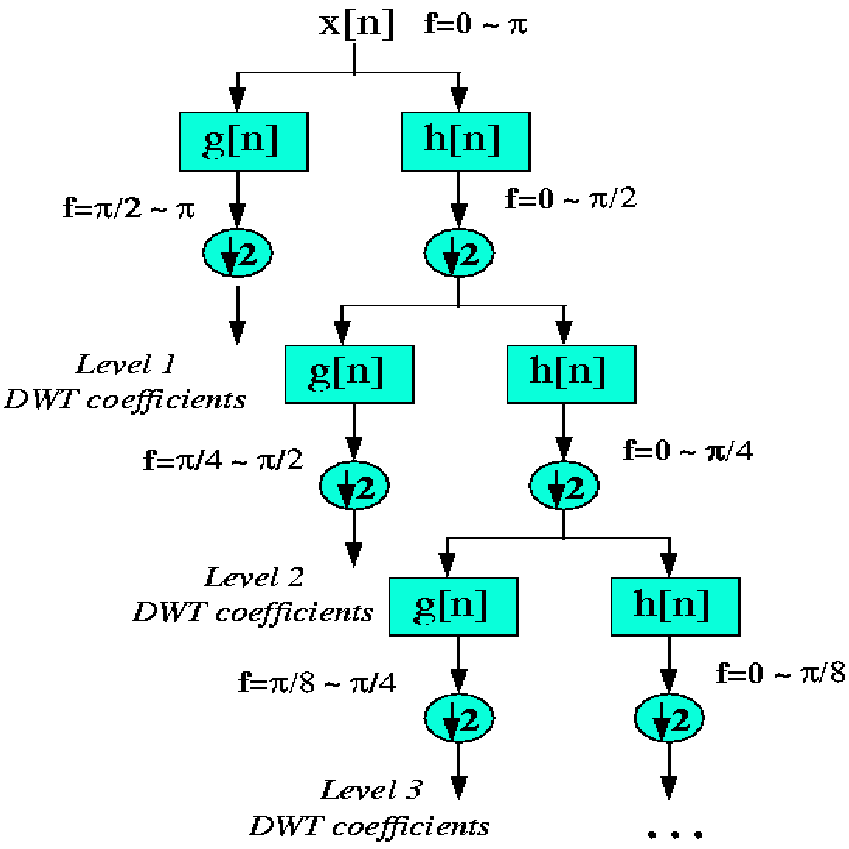


Fig. 6.8 Sub-band coding

6.11 Wavelet families:

There are various families of wavelet namely morlet, mexican hat,Meyer, haar, daubechiesN, symlets, coiflets and so on. Selection criteria varies with application to application depending upon properties like symmetry,vanishing moments,type of analysis orthogonal or bi-orthogonal,reconstruction criteria and so on. Here daubechies-4 tap wavelet is used because of it's exact reconstruction capability as compared to others,can be used for orthogonal as well as bi-orthogonal analysis & it's fast working mechanism.

6.12 Features taken into consideration for extraction:

Preprocessed signal first decomposed at certain levels by use of low pass and high pass filters given in section of 'sub-band coding & multi-resolution analysis'.At each level,we got coefficient for approximation & coefficient for details. After that approximation is further decomposed to get another set of coefficient for approximation and coefficient for details. This procedure is repeated up to the predefined levels. For each coefficient of detail (sub-band) power spectra is then computed to observe amount of signal at particular frequency. A useful way to determine the distribution of energy within data array (values of each sub-band) is to plot wavelet power spectra. For given wavelet transform

W_i ,wavelet power spectrum is defined as absolute value squared of wavelet coefficients $(|(W_i)|)^2$ After calculating power spectrum for each sub-band,following features are calculated for each sub-band:

6.12.1 Spectral centroid:

This is amplitude-weighted average of power spectrum which is related to perception or brightness of audio signal. It is calculated by multiplying the value of each frequency by it's power,then taking sum of all of these. The value is then normalized by dividing it by sum of all power values.

$$SC=\frac{(\sum_{k=1}^{k} P(f_k)*f_k)}{(\sum_{k=1}^{k} P(f_k))}$$

where $P(f_k)$ power spectrum

6.12.2 Spectral flux:

This is measure of amount of local spectral change and used to determine timbre of an audio signal. This is defined as squared difference between normalized power spectra of successive sub-bands. It is measure of how quickly the power spectrum of signal is changing.

$$SF=\sum_{k=2}^{k}(|(P(f_k)-P(f_{(k-1)}))|)^2$$

6.12.3 Spectral spread:

It is measure of spread of spectrum around the spectral centroid. It is defined as:

$$SS=\frac{\sqrt{(\sum_{k=1}^{n}(P(f_n)-SC)^2)}}{\sqrt{(\sum_{k=1}^{n} P(f_n)^2)}}$$

6.12.4 Spectral skewness:

Skewness is measure of the asymmetry of distribution around the centroid value. Skewness is calculated by-

$$SK=\frac{\sum(f_k-SC)^2*mag}{\sum mag}$$

Chapter 7. Classification & Pattern Recognition

There are various approaches to pattern classification such as vector quantization,dynamic time warping, hidden Markov model,artificial neural network. Selection of pattern classification method depends upon the type of speech recognizer to be implemented. As studied in the introduction,speech recognition systems are classified according to size of vocabulary,according to speaker dependency,and according to continuousness of speech. Vector quantization and dynamic time warping are the two methods of template matching approach. Dynamic time warping method is suitable for small vocabulary speech recognizer and due to it's accuracy and simplicity,it is studied here.

Template matching is simplest and earliest approach to pattern classification. Matching is used to determine the similarity between two observations of the same type. In template matching,a template (feature vectors of a pre-recorded word to be recognized in case of word recognition) of pattern to be recognized is already available to the system. The pattern to be matched is compared with the stored template according to the some distance (similarity) measure. Dynamic time warping method is discussed here.

7.1 Dynamic time warping (DTW):

Dynamic time warping is based on “dynamic programming which is defined as “an algorithmic technique in which an optimization problem is solved by caching sub problem solutions rather than recomputing them”.It includes backtracking process which guarantees the best solution for the problem.

DTW is template matching method for classification of patterns (acoustic vector in speech recognition).DTW based acoustic modelling includes creation of templates for each class. Template can include the feature vectors of pre-recorded word to be recognized in case of word recognition or can include mean feature vectors linked to some phonemes if the task is phone recognition.

Two pronounciation of same word in different times are not the same because of speed and duration changes in the speech. This nature of speech needs time alignment when it is being compared to a pre-recorded version which has same phonetic content. DTW is one of the techniques to make this time alignment.

DTW tries to find the minimum distance between reference and test templates. The computation of minimum distance is based on following formula:

$$D(i,j)=d(i,j)+min(D(i,j),D(i,j-1),D(i-1,j))$$

where i=1.....n & j=1....N

Here i and j are the indices of feature vectors to be compared, d(i, j) is the Euclidean distance between two feature vectors and D(i , j) is the global distance until the current point. At the starting point the distance is -

$$D(0,0)=d(0,0)$$

The global distance between a reference template of length N and the observed test sequence of length n is D(N,n) and can be calculated recursively.

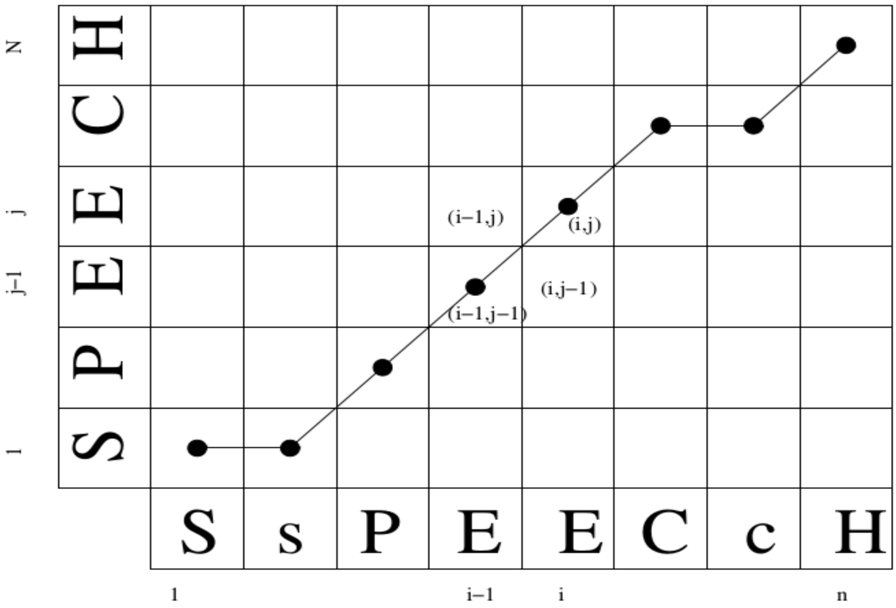


Fig. 7.1 Dynamic time warping, N=number of frames in reference template, n=number of frames in the pattern to be matched

Illustration:

Suppose that we have two feature vectors ,one reference template y(t) and other test template x(t).

a) (Test) template x(t): 1 1 2 3 2 0

b) (Reference) template y(t): 0 1 1 2 3 2 1

Euclidean distance between them is calculated by:

$$\sqrt{(x(t)-y(t))^2}$$

Distance matrix D becomes:

y(t)

1	0	0	1	2	1	**1**
2	1	1	0	1	**0**	2
3	2	2	1	**0**	1	3
2	1	1	**0**	1	0	2
1	0	**0**	1	2	1	1
1	**0**	0	1	2	1	1
0	**1**	1	2	3	2	0
	1	1	2	3	2	0

x(t)

We can see that there is sequence of low numbers ,close to the diagonal ,indicating that which samples of x(t) are closest in value to those of y(t) .These are shown in bold form.

Chapter 8. Results & Conclusions

8.1 Results of speech acquisition & preprocessing step:

8.1.1 Designed Butter-worth Filter in this project:

Hearing range usually describes the range of frequencies that can be heard by humans or other animals. The human range is on average from 20 to 20,000 Hz .The microphone that we have selected is able to pick up the frequencies from 100 Hz to 16,000 Hz. Generally,for speech recognition,we do not require such lower and higher frequencies. That's why I have designed butter-worth band-pass filter having cut-off frequencies 100 Hz and 10,000 Hz. Ideally, a filter has unit gain (0 dB) in pass-band and gain of zero (- ∞) in the stop-band. As stated earlier,it is impossible to realize ideal band-pass filter. Therefore,there exists finite transition band between the pass-band and stop-band. In the transition band, the gain of the filter changes gradually from one (0 dB) in pass-band to zero (- ∞) in stop-band (see fig. 5.13).Practical filters might have pass-band ripple and the stop-band attenuation of filter cannot be infinite.

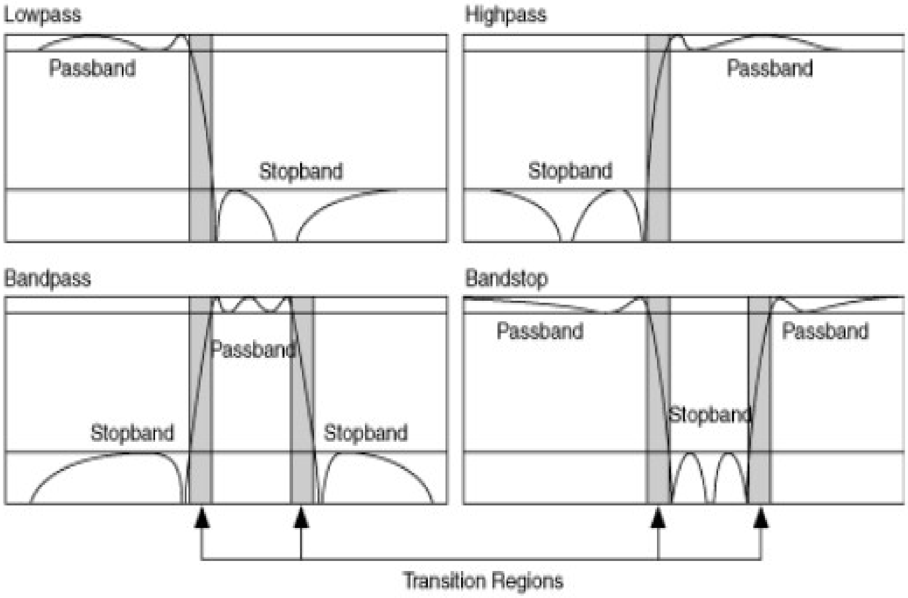


Fig. 8.1 Pass-band, stop-band and transition band for low,high,band-pass &band-stop filter

Pass-band Ripple and Stop-band Attenuation:

In many applications, you can allow the gain in the passband to vary slightly from unity. This variation in the passband is the passband ripple, or the difference between the actual gain and the desired gain of unity. In practice, the stop-band attenuation cannot be infinite, and you must specify a value with which you are satisfied.

By taking above factors into the consideration,the parameters-pass-band corner frequency W_p the cut-off frequency,stop-band corner frequency W_s (the scalar or two element vector with values between 0 and 1, with 1 corresponding to the normalized Nyquist frequency.),pass-band ripple R_p (maximum permissible pass-band loss in decibel) in decibel,stop-band attenuation R_s (number of decibels the stop-band is down from the passband) in decibel are passed to 'buttord' function of MATLAB which has following syntax. It returns the minimum order of a digital butter-worth filter required to meet a set of filter design specifications. The scalar (or vector) of corresponding cut-off frequencies, W_n is also returned.

$$[n,W_n]=buttord(W_p,W_s,R_p,R_s)$$

We use these output arguments in “butter” function of MATLAB.”butter” function designs low-pass,band-pass,high-pass and band-stop digital Butter-worth filter. Butter-worth filter is characterized by a magnitude response that is maximally flat in pass-band. Butter-worth filters sacrifice roll-off steepness in the pass-band and stop-band. Unless smoothness of the Butter-worth filter is needed, an elliptic or chebyshev can generally provide steeper rolloff characteristics with lower filter order.

$$[b,a]=butter(n,W_n,'ftype')$$

Function “butter” designs an order n low-pass,high-pass or band-pass filter with normalized cut-off frequency W_n (must be a number between 0 and 1, where 1 corresponds to Nyquist frequency) where string 'ftype' is 'high','low','bandpass' or 'bandstop'.It returns the filter coefficients in length n+1 row vectors b and a,with coefficients in descending power of z. Refer equation (3) in chapter-5.

The function “butter” in the following form returns the zeros and poles in length n column vectors z and p , and the gain in the scalar k .The function “freqz(b,a)” plot the magnitude and phase response of the designed Butter-worth filter.

In this project,Butter-worth band-pass filter having following design specifications is designed:

1) Pass-band of 100 Hz to 10,000 Hz.

Normalized by dividing (sampling frequency/2) also known as Nyquist frequency.

$$W_p=[100,10000]/(sampling\,frquency/2)$$

2) Less than 3dB of ripple in pass-band i.e. $R_p=3$

3) 11dB attenuation in stop-bands that are 10 Hz wide on both sides of pass-band(transition band of width 10 Hz)

$$R_s=11$$

$$W_s=[110,10010]/(sampling\,frequency/2)$$

Results:

$$[n,W_n]=buttord(W_p,W_s,R_p,R_s)$$

returns order of filter n=6,vector of cut-off frequencies W_n

$$[b,a]=butter(n,W_n,'bandpass')$$

returns filter coefficients b (numerator's polynomial coefficients) and a (denominator's polynomial coefficients).

b=0.5676, 0, -3.4057, 0, 8.5143, 0, -11.3524, 0, 8.5143, 0, -3.4057, 0, 0.5676

a= 1, -1.1051, -4.2794, 4.1184, 8.2498, -6.4055, -9.0267, 5.1425, 5.8392, -2.1174, -2.0940, 0.3561

$$[z,p,k]=butter(n,W_n,'bandpass')$$

returns poles p, zeros z,and gain k as

p= - 0.7515+0.0579 i, -0.7515-0.0579 i,
- 0.7978+0.0167 i, -0.7978-0.0167 i,
- 0.8927+0.2564 i, -0.8927-0.2564 i,
0.9973+ 0.0007 i, 0.9973-0.0007 i,
0.9980+0.0020 i, 0.9980-0.0020 i,
0.9993+0.0027 i, 0.9993-0.0027 i,

z=1, 1, 1, 1, 1, 1, -1, -1, -1, -1, -1, -1

k=0.5676

$$zplane(z,p)$$

plots pole-zero diagram for filter's transfer function.

The coefficients b and a are then used in function “filter” in order to filter our signal as:

$filtered\ samples=filter(b,a,original\,speech\,samples)$ Now, we have filtered speech samples.

Graphs and Conclusions:

Conclusion-1 From graphical view of speech sample:

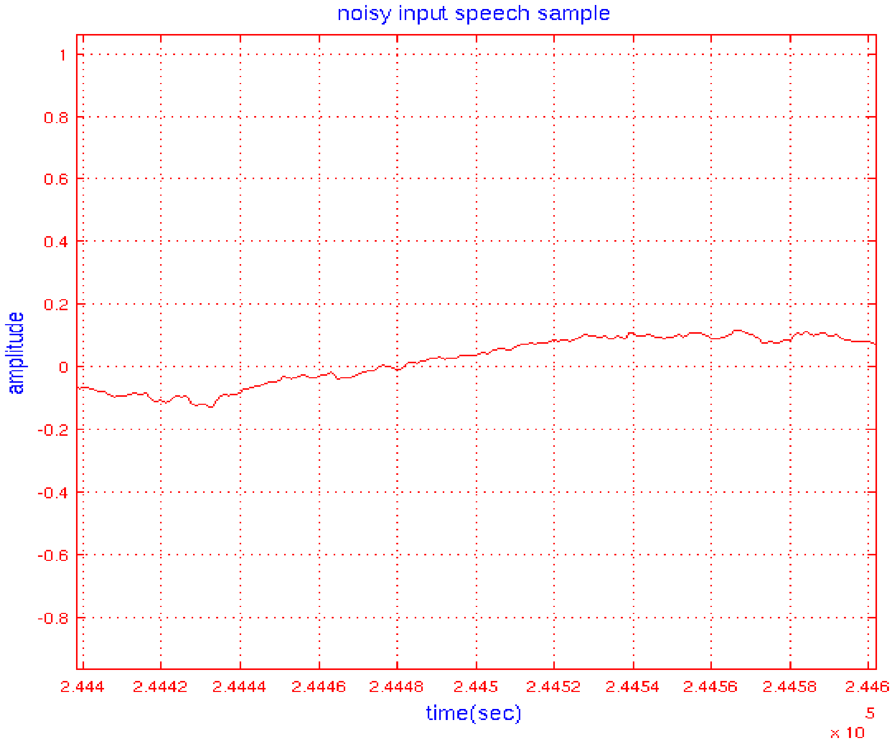


Fig. 8.2 Noisy Speech Signal

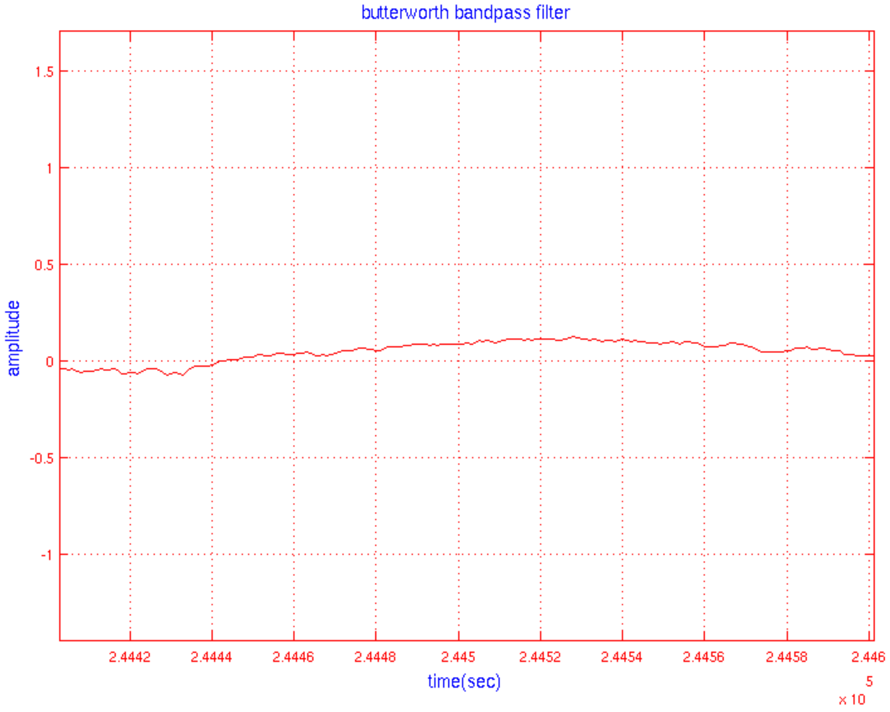


Fig. 8.3 Filtered speech using butter-worth band-pass filter

In fig. 8.2 we can observe that,there is sharpness in the curve. After applying filter,we can observe that curve get smoothed i.e. the speech that we have acquired get filtered in some extent(see fig. 8.3).

Conclusion-2 From pole-zero plot of filter's transfer function:

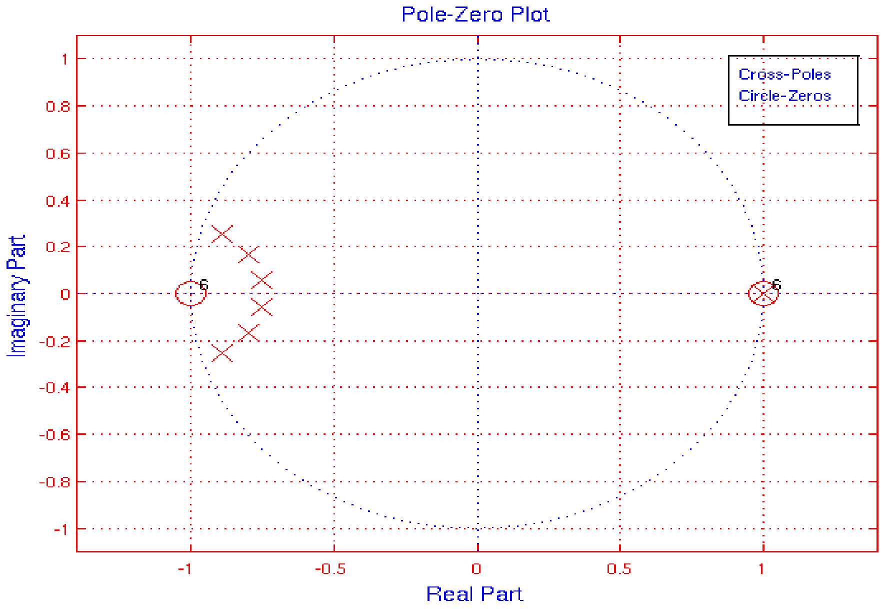


Fig. 8.4 Pole-zero plot

From pole-zero plot,we can observe that all poles and zeros are inside the unit circle which means that the filter that we have designed is stable.

Conclusion-3 From magnitude and phase response characteristics of filter:

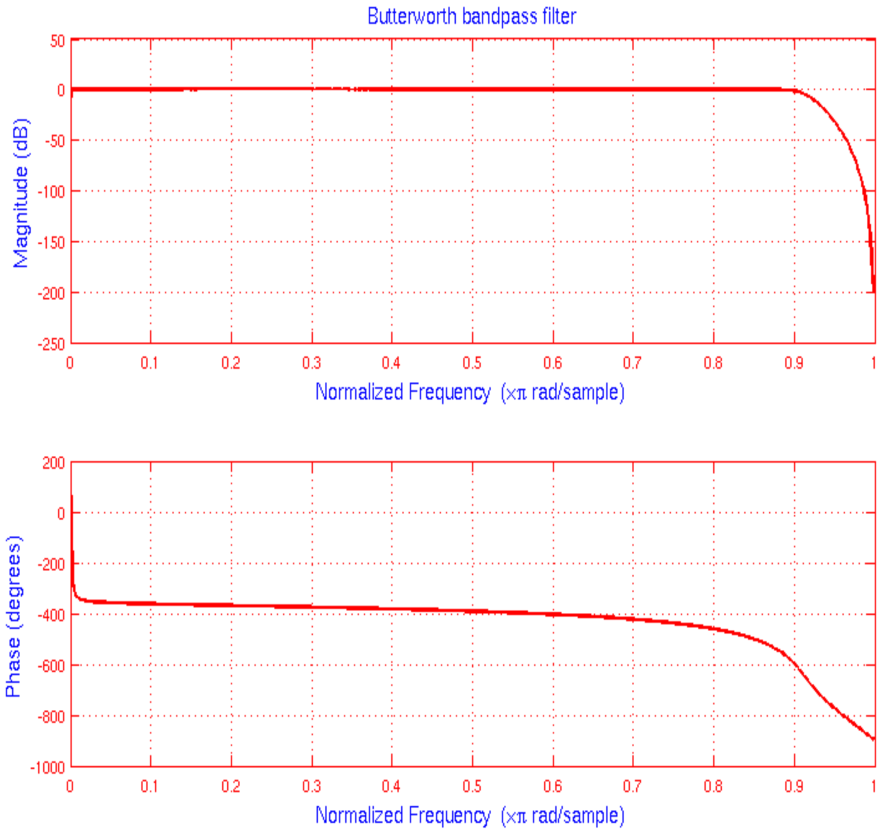


Fig. 8.5 Magnitude and phase response of filter

Butter-worth is a type of signal processing filter designed to have as flat a frequency response as possible in the passband. As shown in upper plot,frequency response of our designed butter-worth filter is somewhat flat in passband,meaning that there is minimum attenuation of frequency band. On other hand,in lower plot, phase response is not linear (phase response is not linear function of frequency) meaning that all frequency components of the input signal are not shifted in time by same constant amount.

8.1.2 Silence removing-conclusion-4 silence removed from signal:

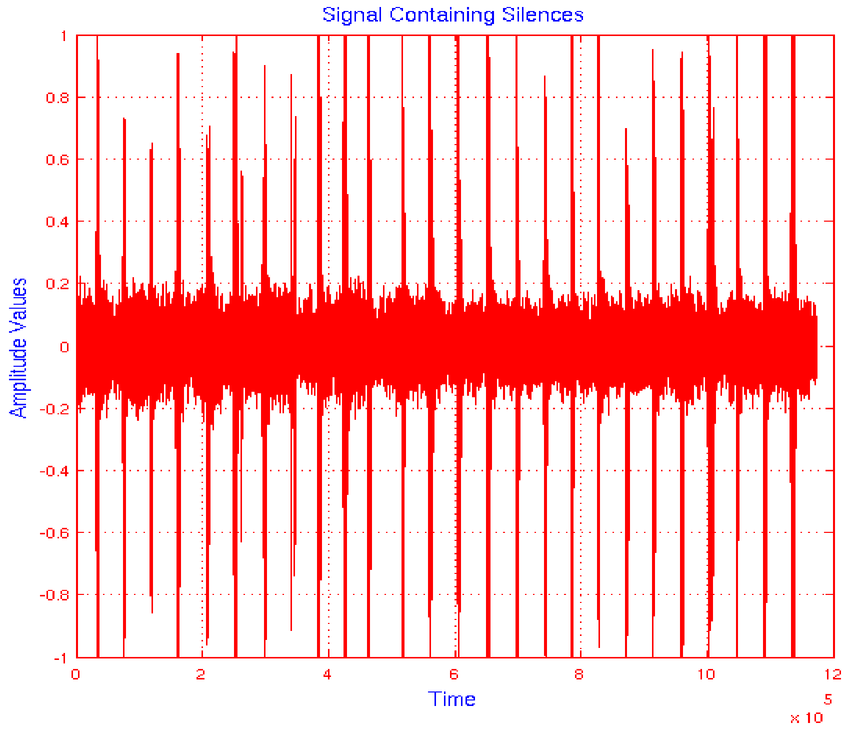


Fig. 8.6 Signal containing silences

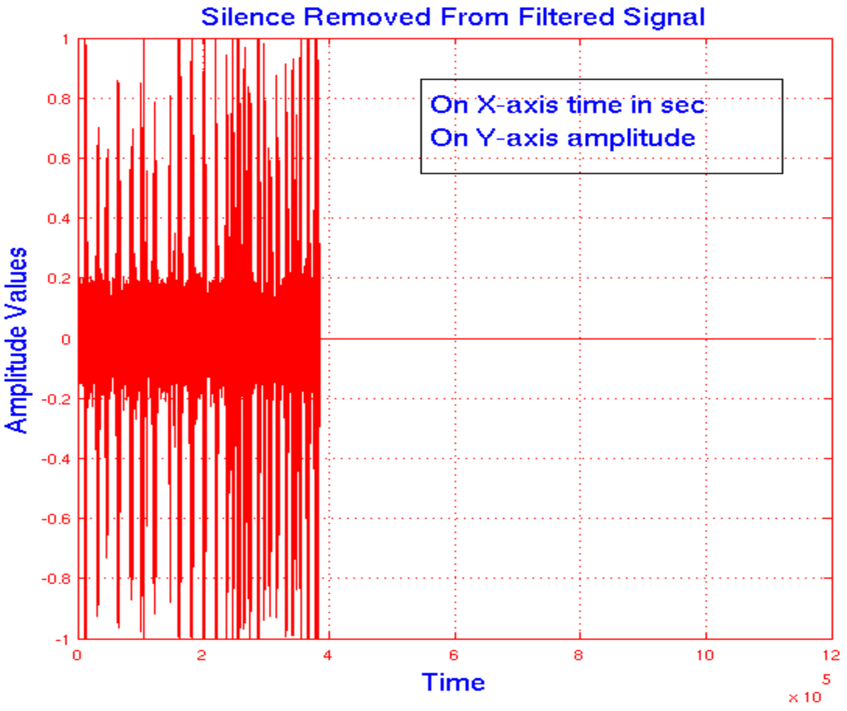


Fig. 8.7 Signal obtained by removing silence

By observing the signal containing silences,non-silent frames are identified and extracted by finding frames with max amplitude more than 0.2212.Since,silence doesn't contain any information for analysis,silent frames are removed for further processing.

8.2 Results of feature extraction step:

Steps:

1) I have read the noise removed and silence removed audio file by 'wavread' command in 'MATLAB':

```
samples=wavread('file.wav')
```

'wavread ( )' command returns the amplitude values.

In this project , file named 'new_va.wav' is read by using wavread function , it returns the amplitude values such as: -0.0591,-0.0627,.........

2) Thereafter , I obtain the size of total number of amplitude values in order to use it in 'wmaxlev' command by using:

```
s=size(samples)
```

It returns value for total number of samples for above '.wav' file: 1173567

3) 'wmaxlev' then calculate the minimum number of levels for wavelet decomposition:

```
l=wmaxlev(s,'wname')
```

where 'wname' specifies the name of wavelet family,here wname=db4

In case of above '.wav' file, total number of levels , l=17.

4) The arguments 'samples' , 'l' is then used in 'wavedec' command to decompose the signal up to the levels 'l' which returns the coefficients of details from level 1 to 'l' ,coefficient of approximation at level 'l' and length of each coefficient in 'L':

```
[C,L]=wavedec(samples,l,'db4')
```

In case of our project: C=cA1,cd1,cd2........cd17 and

L=15,15,24,42,78,150,293,580,1153,2299,4591,9175,18343,36680,73354,146702,293397,586787,1173567.

5) The coefficients of details namely cd1,cd2,.....,cdl are extracted by using the 'detcoef' command:

```
[cd1,cd2,.....,cdl]=detcoef(C,L,[1,2,.....,l])
```

Such as, cd17=0.1381,0.0345,-7.3360,2.5220,-1.2655,0.6315,0.0896,-0.0324,0.0066,0.

6) The power spectra for each coefficients of details is computed by using the following formula:

```
pow_spec_cd=abs(cd)^2
```

Results are shown in upcoming pages.

7) The spectral centroid for each sub-band by using the formula given in chapter number-6 section 6.12.1, is then computed & then by computing mean of spectral centroids of all sub-band first feature is formed. Plots are shown in upcoming pages and conclusion also made below the plot.

Similarly, the spectral flux,spectral spread and spectral skew are calculated and their mean values are considered as our other three features.

8) Finally,we have feature vector:

Feat_vect=[mean_spectral_centroid,mean_spectral_flux,mean_spectral_spread,mean_spectral_skew]

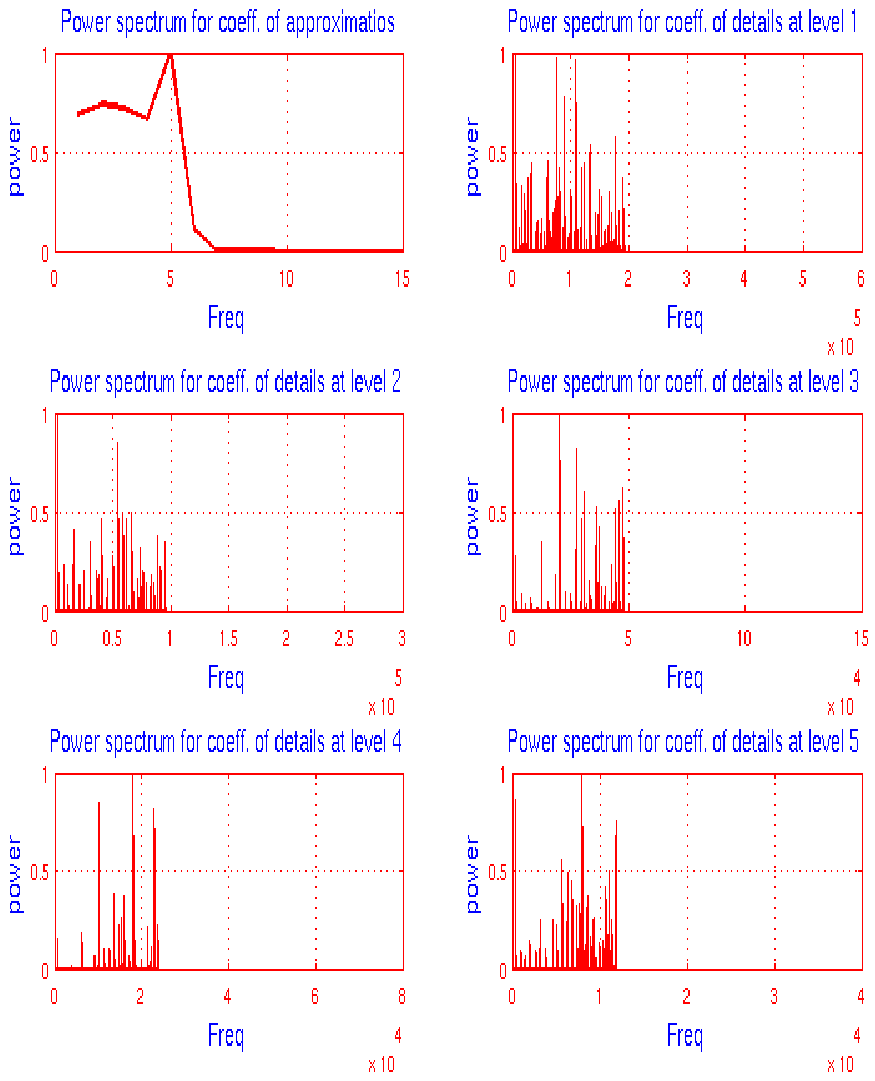


Fig. 8.8 Power spectra of coefficients of details from level 1 to 5

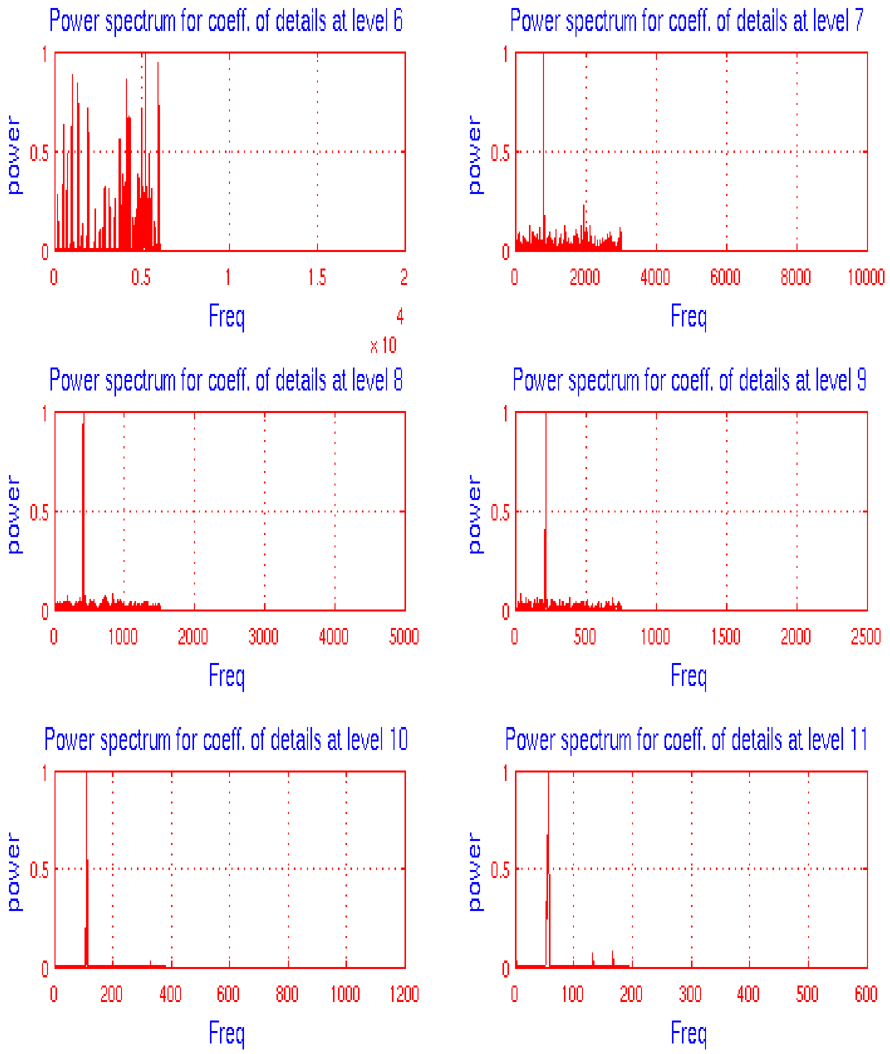


Fig. 8.9 Power spectra of coefficients of details from level 6 to 11

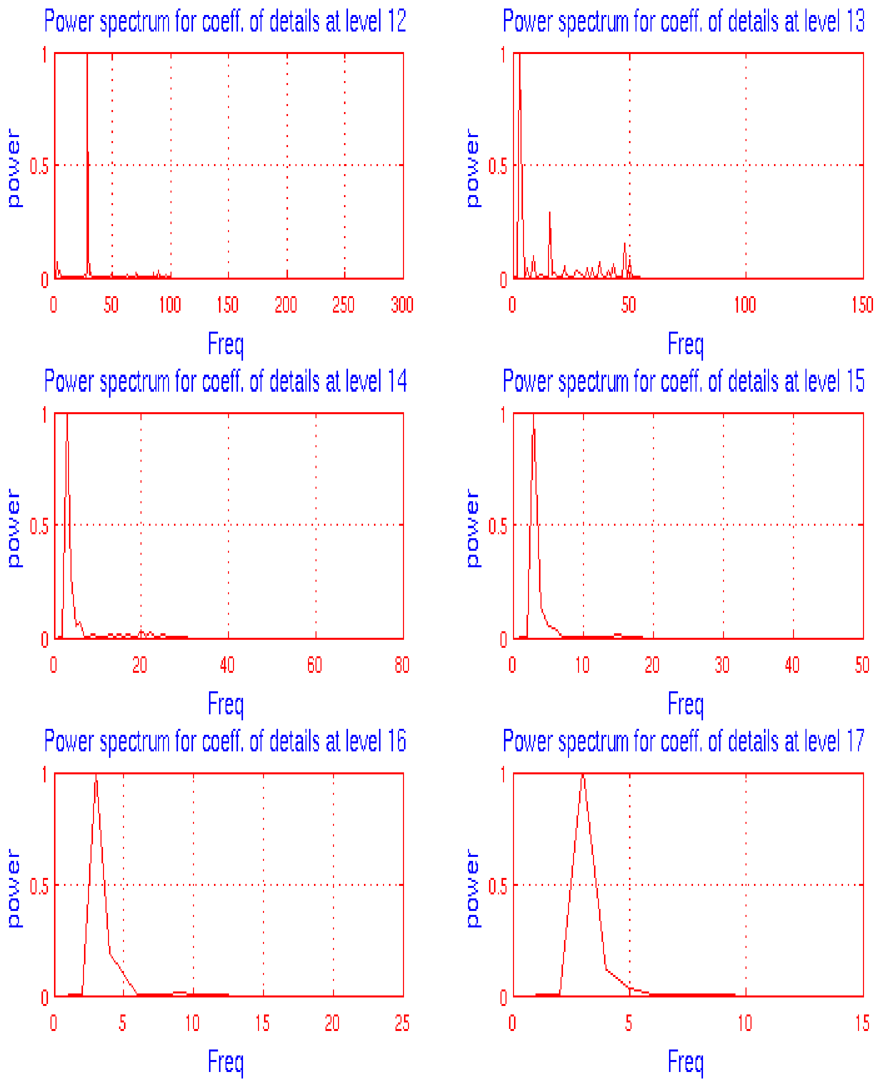


Fig. 8.10 Power spectra of coefficients of details from level 12 to 17

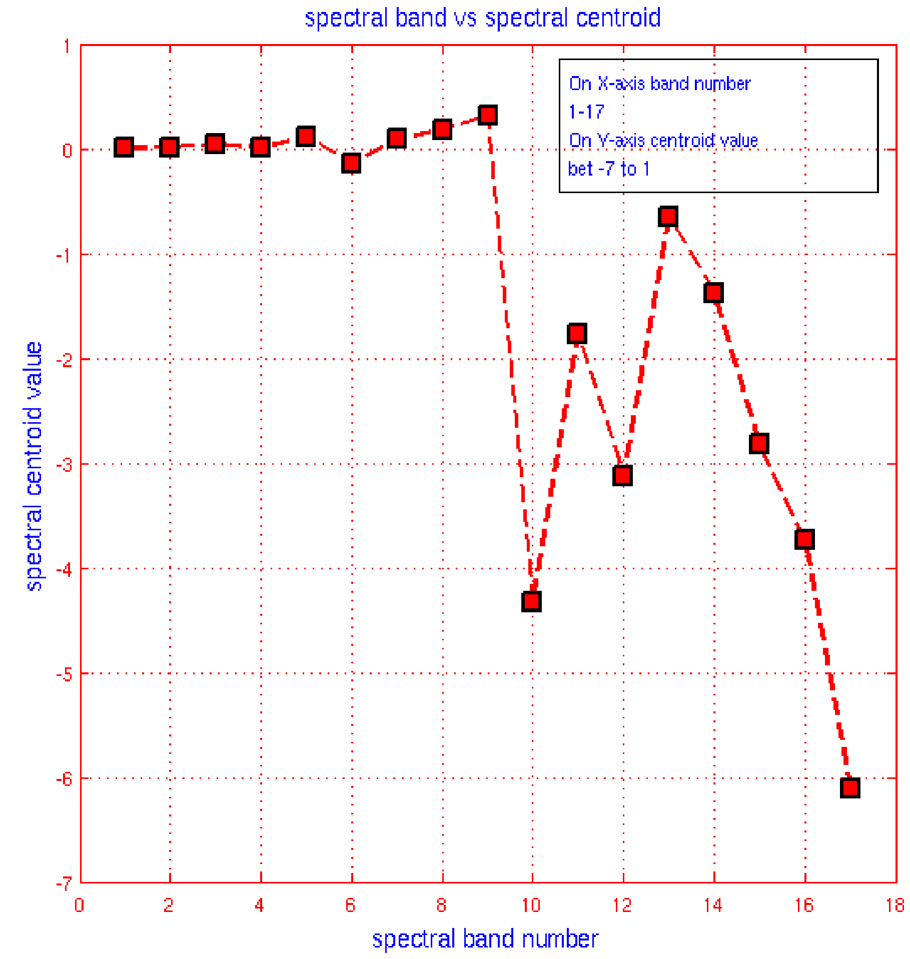


Fig. 8.11 sub-band vs spectral centroid

Negative value of spectral centroid shows that more energy is located in the lower frequency components and vice versa.

Spectr-al Band Numb-er	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Centr-oid Value	0	0.0092	0.0458	0.0021	0.11	-0.13	0.09	0.18	0.31	-4.3	-1.7	-3.1	-0.6	-1.3	-2.8	-3.7	-6.1
Mean Centr-oid	-1.3716																

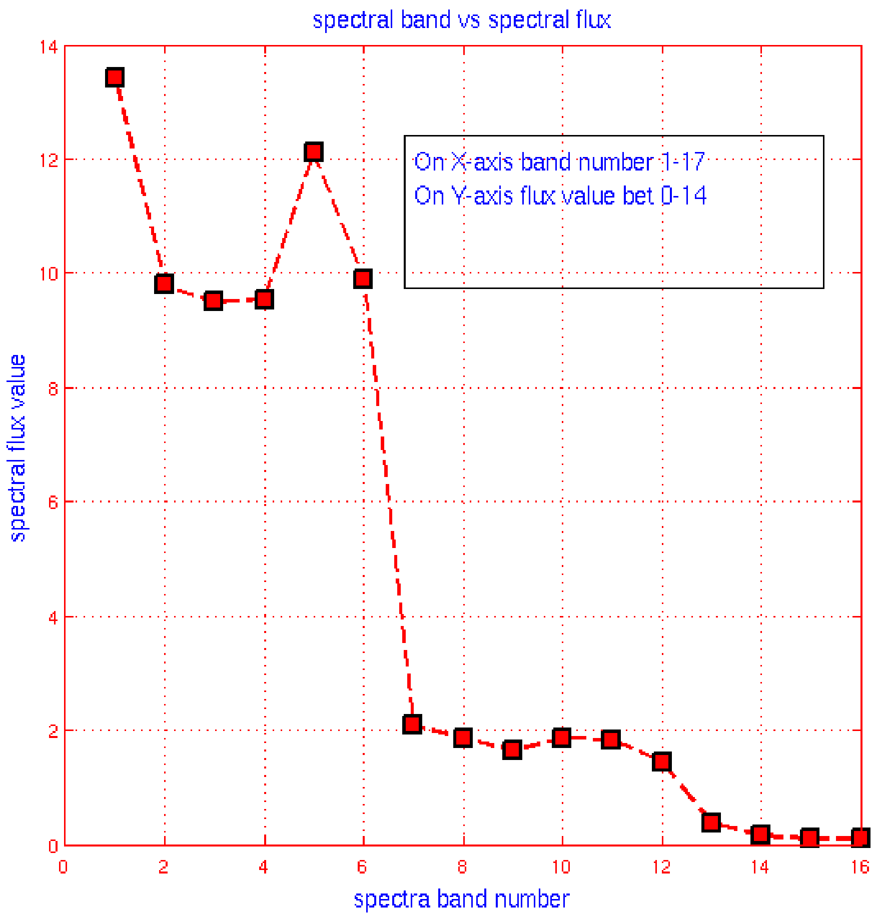


Fig. 8.12 sub-band vs spectral flux

Each flux value shows that 'how quickly the power spectrum changes from previous to next sub-band'.

Spectral Band Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Flux Value	13.4	9.7	9.4	9.5	12.11	9.8	2.08	1.85	1.64	1.86	1.83	1.45	0.37	0.15	0.09	0.09
Mean Flux	4.73															

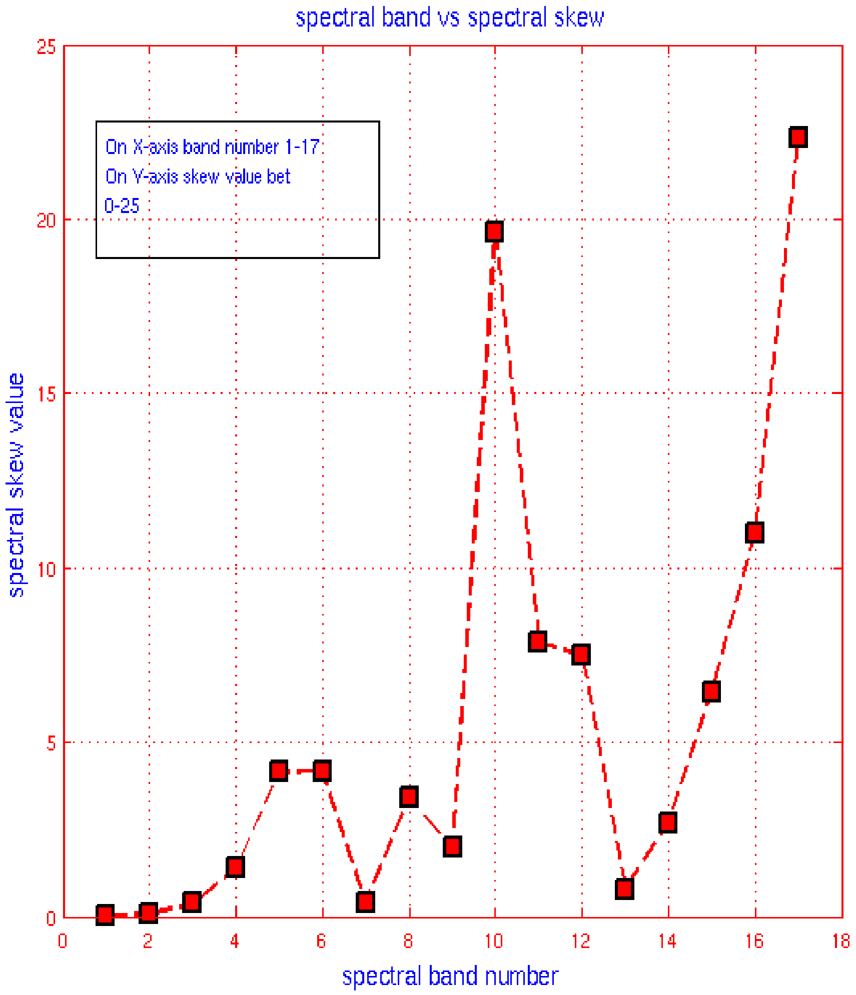


Fig. 8.13 sub-band vs spectral skew

Positive skew values shows that energy of signal in band is more in high frequency components of spectrum and vice versa.

Spectr-al Band Numb-er	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Skew Value	0.0026	0.06	0.37	1.39	4.13	4.16	0.41	3.40	1.96	19.6	7.87	7.48	0.78	2.67	6.45	11.01	22.3
Mean Skew	5.54																

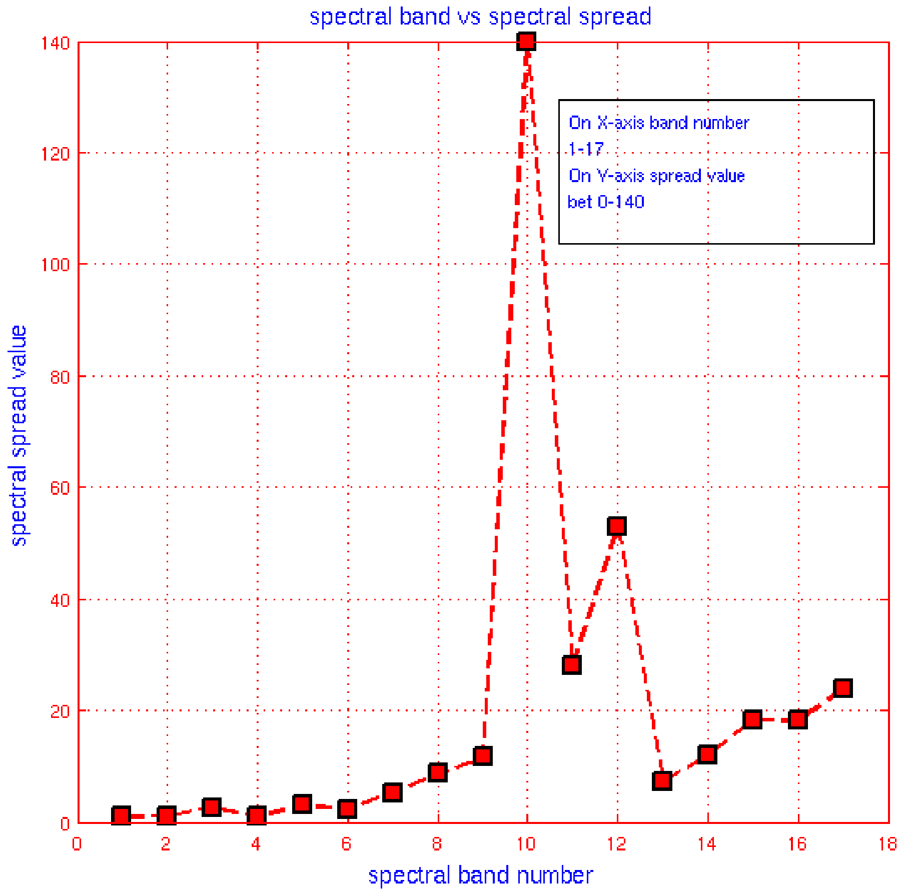


Fig. 8.14 sub-band vs spectral spread

Lower spread values means the spectrum is highly concentrated near the centroid and higher value means that it is distributed across wider range at both sides of centroid.

Spectral Band Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Spread Value	0.99	1.17	2.57	099	3.02	2.33	5.30	8.63	11.64	139	27.9	52.8	7.37	11.8	18.2	18.1	23.8
Mean Spread	19.81																
Feature Vector	[mean_centroid,mean_flux,mean_skew,mean_spread]=[-1.3716,4.73,5.54,19.81]																

Bibliography

[1] John G. Proakis and Dimitris G. Manolakis: “Digital Signal Processing”, Prentice Hall,2012.

[2] Charles K. Chui: “Wavelets : A Mathematical Tool For Signal Analysis” ,SIAM,1997.

[3] John G. Proakis and Vinay Ingale:“Digital Signal Processing Using MATLAB”,Global Engineering,2012.

[4] K.P.Sopan , K.I.Ramchandran and N.G.Resmi : “Insights Into Wavelets”, PHI Learning Pvt. Ltd.,2013.

[5] Journal of computing , vol-2 , issue-3 , march-2010 , ISSN 2151-9617 http://journalofcomputing.org/volume-2-issue-3-march-2010/.

[6] Oppenheim : “Digital Signal Processing” , Prentice Hall , (2002).

[7] International journal of computer science and information security , vol-6 , issue-3 , 2009 , ISSN 1947-5500:https://sites.google.com/site/ijcsis/vol-6-no-3-december-2009.

[8] International journal of engineering trends and technology, vol-4 , issue-2 , 2013 , ISSN 2231 5381:http://www.internationaljournalssrg.org.

[9] Lawrence R. Rabiner,Fellow,IEEE: “A Tutorial on HMM and Selected Applications in Speech Recognition”

[10] Christopher D. Manning & Hinrich Schutze: “Foundations Of Statistical Natural Language Processing”,MIT Press,2003.