CHAPTER 2
LITERATURE SURVEY

As discussed in Chapter 1, two major areas of audio signal processing are Speech and Music signal processing. Speech signal processing includes speech recognition and synthesis, speaker identification and verification, etc (Rabiner and Juang 1993). As far as music signal processing is concerned, the areas of research include music synthesis, transcription, classification, music content analysis, Instrument and voice identification, summarization, etc. Content-based music retrieval forms a major research topic in audio content analysis like rhythm (Lin et al 2009), melody (Schulkind et al 2003) etc. and this area of research is gaining attention (Tzanetakis 2003). Melodic search engines are just a step towards making music documents content-searchable and this necessitates automatic indexing. Instrumentation, harmony, lyrics, meter, Raga, Rhythm, Singer and Genre are other musical structures a listener might wish to find from a given musical piece (Lo and Tsai 2009). Hence it is evident that identifying music content involves significant digital signal processing. It also requires advances in source separation, a process of extracting individual sound sources from a recording of multiple simultaneous sound sources. This Chapter discusses the algorithms that are currently used for Western and Indian music processing.

2.1 BLOCKS OF A SIGNAL PROCESSING SYSTEM

Figure 2.1 gives an overview of a signal processing system. The basic modules include Pre-processing, Segmentation, Feature extraction,
Model Construction and decoding to help identify the components of the input signal. A good set of features together with a robust model, will help in correct content identification of any signal. This chapter discusses the work that has been carried out in various modules of music signal processing towards content identification.

![Signal Processing overview](image)

**Figure 2.1 Signal Processing overview**

### 2.2 PRE-PROCESSING

The pre-processing stage of a signal processing system consists essentially of Noise Removal and Signal separation (Stern 2005).

#### 2.2.1 Noise Removal

Typically, noise is removed from speech signals to increase the performance of the system (Stern 2005). However, performing noise removal for processing music signals generally results in removing important information content, and hence, during music processing noise is not removed from the input (Klapuri and Davy 2006).

#### 2.2.2 Signal Separation

The next module of pre-processing for speech and music processing is Signal separation. Signal separation can also be thought of as source separation, and can be defined as the process of identifying and isolating the
various signals present in a mixture of sound signals. This process of Signal separation can be applied to speech and music. In speech processing, the typical “cocktail” party problem can be defined as isolating an individual speech from a mixture of speeches of many individuals. Several approaches to solve this problem for speech include the Independent Component Analysis (Comon 1994), Statistical independence between signals (Yellin and Weinstein 1996) and Blind Signal Separation (Lee et al 1999). The fundamental idea behind all these approaches is to design a sequence of filter banks to separate the input into individual signals. This is achieved under the assumption that the fundamental frequency of every speaker is unique.

On the other hand, signal separation in a typical music processing system deals with isolating the voice and non-voice components of the signal, again using the principle of how the ear processes an audio signal (Chaffe and Jaffe 1986). In the Spectral filtering approach to signal separation proposed by Every and Szymanski (2004), the separation is aided using a bank of filters. This work attempted to separate signals corresponding to a mix of musical Instruments, into signals consisting of individual Instruments. The spectral filtering approach is based on examining the spectral characteristics, and designing a filter for the same. In the pre-processing stage of this approach, the input signal is split into overlapping frames and the Discrete Fourier transform of each frame is computed and windowed using a Hamming window. Using the characteristics of Western music, where the transcription of a note will be assigned to a constant pitch, a refinement is made of the MIDI pitches for all the time frames of all the notes in each frame. In the second stage, a filter is designed in the frequency domain for each Instrument; its purpose is to remove the harmonics assigned to that Instrument from the spectrum. Then, using the output of each filter, the inverse DFT is used to reconstruct the music signal.
Another algorithm for Western music signal separation proposed by Zhang and Zhang (2005) is based on Harmonic Structure modelling, where the signal is more stable at its harmonic when compared to its monophonic representation. The idea behind this algorithm was to learn a Harmonic Structure model for each music signal in the given musical piece, which consists of the voice and Instrument, and then separate the signals by using these models to distinguish the harmonic structures of the different signals. The mixed music signal is pre-processed by normalizing the mean and energy of the input signal. Then, in the next step, the pitch and the harmonic structures are estimated. Zhang and Zhang have used Terhardt’s algorithm (1979) for estimating the pitch and the harmonics. The spectral peaks of each frame are established, and the number of frequency peaks, which exceeds a threshold, is considered as the frequency of the harmonic component of that particular frame. After determining the harmonic of the frequency, the Average Harmonic value is computed, and using this value, the signal is separated into voice and non-voice components.

All the algorithms presented here for speech and music, are based on designing a bank of filters, and separating them using some distinct characteristics like structure stability, fundamental frequency etc. Therefore, after considering the three algorithms for music signal separation, it has been decided to verify the robustness of these algorithms for Carnatic music processing, rather than attempting for a new algorithm for signal separation.

2.3 SEGMENTATION

In general, audio segmentation algorithms are divided into two categories: Model-based algorithms and Novelty-based algorithms. Model-based algorithms match the trajectory of the feature values with a pre-defined model for identifying and labelling the audio segment, while the Novelty-
based algorithms identify abrupt changes in the trajectory of the feature values alone to decide points of segmentation (Aucouturier et al 2005).

**2.3.1 Model Based Approach**

Herrera et al (2000) proposed several strategies for a music content analysis system, which examined different model-based methods based on supervised learning, like the Support Vector Machines, Neural network, and Bayesian Classifiers. These systems for music segmentation were based on identifying the musical Instruments. In an approach proposed by Raphael, Hidden Markov Models (HMM) are used as the basis for segmentation (1999). The notes pattern of a large corpus of data was collected and analyzed, and used to construct the HMM, which in turn, is used for segmentation. Aucouturier and Sandler (2001) have also used Hidden Markov Models to segment music signals, based on observing the steady statistical property, conveyed by means of the music texture.

Gao et al (2003) also used Hidden Markov Models to segment musical signals into a continuous sequence based on the presence or absence of notes.

**2.3.2 Novelty-based approach**

Novelty-based algorithms for segmentation used general methods of segmenting based on features. Tzanetakis and Cook (1999) implemented some schemes to segment audio streams, using features such as the spectral centroid, spectral flux and the Zero-Crossing Rate (ZCR). An audio texture-based temporal segmentation of the music signal, to be used for music retrieval was also attempted (Tzanetakis and Cook 1999a). Audio texture is identified by determining the sudden change of feature values conveyed by means of the MFCC, LPC, Spectral Centroid etc. These areas in the signal
that reflect this sudden change in features are the segmentation points. Foote (2000) used acoustical parameters in Slaney’s auditory toolbox (1996), to calculate the local self-similarity in music, and also defined a kernel correlation to calculate the audio novelty for music segmentation.

### 2.3.3 Hybrid approach

Several researchers have proposed hybrid segmentation algorithms for Western, music, which are both Model-based and Novelty based. These algorithms are based on the self-similarity matrix (Foote 2000), human perception (Jian et al 2003), harmony (Jensen et al 2005), using rhythm, timbre, harmony etc., (Jensen 2007). The algorithms proposed by Jensen et al (2005) and Jensen (2007) are based on extracting features, and computing a self-similarity to identify the segmentation points. Another hybrid approach used semantic features, such as the beat and phrase detection for segmentation, which is based on segmenting the input signal into fixed length frames, and using the cosine measure to obtain the similarity between the frames, and thereby determine the points of segmentation (Ong and Herrera 2005).

### 2.3.4 Other approaches

Obviously, both modeling and training are time consuming in the model based approach. On the other hand, the extraction of features yielding to the identification of segmentation points is difficult in Novelty based algorithms.

In the work done by Jian et al (2003), a new approach based on human perceptual properties, which is neither model-based nor novelty-based, was proposed for segmentation. In their work, perceptual features like Roughness, Loudness and Periodicity pitch, which determines four musical
perceptual properties like, Timbre, beat, loudness and pitch were extracted. The trajectory of the feature values is identified and a ranking algorithm has been designed to determine the points of segmentation.

Thus, the observation is that Western music processing has been preceded by segmentation, using either a fixed duration or the characteristics of Western music, which are conveyed by signal level features. In this thesis, we try to exploit the characteristics of Carnatic music for the process of Segmentation, rather than using fixed size segmentation.

2.4 FEATURES FOR MUSIC SIGNAL PROCESSING

A good set of features derived from noise free logical segments help in constructing a robust model, to aid error free content identification. In this thesis, we discuss two sets of features – raw signal level features, which include temporal, spectral, and Cepstral features, and music content features, which define the musical characteristics. Some temporal features include energy, amplitude and the Zero-crossing rate. These features are typically used by noise removal algorithms. Spectral features are frequency based features, which are extracted by initially converting the time based signal into the frequency domain using the Fourier Transform. Some spectral features include the Fundamental frequency, Frequency components, spectral centroid, Spectral flux, Spectral density, Spectral roll-off, Energy, etc. These features have information content, and can be used to identify the notes, pitch, rhythm, and melody. The third class of signal level features, namely, Cepstral features are computed by determining the cosine transform of spectrum. Some of these features include the Mel-frequency Cepstral coefficients, Linear Prediction Coefficients, Perceptual linear prediction coefficients, etc. These features typically convey the timbre characteristics, the overall content of the signal, and hence, are used for identifying the Singer, Instrument, Genre or to determine the similarity between signals. Some of the musical content
features include pitch, rhythm, melody, harmony, etc. and some Indian music specific features such as the Raga, Tala, Gamaka, etc.

Let us discuss briefly the various kinds of features that are available for processing.

2. 4. 1 Temporal Features

1. Zero Crossing: It is defined as the number of times a signal crosses the zero level reference. It is a very simple technique, which is used to identify the noise signal, pitch detection, etc.

2. Auto-correlation: It is used to find the similarity between the signal and a shifted version of itself. If the signal is harmonic, the autocorrelation function will have peaks in multiples of the fundamental frequency.

3. Intensity: It is related to the amplitude, and thus to the energy, of the vibration. It is also a measure of the energy flux averaged over a single period.

4. Energy: This feature here is based on the amplitude in the time domain. It is the measure of the RMS amplitude of the frame, determined as the sum of the square of the amplitude over a single period. This feature helps to identify the presence of a dominant component in a frame.

2.4.2 Spectral features

1. Fundamental frequency: It is the lowest frequency of a periodic waveform, often referred to as the formant frequency, or the frequency of the first harmonic.
2. **Tonic:** In Carnatic music the frequency of the middle octave ‘S’ is referred to as the Fundamental Tone (Oke 2004), Tonic (Arthi et al 2011) or ‘Aadhara Sruthi’ (Vidya 2009) all referring to the variable frequency of the Shadja ‘S’. The tonic, the frequency of the Shadja ‘S’ chosen by the performer, is the reference frequency around which other notes are defined (Ashwin et al 2012). In this thesis this frequency is referred to as the tonic.

3. **Spectral Tilt:** It is a measure of a signal’s power distribution vs frequency, which gives the change of frequency between adjacent segments of a music signal (Goncharoff et al 1996) (Kirss 2007).

4. **Spectral Centroid:** It is the balancing point of the subband energy distribution. It is calculated as the first moment of the energy distribution. It determines the frequency area around which most of the signal energy concentrates (Pfeiffer and Vincent 2001) to indicate brightness of sound.

5. **Spectral shape of the residual and sinusoidal component:** This specifies the pitch contour of the signal, which is useful for signal processing (Kirss 2007).

6. **Spectral flux:** It gives a measure of the local spectral change. It is defined as the parameter that determines the change of spectral energy distribution between successive windows (Li et al 2004).

7. **Pulse metric:** It is a novel feature, which uses long-time band-passed autocorrelations to determine the amount of “rhythmicness” in a 5-sec window. It has a property to state
whether there's a strong, driving beat (i.e., techno, salsa, straight ahead rock-and-roll) in the signal. It cannot detect rhythmic pulses in signals with tempo changes.

8. Spectral Roll-off Point: It is the 95th percentile of the power spectral distribution. This measure distinguishes voiced from unvoiced speech as it is a measure of the “skewness” of the spectral shape.

9. Cepstrum Re-synthesis Residual Magnitude: It is the 2-norm of the vector residual after cepstral analysis, smoothing, and re-synthesis.

2.4.3 Cepstral Features

1. MFCC: The Mel-frequency Cepstral coefficients are a perceptually motivated feature set that describes the shape of the spectrum for a short-time audio segment. For each segment, the spectrum is computed by means of DFT. Then mel-scaling is applied using the following equation:

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

   (2.1)

   The frequency components are separated into bins and then the discrete cosine transform (DCT) is applied. From the result the first 13 coefficients are used for processing.

2. PLP: The Perceptually linear prediction coefficients are also a set of perceptually motivated features. For each segment, find the spectrum using the DFT. Calculate the Bark scale frequency. Establish the loudness. The result is a set of coefficients that characterizes the spectral shape and are used as PLP features.
2.4.4 Music Content Features

1. Pitch: This feature is related to the perception of the fundamental frequency of a sound. Pitch is said to range from low or deep to high or acute sounds. It can be derived by estimating the frequency components present in the input signal, which is computed by performing the Fourier transform.

2. Loudness: Intensity is also defined as loudness. This can be identified by determining the spectral energy of the signal.

3. Timbre: This feature is defined as the sound characteristics that allow listeners to perceive the distinction between sounds with the same pitch and same loudness. This feature distinguishes Instruments, and Singers in a given music signal. This is typically conveyed by computing the Cepstral features.

4. Tempo: It is defined as the speed at which a song is played or sung.

5. Tonality: This feature of a song is related to the role played by the different chords of a musical work; tonality is defined by the name of the chord that plays a central role in a musical work. However, the concept of tonality is not applicable to some music Genres.

6. Melody: Melody is a sequence of notes, which constitute the main, most prominent line or voice in a piece of music, or the line that the listener follows most closely. When
accompanied, melody is often the highest line in the piece (voice, violin, flute) and thus stands out clearly. Melody is often the most memorable aspect of a piece of music.

7. Rhythm: It defines how sounds in a piece are grouped and placed in time, often in relation to a pulse. The process in which notes of different durations are organised into patterns.

8. Harmony: Harmony is the succession of chords, or chordal progressions made by two or more parts, or voices, playing or singing together.

9. Raga: It is a special characteristic of Indian music which is conveyed by a pre-defined melodic arrangement of notes. It can be determined by identifying the frequency components followed by the swara or note identification from a given musical piece.

10. Tala: Is yet another important characteristic of Indian music which is the periodic repetitions that accompany a musical piece.

11. Gamakas: These are important characteristic of Carnatic music which is defined as pitch inflexions. The variation from one note to the next is defined as a continuous function, and is not discrete. There are nearly ten ways in which the transition can happen between notes (Sambamurthy 1983). Gamakas essentially characterize a Raga.

12. Meends: They are the variations of frequency given to a note or combination of notes. The variations of notes can be given upto an octave. It is the glide that is given to define the transition between notes.
2.5 FEATURES CURRENTLY USED FOR MUSIC ANALYSIS

The temporal, spectral and Cepstral features that are extracted, can be used for music processing, namely, separation of speech and music, music classification, music content identification, music information retrieval, etc.

2.5.1 Features for Separating Speech and Music

In their work, Scheirer and Slaney used nearly thirteen features (1997) to separate speech and music. Of the thirteen, the authors have used five “variance” features, consisting of the variance in a one-second window of an underlying measure, which is calculated on a single frame. If a feature has the property, that it gives very different values for voiced and unvoiced speech, but remains relatively constant within a window of musical sound, then Scheirer and Slaney have identified that the variance of that feature will be a better discriminator than the feature itself. It is also possible that other statistical analysis of underlying features, such as the second or third central moments, skewness, kurtosis, and so on, might make good features for discriminating classes of sound. They also use the variances of the roll-off point, spectral centroid, spectral flux, zero-crossing rate, and Cepstral re-synthesis residual magnitude as features. In practice, log transformations on all thirteen features have been used, which have been empirically determined to improve their spread and conformity to normal distributions. They also concluded that the multidimensional classifiers, which they have built using, these features provided excellent and robust discrimination between speech and music signals in digital audio. Scheirer (1998) also attempted to understand the tempo and beat of a music signal, based on extracting the spectral and temporal features.

2.5.2 Features for Music Classification and Content Identification
In their work McKinney and Breebaart (2003) performed Music classification by using four sets of features. The features include low-level signal properties, MFCC, psychoacoustic features including roughness, loudness and sharpness; and an auditory model representation of temporal envelope fluctuations. In the work of McKinney and Breebaart (2003), the classification of audio files was performed using quadratic discriminate analysis (Duda and Hart 1973), which provided better preliminary results than linear discriminate analysis. Features were calculated from each file on 10 consecutive 743-msec frames with a 558-msec hop-size. The feature vectors were grouped into classes based on the type of audio and was used to parameterize an \( N \)-dimensional Gaussian mixture model (one Gaussian with its own mean and variance for each class), where \( N \) is the length of the feature vector. Although the size of the feature sets differed, classification was performed using the best nine features, from each set and determined an iterative ranking procedure based on Bhattacharyya distances (Papoulis, 1991). It was concluded by the authors that, combinations of the best features from each set lead to improvements in classification performance. They also concluded that, the features could be ranked across sets or within feature set, and then choose the combination to yield the best performance for classification.

In a work proposed for music segmentation and classification (Zhang and Kuo 2001), the features considered are the pitch, pitch strength, spectral centroid, zero-crossing rate, spectral roll-off frequency, and MFCC coefficients. They have approximately estimated an overall combination of 50 features. In addition to these features, the authors extracted psychoacoustic features, where in the characteristics of Western music was taken into consideration.
Temporal features and Cepstral coefficients were used for Instrument recognition by Eronen and Klapuri (2000). In their work a wide set of features covering both spectral and temporal properties of sounds were investigated and algorithms were designed for their extraction. The authors validated the usefulness of the features using test data that consisted of 1498 samples covering the full pitch ranges of 30 orchestral Instruments from the string, brass and woodwind families, played with different techniques. They concluded with a claim of recognizing the correct Instrument family with 94% accuracy and individual Instruments with 80% accuracy.

In a step towards automatic Instrument recognition, Eronen have used LP coefficients, MFCC and WLP (warped linear prediction) coefficients for the purpose of Instrument recognition (2001). Eleven Cepstral coefficients were calculated separately for the onset and steady state segments based on conventional linear prediction with an analysis order of 9. This resulted in the length of the feature vector calculated for each isolated tone, a total of 44 features. The validation database consisted of the MUMS samples. The sample included 1498 solo tones covering the entire pitch ranges of 30 orchestral Instruments with several articulation styles (e.g. pizzicato, martele, bowed, muted, flutter). The author has taken all tones from the McGill Master Samples collection, except the piano and guitar tones. The author observed that the best accuracy among all features was 33% for individual Instruments (66 % for Instrument families) with WLP Cepstral coefficients (WLPCC) of order 13.

LFCC (log frequency Cepstral coefficients) is used to represent timbre in speech recognition and some music tasks. LFCC is a simplification of MFCC where the full range of Cepstral coefficients are used as against the first 13 to 20 coefficients that is being used in MFCC (Casey and Slaney 2006). The chromagram representation captures the musical qualities of the
sound by collapsing notes across octaves. Features were extracted using 375ms windows at every 100ms. The authors used a constant-Q power spectrum with 1/12th octave resolution, aligned with and corresponding to notes in western tonal music. Each element of this spectrum is compressed to approximate loudness perception using a logarithm. Collapsing of each note in the chromagram representation to the base octave, A1–G#2 (55Hz–104Hz), gave an octave-independent measure of the harmonicity of the music. It was showed that temporal queries were more effective at retrieving musically similar segments from their existing music library.

In addition to the techniques already discussed for segmentation, several music-based features like beat, rhythm, and melody have also been used for segmentation of music signals. In all the work that takes these music-based features the signal characteristics emphasizing these features are considered for analysis (Jensen et al 2005).

Agostini et al have used Spectral features like spectral centroid and spectral bandwidth along with inharmonicity, harmonic weakness for Musical timbre identification (2003). The authors conducted the experiments on real acoustic Instruments with little influence of the reverberant field. A preliminary test with performances of trumpet and trombone has shown that the used features are quite robust against the effects of room acoustics. The only weakness is their dependence on the pitch, which can be estimated from the input consisting of monophonic sources only.

2.5.3 Features for Music Information Retrieval

Tzanetakis et al investigated two audio representations, which emphasize the use of temporal features for music information retrieval (2003). The two representations that were discussed are Symbolic feature conveyed through MIDI and Audio feature conveyed through temporal feature values.
Eronen and Klapuri (2000), extracted MFCC features, computed the mean and covariance of the MFCC frames and used SVMs as features, thereby leading to the process of Music Information Retrieval.

The other features that are normally considered for analysis are spectral tilt, amplitude of sinusoids, amplitude of residual, spectral envelope, spectral shape of residual, vibrato, etc. By analyzing the features set, Zhang and Kuo, claim that frequency based features are more relevant than temporal features for indexing a MIR system (2001).

2.5.4 Fundamental Frequency of Speech and Music

Fundamental frequency estimation of the audio signal is a classical problem in signal processing (Camacho and Harris 2008). The estimation of fundamental frequency has been a research topic for many years both for speech and music signal processing. Fundamental frequency is the physical term for pitch (Camacho and Harris 2008). Pitch is defined as the perceptual attribute of sound which is the frequency of a sine wave that is matched to the target sound in a psychophysical experiment. Fundamental frequency is needed in speech signal processing for determining the speaker in Speaker Verification or Recognition systems. The estimation of fundamental frequency is essential in music signal processing in order to determine pitch pattern, range of pitch frequencies, music transcription, and designing music representation systems.

Fundamental frequency is defined as the lowest frequency at which a system vibrates freely. Fundamental frequency is the reciprocal of the time period between the two lowest peak points of a given signal and hence it can also be determined by looking at the time domain representation of the signal to yield the successive lowest peak points. The features that are used for the determination can be classified as Time Domain features, Spectral features,
Cepstral features and features that are based on the auditory theory. Many of the algorithms that are available for fundamental frequency estimation of speech and music are based on the estimation of frequency domain features and auditory motivated features.

A fundamental frequency estimation algorithm for speech and music, developed by Doval and Rodet (1993), is based on the evolution of the signal by assigning a probabilistic value to the pseudo-periodic signal. This algorithm is based on a HMM using the estimated spectral features to identify the fundamental frequency of the signal and hence it required lot of training to determine the evolution of the signal.

Maher and Beauchamp (1994), used a two way mismatch procedure for estimating the fundamental frequency of music signals. In this algorithm, fundamental frequency is determined by computing quasi-harmonic values for short-time spectra of the input signal. The same value is determined in the neighbouring spectra and then the fundamental frequency is estimated as the least value of the sample input segment considered.

An algorithm was developed by Cheveigne and Kawahara (2002) which is a generalized algorithm for speech and music. It is based on the well-known autocorrelation method which is in turn based on the model of auditory processing. The steps involve determining the autocorrelation value, correcting the errors in the computed value by determining the difference between the autocorrelation values, normalizing the value of the difference function by estimating the mean value, and iterating this correlation value to determine the fundamental frequency of the input signal. The time taken to correct the errors is very high as it is a generic algorithm for speech and music which did not exploit the characteristics of the signal. The accuracy of the algorithm is also not very high for their sample data.
In one of the algorithms, the fundamental frequency of speech and music signal has been estimated based on spectral comparisons (Camacho and Harris 2008). The average peak to valley distance of the frequency representation of the signal is estimated at harmonic locations. This value is determined at several segments of the input and the distance is estimated between successive average peaks to valley value. From these values fundamental frequency value is estimated as the least distance of the average peak to valley values. This work for fundamental frequency estimation was generic for speech and music. The time complexity of this algorithm is very high in the worst case situation since the distance measure needs to be computed between successive segments for all possible combinations in the input signal.

Many algorithms have also been designed for estimating multiple fundamental frequencies corresponding to the Singers and Instruments (Yeh et al 2005), (Klapuri2003). In the algorithm developed by Yeh et al (2005) a quasi-harmonic model is developed to determine the components of harmonicity and spectral smoothness, after which a score value is assigned for the computed harmonicity value and spectral smoothness and based on which the fundamental frequency is estimated.

The algorithm developed by Klapuri (2003), is also based on harmonicity, spectral smoothness and synchronous amplitude evolution of the input signal for estimating the fundamental frequency. They have implemented an iterative approach where the fundamental frequency of the most prominent sound is computed and is subtracted from the mixture of signals. This process of computation and subtraction is iterated to determine the fundamental frequency of the signal.

All the algorithms that have been developed for fundamental frequency estimation were designed for Western music and Speech in general.
and hence has used the signal characteristics of music and the characteristics of human speech for its determination. In addition, these algorithms assume that lowest frequency of the periodic component in the input signal is the fundamental frequency. However this concept cannot be used for Carnatic music signal processing because in Carnatic music we are more particular in estimating the Tonic, to indicate the frequency of the middle octave ‘S’ (Sambamurthy 1983). This tonic is essential and it is the basis for determining an important characteristic of that music – the Raga (Sambamurthy 1983). Therefore an algorithm to estimate this frequency corresponding to the middle octave ‘S’ need to be designed for use as tonic. The use of existing fundamental frequency estimation algorithms could also be explored to verify whether they could be adopted to convey tonic.

2.6 RAGA IDENTIFICATION

Very little work has been done in the area of Raga identification of Indian music and in particular Carnatic music. However Raga identification of Hindustani music is carried out by various researchers. Pandey et al (2003) have constructed a HMM to help in the process of Raga identification of two Hindustani Ragas: Yaman Kalyani and Bhupali. They have constructed a HMM for these two Hindustani Ragas, in which they have defined a probabilistic automata based on the notes, to help in the process of Raga identification. The authors have achieved an accuracy of 87% for these two Ragas. The drawback of this system is the various constraints that are used by the system in terms of the fundamental frequency and monophonic music.

In the work done by Belle et al (2009) an algorithm for Hindustani music Raga identification was performed, which uses the intonation given to individual swaras of the Raga of a given song. The authors have used features at the swara level, which are extracted from the signal as the peak value of a swara, its mean value, the standard deviation of a swara and distribution of a
swara for determining the swara thereby yielding to Raga identification. They were not able to achieve a constant value for the mean and standard deviation of the swaras. In addition the work was carried out to determine only Ragas that have the same swaras.

Chordia and Rae (2007) have done Raga Classification of Hindustani Raga using Pitch Class Distribution (PCD) and Pitch Class Dyad distributions (PCDD). In this work they have divided the input signal into segments and determined the pitch by using the Harmonic Product Spectrum algorithm (Cuadra et al 2001). They estimated the onset of the input signal and determined the frequency component at the place of onset. Then using the detected pitch, the PCD and PCDD are estimated based on the histogram of the pitch contour to determine the Raga (Chordia 2006).

In another work done by Chordia and Rae (2008) the authors have designed a tool to recognize some of the Hindustani Ragas. The Raga identification uses the YIN (Cheveigne and Kawahara 2002) for determining the pitch. Using this estimated pitch, the pitch class distribution is plotted, which is used by a Bayesian Classifier to determine the Ragas of Hindustani music. The Bayesian Classifier used the Pitch Class distribution vectors of each Raga to model a Gaussian probability function and this Gaussian probability function is used to determine the underlying Raga. It has been stated that one of the limitations of the algorithm is the use of only signal level features and have specified the need for a note (swara) model for the process of effective and robust Raga determination. This has motivated us to design a Raga model for Carnatic music based on swaras.

Many of the algorithms for Raga identification are designed for Hindustani music and hence cannot be directly adapted to Carnatic music because of the fundamental difference in the systems of music, the dependence of pitch intensity in Hindustani while Carnatic music is not
completely dependent on intensity but dependent on Gamakas and variation of pitch (Sambamurthy 1983). In addition these algorithms are based on identifying Raga by following the pitch contour variation. However, a Raga has other characteristics that need to be explored for its identification.

2.7 OTHER CONTENT EXTRACTION

The other contents that are typically extracted from a music signal are the Singer, Instrument, Emotion and Genre. The following discussion is on the literature on the non-music content identification that was carried out in the Western music scenario.

2.7.1 Singer Identification

Singer identification is typically carried out by analyzing the voice structure of Singers (Tzanetakis 2004, Tsai and Wang 2006). In an early work for identifying Singers, the authors have designed a set of coefficients called the Octave Space Cepstral Coefficients (OSCC / OFCC) which are based on the octave interval of Western music, and used to construct a model by observing the differences between Singers and Instruments (Maddage et al 2004). A GMM is constructed using the OFCC values to identify the Singers.

Li and Nwe (2006) and Nwe and Li (2007) developed new acoustic features for Singer identification, that extracted information about the Singer’s vibrato characteristics. Applying several banks of filters (triangular, parabolic and cascaded), and transforming the resulting energies into the Cepstral domain, they extracted the Octave Frequency Cepstral Coefficients (OFCC). Their experiments on a 12-Singer database showed that OFCCs outperformed MFCCs and LPCCs.

A Bayesian Information Criterion based approach to Singer identification has been proposed in Western music (Mestres 2007). The
system is based on the idea of using only the vocal segments of a song to build the model of a particular Singer. The most important contribution of the technique is in the manner the vocal segments are located. The borders between vocal and Instrumental parts are first detected using the Bayesian Information Criterion (BIC), and then each segment is classified as vocal or instrumental by a decision tree based on MFCCs. From the vocal segments, a GMM for each Singer is constructed using the Expectation-Maximization algorithm.

In another technique for Singer identification (Mesaros et al 2007, Levy 1982) MFCC has been used for model construction. However they have used new distance measure to perform Singer identification. In general the process of Singer identification that has been carried out for Western music were based on identifying Western music characteristics conveyed using the MFCC and OSCC features.

The features that have been used for identifying Western music Singers need to be explored for determination of Singer in Carnatic music and Tamil film music. The possibility of identifying the duet Singers in a musical piece also need to explored.

2.7.2 Instrument Identification

In section 2.4 we have discussed Instrument identification from the perspective of the type of features used. In this section we discuss other approaches to Instrument identification. In a preliminary approach to Instrument identification, Martin and Kim (1998) performed the identification based on a pattern recognition approach. Spectral features were determined and based on computing a log-lag correlogram, which resembles the human auditory system, a Gaussian model is constructed using which Instruments were recognized. In their work, songs containing only Instruments were
considered and performed identification of 14 orchestral Instruments which includes wind and string Instruments. They concluded that identifying cue phrases in musical signals could result in better Instrument identification.

Essid et al (2006a, 2004), have recognized Instruments by determining a pair of features that distinguishes a pair of Instruments. The features designed by them - Octave band Cepstral coefficients was based on the Octave interval of Western music. Using these features a class based pairwise feature selection is performed using inertia maximization procedure. This pairwise selection of features is performed between every pair of Instruments. The key idea is that rather than choosing the features characterizing an Instrument, the features that distinguishes one Instrument from other is used for identification. Using these features a Gaussian Mixture Model (GMM) is constructed to identify the Instrument. The authors recognized ten Instruments consisting of wind and string Instrument. The disadvantage of the system as explained by them seem to be the process of pairwise feature selection as this process will be very high to consider all groups of Instruments.

Vincent and Rodet (2004) have used Independent Subspace Analysis (ISA) for Instrument identification in musical recordings. They computed short-term log-power spectra of possibly polyphonic music as weighted non-linear combinations of probable note spectra including the background noise. These typical note spectra are computed initially using databases containing isolated notes or on solo recordings from different Instruments. After performing experiments on five Instruments they have concluded that this model had theoretical advantages in not performing music transcription over methods based on GMM and linear ISA. It has been concluded that the drawback of the system is its inability to compute the background noise effectively.
In another approach to Instrument recognition, Kaminskyj and Czaszejko have isolated monophonic musical Instrument sounds using six features: Cepstral coefficients, constant $Q$ transform frequency spectrum, multidimensional scaling analysis trajectories, RMS amplitude envelope, spectral centroid and vibrato (2005). Sounds from nineteen Instruments of definite pitch, covering the note range C3–C6 and representing the major musical Instrument families and sub-families were used to test the system. Nearest neighbor classification was utilized to achieve an identification accuracy of 93%. The drawback is the constraints imposed on the features used.

Instrument recognition has been carried out using non temporal frame level features derived from Line Spectral Frequencies (LSF) (Krishna and Sreenivas 2004). The LSF are used to construct a GMM to identify 14 Instruments belonging to 4 Instrument families.

In the work carried out for Instrument recognition Kitahara et al claim that the fundamental frequency $F_0$ should be incorporated to identify the Instrument (2003). Based on this assumption they have considered nearly 40 spectral features of music signal containing only one Instrument solo and computed a multivariate distribution which is based on the Fundamental frequency. Using this distribution Bayes decision rule is used to classify and identify the Instruments. The authors claim that the pitch dependency on Timbre characteristics has been exploited to improve the recognition accuracy to 75%. After considering this work for Instrument identification it can be concluded that fundamental frequency and Cepstral features plays an important role in Instrument identification. We want to explore this feature for Carnatic music Instrument Identification.
2.7.3 Emotion Identification

In the work carried out for recognizing Emotion in speech, Dellaert et al have explained the need for some useful features that need to be extracted (1996). They claim that the useful features include, fundamental frequency f0, mean, standard deviation of the signal, min and max peak frequency, speaking rate and up / down slopes of the frequency. In addition articulatory features are also responsible for the Emotion conveyed by the speech. Based on this work, many algorithms have been designed for recognizing Emotion in speech by constructing a Gaussian mixture model. In this model, using the signal level features, the mean and standard deviation of the signal is computed and a GMM using k-means clustering or the EM algorithm. Tang et al (2009) have recognized Emotion from speech using the GMM technique and have compared this algorithm with their modified algorithm called “boosted” GMM. They have extracted Mel-frequency Cepstral coefficients to train the GMM and the “boosted” GMM and have recognized Emotion from speech with an accuracy of 85% to 90%.

The characteristics of music are different from those of speech, and hence, researchers began work by modifying the algorithms to suit music. In an early work done for recognizing Emotion in music, Li and Ogihara (2004) have used MFCC as a feature to determine similarity between the input and the trained data. This approach is based on identifying the MFCC values in a musical piece and representing it as a 35-dimension vector. They have identified only three Emotions and have achieved an accuracy ranging from 60% to 80% for the three different Emotions.

The work done for speech (Tang et al 2009) motivated researchers to attempt to recognize Emotion from music using a Gaussian mixture model. Yang et al have done work on identifying and classifying music Emotion (Yang et al 2006) in Western music based on the Thayer’s 2D model for
classifying and recognizing Emotion where they construct a 2-D arousal-valence Emotion plane (Yang et al 2006). The Thayer’s Emotion plane defines the Emotion classes in terms of exciting or calming vs positive or negative (Thayer 1989). Yang et al (2006) have constructed a fuzzy class of Emotions using Fuzzy k-NN classifier based on the Thayer’s model by using the mean and variance of the signal. The fuzzy class is constructed by assigning weights to the Emotion that is conveyed in every musical piece. In the testing phase the input is matched with one of the Fuzzy classes. This method achieved an accuracy of 68%.

Another approach called the “regression approach” was attempted for recognizing Emotion (Yang et al 2007, Yang et al 2008). The fundamental idea behind this approach was to classify Emotion by assuming that the Thayer’s model has a continuously varying value rather than discrete values. Features like loudness, pitch multiplicity, spectral contrast are determined to construct and represent a regression model as a continuous quantity of the arousal and valence values. Based on where the input’s arousal and valence value ranges in the regression model the Emotion is classified by referring to Thayer’s plane and claimed to have achieved Emotion classification accuracy between 68% and 84%.

In another work Han et al (2009) have used Support vector machines for the classification. They have constructed the Thayer’s model by incorporating Juslin’s theory (Juslin et al 2001) to classify eleven Emotions based on constructing a regression function using SVM which is constructed using features like energy, rhythm and the presence of harmonics in a music segment. They claim the use of SVM reduced the Emotion classification error. In another work done for recognizing Emotion, features like intensity, rhythm regularity, and tempo to construct a Gaussian mixture model for Emotion detection have been used (Yeh et al 2009). They have used Thayer’s
Emotion model for grouping the Emotions and have achieved a precision and recall of 80%. All of the above approaches are based on grouping Emotions according to Thayer’s Emotion model and using different features to build a training model. Another approach to Emotion identification is based on identifying the lyrics, in which the lyrics is represented by their feature values (Yang and Lee 2009). Using the features a psychological model, which is an 8-quadrant representation, is constructed and is used for identifying Emotions.

In this thesis, Emotion identification of South Indian music need to be carried out by using the relationship between Raga and the Emotion that a Raga can convey.

2.7.4 Genre Identification

Genre refers to the style of music. Western music has varying Genres like Jazz, Rock, Disco, Pop, Blues, etc. Automatic Genre identification is the process of identifying Genre by means of extracting features. In one of the works for music Genre classification of Western music, Tzanetakis and Cook (2002) have performed hierarchical classification of Genre of Jazz, Classical western music HipHop, Rock, Disco, Pop, blues. They have used three sets of features to convey timbral texture, pitch content and rhythmic content. The MFCC features are extracted which are used to convey Timbral feature, pitch content is conveyed by features like Spectral centroid, Spectral flux, Time-based Zero-crossings, RMS energy, Spectral roll-off are extracted. The MFCC features together with other features are represented as a vector to indicate Timbral texture. Rhythmic texture is computed by determining the beat histogram. Using these features a GMM is constructed to identify 10 Genres with an accuracy of 61%. This gave us an idea of using Timbral features combined with spectral features for Genre identification of South Indian music.
Chen et al (2006) have represented the musical piece in a textual form and using this representation the authors have identified the Genre. They have analysed the limitations of representing music using Timbral and Pitch features for the purpose of Genre identification. They further concluded that these features were not able to capture global statistics of the music and the long-term dependency of musical events to help in Genre identification. They have used HMM to convert the musical piece into a textual representation using the time based features and MFCC. This representation of music is clustered using Latent Semantic Indexing (LSI) to convey semantic information. Using this semantic information the authors have used their proposed classifier MC MFoM to identify the Genre of the underlying music. Using the LSI and the MC MFoM classifiers the authors have identified 10 Genres with a classification accuracy of 70%. From the research that has been carried out, spectral and Cepstral features have been used to construct a classifier for Genre identification.

Grimaldi et al (2006), have used the wavelet transform to extract frequency based features including frequency, spectral energy, mean values of energy, and constructed a Round-robin ensemble classifier to identify 5 Genres with an accuracy of 80%. In another work done by DeCoro et al (2007) the authors have achieved Genre classification using Support Vector Machines, which the authors claim resulted in a better performance than other classifiers. The Support Vector Machine is trained using 111 features extracted from the input signal.

In another work, Genre is used as the basis for Music Classification and Summarization by using SVMs (Xu et al 2005). They are applied to classify music into pure music (Instrument only) and vocal music (voice and Instrument) based on training data. From the input, features like MFCC and LPC are extracted and a clustering algorithm is applied to group the music into pure and vocal music. They claim Support vector machine showed better
performance in music classification than traditional Euclidean distance methods and Hidden Markov Model Methods.

In this thesis we explore South Indian Genre identification like Vayalora Pattu, Therukoothu, Oppari, Thalaattu, Ghana etc. based on the existing features or using a new set of features.

2.8 MODELS USED FOR MUSIC PROCESSING

Gaussian Mixture model and Hidden Markov Models are widely used in music processing. In addition to this, k-NN classifier and SVM have also been used for identification of the various contents available in given musical piece.

The Gaussian Mixture model (GMM) estimates the probability density function (pdf) as a weighted sum of multiple Gaussian densities. This pdf of ‘x’ is defined as a weighted sum of multivariate normal distributions. The standard procedure to train a GMM is the iterative expectation-maximization (EM) algorithm or Maximum a Posteriori (MAP) algorithm from a well-trained prior model. The resulting parameters form an inherently discriminative model of classes. In both the algorithms for constructing GMM, the classification principle is the maximum likelihood classification which is to find the class ‘i’ which maximizes the likelihood $L$ of the set of observations $X = \{x_1, x_2, \ldots, x_m\}$. This criterion assumes that the observation probabilities in successive time frames are statistically independent. GMM computes the probability values using Bayes theorem and is used for clustering. Typically GMM models have been used for Singer and Instrument identification (Vincent and Rodet 2004, Essid et al 2006).

Hidden Markov Models (HMM) are probabilistic statistical models (Rabiner and Juang 1986) based on the principles of Markov chain. The HMM was initially used for speech processing and later applied to
applications which can be based on the principle of Markov chain (Rabiner 1989). This is also being used for Music processing in almost every module from Segmentation to Model Construction. HMM has also been used for segmentation of music (Gao et al 2003) as well as for modeling the Raga in Hindustani music (Pandey 2003) and Genre identification (Tzanetakis and Cook 2002).

Support Vector machines (SVM) are binary classifiers that have found application in music signal processing for Emotion identification (Han et al 2009). SVM’s are supervised classification systems that find a hyperplane to separate two classes of data. The idea behind SVM is that it is a linear method in a high-dimensional feature space which is non-linearly related to input space. Hence all computations are done in the input space. The SVM perform classification by constructing an \( N \)-dimensional hyperplane that separates the data into two categories. Thus the aim of the SVM is to find the optimal hyperplane that separates a given cluster of vectors to the two sides of a hyperplane.

Latent Dirichlet Allocation (LDA), a topic model (Blei et al 2003) has also been used for music processing. In the work proposed by Hu (2009), the author has analysed the performance of LDA for text, images and music. The author has designed the model for text, with the assumption that every document will cater to a set of topics and every topic will have a set of words specifying it. Hence, these words that describe a topic can be assigned a higher probability. Using the probability distribution in a given document, the number of topics covered by that document can be understood. The authors have used Dirichlet distribution for the distribution of words in a given topic and this distribution is governed by the Dirichlet parameters \( \alpha \) and \( \theta \). The parameter \( \alpha \) is a \( K \)-dimensional parameter that is constant over all documents within a corpus. The parameter \( \theta \) is the topic weight vector, indicating the
contribution of each of the $K$ topics to a given document. The initial $\alpha$ is assumed to be a smaller value and this value is re-computed using the value of $\theta$ using Baye’s theorem. By using the values of $\alpha$ and $\theta$, the distribution of topics in a given document is computed thereby determining the topics in a given document. The author has also explored the determination of these values for images and music.

In the work for images, the image is segmented and each segment is viewed as a bag of words. For each of these segment vectors the Dirichlet parameters are identified to determine the topics in the given image. In the work for music, the author has used LDA for determining the harmonic structure available in a given musical piece. The $K$-dimensional parameter $\alpha$ is computed by assuming the value of $K$ corresponding to the number of notes in major key and minor key which is typically 24. Then they have matched document corpus to music corpus by comparing music in the corpus to a document consisting of a distribution of notes. They have established the note pattern catering to major scale and minor scale using LDA. In another work done by Hoffman et al (2008) a modified LDA has been used for determining musical similarity between music files. They have modeled the dirichlet parameters using the MFCC feature set which typically conveys the timbral characteristics of a given music.

In this thesis the use of these three models has been explored in different stages of music content identification.

2.9 MUSIC INFORMATION RETRIEVAL (MIR)

One important application of the components extracted from music is its use for effective Music Information Retrieval. Several algorithms have been proposed by various researchers for the process of music information retrieval. A scalable P2P system for content based music information retrieval
has been developed by Tzanetakis (2003). This algorithm is based on Rendezvous points. They have extracted features based on the Short Time Fourier Transform and Mel-Frequency Cepstral Coefficients to represent sound texture, rhythm and pitch content. The means and variances are computed for the Spectral Centroid, Roll-off, Flux and Zero-Crossings and the first 5 Mel Frequency Cepstral Coefficients (MFCC) over a 1 second texture window are calculated for representing Spectral Texture. These features are called Rendezvous points. The authors have designed the algorithm to search based on only a selected quality or qualities of the musical piece while ignoring all other parameters. This algorithm had a high time complexity for the process of search and retrieval, since focus was not given on the indexing process to aid in efficient retrieval.

It is evident that automatic content based music information retrieval greatly relies on a good indexing algorithm. The typical characteristics of an efficient indexing algorithm are efficiency in time and space while providing good precision and recall. In an algorithm proposed by Rauber, for the automatic indexing of music, the Genre of the music is used for indexing either directly or through the features conveying the Genre of the musical piece (2002). This algorithm used time-invariant features which are extracted based on psychoacoustic models for representing the Genre of the input music signal. Based on the extracted features a clustering algorithm was used to group similar Genre music together which are then used as indices for the given musical piece. This algorithm used only the Genre parameter for retrieval and hence can be modified to include other features of the song.

In another work by Shih et al, they assumed an input file in MIDI representation and have modified a table driven algorithm for indexing using the tempo characteristics of the signal (2001). In this work a bar index table is built based on the tempo characteristics and then using the Lempel – Ziv
algorithm retrieval is performed based on the bar indexes. This algorithm has given good results only when the input is available in MIDI representation.

In another work by Yang a spectral indexing algorithm based on multiple hash tables have been proposed for indexing while retrieval is performed based on Hough transform (2002). The authors have used minimal features like Short time Fourier transform and have tested for a total of 300 minutes of audio input. The drawback of this approach is that it used a minimal feature set and hence did not have a good precision and recall (Yang 2002).

Doraisamy and Ruger have proposed a music information retrieval system that uses an $n$-gram approach for indexing that uses the presence of adjacent and concurrent patterns of the input segment (2003). The input signal is a polyphonic musical piece and is converted to its MIDI representation and using the concept of $n$-grams, overlap $n$-grams are constructed which are used for indexing and later retrieval of the input musical piece. Another $n$-gram based indexing technique was proposed for Indonesian Folk Songs by Marsye and Adriani (2009). They have developed an algorithm to retrieve Indonesian folk songs which is based on pattern matching using the $n$-gram approach. The folksongs are indexed using features like MFCC, and the indexing was done partially for some songs and fully for others. The fully indexed songs gave better results irrespective of the query length and the position of the query fragment.

Lo and Wang have proposed another work for indexing and retrieval to address the problem of variable query length of the music segment during retrieval (2008). Existing indexing techniques, like suffix trees have been used to match the query phrase with a tree representation but have addressed the varying query length of the segment. In this work they have used a concept called Grid Suffix trees for the construct of indices which uses
multiple features for music indexing (Lo and Wang 2008). The grid suffix tree is conceptually visualized as a multi-dimensional suffix tree which has been claimed having a reduced space complexity when compared to other approaches. Another grid-like structure is called dual-ternary indexing in which Jiah and Chang have used a two-dimensional grid and three number notations to represent pitch content (2008).

2.10 OBSERVATIONS FROM THE SURVEY

From the literature survey that has been carried out, and based on the motivations discussed in Chapter 1, the following are the conclusions that have been arrived:

- Noise removal is not recommended for music processing, while signal separation is necessary to isolate the voice and non-voice components of the signal.

- Segmentation for Carnatic music needs to cater to swara identification, and hence, can be based on the typical characteristics of Carnatic music, which would yield such segments.

- The determination of the swara is an important objective of our work, since this is essential for Raga identification. But determining other swaras requires the tonic (frequency of the middle octave ‘S’), and hence, designing an efficient algorithm for the same is essential.

- The possibility of using the MFCC for non-music content identification, namely, the Singer, Instrument, Genre and Emotion need to be explored. However, from the literature
survey that has been carried out, it can be concluded that fundamental frequency plays an important role in identifying this type of content. Hence, using Cepstral features that incorporate the tonic as against the generic fundamental frequency and exploiting Carnatic music characteristics could probably lead to better music content identification.

- The use of existing models like the GMM, HMM, SVM and LDA needs to be explored for identifying typical music content like the swara, Tala and non-music content such as the Instrument, Singer, Genre and Emotion. However the design of a new model to identify Ragas of Carnatic music could be explored as the existing models have not been successful.

- After identifying the contents, the use of these components as key values for indexing a MIR system, needs to be explored using an existing indexing structure. From the literature, it was observed that a new index structure, that would incorporate all the identified contents as key values for indexing, could be designed to enable querying based on the characteristics of the music.