## CONTENT-BASED MUSIC STRUCTURE ANALYSIS

## NAMUNU CHINTHAKA MADDAGE (B.Eng, BIT India)

A THESIS SUBMITTED<br>FOR THE DEGREE OF DOCTOR OF PHILOSOPHY DEPARTMENT OF COMPUTER SCIENCE NATIONAL UNIVERSITY OF SINGAPORE 2005

## Acknowledgement

After sailing for four years on this journey of research, I have anchored at a very important harbour to make a documentary about the experiences and achievements about the journey. My journey of research so far has been full of rough cloudy stormy days as well as bright sunny days.

The journey of research where I am now at this stop could not have been successfully reached, without kind, constructive and courageous advice of two well experienced navigators. My utmost gratitude goes to my supervisors: Dr. Mohan S. Kankanhalli and Dr . Xu Changsheng for giving me precious guidance for more than three years.

My PhD studies would have never started in Singapore without the guidance from Dr . Jagath Rajapaksa, Ms. Menaka Rajapaksa and late Dr. Guo Yan and 4 years of full research scholarship from NUS \& I2R. I am grateful to them for opening the door to success. Wasana, thank you for encouraging me to be successful in the research. I acknowledge Dr. Zhu Yongwei, Prof. Lee Chin Hui, Dr. Ye Wang, Shao Xi and all my friends for their valuable discussions and thoughts during the journey of research.

This thesis is dedicated to my beloved parents and sister. Without their love and courage, I could have sufficiently not strengthened my will power for this journey.
 (Amma, Thaththa and Akka)!

## Table of Contents

Acknowledgement ..... i
Table of Contents ..... ii
Summary ..... v
List of Tables ..... vii
List of Figures ..... viii
1 Introduction ..... 1
2 Music Structure ..... 9
2.1 Time information and music notes ..... 11
2.2 Music scale, chords and key of a piece ..... 15
2.3 Composition of music phrases ..... 19
2.4 Popular song structure ..... 19
2.5 Analysis of Song structures ..... 23
2.5.1 Song characteristics ..... 23
2.5.2 Song structures ..... 25
3 Literature Survey ..... 29
3.1 Time information extraction (Beats, Meter, Tempo) ..... 31
3.2 Melody and Harmony analysis ..... 37
3.3 Music region detection ..... 44
3.4 Music similarity detection ..... 50
3.5 Discussion ..... 51
4 Music Segmentation and Harmony Line Creation via Chord Detection ..... 53
4.1 Music segmentation ..... 57
4.2 Windowing effect on music signals. ..... 61
4.3 Silence detection ..... 64
4.4 Harmony Line Creation via Chord Detection ..... 65
4.4.1 Polyphonic music pitch representation ..... 68
4.4.1.1 Pitch class approach to polyphonic music pitch representation ..... 68
4.4.1.2 Psycho-acoustical approach to polyphonic music pitch representation ..... 71
4.4.2 Statistical learning for chord modelling ..... 73
4.4.2.1 Support Vector Machine (SVM) ..... 74
4.4.2.2 Gaussian Mixture Model (GMM) ..... 75
4.4.2.3 Hidden Markov Model (HMM) ..... 76
4.4.3 Detected chords' error correction via Key determination ..... 76
5 Music Region and Music Similarity detection ..... 79
5.1 Music region detection. ..... 79
5.1.1 Applying music knowledge for feature extraction ..... 83
5.1.1.1 Cepstral Coefficients ..... 83
5.1.1.2 Linear Prediction Coefficients (LPCs) ..... 92
5.1.1.3 Linear Predictive Cepstral Coefficients (LPCC) ..... 97
5.1.1.4 Harmonic Spacing measurement using Twice-Iterated Composite Fourier Transform Coefficients (TICFTC) ..... 99
5.1.2 Statistical learning for vocal / instrumental region detection ..... 105
5.2 Music similarity analysis ..... 105
5.2.1 Melody-based similarity region detection ..... 107
5.2.2 Content-based similarity region detection ..... 108
5.3 Song structure formulation with heuristic rules ..... 112
5.3.1 Intro detection ..... 113
5.3.2 Verses and Chorus detection ..... 113
5.3.3 Instrumental sections (INST) detection ..... 116
5.3.4 Middle eighth and Bridge detection ..... 116
5.3.5 Outro detection ..... 116
6 Experimental Results ..... 117
6.1 Smallest note length calculation and silent segment detection ..... 117
6.2 Chord detection for creating harmony contour ..... 118
6.2.1 Feature and statistical model parameter optimization in synthetic environment ..... 119
6.2.2 Performance of the features and the statistical models in the real music environment ..... 122
6.3 Vocal/instrumental region detection ..... 124
6.3.1 Manual labelling of experimental data for the ground truth ..... 125
6.3.2 Feature and classifier parameter optimization ..... 126
6.3.3 Language sensitivity of the features ..... 128
6.3.4 Gender sensitivity of the features ..... 129
6.3.5 Overall performance of the features and the classifiers ..... 130
6.4 Detection of semantic clusters in the song ..... 133
6.5 Summary of the experimental results ..... 139
7 Applications ..... 141
7.1 Lyrics identification and music transcription ..... 141
7.2 Music Genre classification ..... 143
7.3 Music summarization ..... 144
7.3.1 Legal summary making. ..... 145
7.3.2 Technical summary making ..... 146
7.4 Singer identification system ..... 148
7.4.1 Singer characteristics modelling at the music archive ..... 150
7.4.2 Test song identification ..... 151
7.5 Music information retrieval (MIR) ..... 153
7.6 Music streaming ..... 156
7.6.1 Packet loss recovery techniques for audio streaming ..... 157
7.6.2 Role of the music structure analysis for music streaming ..... 161
7.6.3 Music compression ..... 163
7.7 Watermarking scheme for music ..... 164
7.8 Computer aid tools for music composers and analyzers ..... 166
7.9 Music for video applications ..... 167
8 Conclusions ..... 168
8.1 Summary of contributions ..... 168
8.2 Future direction ..... 171
References ..... 172
Appendix - A ..... 191

## Summary

This thesis proposes a framework for popular music structure detection, which incorporates music knowledge with audio signal processing techniques.

The important components of the music structure are modelled hierarchically in the layers of the music structure pyramid. The bottom layer of the pyramid is the time information (Tempo, Meter, Beats) of the music. The second layer is the harmony/melody, which is created by playing music notes. Information about the Music regions i.e. Pure instrumental region, Pure vocal region, Instrumental mixed vocal region and Silence region are discussed in the third layer. The fourth layer and the higher layers in the music structure pyramid discusses semantic meaning(s) of the music which are formulated based on the music information in the first, second and third layers. The popular song structure detection framework discussed in this thesis covers methodologies for the layer-wise music information in the music pyramid.

The process of any content analysis consists of three major steps. They are signal segmentation, feature extraction, and signal modelling. For music structure analysis, we propose a rhythm based music segmentation technique to segment the music. This is called Beat Space Segmentation. In contrast, the conventional fixed length signal segmentation is used in speech processing. The music information within the beat space segment is considered more stationary in its statistical characteristics than in the fixed length segments. The process of beat space segmentation covers the extraction of bottom layer information in the music structure pyramid.

Secondly, to design the features to characterize the music signal, we consider the octave varying temporal characteristics in the music. For harmony/melody information extraction (information in the $2^{\text {nd }}$ layer), we use the psycho acoustic profile feature and obtain a better performance compared to the existing pitch class profile feature. To capture the octave varying temporal characteristics in the music regions, we design a new filter bank in the octave scale. This octave scale filter bank is used for calculating cepstral coefficients to characterise the signal content in music regions (information in the $3^{\text {rd }}$ layer). This proposed feature is called Octave Scale Cepstral Coefficients and its performance for music region detection is compared with existing speech processing features such as linear prediction coefficients (LPC), LPC derived cepstral coefficients, Mel frequency cepstral coefficients. This feature is found to perform better than speech processing features.

Thirdly, existing statistical learning techniques (i.e. HMM, SVM, GMM) in the literature are optimized and used for modelling the music knowledge influenced features to represent the music signals. These statistical learning techniques are used for modelling the information in the second and third layers (Harmony/melody line and the music regions) of the music structure pyramid.

Based on the extracted information in the first three layers (time information, harmony/melody, music regions), we detect similarity regions in the music clip. We then develop a rule based song structure detection technique based on detected similarity regions. Finally, we discuss music related applications, based on proposed framework of popular music structure detection.

## List of Tables

Table 2-1 : Music note frequencies (F0) and their placement in the Octave scale sub-bands. ..... 13
Table 2-2: Distance to the notes in the chord from the key note in the scale ..... 16
Table 2-3: Names of the English and Chinese singers and their album used for the survey ..... 23
Table 5-1: Filter distribution for computing Octave Scale Cesptral Coefficients ..... 91
Table 5-2: Parameters of the Elliptic filter bank used for sub-band signal decomposition in octave scale ..... 96
Table 6-1: Technical details of our method and the other method ..... 123
Table 6-2: Details of the Artists ..... 125
Table 6-3: Optimized parameters for features ..... 127
Table 6-4: Evaluation of identified and detected parts in a song. ..... 135
Table 6-5: Technical detail comparison of other method with ours. ..... 136
Table 6-6: Accuracies of semantic cluster detection and identification of the song "Cloud No 9 by Bryan Adams" based on beat space and fixed length segmentations ..... 138

## List of Figures

Figure 1-1: Conceptual model for song music structure. ..... 2
Figure 1-2: Thesis Overview ..... 6
Figure 2-1: Information grouping in the music structure model. ..... 10
Figure 2-2: Correlation between different lengths of music note ..... 11
Figure 2-3: Ballad \#2 key-F major. ..... 12
Figure 2-4: The variation of the F0s of the notes in C8B8 octave when standard value of A4 $=440 \mathrm{~Hz}$ is varied in $\pm$ percentage ..... 14
Figure 2-5: Succession of music notes and music Scale. ..... 16
Figure 2-6: Chords that can be derived from the notes in the four music scales types ..... 17
Figure 2-7: Overview of top down relationship of notes, chords and key. ..... 18
Figure 2-8: Rhythmic groups of words ..... 19
Figure 2-9: Semantic similarity clusters which define the structure of the popular song ..... 20
Figure 2-10: Two examples for verse- chorus pattern repetitions. ..... 22
Figure 2-11: Percentage of the average vocal content in the songs ..... 24
Figure 2-12: Tempo variation of songs ..... 25
Figure 2-13: Percentage of the smallest note in songs. ..... 25
Figure 3-1: MIDI music generating platform in the Cakewalk software (top) and MIDI file information representation in text format (bottom). ..... 30
Figure 3-2: Instrumental tracks (Drum, Bass guitar, Piano) and edited final tract (mix of all the tracks) of a ballad (meter- $4 / 4$ and tempo - 125 BPM) "I Let You Go" sung by Ivan. First 6 seconds of the music is considered. ..... 32
Figure 3-3: Basic steps followed for extracting time information. ..... 33
Figure 4-1: Spectral and time domain visualization of $(0 \sim 3667) \mathrm{ms}$ long clip played in " 25 Minutes" by MLTR. Quarter note length is 736.28 ms and note boundaries are highlighted using dotted lines. ..... 54
Figure 4-2: Notes played in the $6^{\text {th }}, 7^{\text {th }}$, and $8^{\text {th }}$ bars of the rhythm guitar, bass guitar, and electric organ tracks of the song "Whose Bed Have Your Boots Been Under" by Shania Twain. Notes in the electric organ track are aligned with the vocal phrases. Blue solid lines mark the boundaries of the bars and red solid lines mark quarter note boundaries. Grey dotted lines within the quarter notes mark eighth and sixteenth note boundaries. Some quarter note regions which have smaller notes are shaded with pink colour ellipses.55
Figure 4-3: Rhythm tracking and extraction ..... 58
Figure 4-4: Beat space segmentation of a 10 second clip ..... 61
Figure 4-5: The frequency responses of Hamming and rectangular windows. ..... 63
Figure 4-6: Silence region in a song ..... 64
Figure 4-7: Concept of sailing music regions on harmony and melody flow ..... 65
Figure 4-8: Section of both bass line and treble line created by a bass guitar and a piano for the song named "Time Time Time". The chord sequence, which is generated using notes played on both the bass and treble clefs, is shown at the bottom of the figure. ..... 66
Figure 4-9: Chord detection steps ..... 67
Figure 4-10: Music notes in different octaves are mapped into 12 pitches ..... 69
Figure 4-11: Harmonic and sub-harmonics of C Major Chord is visualized in terms of closest music note ..... 71
Figure 4-12: Spectral visualization Female vocal, Mouth organ and Piano music ..... 72
Figure 4-13: Chord detection for the $i^{t h}$ beat space signal segment. ..... 74
Figure 4-14: The HMM Topology ..... 76
Figure 4-15: Correction of chord transition ..... 78
Figure 5-1: Regions in the music ..... 80
Figure 5-2: The steps for vocal instrumental region detection ..... 83
Figure 5-3: Steps for calculating cepstral coefficients ..... 84
Figure 5-4: The filter distribution in both Mel scale and linear scale ..... 87

Figure 5-5: Music and speech signal characteristics in frequency domain. (a) - Quarter note length ( 662 ms ) instrumental (Guitar) mixed vocal (male) music, (b) - Quarter note length ( 662 ms ) instrumental
(Mouth organ) music, (c) - Fixed length (600ms) speech signal, (d) - Ideal octave scale spectral envelopes ..... 88
Figure 5-6: Log magnitude spectrums of bass drum and side drum ..... 89
Figure 5-7: The filter band distribution in Octave scale for calculating cepstral coefficients ..... 90
Figure 5-8: Plot of the 20 Singular values, which are computed from OSCCs and MFCCs for vocal and instrumental music frame. ..... 91
Figure 5-9: Average of singular values ..... 92
Figure 5-10: Computation of selective band linear predictive coefficients (LPCs) ..... 95
Figure 5-11: Selective-band power spectrum approximation using all pole speech model $H(z)$ ..... 97
Figure 5-12: Harmonic structures of vocal and instrumental signal segments ..... 100
Figure 5-13: Twice - iterated composite Fourier transform of $i^{t h}$ signal frame ..... 101
Figure 5-14: The $1^{\text {st }} \& 2^{\text {nd }}$ FFT of instrumental and vocal frames. Frame size is a quarter note length $(735 \mathrm{~ms})$ ..... 102
Figure 5-15: The mean-removed bin $B_{l}$ (.) with beat space ( 662 ms ) frames of "Sleeping child" by MLTR ..... 104
Figure 5-16: Twice -iterated composite Fourier transform coefficients ..... 104
Figure 5-17: Classification. ..... 105
Figure 5-18: Similarity regions in the music ..... 106
Figure 5-19: Melody based similarity region detection by matching chord patterns ..... 107
Figure 5-20: 8 and 16 bar length chord pattern matching results ..... 108
Figure 5-21: Vocal similarity matching in the $i^{\text {th }}$ and $j^{t h}$ MBSRs ..... 108
Figure 5-22: The response of the $9^{\text {th }}$ OSCC, MFCC and LPC to the Syllables of the three words 'clue number one'. The number of filters used in OSCC and MFCC are 64 each. The total number of coefficients calculated from each feature is 20 ..... 109
Figure 5-23: Vocal sensitivity analysis of OSCCs and MFCCs using SVD. ..... 110
Figure 5-24: The normalized content-based similarity measure between regions $\mathrm{R}_{1}$ through $\mathrm{R}_{8}$ computed from melody-based similarity regions of the song as shown in Figure 5-20 (Red dash line) ..... 112
Figure 6-1: Actual and computed $16^{\text {th }}$ note lengths of songs ..... 118
Figure 6-2: Note mixing procedure for creating a synthetic chord. ..... 120
Figure 6-3: Average chord classification accuracy of the statistical models ..... 122
Figure 6-4: Manually annotated the intro and the verse 1 of the song "Cloud No 9 by Bryan Adams". ..... 123
Figure 6-5: This manual annotation describes the time information of the vocal and instrumental boundaries in the first few phrases of the song "On a day like today" by Bryan Adams. The frame length is equal to the $16^{\text {th }}$ note length beat space segment (182.49052 $\mathrm{ms})$.It is the smallest note length that can be found in the song. ..... 125
Figure 6-6: Average classification accuracies of the features in the language sensitivity test ..... 129
Figure 6-7: Average classification accuracies of the features in their gender sensitivity test ..... 130
Figure 6-8: Overall classification accuracy of features with HMM ..... 131
Figure 6-9: Classifier performances in vocal / instrumental classification. ..... 132
Figure 6-10: Effect of classification accuracy with frame size. ..... 133
Figure 6-11: The average detection accuracies of different sections ..... 135
Figure 6-12: A failure case of our semantic clusters detection algorithm. Figure (a) shows the manually annotated positions of the components in the song structures. Figure (b) shows the detected components and their positions. Figure (c) shows the identification and detection accuracy of the components in the semantic clusters. ..... 137
Figure 7-1: Primary information required for lyrics identification and music transcription ..... 142
Figure 7-2: Illustration of music summary generation using music structure analysis ..... 146
Figure 7-3: Technical summary making steps ..... 147
Figure 7-4: Vocal and the relative instrumental section modelling of songs of same singer ..... 150
Figure 7-5: Singer identification of the test song ..... 152
Figure 7-6: Singer information retrieval comparison when original album and converted wave files are played on Windows Media Player (version 10).Test album is Sings the Standards by Cliff Richard. ..... 154
Figure 7-7: Architecture of music information retrieval system ..... 155
Figure 7-8: Music streaming software "Yahoo Music Launchcast Radio" given in Yahoo messenger for listening to the songs played at different music stations. ..... 157
Figure 7-9: Forward error correction (FEC) mechanism for packet repair ..... 160
Figure 7-10: Interleaving mechanism for packet repair ..... 160
Figure 7-11: Sender-receiver based music information embedded packet loss recovery scheme ..... 162
Figure 7-12: MP3 codec architecture ..... 163
Figure 7-13: Design platform for content specific watermarking scheme ..... 165

## 1 <br> Introduction

Recent advances in computing, networking and multimedia technologies have resulted in a tremendous growth of music-related data and have accelerated the need for both analysis and understanding of the music content. Because of these trends, music content analysis has become an active research topic in recent years.

Music understanding is the study of the methods by which computer music systems can recognize patterns and structures in the musical information. One of the research difficulties in this area is the general lack of formal understanding of music. For example, experts disagree over how music structure should be represented, and even within a given system of representation, the music structure is often ambiguous. Considerable amounts of research have been devoted to music analysis, yet we do not appear to be appreciably closer to understanding the properties of musical signals which are capable of evoking cognitive and emotional responses in the listener. It is the inherent complexity in the analysis of music signals which draws so much attention from such diverse fields as engineering, physics, artificial intelligence, psychology, and musicology.

One of the main attractions of digital audio is the ability to transfer and reproduce it in the digital domain without degradation. Many hardware and software tools exist to replace the array of traditional recording studio hardware, performing duties such as adding effects, reducing noise, and compensating for other undesired signal characteristics, all without introducing loss in the signal paths between the processing
components. The digital environment has opened up opportunities for researchers of different expertise to collaborate with each other to analyze and characterize the music signals in high dimensional space.

We believe that music relationships (beats arrangement with tempo, music notes, chord progression, vocal alignment with the instrumental music etc) form the basis of music. The degree of understanding of these relationships is reflected by the depth levels of the music structure. This basic music structure is shown in Figure 1-1.


Figure 1-1: Conceptual model for song music structure
The foundation of music structure is the timing information (rhythm structure), which is the bottom layer of the music structure pyramid. Music signals are characteristically
very structured: at the lowest level, sinusoids are grouped together to form music notes of particular pitches. Notes are grouped to form chords or harmonies (the $2^{\text {nd }}$ layer in the pyramid). Even higher levels of structure (the $3^{\text {rd }}$ layer) may establish themes through repetition and simple transformations of smaller elements. This successive abstraction to higher levels can be called music context integration.

It is difficult to understand how the human brain decodes embedded information from perceived music. At the very basic level, listeners are capable of identifying melody fluctuations and contours in the music in terms of note level discrete steps. For example, even listeners who have had very little music training still snap their fingers or clap their hands to the temporal structure they perceive in music with little effort. Usually, music phrases describe messages which are delivered by the performer. How these messages are embedded within the music structure and the level at which the brain decodes such information would generate auditory sensations in the listener's mind. At a high-level, these sensations may be the reflections of sensations generated in the composer/performer's mind or may be very different. However, we have not attained the level of modelling of those aspects of the mind yet.

The analysis of basic components of music structure is important for many applications such as lyrics identification, music transcription, genre classification, music summarization, singer identification, music information retrieval (MIR), music streaming, music watermarking and computer aided music tools for composers and analyzers. The importance of music structural analysis for these applications is detailed in chapter 7.

In this thesis, we propose methodologies for extracting and analyzing different layers of music structure information. Figure 1-2 explains the overview of this thesis. In contrast with conventional fixed length audio segmentation (Rabiner and Juang 1993 [94]), an alternate segmentation technique, in which the length of the signal segment is proportional to the rhythm of the music (i.e. inter beat intervals) is proposed for music segmentation. Thereafter, dynamic behaviour of music signal properties such as octave-based spectral behaviours is studied for designing features and their performance is compared with that of existing speech signal characterizing features.

Music is a way of expressing both the depth and height of human thoughts in a creative manner. Based on its content, we can categorize music into different genres such as popular (POP), rock, classic and jazz. Creation of music is highly influenced by different cultures, communities, and societies, which has its own way of making and breaking rules. Thus, it is difficult to judge what music belongs to which genre. Figure $1-1$ is a simple way of visualizing the underlying layers of music content, which helps to decode important information for designing music applications. In this thesis we have narrowed down the scope of music structural analysis to popular music with 4/4 time signature, which is the most commonly used meter in popular (mostly in POP music) music (Goto 2001 [48]) in this thesis.

Music theory reveals that the temporal properties in music change in the steps of music notes (chapter 2). In our proposed approach, we first extract rhythm information such as the length of inter-beat intervals. Since the song's meter is assumed to be $4 / 4$, the length of the inter-beat interval is equal to the duration of the quarter note, which reveals the tempo of the song. Further analysis of the note
structure using onset detection indicates the appearance of smaller notes such as eighth, sixteenth, and thirty-second notes in the song (see chapter 4). The music signal is then segmented according to the length of the smallest note (eighth, sixteenth or thirty-second) that can be seen in the music, unlike the conventional fixed length segmentation in speech processing. This new acoustic segmentation method is called beat space segmentation (BSS) in this thesis. Spectral domain analysis shows that signal section is harmonically quasi-stationary within the beat space segment (BSS). After a song is segmented, musically inspired features are extracted to characterize the music content. To detect both pitch fluctuations and melody / harmony contours in the song, pitch class profile features (PCP) and psycho-acoustic profile features (PAP) are extracted from the beat space segmented frames. Chapter 4 discusses melody/ harmony detection and chord progression in detail.

A music signal's complexity varies with the source mixtures, which clearly defines four regions in the music signals. They are pure vocal regions (vocal only)-PV, instrumental mixed vocal regions-IMV, pure instrumental regions-PI, and silence regions -S. In our survey, we noticed that the appearance of pure vocal regions in popular music is very rare. Thus, PV and IMV regions are merged into a general class called vocal regions. Chapter 5.1 discusses the identification procedures of these regions. For the characterization of vocal/instrumental regions, feature extraction technique in octave scale is proposed and compared against existing Mel-scale cepstral features. In addition, an octave scale linear predictive coefficients (OSLPCs), octave scale linear predictive cepstral coefficients (OSLPCCs) and Twice-Iterated Composite Fourier Transform Coefficients (TICFTC) have been explored for the vocal / instrumental region detection problem.


Figure 1-2: Thesis Overview
The performance of statistical models i.e. Hidden Markov Model (HMM), Gaussian
Mixture Model (GMM), and Support Vector Machine (SVM), has been compared for both chords detection and vocal/instrumental region detection in music. Music structure formulation is discussed in chapter 5.3. Based on the existence of similar chord transition patterns, melody based similarity regions are identified. Using a more
detailed similarity analysis of the vocal content in these melody based similarity regions, content-based similarity regions can be identified. Using heuristic rules which are commonly employed by music composers, music structure has been defined.

## Contributions of the thesis

The scope of this thesis has been limited to the analysis of popular music structure where the meter of the songs is $4 / 4$. The important information in the music structure is conceptually visualized in the layers of the proposed music structure pyramid (Figure 1-1).

Incorporation of music knowledge into audio signal processing for music content analysis is the main contribution of this thesis. We propose a novel rhythm based music segmentation technique for music signal analysis, whose performance has been shown to be superior to that of the conventional fixed length segmentation that has been used in speech processing.

Two features, pitch class profile (PCP) feature and psycho acoustic profile (PAP) feature, are studied for polyphonic music pitch representation. It is found that the PAP feature can more effectively characterize polyphonic pitches than the commonly used PCP feature. Thus, we use the PAP feature for our harmony line creation via music chord detection.

We studied the octave varying temporal characteristics of the music signals and applied these characteristics to various speech processing features such as linear
prediction coefficients (LPC), LPC derived cepstral coefficients, and Mel frequency cepstral coefficients. Then, we proposed the Octave Scale Cepstral Coefficient (OSCC) feature and the Twice-Iterated Composite Fourier Transform Coefficient (TICFTC) feature for music region (vocal/instrumental) detection in music. The comparison between all features showed that OSCC can detect vocal/instrumental regions more accurately than other features.

We studied the existing statistical learning techniques, i.e. SVM, GMM and HMM, and optimized the models' parameters for both the chord detection task and the music region detection task. It is found that HMM can model temporal properties of the music signals better than GMM or SVM. We conducted a survey to analyse the characteristics of popular song structures. Based on the analysis results, we designed a rule-based algorithm to detect the song structures of the popular music genre.

## Overview of the thesis

The overview of this thesis is depicted in Figure 1-2. We incorporate music knowledge with signal processing techniques in order to extract music information. Chapter 2 discusses the music knowledge. Existing music processing techniques are surveyed in chapter 3. Chapter 4 details our proposed methods for rhythm based signal segmentation and harmony line detection. Detection of music regions, music similarity regions, and semantic clusters are explained in chapter 5. From the experimental results, we analyse the strength and weakness of the proposed music information extraction techniques in chapter 6. Chapter 7 discusses the possible music applications, which can benefit using our proposed music structure analysis techniques. Finally, we conclude the thesis in chapter 8.

## 2

## Music Structure

Music is universal language for sharing information among the same or different communities. The amount of information embedded in music can be huge and designing computer algorithms for decoding semantic level information is an extremely complex task. The human mind is superior in such refined decoding tasks.

In this thesis, we extract basic ingredients which have been used in the music composition and which are useful for developing important applications. Figure 2-1 explains the conceptual model of music structure. The foundation of music structure is the timing information (i.e. Time signature and Tempo), which is the bottom layer of the music structure pyramid. The harmony /melody (the second layer) is created by playing music notes together at different scales according to the beats. The vocal line is then embossed on the surface of the melody, which creates two important regions in the music, the instrumental region and the vocal region. The layout of these regions in the harmony / melody contours is conceptually visualized in Figure 4-7. The top layer of the music pyramid depicts the semantics of the song structure, which describes the events or messages to the audience [28]. Understanding the information in the top most layer is the most difficult and is too complex for current technologies. The information in popular songs can be semantically clustered as Intro, Verse, Chorus, Bridge, Middle eighth and Outro. When we think of the semantic meaning of music, these clusters can be considered the least complex level of semantics in the song. However, it is challenging to detect even these clusters.


Figure 2-1: Information grouping in the music structure model

The scope of this thesis encompasses the extraction of the layer-wise information of the music structure pyramid, which is useful for developing music related applications (detailed in chapter 7). We have simplified the task of mining semantic meanings for the top of identifying semantic clusters, i.e. Intro, Verse, Chorus, Bridge and Outro, of the song. The following sections of this chapter discuss music terms, different units, and entities that are used for composing music information at the different layers of the music structure pyramid.

### 2.1 Time information and music notes

The duration of a song is measured in number of bars [100]. The term bar is explained with the other music terms below. While listening to music, the steady throb to which one could clap is called the Pulse, or the Beat, and the Accents are the beats which are stronger than others. The number of beats from one accent to an adjacent one is equal and divides the music into equal segments. Thus, these segments of beats from one accent to another are called the bar (see Figure 2-8).

The music note length can be changed by varying attack, sustain and decay characteristics of the note. Figure 2-2 discuses the correlation between different lengths of music note. In the $1^{\text {st }}$ column, Semibreve, Minim, Crotchet, Quaver, Semiquaver and Demisemiquaver are the names of the notes played in western music, and are respectively classified as Whole, Half, Quarter, Eighth, Sixteenth and Thirtysecond notes according to their durations (onset to offset), which are the fractions of the Semibreve. In the third column, the durations of silence (Rests) are also equal to the note length.

| Note | Shape | Rest | Value in terms of a Semibreve | Corresponding names commonly use in U.S.A and Canada |
| :---: | :---: | :---: | :---: | :---: |
| Semibreve | $\bigcirc$ | 5 | 1 | Whole Note |
| Minim | d | $\Omega$ | 1/2 | Half Note |
| Crotchet | d | \& or $p^{-}$ | 1/4 | Quarter Note |
| Quaver | $\downarrow$ | 9 | 1/8 | Eighth Note |
| Semiquaver | $\delta$ | \% | 1/16 | Sixteenth Note |
| Demisemiquaver | d | \% | 1/32 | Thirty-second Note |

Figure 2-2: Correlation between different lengths of music note

Time signature (TS) (alternatively called Meter) indicates the number of beats per bar in a music piece. TS is $4 / 4$ indicates four crotchet beats in each bar. Similarly, $3 / 8$ means three quaver beats in a bar, and $2 / 2$ means two minim beats in a bar. The
frequency of the beats is known as the Tempo and is measured at BPM (Beats per Minutes). At TS equals to $3 / 8$, the tempo is the number of quaver beats per minutes.

As an example, Figure 2-3 shows the first three bars of the music sheet. Vertically aligned notes in the Staff (treble clef or bass clef) means that they are played simultaneously. The staff consists of a series of five parallel lines. The red coloured horizontal dashed line marks the position of the C 4 (middle ' C ') note, which appears on neither the bass clef nor the treble clef. The boundaries of the bars are marked in red colour vertical lines. The TS is four crotchet beats per bar (4/4). In the treble clef, the first and third bars are constructed by 4-quarter notes and 2-half notes respectively. However, the second bar is constructed by 3-quarter notes and 2-eighth notes. All three bars of bass clef contain whole notes. In the first bar of the Treble clef, the C, F, and A Crotchet notes are played simultaneously in the first quarter note, which formulates the F major chord.


Figure 2-3: Three bars of a staff

Melody is constructed by playing solo notes according to TS and Tempo. Melody is monophonic in nature. In contrast, harmony, which creates the polyphonic music nature, is generated by playing more that a note at a time, i.e. Chords. Note that $\mathrm{A} 4=440 \mathrm{~Hz}$ is commonly used as the reference pitch in concerts and is the American
standard pitch (Zhu et al 2005 [144]). Based on this reference pitch, the fundamental frequencies of the 12 pitch class notes with their octave alignments are noted in Table 2-1. The frequency ranges shown in row number 3 are calculated using $\log _{2}$ scale and all the fundamental frequencies (F0s) of the 12 pitch class notes in the octaves fall within these frequency ranges. Thus, these frequency ranges can be considered the limits of Octave envelopes (see Figure 5-5). The F0s of the notes in the C0B0 and C1B1 octaves are spaced narrowly than those of the other higher octaves. In order to differentiate these notes, we need a very high frequency resolution $(\leq 1 \mathrm{~Hz})$. Also very few percussion instruments play in those lower octaves. Thus, C0B0, C1B1, and C2B2 are merged together and considered a single band i.e. sub-band 01.

Table 2-1: Music note frequencies ( F 0 ) and their placement in the Octave scale subbands.

| Sub-band No |  |  | 01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Octave scale |  | $\sim \mathrm{B} 1$ | C2 ~ B2 | C3 ~ B3 | C4 ~ B4 | C5 ~ B5 | C6 ~ B6 | C7 ~ B7 | C8 ~ B8 | $\begin{array}{\|cc\|} \hline l \\ 2 \\ \frac{2}{\infty} & 0 \\ \hline \end{array}$ |
| Freq-range (Hz) |  |  | 64~128 | 128~256 | 256~512 | 512~1024 | 1024~2048 | 2048~4096 | 4096~8192 |  |
|  | C | $\begin{aligned} & \text { J } \\ & \text { l } \end{aligned}$ | 65.406 | 130.813 | 261.626 | 523.251 | 1046.502 | 2093.004 | 4186.008 |  |
|  | C\# |  | 69.296 | 138.591 | 277.183 | 554.365 | 1108.730 | 2217.460 | 4434.920 |  |
|  | D |  | 73.416 | 146.832 | 293.665 | 587.330 | 1174.659 | 2349.318 | 4698.636 |  |
|  | D\# |  | 77.782 | 155.563 | 311.127 | 622.254 | 1244.508 | 2489.016 | 4978.032 |  |
|  | E |  | 82.407 | 164.814 | 329.628 | 659.255 | 1318.510 | 2637.02 | 5274.04 |  |
|  | F |  | 87.307 | 174.614 | 349.228 | 698.456 | 1396.913 | 2793.826 | 5587.652 |  |
|  | F\# |  | 92.499 | 184.997 | 369.994 | 739.989 | 1479.978 | 2959.956 | 5919.912 |  |
|  | G |  | 97.999 | 195.998 | 391.995 | 783.991 | 1567.982 | 3135.964 | 6271.928 |  |
|  | G\# |  | 103.826 | 207.652 | 415.305 | 830.609 | 1661.219 | 3322.438 | 6644.876 |  |
|  | A |  | 110.000 | 220.000 | 440.00 | 880.000 | 1760.000 | 3520.000 | 7040.000 |  |
|  | A\# |  | 116.541 | 233.082 | 466.164 | 932.328 | 1864.655 | 3729.310 | 7458.62 |  |
|  | B |  | 123.471 | 246.942 | 493.883 | 987.767 | 1975.533 | 3951.066 | 7902.132 |  |

ISO 16 standard specifies A4 $=440 \mathrm{~Hz}$ and it is called as concert pitch

Though the common practice pitch standard value of A4 is 440 Hz , the old instrument pitch standard was $\mathrm{A} 4=435 \mathrm{~Hz}$. In general, music instruments may not be exactly tuned to the standard reference pitch due to the physical conditions of the instruments. Thus, there is a tendency for the music pitches to fluctuate due to the physical conditions of the instruments. The idea we elaborate in this thesis is the octave
behaviours of the music signals. We consider octave behaviours for music signal analysis and modelling. Therefore, it is important to measure the music pitch fluctuation within an octave. The upper and lower limits of an octave are noted in Table 2-1 row 3. These frequency ranges are called Octave envelopes, where 12 pitch class notes fluctuate with the octave envelope. It is found that $+3.6 \%$ and $-2.2 \%$ are the upper and lower limits of the $\mathrm{A} 4=440 \mathrm{~Hz}$ variations $(430 \mathrm{~Hz} \sim 456 \mathrm{~Hz})$ which allow the F0 of the 12 music notes to vary within their respective octave envelopes. Figure 2-4 shows the 12 notes' pitch variations within the octave envelope in sub-band 07 with respect to the pitch variation of A4.


Figure 2-4: The variation of the F0s of the notes in C8B8 octave when standard value of A4 $=440 \mathrm{~Hz}$ is varied in $\pm$ percentage

Though the physical octave ratio is $2: 1$, cognitive experiments have revealed that this ratio increases by $3 \%$ at about 2 kHz (see Chapter 3.2). Such an octave enhancement effect would not exceed the limits of its respective octave envelope. We have considered these frequency limits of octave envelopes explicitly in our algorithms for music signal analysis.

### 2.2 Music scale, chords and key of a piece

A set of notes, which forms a particular context and note pitches arranged in ascending or descending order, is called a music scale. The eight basic notes (C, D, E, F, G, A, B, C), the white notes on the keyboard, can be arranged in an alphabetical succession of sounds ascending or descending from the starting note. This note arrangement is known as the Diatonic Scale [100] and is the most common scale used in traditional western music (Krumhansl 1979 [66]). Psychological studies have suggested that the human cognitive mechanism can effectively differentiate the tones of the diatonic scale (Krumhansl 1979 [66]). Chromatic scale, which is the cyclic nature in octave periodicities, shares the same symbol/value for two tones separated by an integral number of octaves (see Figure 2-5 left top).

In a music scale, the pitch progression for one note to the other is either the half step (a Semitone-S) or the whole step (a Tone -T). Thus, this expands the eight basic notes into 12 pitch classes. The first note in the scale is known as Tonic and is the keynote (tone-note) from which the scale takes the name. Music scales are divided into four scale types, one Major scale and three minor scales (Natural, Harmonic and Melodic), according to the pitch progression patterns. These four scale types are commonly practiced in western music [100]. The Major scale, Natural Minor scale, Harmonic Minor scale and Melodic Minor scale follow the pattern of "T-T-S-T-T-T-S", "T-S-T-T-S-T-T", "T-S-T-T-S-(T+S)-S", and "T-S-T-T-T-T-S" respectively. Figure 2-5(leftbottom) shows the note progression in the G scale. The Table in the Figure (right) lists the notes that are present in the Major and Minor scales for the G pitch class. Music chords are constructed by selecting notes from the corresponding scales. Types of commonly used chords are Major, Minor, Diminished, and Augmented.


| G Scale | Notes in the C - Scale |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | I | II | III | IV | V | VI | VII | I |
| Major | G | A | B | C | D | E | F\# | G |
| Natural Minor | G | A | A\# | C | D | D\# | F | G |
| Harmonic Minor | G | A | A\# | C | D | D\# | F\# | G |
| Melodic Minor | G | A | A\# | C | D | E | F\# | G |

G Scale

Figure 2-5: Succession of music notes and music Scale

The first note of the chord is the key-note in the scale and Table 2-2 shows the note distances to the second and third notes of the chord from the key note. Since three notes in the scale are used to generate the chord, these chords are called Triads.

Table 2-2: Distance to the notes in the chord from the key note in the scale

| Notes | Distance in whole step (T) to the notes from Key note |  |  |
| :---: | :--- | :---: | :--- |
|  | $1^{\text {st }}$ note | $2^{\text {nd }}$ note | $3^{\text {rd }}$ note |
| Major (maj) | 0.0 T | 2.0 T | 3.5 T |
| Minor (min) | 0.0 T | 1.5 T | 3.5 T |
| Diminished (dim) | 0.0 T | 1.5 T | 3.0 T |
| Augmented (aug) | 0.0 T | 2.0 T | 4.0 T |

T - Implies a Tone / whole step in music theory

When we know the notes that are in the different scales, the note distance relationship in the Table 2-2 can be used to find all the possible chords that can be derived from the scale. Figure 2-6 illustrates all possible chords in the different music scales. The scale's name is derived from its key note (first note) and 12 scales appear in one type of music scale. All four chord types (Major, Minor, Diminished and Augmented) appear in both the Melodic Minor and the Harmonic Minor scale types. In contrast, the Augmented chord type doesn't appear in both the Major and the Natural Minor
scale types. We can see from Figure 2-6 that the chords in a particular major scale appear in a different natural minor scale. For example, chords in the C major scale appear in the A natural minor scale. It implies that notes in both the C major scale and the A natural minor scale are the same. This cyclic scale equality in both the Major scale and the Natural Minor scale can be formulated as \{C C\# D D\# E F F\# G G\# A A\# B $\}_{\text {Major scale }}=\{\text { A A\# B C C\# D D\# E F\# G G\# }\}_{\text {Natural Minor Scale }}$.

| Major Scale |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | C | Cmaj | Dmin | Emin | Fmaj | Gmaj | Amin | Bdim |
|  | C\# | C\#maj | D\#min | Fmin | F\#maj | G\#maj | A\#min | Cdim |
|  | D | Dmaj | Emin | F\#min | Gmaj | Amaj | Bmin | C\#dim |
| ${ }^{\text {M }}$ | D\# | D\#maj | Fmin | Gmin | G\#maj | A.tmaj | Cmin | Ddim |
| 4 | E | Emaj | F\#min | G\#min | Amaj | Bmaj | C\#min | D\#dim |
| Z | F | Fmaj | Gmin | Amin | A.tmaj | Cmaj | Dmin | Edir |
| [1] | F\# | F\#maj | G\#min | A\#min | Bmaj | C\#maj | D\#min | Fdim |
| k | G | Gmaj | Amin | Bmin | Cmaj | Dmaj | Emin | F\#dim |
| U | G\# | G\#maj | AFmin | Cmin | C\#maj | D\#maj | Fmin | Gdim |
|  | A | Amaj | Bmin | Cfmin | Dmaj | Emaj | F\#min | G\#dim |
|  | A\# | Atmaj | Cmin | Dmin | D\#maj | Emaj | Gmin | Adim |
|  | B | Bmaj | Cfmin | D $\quad$ min | Emaj | F\#maj | G\#min | A\#dim |


| Natural Minor Scale |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | C | Cmin | Ddim | D\#maj | Fmin | Gmin | G\#maj | A\#maj |
|  | C\# | C\#min | $\mathrm{D} \# \mathrm{dim}$ | Emaj | F\#min | G\#min | Amaj | Bmaj |
|  | D | Dmin | Edim | Fmaj | Gmin | Amin | A\#maj | Cmaj |
| ¢ | D\# | D\#min | Fdim | F\#maj | G\#min | A\#min | Bmaj | c\#maj |
| $\sum$ | E | Emin | F\#dim | Gmaj | Amin | Bmin | Cmaj | Dmaj |
| Z | F | Fmin | Gdim | G\#maj | A\#min | Cmin | C\#maj | D\#maj |
| [1 | F\# | F\#min | G\#dim | Amaj | Bmin | C\#min | Dmaj | Emaj |
| < | G | Gmin | Adim | A\#maj | Cmin | Dmin | D\#maj | Fmaj |
| $\sim$ | G\# | G\#min | A\#dim | Bmaj | C\#min | D\#min | Emaj | F\#maj |
|  | A | Amin | Bdim | Cmaj | Dmin | Emin | Fmaj | Gmaj |
|  | A\# | A\#min | Cdim | C\#maj | D\#min | Fmin | F\#maj | G\#maj |
|  | B | Bmin | C\#dim | Dmaj | Emin | F\#min | Gmaj | Amaj |
| Harmonic Minor Scale |  |  |  |  |  |  |  |  |
|  | C | Cmin | Ddim | D\#aug | Fmin | Gmaj | G\#maj | Bdim |
|  | C\# | CHmin | D\#dim | Eaug | $\mathrm{F} \% \mathrm{~min}$ | G\#maj | Ama | Cdim |
|  | D | Dmin | Edim | Faug | Gmin | Amaj | A\#maj | c\#dim |
|  | D\# | D\#min | Fdim | F\#aug | G\#min | A\#maj | Bmaj | Ddim |
| $\sum$ | E | Emin | F\#dim | Gaug | Amin | Bmaj | ¢ Cmaj | D\#dim |
| Z | F | Fmin | Gdim | G\#aug | A\#min | Cmaj | C\#maj | Edim |
| H | F\# | F\#min | G\#dim | Aaug | Bmin | c\#maj | Dmaj | Fdim |
| S | G | Gmin | Adim | A\#aug | Cmin | Dmaj | D\#maj | F\#dim |
| - | G\# | G\#min | A\#dim | Baug | C\#min | D\#maj | Emaj | Gdim |
|  | A | Amin | Bdim | Caug | Dmin | Emaj | Fmaj | G\#dim |
|  | A\# | A\#min | Cdim | C\#aug | D\#min | Fmaj | F\#maj | Adim |
|  | B | Bmin | C\#dim | Daug | Emin | F\#maj | Gmaj | A\#dim |

Figure 2-6: Chords that can be derived from the notes in the four music scales types

The set of notes on which the piece is built is known as the Key. Furthermore, by grouping these notes we can identify the set of chords which belong to the key. These top-down relationships of notes, chords, and keys are illustrated in Figure 2-7. In Figure 2-7, the top layer represents the music notes in different octaves. In the second layer, chords are formulated by combining notes according to the note relationships, which are described in Table 2-2. Based on the different chord combinations we derive 12 music scales, each in four different types of scales (the $3{ }^{\text {rd }}$ layer). Major and Minor are the two possible types of keys derived from the major and the natural minor scales respectively (Shenoy et al 2004 [109]). For example, all the chords in
the D Major scale (i.e. Dmaj Emin F\#min Gmaj Amaj Bmin C\#dim) belong to the D Major key, and all the chords in the C Natural Minor scale (i.e. Cmin Ddim D\#maj Fmin Gmin G\#maj A\#maj) belong to the D Minor key. The set of chords derived in a Natural Minor scale can be found in a different Major scale. Thus, a Minor key (chords in natural minor scale) which has the same set of chords as a Major key is called relative Minor key of the Major key. For example, the relative Minor key of the C major is A minor. Since notes in the major scale and the minor scale are arranged differently, music of these scales generates different feelings altogether. Sad feelings may be developed upon hearing music in a minor key. Although the Minor key is derived from notes in the natural minor scale, musicians usually play notes in both Harmonic and Melodic minor scales to harmonize their piece.


Figure 2-7: Overview of top down relationship of notes, chords and key

The Key identification in music is useful for error correction in chord detection algorithms because the key indicates the possible fluctuation of the set of chords in the harmony line (see chapter 4.4 .3 for more details).

### 2.3 Composition of music phrases

The rhythm of words can be made to fit into a music phrase [100]. The vocal regions in music are constructed using words and syllables, which are spoken according to a time signature(TS). Figure 2-8 shows how the words "Little Jack Horner sat in the Corner" form themselves into a rhythm, and the music notation of those words. The important words or syllables in the sentence fall onto accents to form the rhythm of the music. Typically, these words are placed at the first beat of a bar. When TS is set to two Crotchet beats per bar, we see that the duration of the word "Little" is equal to two Quaver notes and the duration of the word "Jack" is equal to a Crotchet note.


Figure 2-8: Rhythmic groups of words

The durations of music phrases in popular music are commonly two or four bars [100] [120]. However, accents are still placed on the first beat of the bar even though the rhythmic effect is different. The incomplete bars are filled with rests (Figure 2-3 the $2^{\text {nd }}$ and $3^{\text {rd }}$ bars) or humming (duration of humming is equal to the length of a note).

### 2.4 Popular song structure

Popular song structure often contains Intro, Verse, Chorus, Bridge, Middle eighth, INST-instrumental sections and Outro [120]. As shown in Figure 2-1, these parts are built upon melody-based similarity regions and content-based similarity regions. Melody-based similarity regions are defined as the regions which have similar pitch
contours constructed from the chord patterns. Content-based similarity regions are defined as the regions which have both similar vocal content and melody. Corresponding to the music structure, the Chorus sections and Verse sections in a song are considered the content-based similarity regions and melody-based similarity regions respectively. These parts can be considered semantic clusters and are shown in Figure 2-9. All the chorus regions in a song can be clustered into a chorus cluster. All the verse regions in the song can be grouped into a verse cluster and so on.


Figure 2-9: Semantic similarity clusters which define the structure of the popular song

The intro may be $2,4,8$ or 16 bars long, or there maybe no intro in a song. The intro is usually composed of instrumental music. Both verse and chorus are 8 or 16 bars long. Typically, the verse is not as strong melodically as the chorus. However, in some songs they are equally strong and most people can hum or sing both. A bridge links the gap between the verse and chorus, and may be only two or four bars. Silence may also act as a bridge between the verse and chorus of a song, but such cases are rare. Middle eighth, which is 4,8 or 16 bars long, is an alternate version of a verse with a new chord progression possibly modulated by a different key. Many people use the term "middle eighth" and "bridge" synonymously. However, the main difference is that the middle eighth is longer (usually 16 bars) than the bridge and usually
appears after the third verse in the song. There are instrumental sections (i.e. INST) in the song and they can be instrumental versions of the chorus, verse, or entirely different tunes with a set of chords together. Outro, which is the ending of the song, is usually a fade-out of the last phrases of the chorus. We have described the parts of the song which are commonly arranged according to the simple verse-chorus and repeat pattern. Two variations on the themes are listed below:
(a). Intro, Verse 1, Verse 2, Chorus, Verse 3, Middle eighth, Chorus, Chorus, Outro (b). Intro, Verse 1, Chorus, Verse 2, Chorus, Chorus, Outro

Figure 2-10 illustrates two examples for the above two patterns. Song "25 minutes" by MLTR follows the pattern (a) and "Can’t Let You Go" by Mariah Carey follows the pattern (b). For a better understanding of how artist have combined these parts to compose a song, we conducted a survey on popular Chinese and English songs. Details of the survey are discussed in the next section.

Song: 25 Minutes
Artist / Band: MLTR
Song structure: Intro (INST) Verse 1 Verse 2 Bridge Chorus 1 Verse 3 Bridge Chorus 2 Middle eight Chorus 3 Chorus 4 Outro (INST)

| Intro: Instrumental (INST) | Verse 3 | Chorus 3 |
| :---: | :---: | :---: |
|  | Against the wind | Boy I've missed your kisses |
| Verse 1 | I'm going home again | All the time but this is |
| After some time | hummma (humming) | Twenty five minutes too late |
| I've finally made up my mind | Wishing me back | Though you traveled so far |
| She is the girl | To the time when we were more than friends | Boy I'm sorry you are |
| And I really want to make her mine |  | Twenty five minutes too late |
|  | Bridge |  |
| Verse 2 | But still I see her in front of the church | Chorus 4 |
| I'm searching everywhere | The only place in town where I didn't search | Boy I've missed your kisses |
| To find her again | She looked so happy in her wedding dress | All the time but this is |
| To tell her I love her | But she cried while she was saying this | Twenty five minutes too late |
| And I'm sorry about the things I've done |  | its too late |
|  | Chorus 2 | Though you traveled so far |
| Bridge | Boy I've missed your kisses | Boy I'm sorry you are |
| I find her standing | All the time but this is | Twenty five minutes too late |
| in front of the church | Twenty five minutes too late | I can still hear her say... |
| The only place in town where I didn't search | Though you traveled so far |  |
| She looks so happy in her wedding dress | Boy I'm sorry you are | Outro - Instrumental (INST) |
| But she's crying while she's saying this | Twenty five minutes too late |  |
| Chorus 1 | Middle Eight |  |
| Boy I've missed your kisses | Out in the streets |  |
| All the time but this is | Places where hungry hearts have nothing to eat |  |
| Twenty five minutes too late | Inside my head |  |
| Though you traveled so far | Still I can |  |
| Boy I'm sorry you are | hear the words she said |  |
| Twenty five minutes too late |  |  |

Song : Cant let go
Artist / Band: Mariah Carey
Song structure : Intro (INST) Verse 1 Chorus 1 Verse 2 Chorus 2 Bridge Chorus 3 Outro (INST)

| Intro : Instrumental (INST) | Verse 2 <br> just cast aside, | Bridge do you ever realise, |
| :---: | :---: | :---: |
| Verse 1 | u dont even know i'm alive | the sorrow i have inside, |
| There you are, | u just walk on by, | everyday of my life. |
| holding her hand | don't care to see me cry | do you know the way it feels |
| i am lost, | and here i am | when all your love just dies, |
| dying to understand | still holding on, | i try and try to deny that i need you |
| didnt i , | I cant accept | but still you remain on my mind |
| cherish u right | my world is gone |  |
| dont u know, | no no... |  |
| u were my life. |  | Chorus 3 |
|  | Chorus 2 | even though i try, |
| Chorus 1 | even though i try, | i cant let go |
| even though i try, I cant | i cant | something in your eyes |
| let go | let go, | captures my soul |
| something in your eyes | something in your eyes captures my soul. | and everynight i see you |
| captures | and everynight i see you in my dreams, | in my dreams |
| my soul | you're all i know, | you're all i know |
| and everynight i see you in my dreams | i cant let go | i can't let go... |
| you're all i know |  |  |
| i cant let go |  | Outro - Instrumental (INST) |

Figure 2-10: Two examples for verse- chorus pattern repetitions.

## 2．5 Analysis of Song structures

We have conducted a survey using popular English and Chinese songs to better understand song structures．One aspect of the survey is to discover characteristics of the songs such as tempo variation，total vocal signal content variation，and the different smallest notes（Quarter note，Eighth note，Sixteenth note，or Thirty second note）．The other aspect is to find out how the components of the popular song structure［120］（i．e．Intro，Verse，Chorus，Bridge，INST，Middle eighth and Outro ［120］）have been arranged to formulate the song．A total of 220 songs，consisting of 10 songs from each singer，have been used in the survey．They are listed in Table 2－3．

Table 2－3：Names of the English and Chinese singers and their album used for the survey

| Language | English |  | Chinese |  |
| :---: | :---: | :---: | :---: | :---: |
| Gender | Female | Male | Female | Male |
| 1 | Celine Dion <br> （Falling into you） | Ben Jelen （Give it all away） | FIR（Fly）飛兒樂團（飛） | Zhou Jie Lun（Qi Li Xiang）周杰倫（七里香） |
| 2 | Dido（Life for rent） | Bryan Adams （On a Day Like Today） | Liang jinru（Lian ai de li liang）梁静茹（樂愛的力量） | A du（Tianhei） <br> 阿杜（天黑） |
| 3 | Faith Hill （Love will always win） | Elton John （Greatest Hits 1970－2002） | Rene Liu（Full Bloom）刘若英 | Zhou Chuan Xiong（Transfer）周傳雄 |
| 4 | Mariah Carey （Greatest Hits） | Michael Bolton（Vintage） | Pei Ni（hao）佩妮（好） | Liu Wen Zheng（Collection）刘文正 |
| 5 | Shania Twain （Come on Over） | Michael Learns to Rock－ MLTR（Paint My Love） | Zhao Wei（Piao）趙薇（飘） | Huang Pinyuan（Xiao－Wei）黄品源（小薇） |
| 6 |  | Richard Marx （Greatest Hits） |  |  |
| 7 |  | West life（West life） |  |  |

## 2．5．1 Song characteristics

To find out the vocal content variation of the songs，we first manually annotate the vocal and instrumental regions in the songs by conducting listening tests．The song annotation procedure is detailed in chapter 6．3．1．Figure 2－11 shows the percentage of
the vocal signal content of the 200 songs. It is found that the average vocal signal content of a song is around $60 \%$. The vocal content of the songs vary between 50 to $75 \%$.


Figure 2-11: Percentage of the average vocal content in the songs

The details of the songs such as tempo, meter and note are collected from the music sheets. Figure 2-12 shows the tempo variation of the songs. All the songs have a $4 / 4$ meter. Thus, the tempo is the number of quarter notes per minute. The songs have tempo variations of between 30 to 190 BPM (Beats per minutes). The average tempo of a song is around 80 BPM , which implies that the quarter note is 750 ms long.

We then look for the smallest note that appears in a song. Figure 2-13 shows the percentage of different notes which appear as the smallest note in a song. According to the results, the sixteenth note is the smallest note for around $50 \%$ of Chinese and

English songs. Overall, the eighth note or the sixteenth note appears most frequently as the smallest note in popular songs.


Figure 2-12: Tempo variation of songs


Figure 2-13: Percentage of the smallest note in songs

### 2.5.2 Song structures

We performed further analysis on popular song structures. The analysis uncovered 12 facts and which are explained below. Based on these findings, we formulated some heuristic rules to automatically detect popular song structures. Our music structure detection algorithm is explained in chapter 5.3.

1. Songs which have INTRO:

Instrumental INTRO:

2 Songs which do not have INTRO
Songs which start with the VERSE
Songs which start with the CHORUS

3 Songs which have instrumental OUTRO
Songs which have chorus melody as instrumental OUTRO

Chinese English Overall

| $97.9 \%$ | $95.0 \%$ | $96.3 \%$ |
| :--- | :--- | :--- |
| $94.8 \%$ | $85.0 \%$ | $89.4 \%$ |

94.8\%
85.0\%
89.4\%

| $1.0 \%$ | $0.0 \%$ | $0.5 \%$ |
| :--- | :--- | :--- |

72.9\%
31. 7\%
50. 0\%
5.2\%
17.1\%
19.4\%

4 Songs which don not have Instrumental OUTRO
Songs with fading CHORUS (vocals or /and humming)
$31.3 \% \quad 80.8 \%$
58.8\%

5 Songs which have MIDDLE-EIGHTH
25.0\%
$44.2 \%$
$35.6 \%$

6 Number of VERSEs and CHORUSes

| Number of <br> Verses (V) or <br> Choruses (C) | English (V) | Chinese (V) | English (C) | Chinese (C) | Total (V) | Total (C) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 0 |  |  | $1.7 \%$ |  |  | $0.9 \%$ |
| 1 | $1.7 \%$ | $6.3 \%$ | $1.7 \%$ |  | $3.7 \%$ | $0.9 \%$ |
| 2 | $65.0 \%$ | $40.6 \%$ | $19.2 \%$ | $31.3 \%$ | $54.2 \%$ | $24.5 \%$ |
| 3 | $20.8 \%$ | $35.4 \%$ | $50.8 \%$ | $47.9 \%$ | $27.3 \%$ | $49.5 \%$ |
| 4 | $10.0 \%$ | $12.5 \%$ | $21.7 \%$ | $9.4 \%$ | $11.1 \%$ | $16.2 \%$ |
| 5 | $0.8 \%$ | $4.2 \%$ | $4.2 \%$ | $8.3 \%$ | $2.3 \%$ | $6.0 \%$ |
| 6 | $0.8 \%$ |  | $0.8 \%$ | $2.1 \%$ | $0.5 \%$ | $1.4 \%$ |
| 7 |  | $1.0 \%$ |  | $1.0 \%$ | $0.5 \%$ | $0.5 \%$ |
| 8 | $0.8 \%$ |  |  |  | $0.5 \%$ |  |

7 CHORUS and VERSE combinations

| $\begin{aligned} & > \\ & \dot{1} \\ & 0.0 \\ & >0 \end{aligned}$ | Chorus - C |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|  | 0 |  |  |  |  |  |  |  |  |
|  | 1 |  |  | 2.4\% | 0.9\% |  | 0.5\% |  |  |
|  | 2 |  |  | 10.2\% | 30.1\% | 19.6\% | 2.8\% |  | 0.5\% |
|  | 3 |  | 0.5\% | 6.5\% | 12.0\% | 5.1\% | 2.3\% | 0.9\% |  |
|  | 4 | 0.5\% | 0.5\% | 3.7\% | 5.6\% | 0.5\% | 0.5\% |  |  |
|  | 5 |  |  | 1.4\% | 0.9\% |  |  |  |  |
|  | 6 |  |  |  |  |  |  | 0.5\% |  |
|  | 7 |  |  | 0.5\% |  |  |  |  |  |
|  | 8 | 0.5\% |  |  |  |  |  |  |  |

8. Length of the Chorus and Verse in bars

For Chinese songs

|  | Length of Chorus in Bars |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | 4 | 8 | 16 | Others |
|  | 4 | 7\% | 8\% | 1\% | 3\% |
|  | 8 | 2\% | 46\% | 13\% | 2\% |
|  | 16 | 0\% | 6\% | 12\% | 0\% |
|  | Others |  |  |  |  |

For English songs

|  | Length of Chorus in Bars |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | 4 | 8 | 16 | Others |
|  | 4 | 2\% | 3\% | 0\% | 0\% |
|  | 8 | 3\% | 50\% | 11\% | 7\% |
|  | 16 | 0\% | 10\% | 14\% | 1\% |
|  | Others |  |  |  |  |

For all the songs

|  | Length of Chorus in Bars |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | 4 | 8 | 16 | Others |
|  | 4 | 4\% | 5\% | 1\% | 2\% |
|  | 8 | 3\% | 48\% | 12\% | 5\% |
|  | 16 | 0\% | 8\% | 13\% | 1\% |
|  | Others |  |  |  |  |

## Song structures

9. Songs which have V1-C1-V2-C2 pattern Songs which have MIDDLE-EIGHTH
Chinese English Overall
$38.5 \% \quad 68.3 \%$
55.1\%
$6.3 \%$
36.7\%
23.1\%
(V1-Verse 1 and C1-Chorus 1)
10. Songs which have V1-V2-C1-V3-C2 patten Songs which have MIDDLE-EIGHTH
25.0\%
11. 0\%
19.4\%
4.2\%
12. 3\%
$6.5 \%$
13. Other V-C repetitions
V1-V2-C1-V3-V4-C2
$10.4 \%$
6.7\%
8.3\%
V1-V2-C1-C2
10.4\%
5.8\%
7.9\%
14. The rest of the song structure followed by $\mathrm{V} 1-\mathrm{C} 1-\mathrm{V} 2-\mathrm{C} 2$ and $\mathrm{V} 1-\mathrm{V} 2-\mathrm{C} 1-\mathrm{V} 3-\mathrm{C} 2$

| Patterns followed by the pattern P1 and P2 | Pattern P1 (V1-C1-V2-C2) | Pattern P2 (V1-V2-C1-V3-C2) |
| :---: | :---: | :---: |
| C | $17.6 \%$ | $6.9 \%$ |
| C-C | $7.9 \%$ | $1.4 \%$ |
| MIDDLE EIGHTH-C | $9.7 \%$ | $3.2 \%$ |
| MIDDLE EIGHTH-C-C | $3.2 \%$ | $0.9 \%$ |
| V | $1.4 \%$ | $1.9 \%$ |
| V-C | $1.4 \%$ | -- |
| V-C-C | $0.9 \%$ | -- |

## Summary of the statistical analysis

- It is found that over $95 \%$ of both the Chinese and the English popular songs have an introduction (i.e. Intro) and over $85 \%$ of the songs of each language have instrumental Intros.
- Around $40 \%$ of the songs have either an instrumental or a mixed vocal Outro, which is composed from the chorus melody.
- The Middle eighth part is more commonly used in the English songs than in the Chinese songs.
- The chorus and verse are always four bar, eight bar and sixteen bar long. Most songs have the same chorus and verse length either eight bars or sixteen bars.
- Over $80 \%$ of English songs have either V1-C1-V2-C2 or V1-V2-C1-V3-C2 patterns.
- Over $55 \%$ of the songs have a chorus before the Outro.

We will utilize these popular music structure characteristics in the design of our algorithms in Chapter 5.3.

## Literature Survey

In Chapter 2, we have briefly described information embedded in the music structure and laid out the information in the music structure pyramid according to bottom-up hierarchy. The music research community has been exploring different methodologies to extract music structure information. These methodologies incorporate perceptual, statistical and psychological characteristics of music signals. Past music research has been mainly focused on symbolic music like MIDI (Music Instruments for Digital Interface), which requires a small amount of storage space and has access to music score information (beat structure, melody, harmony, music sources- tracks, tempo, etc.). Figure 3-1 (top) describes the MIDI song generating platform of the Cakewalk software. Figure 3-1 (down) shows the text format of the created MIDI information. Most music information retrieval (MIR) systems have been implemented on symbolic databases (Zhu 2004. [143]) and key research challenges have been in sequential music content matching and the extraction of music symbols which are mainly in text format. However, present computing algorithms are mature enough to handle such challenges.

Recent developments in high bandwidth data transmission, large data storage, and high computing power have allowed the research efforts to shift to real sound recordings. Past research publications have revealed that understanding the ingredients on which music structure is formed is necessary for developing music applications. Some of the focuses of interest are error concealment in music streaming, music protection (watermarking), and real music representation techniques
such as music summarization, compression, genre classification and artist identification. In chapter 7, we discuss how the proposed music structural analysis framework can help develop these applications. The following discussions in this chapter give general idea of where researchers stand in trying to solve real world music problems. Our level of awareness of music structure leads to confidence in the solving of real world music problems. The following sections, which discuss previous research contributions to music research, are organized according to the order of information layers in the music structure pyramid (shown in Figure 2-1).


Figure 3-1: MIDI music generating platform in the Cakewalk software (top) and MIDI file information representation in text format (bottom).

### 3.1 Time information extraction (Beats, Meter, Tempo)

Time information extraction is the primary step in achieving deep understanding of music content. Researchers have been actively working on designing algorithms for detecting beats, the meter, and the tempo of music. In chapter 2 , we have pictorially explained (Figure 2-1) how music chords, harmony, music phrases, and semantic regions are built upon this time information.

Rhythm tracking research was initially focused on symbolic music data, i.e. MIDI. Allen and Dannenberg (1990) [3] have proposed a real time beat tracking / prediction technique for a MIDI sequence using a beam search tool. Since all beats and music notes are represented in symbolic format, the rhythm tracking problem turns out to be pattern matching exercise. However, real sound recordings do not provide such symbolic stampings on the signal. Therefore, low-level signal analysis steps have been introduced to extract the time information (beats, meter and tempo) from the music signals.

To give a good idea of the target problem, Figure 3-2 shows 6 seconds of the drum, bass guitar, piano, and edited track (mixture of 3 instrumental tracks) of the song "I Let you Go" by Iven. Dotted vertical lines denote the inter-beat intervals (IBIs), quarter notes in this case. As seen in the Figure, inter-beat intervals are very clear in the time domain signal of the drum track. However, after mixing all the tracks it is hard to find the positions of the beats. Successful detection of beat positions and interbeat intervals can define the tempo and meter of the music. The drum track is considered the time stamp in a song because, like most the percussion instruments, the drums produces high energy impulses, which are significant in the signal.


Figure 3-2: Instrumental tracks (Drum, Bass guitar, Piano) and edited final tract (mix of all the tracks) of a ballad (meter- $4 / 4$ and tempo - 125 BPM) "I Let You Go" sung by Ivan. First 6 seconds of the music is considered.

In the drum track, strong ${ }^{1}$ and weak ${ }^{2}$ beat pattern repetition helps to detect the meter and tempo of the music (Goto 2001 [48]). Thus, the absence of a drum track in the music makes meter and tempo detection much harder. To reduce the time information extraction to a simpler task, many of the previous works have been carried out with the assumptions given below.
> Meter doesn't change in the music
$>$ Tempo doesn't change in the music
$>\quad$ Expected tempo range is fixed

Commonly discussed basic steps in the time information extraction systems are shown in Figure 3-3. Onset detection has been considered the first step towards detecting beats, meter, and tempo. Energy transient analysis in both the time domain

[^0]and the frequency domain has been considered the main step for onset detection in music signals.


Figure 3-3: Basic steps followed for extracting time information.

Duxburg et al. (2002) [34] detected energy fluctuations in both the time and the frequency domains. In order to increase energy fluctuation resolution, the signal is first decomposed into 5 sub-bands whose frequency ranges are $\{(0 \sim 1.1),(1.1 \sim 2.2)$, (2.2~5.5), (5.5~11.0) and (11~22)\} KHz. Thresholds were set in each sub-band to detect onsets. Bello and Sandler (2003) [9] only analyzed frequency fluctuations (Phase) to detect onsets in the music. Scheirer (1998) [102] decomposed the signal into 6 sub-bands $(0 \sim 200,200 \sim 400, ~ 400 \sim 800, ~ 800 \sim 1600, ~ 1600 \sim 3200$, and 3200~higher freq Hz ) and passed the decomposed signals through a combo filter bank to analyze the tempo and beats of the signal. The algorithm, known as the "perceptual model", was tested on 60 different classes of genres. Although beat detection reveals the meter of the music, no direct information could be found in that paper about meter estimation. Klapuri (1999) [64] described a psycho acoustical model which consists of critical filter bank (21 filters) in $44 \mathrm{~Hz} \sim 17 \mathrm{kHz}$ frequency range to decompose the
music signal. Then sub-band onsets are detected by operating first order differentiation on detected time domain energy transients in each sub-band signals. By summing up all the sub-band onsets and using a global threshold, genuine onsets are detected.

Dixon (2001) [29] used the time-domain local peak tracking technique to detect onsets and cluster inter-onset intervals (IOI) in order to detect tempo and meter in classical, jazz and popular music. Tzanetakis et al. (2001) [124] proposed computation of a beat histogram to characterize the different music genres. The beat histogram was generated by decomposing the signal in to sub-bands by wavelets. However, they didn't describe how meter or tempo can be detected from the beat histogram.

Cemgil et al. (2001) [18] proposed a stochastic dynamic system for tempo tracking. Tempo was modelled as a hidden state variable of the system. Their simulation results on two Beatles songs illustrated a tracking accuracy of $90 \%$ of the beats. Since the authors have not tested their approach on real music data such as vocal mixed instrumental lines, there is no guarantee that the algorithm will work well on complex music data.

Gouyon et al (2002) [50] discussed the extraction of rhythmic attributes from percussive signals. They analyzed energy transients to construct an IOI histogram and computed the position of the smallest rhythmic pulse units also known as "Ticks", from the peak tracker. Since the system was only tested on synthetically generated samples, which consist of 5 -second drum clips, there is no guarantee that the system
will work well with polyphonic music. Jensen and Andersen (2003) [55] introduced a beat probability vector to keep track of previous beat intervals and enhance beat prediction capabilities. For beat tracking, they constrained the tempo to be within the (50~200) BPM range. Alonso et al (2003) [2] used the noise/harmonic decomposition technique to estimate the tempo of 54 excerpts from different music genres, and claimed an averaged accuracy, of $96 \%$. The signal was first decomposed into 12 uniform non-overlapping sub-bands using a $200^{\text {th }}$ order FIR filter bank with 80 dB stop band ripples. Then the sub-band signals were projected into a noise subspace to detect periodicities. These detected periodicities indicate the tempo.

To detect tempo, micro time, and time signature, Uhle and Herre (2003) [128] first decomposed the signal into 7 sub-bands whose frequencies are in the range of $\{(0 \sim 125),(125 \sim 250),(250 \sim 500),(500 \sim 1000),(1000 \sim 2000),(2000 \sim 4000)$ and (4000~8000) \} Hz. Sub-band onsets were detected from half wave rectified amplitude envelopes of each sub-band. Based on inter-onset interval (IOI) analysis, the rhythm characteristics of the signal were detected. Goto (1994 [46] \& 2001 [48]) assumed the music's meter to be $4 / 4$ and its tempo to be between $61 \sim 185$ BPM. His initial work, Goto (1994), was focused on beat tracking for signals with drum tracks. Onsets were detected by analyzing frequency transients. Based on the inter-onset interval (IOI) histogram, beats were predicted in the song. Later, Goto (2001) [48] proposed a real time beat tracking system which deals with music without drum sounds. In this system, information about onset times, chord change, and drum patterns were taken into consideration for identifying the inter-beat intervals (quarter-note level, half-note level, and bar level). However, the frequency resolution $(21.53 \mathrm{~Hz})$ used for detecting
chord changes is arguable since the F0 (fundamental frequency) differences of notes in lower octaves (see Table 2-1) are as low as 1 Hz .

Gao and Lee (2004) [42][41] extracted 12-MFCCs from 23.3 ms , $50 \%$ overlap music segments and fed them into a maximum a posterior algorithm to detect both onsets and beats in the music. However, this prediction method inadequately accounted for the beats and meter correlation in the music. This is because without knowing the meter of a song, it is difficult to identify beats from the detected onsets. Scaringella and Zoia (2004) [103] proposed a real time beat tracking system based on the combination of lossy onset detection, note accentuation evaluation (to estimate metrically essential events), and a multi-agent mechanism. For lossy onset detection, they used a signal energy fluctuation tracking technique in the time domain, similar to that used by Dixon (2001) [29]. To track tempo, Davies and Plumbley (2004) [26] first detected onsets in music using a similar method as described in Bello et al (2004) [9]. Beats were then predicted by running an auto-correlation function (ACF) over detected onsets. However, the evaluation of the performance of the proposed tempo tracking system is weak because they didn't mention the characteristics (time signature, actual, and tempo variation) of the data set used for experiments and how these characteristics would effect the results.

Pikrakis et al (2004) [92] proposed a method to extract meter and tempo for 300 audio recordings of Greek dance folklore music and neighbouring Eastern music traditions. They assumed that the meter remains constant throughout the music and the tempo varying in $40 \sim 330$ BPM range. The algorithm is based on the analysis of selfsimilarity of Mel frequency cepstral coefficient features, which were extracted from

100 ms windows with $97 \%$ overlap. Sethares and Staley (2001) [104] measured the periodicities and meter of music by projecting octave based decomposed music signals onto a set of non-orthogonal periodic sub spaces. Sethares et al. (2005) [105] discussed two beat tracking methods for real audio data based on both the Bayesian decision framework and the gradient strategy. They claimed that the gradient approach is numerically less complex than the Bayesian framework. However, authors were disappointed by the overall performance of the gradient algorithm due to the unacceptable user interaction over tuning both window and step sizes. Since they discussed neither performance evaluation of the algorithms based on ground truth data sets nor characteristics of the dataset, the algorithm couldn't be validated.

Wang and Vilermo (2001) [131] proposed a compressed domain beat tracking system based on full band and sub-band analysis of the Modified Discrete Cosine Transform (MDCT) coefficients. An inter-onset interval (IOI) histogram was used to select the correct beats. Out of 6 popular songs (all the songs have $4 / 4$ meter), beats in 4 songs were correctly tracked. However, the test dataset is too small to validate the algorithm.

### 3.2 Melody and Harmony analysis

Playing music notes according to time step (meter and tempo) forms complex tonal structures such Chords and Keys. Figure 2-7 illustrates the correlations between music notes, chords, and keys (chapter 2.2). Different approaches have been discussed in the literature to detect notes, chords and keys based on the detection of F0 and harmonic components of the notes. The foundations of these approaches are based on early $20^{\text {th }}$ century research on the psychological representation of tones.

Stevens et al. (1937) [114] and Stevens and Volkmann (1940) [115] described pitch perception as a continuous psychological effect proportional to the magnitude of the frequency (i.e. pitch height). The octave, which is the basis of the music tonal system, has been studied by many researchers (Dowling and Harwood 1986 [31]). It was proved that tones which differed by an octave interval are psychologically closely related (Allen 1967 [1], Bachem 1954 [6], Deutsch 1973 [26] \& [28]). Based on this evidence, circular representation of the music pitch over octaves (see Figure 2-5) was proposed (Bachem 1950 [5], Shepard 1964 [110], Krumhansl 1979 [66] , Bharucha and Stoeckig 1986 [12] \& 1987 [13]). This circular representation is known as the chroma cycle (Rossing et al 2001 [99]).

The frequency ratio of a physical octave is 2:1 (Rossing et al 2001[99] and see Table 2-1). However, cognitive experiments have highlighted that this subjective octave ratio is close to $2: 1$ at lower frequencies but increases with the frequency and exceeds the physical octave ratio by $3 \%$ at about 2 kHz . (Ward 1954 [135], Terhardt 1974 [121], Sundberg and Lindqvist 1973 [117], Ohgushi 1978 [89] \& 1983 [90]). Musically trained/untrained listeners (Ward 1954 [135]) and number of music cultures (Dowling and Harwood 1986 [31]) presented this octave enlargement effect. In order to study this octave enlargement, Ohgushi (1978 [89] \& 1983 [90]), Hartmann (1993) [53], McKinney and Delgutte (1999) [82] suggested an octave matching scheme based on a temporal model which predicts the octave enlargement effect. However, the constant $2: 1$ physical octave ratio is commonly practiced to simplify the complexity of musicology. All algorithms proposed in this thesis are also based on the same constant ratio.

Listening tests have also revealed that octave judgments for music tones over 2 kHz is difficult. Pitch perception experiments conducted by Ritsma (1967) [96] concluded that fundamental frequencies in the $100-400 \mathrm{~Hz}$ range and their $3^{\text {rd }}, 4^{\text {th }}$, and $5^{\text {th }}$ harmonics, which cover up to 2 kHz in frequency range, produce well-defined pitch perception by human ears. Biasutti (1997) [14] conducted hearing tests using 12 subjects to find the frequency limits (lower and upper) of musicians in identifying major and minor triads. These lower and upper limits were found in the (120~3000) Hz frequency range. Ward (1954) [135], Attneave and Olson (1971) [4] have also acknowledged that the upper limit of music pitch is in the range of $4-5 \mathrm{kHz}$. Thus, the useful upper limit of the fundamental frequencies of tones produced by music instruments is set below 5 kHz . The highest tone (C7) of the piano has a frequency of 4186 Hz.

There are two approaches discussed in the literature for representing music pitch. Goldstein (1973) [45] and Terhardt (1974 [121] \& 1982 [122]) proposed two psychoacoustical approaches, harmonic representation and sub-harmonic representation of complex tones, respectively. In Goldstein's pitch representation, music tone is characterized by the fundamental frequency (F0) with harmonic partials. Terhardt proposed that each separable component of a complex tone generates eight subharmonics and that the frequency of the most commonly generated sub-harmonic determines the perceived pitch. Houtgast (1976) [52] also claimed that listeners can discriminate the sub-harmonics of the higher frequencies. Laden and Keefe (1989) [67] examined representation of music pitch in a neural net designed to classify music chords (Major, Minor, and Diminished). They implemented Goldstein's harmonic perceptual model and Terhardt's sub-harmonic perceptual model in terms of the pitch
classes in octaves. Five harmonic components and six sub-harmonic components were considered in their neural net classification models. They claimed that the psychoacoustical representation of music pitches has advantages in encoding information concerning chord inversions and spectral content over the Pitch Class representation. In the higher level, they concluded that the merger of psycho acoustical and cognitive approaches in a neural net paradigm might offer a way to model a musician's cognitive processes from ear to brain.

Moorer (1975) [85], in his thesis, discussed techniques for the harmonic content analysis of continuous music signals for transcription. His work was limited to the mixture of duets. In the initial round, an optimum-comb periodicity detector was used to estimate the loose boundaries of different harmonic content. Then, band pass filtering was used to eliminate unwanted sub-harmonics. Based on the detected harmonics, music notes are identified. Moorer's fundamental research on music signal analysis for transcription is useful for further research by the music community.

Most music communities see the set of music tones as consisting of a finite set of pitches (Deutsch 1999 [28] and Rossing et al. 2001 [99]). An octave is divided into 12 tones (dodecaphonic notes ${ }^{3}$ or 12 pitch class notes) which are about equally spaced in terms of log frequency, and the interval between adjacent pitches is called a half step or a semitone. Two tones separated by 12 half-steps form an octave interval, with a frequency ratio of approximately $2: 1$ (see Table 2-1). Krishnaswamy (2003) [65] investigated one of the claims in Indian classical (Carnatic) music, "there are some musicologists who maintain more than 12 intervals per octave". He used 3 pitch

[^1]trackers i.e. the short time Fourier transform (STFT), the time domain autocorrelation function, and a harmonic source separation technique, with a frame size of 25 to 50 ms to analyze the note pitch transitions. In his examination, he found only 12 distinctive pitches per octave.

The twelve-pitch class profile arrangement has been commonly used to characterize music notes and chords in the literature. Krumhansl (1979) [66] and Bharucha (1986 [12] \& 1987 [13]) discussed the Pitch-Class approach where music tones are characterized by vectors of F0s of dodecaphonic notes. In his real time chord recognition system, Fujishima (1999) [40] compared the nearest neighbour classifier and the weighted sum matching methods to identify the chords from music frames, which are characterized by 12 dimension chroma vectors. Experimental audio data were sampled at 8 kHz and framed into 256 ms (2048-point) non-overlapping frames. For extracting chroma feature vectors from frames, C was set to 12 (12 pitch class) and $f_{\text {ref }}$ was set to 65.4 Hz , which is the frequency of C 2 notes (see Table 2-1) in equation (5-7). Experiments were performed to identify 27 chords (sub-sets of G and F chords) generated by the Yamaha synthesizer and in CD recordings, and claimed over $90 \%$ accuracy. Although he reasoned that previous methods suffered recognition errors because of both noise and the overlapping harmonics of notes due to the polyphonic nature of the signals, it is not clear how he overcame these difficulties by using 12 dimensional chroma vectors to characterize the signal frames. Su and Jeng (2001) [115] represented the harmonic content of chords types (i.e. Major, Minor, Argument, and Diminished) in the time-frequency map using wavelets and modelled them in a self-organizing map. The performance of the system was evaluated on only

480 isolated chord samples. Thus, this setup may not work well in identifying chord types with continuous chord progression.

Sheh and Ellis (2003) [108] tested 24 dimensions of both pitch class profile (PCP) features and MFCCs for chord detection. Features were extracted from 100ms frames with nearly 2.7 Hz frequency resolution in the ( $0 \sim 5$ ) kHz frequency range and were modelled by HMMs. A test was conducted on 20 songs with known chord boundaries. Though the authors claimed that PCP features performed better than MFCCs, they have not validated their system for detecting the correct chord boundaries. Shenoy et al. (2004) [109] discussed a rule based technique to detect the key of a song (only $4 / 4$ music) by identifying Major and Minor chords. Inter-beat interval frames were characterized by 12 dimensional chroma vectors which accounted for 5 octaves (C2~B6). Then authors ran a 16 bar window to detect the key and reported $90 \%$ accuracy over 20 English songs with the key assumed to be constant. However, their analysis procedures were inadequate to justify the selection of inter-beat frame size and did not mention the minimum percentage of chords required to identify the key.

The twelve dimensional chroma based signal characterizing system was proposed by Yoshioka et al. (2004) [141] for chord analysis assuming that songs have only major keys. The authors claimed that the mutual dependency of chords causes errors in detecting chord boundaries and identifying chord symbols. However, their claims were inadequately supported by the experimental steps. Melody of the music is created by playing solo notes with time. Eggink and Brown (2004) [35] proposed a method to extract the melody line from complex audio using knowledge of the signal source and fundamental frequency (F0) detection. Multiple F0s were computed for

70 ms with a $33 \%$ overlapping of frames by setting a threshold on the frequency spectrums, which were computed on signal frames. However, the detection accuracy of the melody boundaries was not discussed in the paper.

Goto (2001) [47] proposed a method to detect both the melody line and the bass line, which is independent of the number of sources in the CD recording. In experiments on jazz and popular music using an adaptive tone model whose parameters were estimated using the EM algorithm, he achieved $88.4 \%$ and $79.9 \%$ frame based accuracies on melody and bass line detection respectively. However, it is not clear in the paper how well melody and bass line boundaries were detected.

Szczerba and Czyżewski (2002) [119] ran autocorrelation over the signal section within the frame to calculate the pitch of the music. Since experiments were conducted on solo instrumental lines, it is difficult to judge how well the system may perform on polyphonic music. Klapuri (2003) [63] used a recursive algorithm for multiple F0 frequency estimation. For frequency analysis, music signals were segmented into 93 ms and 190 ms frames. Then, these frames were transformed into the frequency domain using the Discrete Fourier Transform (DFT). The proposed algorithm claimed to detect both in-harmonic and harmonic sounds in the polyphonic environment. However, the author did not justify his selection of the fixed frame size. Thus, this frame size may not provide good performance detection on both harmonics and in-harmonics when the music's rhythm is changed.

Previous methods for harmonic structure analysis have commonly utilized linear frequency transformation techniques such as Discrete Fourier Transform (DFT).

However, Brown (1991) [15] discussed the importrance of the music signal analysis based on their octave fashion temporal behaviours. She highlighted that non-linear frequency analysis is more sensitive to the frequency components of music tones. In her method, she used a wider window to calculate lower octave frequencies and a smaller window for higher octave frequency. Results revealed that constant Q transformation performs better in identifying F0s and harmonics in music tones than DFT does.

Zhu et al. (2005) [144] proposed a music scale root $^{4}$ and key determination method based on pitch profile features and a tone clustering algorithm. Constant Q transformation was used instead of Fourier transformation, to extract the 12 dimensional pitch class profile feature from 11.6 ms frames covering 7 octaves (27.5 $\sim 3520 \mathrm{~Hz}$ ). The experiments only considered only two minutes of the song, with the assumption that the key remains constant throughout the song.

### 3.3 Music region detection

Information about music regions is placed in the $3^{\text {rd }}$ layer of the music structure pyramid (Figure 2-1). Music regions that can be seen in a song can be classified into pure instrumental (PI), pure vocal (PV), instrumental mixed vocal (IMV), and silence (S) regions. The detection of these regions' boundaries aids in the further analysis of their signal content. Therefore, music region detection is an important step in many applications such as singer identification, music synthesis, transcription, etc. Generally, PV regions are rare in popular music. Therefore, the PV and IMV regions

[^2]can be considered together as Vocal (V) regions. Since silence doesn't technically carry meaningful information, our focus is to effectively detect the PI and V regions in music. Music region detection is useful for semantic information analysis in the $4^{\text {th }}$ and above layers. Some previous works have focused mainly on the analysis of content in the instrumental region. Those methods assumed the entire music content to be solely instrumental music. Therefore, the focus was to identify the types and the names of the instruments (string, bowing, blowing, brass, percussion etc.) played. In this section, we survey past research for both content analysis in the instrumental region and the techniques used for detecting music regions (PI, PV, IMV, and S) in the music.

For instrument identification, Coei et al. (1994) [23] trained a Self-Organizing-Map (SOM) with Mel frequency Cepstral coefficients (MFCC) extracted from the isolated music tones of 40 different music instruments for the purpose of timbre classification. Brown and Cooke (1994) [16] built a system to recognize instruments, in which note similarity and onset asynchrony were used to group duets played by synthesized brass and clarinet. Kaminskyj and Materka (1995) [59] extracted amplitude envelope features from an octave isolated tones to identify the guitar, piano, marimba, and accordion. The instrument classification abilities of a feed-forward neural network with a K-nearest neighbour classifier were compared. They found that both classifiers achieved nearly $98 \%$ accuracy. Kashino and Murase (1998) [60] designed a note transcription (pitch and instrument name) system, which initially matched the input note with notes in the database. A probabilistic network was then employed for music context integration for the purpose of instrument identification. Their system was evaluated with single pitch music, consisting of piano, violin and flute notes. The
authors claimed that instrument identification improved from $67.8 \%$ to $88.5 \%$ after integration of the music context block. To identify music instruments, Fujinaga (1998) [39] trained a K-nearest neighbour classifier with features extracted from 1338 spectral slices representing 23 instruments playing a range of pitches.

Martin (1999) [79] trained a Bayesian network with different types of features, including spectral feature, pitch, vibrato, tremolo features, and note characteristic features, in order to recognize non-percussive music instruments. Eronen et al. (2000) [36] proposed a system for instrument recognition using a wide set of features to model both the temporal and spectral characteristics of sounds. They used the hierarchical classification approach proposed earlier by Martin (1999) [79]. Few instrument recognition systems have been proposed to operate on real music recordings. Brown (1999) [17] trained two Gaussian Mixture Models (GMMs) with constant-Q Cepstral coefficients extracted from the oboe and the saxophone, using approximately one minute of music data for each. Dubnov and Rodet (1998) [32] used quantized MFCC feature vectors to characterize the music played on instruments. Then, they used a clustering algorithm to group similar vectors and measured the similarity of the instruments. Marques (1999) [80] built a classifier to recognize flute, clarinet, harpsichord, organ, piano, trombone, violin and bagpipes. MFCCs were extracted from the recordings of solo instruments to train the classifier.

For singing voice detection, Berenzweig and Ellis (2001) [10] used probabilistic features generated from Cepstral coefficients using a multilayer perceptual neural network acoustic model with 2000 hidden units. Two Hidden Markov Models (vocalHMM and non-vocal-HMM) were trained with the above mentioned features
originally extracted from 61 fragments ${ }^{5}$ of training data to classify the vocal and instrumental sections of a given song. The reported accuracy was $81.2 \%$ with a $40-$ fragment training dataset. Kim and Brian (2002) [62] first filtered the music signal using an IIR band-pass filter ( $200 \sim 2000 \mathrm{~Hz}$ ) to highlight the energy of the vocal region. The vocal regions were detected by analysing the high amount of harmonicity of the filtered signal using an inverse comb filter bank. They achieved 54.9\% accuracy with a test set of 20 songs. Zhang and Kuo (2001) [145] used a simple threshold, calculated using energy, average zero crossing, harmonic coefficients and spectral flux features, to find the starting point of the vocal part of the music. The same technique was applied to detect the semantic boundaries of the online audio data, i.e. speech, music and environmental sounds, for classification. However, the vocal detection accuracy was not reported. Tsai et al. (2003) [123] trained a $64-$ mixture vocal GMM and a 80 -mixture non-vocal GMM using MFCCs extracted from 32 ms signal frames which are 10 ms overlapped with each other. In total they used 216 song tracks for the experiments. They achieved $79.8 \%$ accuracy with 200 test tracks. Bartsch and Wakefield (2004) [8] extracted spectral envelopes to characterize 10 ms frames of vocal notes sung by 12 singers. Singers were then identified using a quadratic classifier. However, authors only used vocal music data in their experiments.

Tzanetakis (2004) [127] simplified the vocal detection problem by applying a songspecific bootstrapping technique which he claimed outperformed normal statistical learning techniques (a classifier trained with bulk data being used to classify an unknown song). He collected 8 or 16 random vocal clips (snippets) of two seconds in

[^3]length to train the classifier. He detected all the vocal regions in the song using spectral shape features (Centroid, Roll off, Relative Sub-band Energy). These features were previously proposed in his PhD thesis [125] for genre classification. Though he claimed that spectral shape features are better than speech processing features (MFCC and LPC), there were no comparison results noted in the paper to support this claim. Test results (on 10 jazz songs) mentioned in the paper indicate that Logistic Regression and Neural Net classifiers both perform better than SVM, Nearest Neighbours, Bayes, and J48 classifiers. However, how the classifier results were compared based on classifier parameter optimization was not discussed in the paper. The major drawback is that music domain knowledge has not been applied to give a solid ground truth for vocal/instrumental boundary detection. The previous methods have borrowed mature speech processing ideas, including fixed frame size segmentation (usually of $20 \sim 100 \mathrm{~ms}$ frame size with frame overlap), speech processing/coding feature extraction, and statistical learning procedures or linear threshold for segment classification, and applied them to detect the vocal/non-vocal boundaries of music signals. Although such methods have achieved up to $80 \%$ in terms of frame level accuracy, their performance is still limited because available music knowledge has not been effectively exploited. For example, the desired regions are aligned with beats and the durations of the regions are multiples of music notes (see chapter 2.3).

Many current research efforts have considered rhythm-based segmentation for music content analysis. Nwe et al (2004) [87], [88] used similar quarter note length segmentation techniques similar to those initially proposed by Maddage et al (2004) [73]. The signal within a quarter note frame is sub-framed into 20 ms durations with

13 ms overlap with each other in order to build a 2-D feature matrix. They claimed that Log Frequency Power Coefficients (LFPCs), which tap spectral strengths from 12 logarithmically spaced band pass filters in the 130 Hz to 16 kHz frequency range, perform better than MFCCs. An HMM based high computational multi-model approach based on music structure (Intro, Verse, Chorus, bridge, and Outro) was discussed for vocal detection. They proposed that signal strengths are different in intro, verse, etc., and can be used as in the multimodality scenario. Although the authors claimed using that their proposed approach worked better, we have no basis to believe that studio recorded music would always exhibit different signal strengths in their song structures. In addition, quarter note level resolution for signal segmentation has been found to be inadequate for vocal boundary detection (Maddage et al 2004 [75]).

Leung and Ngo (2004) [68] characterized music segments using 39 dimensional perceptual linear predictive coding (PCP) features and classified them into either the vocal class or the instrumental class using an SVM classifier. In the first step, PCP features were extracted from 500 ms signal frames and the generalized likelihood ratio (GLR) distance between adjacent frames was calculated. The local maximums were then detected on the frame-GLR distance plot and the authors claimed that the signal section between adjacent maximums belongs to either the vocal class or the instrumental class. These signal sections, known as signal segments, were further framed into smaller 16 ms sections and PCP feature extraction was repeated. However, the feature dimensions were reduced through the ICA-FX. Five pop songs were used for training the SVM and 25 songs were used for testing. Results were verified with manual annotations. A frame level accuracy of over 75\%was reported. However,
boundary detection is inefficient because the frame level resolution or error rate will be in the multiples of 500 ms .

### 3.4 Music similarity detection

We have discussed earlier in this chapter the previous methods that have been used for analyzing information in the different layers (beats, melody, harmony, chords and music regions) of the music structure pyramid (Figure 2-1). Based on similarities between information in the different layers, we can group the music content into different similarity regions. For example, some similarity regions may consist of similar vocals and some regions can be grouped based on similar harmonies. Techniques have been proposed in the literature for music similarity analysis.

Previous works on music structure analysis have focused on feature-based similarity matching for the detection of repeated patterns in music. Dennenberg and Hu (2002) [25] proposed using chroma based and autocorrelation based techniques to detect the melody line in music. Repeated segments in the music were identified using Euclidean distance similarity matching and clustering of the music segments. Goto (2003) [49] and Bartsch and Wakefield (2001) [7] used pitch sensitive chroma-based features to detect repeated sections (i.e. chorus) in the music. Foote et al. (2002) [38] constructed a similarity matrix and Cooper and Foote (2002) [22] defined a global similarity function based on extracted MFCCs to find the most salient sections in the music. Logan and Chu (2000) [69] used clustering and hidden Markov models (HMM) to detect key phrases in the choruses. Lu and Zhang (2003) [70] estimated the most repetitive segments of the music clip using MFCC and octave-based on the spectral contrast. Pikrakis et al. (2003) [92] used a context dependent dynamic time
warping algorithm to find the music patterns in a monophonic environment. Although the authors claimed an accuracy of over $95 \%$ in pattern recognition of Greek traditional music, it is not clear what ground truths they used for recognition. Xu et al. (2005) [140] used an adaptive clustering method based on the features, linear prediction coefficients (LPC) and MFCCs, to create a music summary. Chai and Vercoe (2003) [19] characterized music with pitch, spectral, and chroma based features and then analyzed recurrent structures to generate a music thumbnail. Gao et al. (2004) [43] proposed a rhythm based clustering technique to detect certain kinds of music structures. They used their unsupervised beat detection technique [42] for onset based segmentation. The music frames were then characterized with 12 dimensional MFCCs and clustered into similar music groups. The authors haven't given any justification as to how they have assessed their music grouping or what kind of structural information they were interested in their algorithm.

### 3.5 Discussion

Overall, music structure analysis can be considered the extraction of information in the different layers of the music structure pyramid. Previous methods for music structure analysis commonly followed fixed length signal segmentation, feature extraction, and statistical modelling techniques. These general steps are similar to the steps being followed in speech signal processing. However, the important argument is how well these speech processing techniques are suitable for music signal processing, because speech and music differ significantly from each other in both production and perception. Unlike speech signals, music signals' sources are heterogeneous in nature and change over time.

The main question that we address in this thesis is how well conventional fixed length signal segmentation is suited for music content analysis. From the music composition point of view, the transition of melody, harmony, music regions and semantic music regions are proportional to inter-beat time intervals. We argue that information in the music signal change in discrete inter-beat interval steps. Thus, information within inter-beat intervals can be considered to be stationary rather than fixed with regards to signal frame length. The proposed beat proportional signal segmentation for layerwise music structure analysis, which incorporates music domain knowledge, is the major contribution of this thesis.

The other question that we aim to answer in the thesis is how well signal characterization techniques from speech processing are suitable for music content analysis tasks such as the detection of music regions and the identification of harmony / melody contours. In the chapter 6, we experimentally prove that the incorporation of music domain knowledge can increase the efficiency of signal processing for music content analysis. The rest of the chapters in this thesis are organized as follows. Chapter 4 discusses the proposed rhythm based segmentation technique for music segmentation. Then, chord detection and error correction methods for creating the harmony line of the music are discussed. The steps for detecting music regions, music similarity regions and song structure are detailed in chapter 5. Experimental results are explained in chapter 6, and proposed music applications based on music structure analysis are discussed in chapter 7 . We conclude the thesis in chapter 8 .

## 4 <br> Music Segmentation and Harmony Line Creation via Chord Detection

The fundamental step in audio content analysis is signal segmentation, where information within the segment, can be considered fairly quasi-stationary. Feature extraction and statistical learning followed by music segmentation are essential steps in music content analysis tasks such as music region detection and hamony contour extraction (via chord detection). High accuracy in the above steps would lead to better performance in music applicaions such as music transcription, music information retrieval, lyrics identification, polyphonic music transcription, and music summarization, as discussed in chapter 7 .

The determination of segment size, which is needed to extract certain levels of information, requires a better understanding of the rates of information flow in the audio data. Over three decades of speech research has revealed that $20 \sim 40 \mathrm{~ms}$ of fixed length signal segmentation is appropriate for speech content analysis (Rabiner et al 1993 [94]). The composition of music (see chapter 2) reveals that the rate of information (notes, chords, key, vocal phrases, instruments) flow is proportional to inter-beat ${ }^{6}$ time intervals. Figure 4-1 shows the quarter, eighth and sixteenth note boundaries in a song clip. We see that the fluctuations of signal properties in both spectral domain and time domain are aligned with note boundaries. Usually, smaller notes (Quaver, Semiquaver and Demisemiquaver) or rests are played in the bars to

[^4]align the harmony contours with the rhythmic flow of the lyrics and fill the gaps between lyrics.


Figure 4-1: Spectral and time domain visualization of ( $0 \sim 3667$ ) ms long clip played in " 25 Minutes" by MLTR. Quarter note length is 736.28 ms and note boundaries are highlighted using dotted lines.

The conceptual model of music structure shown in Figure 2-1 emphasizes that information in different layers is constructed by different music notes. Figure 4-2 illustrates the different notes played in the $6^{\text {th }}, 7^{\text {th }}$ and $8^{\text {th }}$ bars of 3 tracks. We see that only quarter notes are played in these 3 bars of the bass guitar track. The note resolution increases to eighth note and sixteenth note levels in the electric organ and rhythm guitar tracks, respectively. When these tracks are mixed together, the time domain signal representation is shown at the bottom of the Figure 4-2.

As seen in Figure 4-1, the fluctuation of temporal properties in the time domain and the spectral domain is proportional to the inter-beat interval. This behaviour tells us
that the ideal segmentation for music information extraction (melody/harmony contour extraction, vocal/instrumental region detection, etc) is to segment the music according to the length of the individual music note. However, after mixing all the tracks together, it is extremely difficult to detect individual note boundaries in the final track using existing onset detection techniques (see time domain signal representation in Figure 4-2 ).


Figure 4-2: Notes played in the $6^{\text {th }}, 7^{\text {th }}$, and $8^{\text {th }}$ bars of the rhythm guitar, bass guitar, and electric organ tracks of the song "Whose Bed Have Your Boots Been Under" by Shania Twain. Notes in the electric organ track are aligned with the vocal phrases. Blue solid lines mark the boundaries of the bars and red solid lines mark quarter note boundaries. Grey dotted lines within the quarter notes mark eighth and sixteenth note boundaries. Some quarter note regions which have smaller notes are shaded with pink colour ellipses.

As detailed in Figure 2-2, longer notes (Semibreve, Minim, Crotchet) are integer multiples of the smaller notes (Quaver, Semiquaver, Demisemiquaver). Therefore, the smallest note length segmentation (for the song in Figure 4-2, it is a sixteenth note level segmentation) is a better solution over segmenting the music according to the individual notes.

Our proposed inter-beat time proportional segmentation (at the smallest note level) is called Beat Space Segmentation (BSS). Reasons why BSS is used instead of fixed length segmentation, which has been widely used in the speech processing over 3 decades, are summarized below.

## From the music composition point of view:

- A careful analysis of the song structure reveals that the time lengths of the music regions (PV, IMV, \& PI) are proportional to the inter-beat time interval of the music (Chapter 2).
- Harmony /melody changes in the music occur at note boundaries. This is more likely to occur at the half note (meter is $4 / 4$ ) in popular music (Goto 2001 [48]).


## From the signal processing point of view:

- Based on above 2 points, we can consider the information within the beat space segment to be quasi-stationary.

Inter-beat level signal analysis and correction was initially discussed in the correction algorithm for music sequence (CAMS) in Maddage et al 2003 [72]. To carry out BSS, it is necessary to compute the smallest note length that appears in a song. The smallest note length detection via time information extraction is discussed in the next section. Section 4.2 discusses the windowing effect on music signals. After the music is segmented into the smallest note-length frames, silent (rest) frames are identified in section 4.3. This chapter concludes with section 4.4, which discusses chord detection for harmony contour creation.

### 4.1 Music segmentation

Time information (meter, tempo) is the foundation layer of the music structure pyramid (see Figure 2-1). Meter and tempo illustrate the beat structure, i.e. how many different music notes per bar and duration of the note (detailed in chapter 2.1). The literature survey in chapter 3.1 reveals that many research efforts have been carried out to find meter and tempo. Yet, this task is a challenging problem due to the polyphonic nature of the music signals. The ultimate goal of this thesis is to combine the layered information in the music structure pyramid towards understanding the semantic meaning(s) of the music. To achieve this, the accuracy of the time information extraction task should be high. To maintain a high accuracy level in the analysis and reduce algorithm complexity, we narrow down the scope of the time information analysis by applying the below constraints

- Meter is assumed to be $4 / 4$. This is the most frequently used meter of popular songs.
- Meter is constant throughout the song.
- Tempo is constrained between 30-190 BPM. Since meter is $4 / 4$, tempo implies the number of quarter notes per minute (see Figure 2-12).

Since the meter is assumed to be $4 / 4$, the main task is to find the length of the quarter note and the smallest note (eighth note, sixteenth note or thirty-second note) in a song. In our survey, we found that most popular songs have the eighth note or the sixteenth note as the smallest note. Therefore, the $3^{\text {rd }}$ constrain implies that the minimum duration of the sixteenth note is roughly 79 ms .

Our proposed approach for smallest note length detection is shown in Figure 4-3. Since the music harmonic structures are in octaves (Rossing et al. 2001 [99], chapter 3.2), we decompose the signal into 8 sub-bands using the discrete wavelet transform technique. The frequency ranges of the sub-bands are shown in Table 2-1. Wavelets analyze the signal with multi-frequency resolutions and are more capable of describing discontinuities and sharp spikes in the signal than single resolution traditional Fourier transform technique. Therefore, wavelets are a good choice for analysing beats and onsets, which describes the starting of different events in the music signals. Previous research by Tzanetakis et al. (2001) [124] has also discussed the strengths of wavelets for rhythm analysis (formulation of beat histograms) of music signals. We selected $5^{\text {th }}$ order the Daubechies wavelet for decomposing music signals because of its better performance of reducing correlated information in the decomposed signals (Rao and Bopardikar 1998 [97]).


Figure 4-3: Rhythm tracking and extraction

The sub-band signals are segmented into 40 ms with $50 \%$ of overlap. This frame size is smaller than the minimum length of the sixteenth note within the $30-190 \mathrm{BPM}$ tempo range. Both the frequency and the energy transients are then analyzed using a method similar to the one described in Duxburg et al (2002) [34]. Duxburge used
quadrature filter bank whose frequency ranges are $\{(0 \sim 1.1)$, (1.1~2.2), (2.2~5.5), (5.5~11.0) and (11~22) \} KHz, whereas we use octave scale filter bank for decomposition of the music signals. Then exponential weighting function is used as the detection function, which emphasises both frequency and time domain energy transitions. We measure the frequency transients in terms of the progressive distances between the spectrums in sub-band 01 to 04 because the fundamental frequencies (F0s) and harmonics of the music notes in popular music are strong in these subbands. Equation (4-1) explains the progressive distance calculation between the $(n+1)^{\text {th }}$ frame and the $(n)^{\text {th }}$ frame where $f_{(n+l)}$ is the vector representation of frequency strengths in the $(n+1)^{\text {th }}$ frame.

$$
P D(n+1)=\left|f_{(n+1)}-f_{(n)}\right| \quad \begin{gather*}
\text { where } P D(n+1) \text { is the Progressive distance }  \tag{4-1}\\
\text { between the }(\mathrm{n}+1)^{\text {th }} \text { frame and the } \mathrm{n}
\end{gather*}
$$

The progressive distance calculated from the frequency components in the frames of sub-band 01 to 04 indicate the chord changes in the music. The point where the distance is high indicates a chord change. We apply the following chord knowledge (Goto 2001 [48]) towards estimating the length of the bar or the half note.

- Chords are more likely to change on beat times than at other positions.
- Chords are more likely to change on half note times than at the other positions of beat times.
- Chords are more likely to change at the beginning of the measures (bars) than at the other positions of half note times.

The above chord knowledge reveals that the progressive distance is very high at the start of the bar and the start of the half note. To estimate the length of the bar, we normalized the progressive distance calculated in sub-band 01 to 04 . Then, we filter out the normalized distances below $80 \%$ of each sub-band. We measure the gap
between two adjacent progressive distance values in the sub-bands. These gaps indicate the bar length or the half note length, which is the half of the bar measure. We estimate the length of the bar by taking the average of the computed bar lengths. After estimating the bar length we continue to detect sub-band onset using similar method described in Duxburg et al (2002) [34].

The progression of bass clef notes, which are mainly generated from bass guitars, is highlighted in sub-bands 01 and 02 . The onsets, which are computed on these two sub-bands, describe mostly the half note level or the bar level rhythm transition. The onsets of the quarter, eighth, or sixteenth notes appear mainly in sub-bands 03 and 04 . The energy transients, computed from sub-bands 05 to 08 , are more sensitive to quarter note, half note, and full note (or bar) level rhythm progressions. In order to consider different note level rhythm progressions, we take the weighted summation of the onsets, detected in each sub-band as described in Equation (4-2). The value $\operatorname{On}(n)$ is the weighted sum of the sub-band onsets detected in all eight sub-bands $S b_{i}(n)$ at frame no -n- of the music signal. In our experiments, we noticed that the onsets of bass drums, bass guitars and the bass notes of pianos are found in sub-bands 01 and 02. The timing of snares and side drums are highlighted in sub-bands 06 to 08 . The weight matrix $\mathrm{w}=\{0.6,0.9,0.7,0.9,0.7,0.5,0.8,0.6\}$ is empirically found to be the best set for calculating the final onset $O n($.$) in the music.$

$$
\begin{equation*}
O n(n)=\sum_{i=1}^{8} w(i) S b_{i}(n) \tag{4-2}
\end{equation*}
$$

The positions of local maxima are computed by differentiating the auto correlated signal $O n($.$) . Since we already estimated the length of the bar from the progressive$ distance in sub-bands 01 to 04 , we can further estimate the length of either the
quarter, eighth, or sixteenth note by measuring the distance between adjacent local minima. Then, we employ dynamic programming to check for patterns of equally spaced strong and weak beats from among the detected dominant onsets Ont(.), and compute both the inter-beat length and the smallest note length. Figure 4-4 (a) illustrates a 10 second song clip. The detected onsets are shown in Figure 4-4 (b). The autocorrelation of the detected onsets is shown in Figure 4-4 (c). Both the eighth note level segmentation and the bar measure are shown in Figure 4-4 (d) and computed eighth note length is 183.106065 ms .


Figure 4-4: Beat space segmentation of a 10 second clip

### 4.2 Windowing effect on music signals

The song is first segmented into the smallest note length frames. Then, different features are extracted from the signal section to find both the chord and the music region (vocal/instrumental) that the signal segment belongs to. In speech processing, the sliding window technique is used for signal segmentation. The $10-20 \mathrm{~ms}$ length Hamming window and rectangular window with 10 kHz sampling frequency are
commonly used for extracting speech information such as voice/un-voice regions, vowels, consonants, etc (Rabiner et al 1993 [94]). However, the question is how to find the suitable window for music signal analysis. Figure 4-5 shows the time domain and frequency domain characteristics of Hamming windows and rectangular windows. Figure 4-5 (a) and (d) respectively show time domain 10000 points length Hamming and rectangular windows, which are generated at a 44.1 kHz sampling frequency. Their frequency domain characteristics are shown in Figure $4-5$ (b) and (e), respectively. Figure 4-5 (c) and (f) show the frequency domain characteristics of when the window lengths are 100 points each.

Both rectangular and Hamming windows act like low-pass filters. The Hamming window (Figure $4-5$ (c)) has nearly twice the passband cuttoff $(560 \mathrm{~Hz}$ ) than of rectangular window passband cuttoff ( 388 Hz ) (Figure 4-5 (f)). However, the bandwidth of the window decreases directly with an increase in the window length. This can be seen in Figure 4-5 (b) and (e). When the number of points increases to 10000 from 100 (Figure 4-5 (c) and (f)) i.e. increases $100^{\text {th }}$ fold, the bandwidth decreases $100^{\text {th }}$ fold. In terms of stopband characteristics, the Hamming window has sharper cutoff than the rectangular window. The Hamming window has nearly 30 dB more attenuation than the rectangular window (Figure 4-5 (b) \& (e), and Figure 4-5 (c) \& (f)).

Unlike speech signals, music signals are wide band signals (Everest 2001 [37], Rossing et al [99]). Thus, music information is spread widely over 15 kHz in the frequency spectrum, whereas speech information is spread below 5 kHz in the frequency spectrum. The bandwidths of both the windows are significantly small for
music signals. Compared to the rectangular window, the Hamming window has a sharp attenuation and suppresses valuable information in the higher frequencies by nearly 3 fold over the rectangular window. From that point of view, the rectangular window is better than the Hamming window for music signal analysis. However, when we take the short time Fourier transform, the rectangular window frequency response comes to the picture due to the properties of short time Fourier transform, which is neither avoidable nor as bad as the Hamming window effect. Thus, the rectangular window is considered in all our feature extraction processes.


Figure 4-5: The frequency responses of Hamming and rectangular windows.

Beat space segmentation is utilized to find the harmony contour and music regions (pure vocal-PI, pure instrumental-PI, instrumental mixed vocal-IMV, and silence-S). Silence frames (regions) are detected during pre-processing in order to avoid further
analysis of those frames as they carry no valuable information. Silence detection is explained in the next section. Harmony contour detection via chord detection is explained in section 4.4. Chapter 5 discusses the music region detection.

### 4.3 Silence detection

After music is segmented into beat spaced segments, we need to detect silent segments and remove them from the music sequence.

Clip of the song name Duo Hao (217.096~220.293)ms


Figure 4-6: Silence region in a song

Silence is defined as a segment of imperceptible music, including unnoticeable noise and very short clicks. A silence region in a song is shown in Figure 4-6. We use the short-time energy feature to detect silence. If the short-time energy function is continuously lower than a certain set of thresholds the segment is indexed as silence. Note that there may be durations in which the energy is higher than the threshold, but the durations should be short enough and far apart from each other. Equation (4-3) describes how to calculate short time energy in the beat space segmented frame, where $x(m)$ is the discrete time music signal, $n$ is the time index of the short-time energy, and $w(m)$ is a rectangle window. $N$ is the length of the window and is the length of beat space segment.
$E_{n}=\frac{1}{N} \sum_{m}[x(m) w(n-m)]^{2} \quad$ where $\quad w(n)=\left\{\begin{array}{cc}1, & 0 \leq n \leq N-1 \\ 0, & \text { Otherwise }\end{array}\right.$

### 4.4 Harmony Line Creation via Chord Detection

The concept of music regions floating on the harmony/melody flow can be visualized as boats sailing on the sea waves, as shown in Figure 4-7. Detection procedures of both harmony and music regions are independent from each other, even though they share the same music segments (see Figure 1-2). Chapter 5.1 discusses the music region detection.


Figure 4-7: Concept of sailing music regions on harmony and melody flow
Playing music notes according to time steps described by both meter and tempo forms complex structures such as chords, melodies, and large music units. Harmony, in music is created by playing several music notes simultaneously. When more than 2 notes are played together, a chord is created. Thus, chord identification of a given music piece can stamp the harmonic line with chord symbols. Chord composition and chord notations are described in chapter 2. Detection of a harmony line in the music is useful for many music applications such as query by humming for music information retrieval systems, music transcription, music synthesis, error concealment, watermarking, etc. The harmony line created by a chord contour is depicted in Figure 4-8. The smallest blocks/units shown in the music sheet denote the quarter note
length, since the meter of the music is $4 / 4$. Figure $4-8$ shows the notes played on the bass clef and the treble clef by a guitar and a piano. Music is composed such that all the notes played at a particular time belong to a chord. The bottom of the figure shows the chords flow. Notes played on the first bar belong to the F major ( F ) chord. In the twelfth bar, the first half note duration belongs to the F major chord and the second half note duration belongs to the A diminished (Adim) chord.


Figure 4-8: Both bass line and treble line created by a bass guitar and a piano .The chord sequence, which is generated using notes played on both the bass and treble clefs, is shown at the bottom of the figure.

This section discusses the incorporation of acoustic signal processing techniques into music knowledge for chord detection and error correction. Different techniques for characterizing the polyphonic music pitch are discussed in the following sub sections. Then, statistical modelling methods are investigated to classify these polyphonic pitches into chords. Finally, chord detection errors are corrected via Key determination of the music piece. The chord detection steps are shown in Figure 4-9. The first step is the beat space segmentation of the music signal. Thereafter, features are extracted from the smallest note length signal frames (BSS frames) to represent
the polyphonic music pitches embedded in the signal frames. In section 4.4.1.1, we discuss two polyphonic pitch representation techniques. Statistical learning followed by the polyphonic pitch representation enables us to detect the chords in the music. Only four popular chord types (Major, Minor, Augmented, and Diminished) are considered in this thesis. Therefore, there are only twelve chords per each chord type to be modelled resulting 48 models. Statistical learning techniques for modelling chords are detailed in section 4.4.2.


Figure 4-9: Chord detection steps

### 4.4.1 Polyphonic music pitch representation

Detecting the fundamental frequencies (F0s) of notes, which comprise the chord, is the main idea towards identifying the chord. Many previous chord detection systems utilized a pitch class profile (PCP) representation of music signals (Krumhansl 1979 [66], Bharucha and Stoeckig 1986 [12] \& 1987 [13] and see chapter 3.2). In the following sub-section, we describe the procedure for PCP feature derivation. Experimental results of the analysis of PCP feature behaviour are discussed in chapter 6.2.

Psychological experiments conducted in 1970s and 1980s have revealed that human ear is sensitive to not only the fundamental frequency (F0) of pitches but also several harmonics and sub-harmonics (Goldstein 1973 [45] and Terhardt 1974 [121] \& 1982 [122]). However, there is no adequate information in the literature discussing how well these psychological pitch representations characterize the polyphonic music signals. The functionality of the psychological polyphonic pitch representation is also discussed in this section, and its capabilities for pitch representation are experimentally analyzed in chapter 6.2 .

### 4.4.1.1 Pitch class approach to polyphonic music pitch representation

In musicology, the octave is divided into 12 tones, and the frequency ratio of two consecutive octaves is assumed to be 1:2 (see chapter 2 and chapter 3.2). As shown in Figure 4-10 (upper), the pitch class profile ( PCP ) feature is derived by projecting all the fundamental frequencies (F0s) of the notes in each octave to 12 pitches. Therefore, this pitch class representation is called the 12 pitch class profile feature. In a such projection, only the effects of the fundamental frequencies of the music notes are accounted for, and harmonic or sub-harmonic effects are inactive. In order to
construct the 12 Pitch Class Profile (PCP) feature vector (12 components), the music signal is first transformed into the frequency domain. Secondly, the squared strengths of the fundamental frequencies of the music note in different octaves (see Table 2-1) are summed together to compute the coefficient of the 12 PCP feature vector, which is depicted in Figure 4-10 (lower).


Strengths of all the note frequencies are mapped in to 12 pitches


Figure 4-10: Music notes in different octaves are mapped into 12 pitches

The polyphonic music contains the signals of different music notes played at lower and higher octaves. Some instruments, for example string instruments have a strong $3^{\text {rd }}$ harmonic component (Rossing et al 2001 [99]) which nearly overlaps with the $8^{\text {th }}$ semitone of the next higher octave. This is problematic in lower octaves and leads to wrong chord detection. For example, the $3^{\text {rd }}$ harmonic of the note C 3 and the F 0 of note G4 nearly overlap (see Table 2-1). To overcome such situations, our implementation begins by transforming the music frames into the frequency domain using FFT with a 2 Hz frequency resolution (i.e. [sampling frequency (Fs) / number of

FFT points $(\mathrm{N})] \approx 2 \mathrm{~Hz}$ ). Secondly, the value of C in equation (4-4), which maps linear frequencies into the octave scale, is set to 1200 , where the pitch of each semitone is represented with as high a resolution as 100 cents. This frequency mapping is shown in Figure 4-10 and Figure 5-7. We consider only the 128 to 8192 Hz frequency range (sub-bands 02 to 07 in Table 2-1) for constructing the PCP feature vectors in order to avoid adding percussion noise, i.e. base drums in frequencies below 128 Hz and cymbal and snare drums in frequencies over 8192 Hz , to these features. By setting $F_{\text {ref }}$ to 128 Hz , the lower frequencies can be eliminated. The initial 1200-dimensional $P C P_{I N T}($.$) vector is constructed based on equation (4-5), where X($. is the normalized linear frequency profile, computed from the beat space segment using FFT.

$$
\begin{align*}
& f_{\text {octave }}=\left[C * \log _{2}\left(\frac{F s * f_{\text {linear }}}{N * F_{\text {ref }}}\right) \bmod C\right]  \tag{4-4}\\
& P C P_{\text {INT }}\left(f_{\text {octave }}\right)=\sum_{f_{\text {linear }} * f_{\text {octave }}}\left|X\left(f_{\text {linear }}\right)\right| \quad f_{\text {octave }}=1,2,3 \ldots . .1200 \tag{4-5}
\end{align*}
$$

In order to obtain an optimal balance between computational complexity and efficiency, the original 1200 dimensions of the PCP feature vector is reduced to 60 using equation (4-6). Thus, each semitone is represented by summing 100 cents into 5 bins according to equation (4-6). In equation (4-6), $p$ denotes the $p^{\text {th }}$ coefficient of the 60 dimensional vector. Finally, each segmented music frame is represented by a 60 coefficient feature vector.
$P C P(p)=\sum_{i=[20(P-1)+1]}^{200 * P} P C P_{I N T}(i) \quad p=\begin{gathered}1,2,3 \cdots 60 \\ i \leq 1200\end{gathered}$

### 4.4.1.2 Psycho-acoustical approach to polyphonic music pitch

## representation

Pitch class representation only characterizes the effects of the F0 of music pitches and maps all the harmonic and sub-harmonic effects of polyphonic pitches into the same space of F0s of the 12 music tones. Earlier experiments on pitch perception reveal that the central processing mechanism of the human auditory system responds to not only the fundamental frequencies of the pitch but also its harmonics and subharmonics (Goldstein 1973 [45], Houtgast 1976 [52], Laden and Keefe 1989 [67], Terhardt 1974 [121] \& 1982 [122] and Rossing et al 2001 [99]). Thus, Goldstein modelled the pitch with F0 plus its harmonic partials, and Terhardt modelled the same pitch with up to the $7^{\text {th }}$ sub-harmonic. Sub-harmonic derivation is described in equation (4-7).

$$
\begin{equation*}
N^{t h} \text { Sub-harmonic }=\frac{\text { Fundermental frequency }(F 0)}{N} \text { where } N \text { is an integer } \tag{4-7}
\end{equation*}
$$

Figure 4-11 shows the harmonics and sub harmonics of the C Major Chord, which are approximated up to the nearest music notes.

Harmonic and sub harmonic progression of C Major Chord in two octaves. Frequencies of the harmonics and sub-harmonics are represented in terms of closest music notes.


Figure 4-11: Harmonic and sub-harmonics of C Major Chord is visualized in terms of closest music note.

Overlapping harmonics and sub-harmonics are marked by a grey background. The F0s of the music notes are noted in Table 2-1. We incorporate this psycho-acoustical effect of music perception i.e. harmonic and sub-harmonic strengths, towards characterizing the polyphonic pitch of the music. Another reason for harmonic complex analysis is that singing voice and most music instruments (wind type - flute and mouth organ, bow type - violin and cello, string type piano and guitar, and some percussion instruments) produce harmonic spectra. Therefore, the effect of these harmonic spectra of complex tones ${ }^{7}$ is useful for identifying music chords. Figure shows the harmonic spectra of pure female vocal, mouth organ, and piano music. In order to take the harmonics and sub-harmonics into account, the beat space segmented music frames are first transformed into the frequency domain with a nearly 2 Hz frequency resolution.


Figure 4-12: Spectral visualization Female vocal, Mouth organ and Piano music

[^5]As shown in Figure 4-11, harmonics and sub-harmonics can be approximated to the F0 of a note in another octave. Thus, the strengths of the F0s of music notes in the 128 to 8192 Hz frequency range (i.e. sub-band 02 to 07 in Table 2-1) are mapped into the feature vector in order to represent the psycho-acoustic effect of the polyphonic pitch. Equation (4-8) describes the psycho-acoustic feature vector for the $i^{t h}$ music segment. This feature is called a Psycho-acoustic profile (PAP) feature.

$$
\begin{equation*}
V_{i}=\langle\mathrm{F} 0 \mathrm{~s} \text { of the notes }(\mathrm{C} 3 \sim \mathrm{~B} 8)\rangle \ldots . .72 \text { dimention feature vector } \tag{4-8}
\end{equation*}
$$

### 4.4.2 Statistical learning for chord modelling

Different statistical learning techniques have been surveyed for modelling music chords. Figure 4-13 shows the steps for detecting the chord of the $i^{\text {th }}$ beat space segment. In this thesis, four types of chords are considered, namely, Major, Minor, Augmented, and Diminished. Since each chord type has 12 different chords, we have a total of 48 statistical models in our chord detection system. In Figure 4-13, $M_{j}$ denotes the $j^{\text {th }}$ chord statistical model, which is learned from the training data of the $j^{t h}$ chord.

In this thesis, we employ three statistical learning techniques, Support Vector Machine (SVM), Gaussian Mixture Model (GMM), and Hidden Markov Model (HMM), to model these 48 chords. These techniques are discussed in the following sub-sections. The experimental results discussed in chapter 6.2 shows that HMM is more capable of modelling music chords than SVM or GMM. After the detection of music chords, we apply knowledge of the "Key of the music" to correct the detected error chord. The error correction procedure is detailed the next sub-section.


Figure 4-13: Chord detection for the $i^{\text {th }}$ beat space signal segment

### 4.4.2.1 Support Vector Machine (SVM)

The mechanism of this principle learning method follows the structural risk minimization method which is rooted in VC (Vapnik-Chervonenkis) dimension theory (Vapnik 1998 [129]). There are two major differences between this popular back-propagation (BP) algorithm and the SVM learning algorithm. Firstly unlike BP, SVM operates only in the batch mode. Secondly, the BP algorithm minimizes a quadratic loss function regardless of the learning task. In contrast, the SVM algorithm for pattern recognition minimizes the number of training samples that fall inside the margin of separation between positive and negative samples. It is considered more appropriate than the mean-square error criterion used in the BP algorithm for classification tasks. SVM handles the quadratic programming problem. It is attractive because it guarantees finding the global maximum of the error surface. In terms of run time, SVMs are currently slower than other neural networks (e.g. MLP trained with BP algorithm) for similar generalization performance. The SVM algorithm can construct a variety of learning machines by use of different kernel functions. Three kinds of kernel functions are usually used. They are:
> Polynomial kernel of degree d

$$
\begin{equation*}
k(x, y)=(<x, y>+1)^{d} \tag{4-9}
\end{equation*}
$$

$>$ Radial basis function with Gaussian kernel of width $\mathrm{C}>0$

$$
\begin{equation*}
K(x, y)=\exp \left(-|x-y|^{2} / c\right) \tag{4-10}
\end{equation*}
$$

$>$ Neural networks with tanh activation function, gain $k$ and shift $\mu$

$$
\begin{equation*}
k(x, y)=\tanh (k<x, y>+\mu) \tag{4-11}
\end{equation*}
$$

Where x is input vector and y is input pattern.

Our earlier experimental results have shown that equation (4-10) performed better for music signal classification (Maddage et al 2003 [72] \& 2004 [75], Xu et al 2003 [139]). In the implementation, we used the SVM Torch II package (Collobert and Bengio 2001 [21]).

### 4.4.2.2 Gaussian Mixture Model (GMM)

In recent years, GMM has been used for speaker identification (Reynolds and Rose 1995 [95]) and singer identification (Kim and Brian 2002 [62]). The multiple weighted Gaussians, which attempt to model each class of training data, are beneficial when analyzing data that has a distribution not well modelled by a single cluster. Based on the calculated distances between test points and the multiple Gaussians distributed in the class distributions, the maximum likelihood discriminant function, classifies the test point to the closest class. The Expectation-Maximization (EM) algorithm is used to estimate the parameters (mixture weights, mean, and variance) of the GMMs. Since GMM can be considered a one state HMM, we use the continuous density Hidden Markov Model (CDHMM) in the HTK package (Young et al 2002 [142]) in the implementation.

### 4.4.2.3 Hidden Markov Model (HMM)

The HMM is also a powerful statistical tool which has been widely used in speech recognizing systems (Rabiner and Juang 1993 [94]). It is a finite state network with $N$ states, with each state modelled by a probability density function. Figure 4-14 illustrates a HMM topology with five states, where the first is an entry state, and the last is an exit state. Unlike these two states, each middle state has a probability distribution $b_{j}$ where, in speech recognition, mixture Gaussian densities are chosen. For vocal/instrumental region detection and harmony detection, the music data distribution is also assumed to be a mixture of Gaussians. The mixture weights, mean, and covariance of the Gaussian mixtures in each state and both initial and state transition probabilities are estimated using either the Baum-Welch or the segmental $K$-means algorithm.


Figure 4-14: The HMM Topology

### 4.4.3 Detected chords' error correction via Key determination

The pitch difference between the notes of the two chord pairs Major and Augmented chord) and Minor chord and Diminished is small. In our experiments, we sometimes find that the observed final state probabilities of HMMs corresponding to these chord pairs are high and close to each other. This may lead to a wrong chord detection. Thus, we apply a rule-based method (key determination) to correct these wrongly
detected chords. Then, we apply heuristic rules based on popular music composition to correct the chord transition.

The key is defined by a set of chords, as detailed in chapter 2.2. Songwriters sometimes use relative major and minor key combinations in different sections, such as a minor key for the Middle eighth and a major key for the rest, in order to break up the perceptual monotony of the song (Shenoy et al 2004 [109]). However, songs with multiple keys are rare [120]. Therefore, a 16-bar length window is run over the detected chords to determine the key of that section. A key is assigned to that section according to the corresponding to a majority of the chords. The 16-bar length window is sufficient to identify the key (Shenoy et al 2004 [109]). If Middle eighth (chapter 2.4 ) is present, we can estimate the region where it appears in the song by detecting the key change. Once the key is determined, the error chord is corrected as follows:
> First, we normalize the observations of the 48 HMMs representing 48 chords according to the highest probability observed from the error chord.
$>$ The next highest observed chord whose observation is above a certain threshold (THchord) and belongs to the same key replaces the error chord.
$>$ We replace the error chord with the previous chord if there is no observation which is above the THchord and which belongs to the chords of same key.

The value THchord $=0.64$ is empirically found to be good for correcting chords. The music signal is assumed to be quasi stationary between the inter-beat intervals, because the melody transition occurs on beat time. Thus, we apply the following chord knowledge (Goto 2001. [48]) to correct the chord transition within the window.

- Chords are more likely to change on beat times than at other positions.
- Chords are more likely to change on half note times than at other beat times.
- Chords are more likely to change at the beginning of the measures (bars) than at other positions of half note times.

The above three points are explained in Figure 4-15. In Figure 4-15, the size of the beat space segment is the eighth note and the size of the half note is four beat space segments. After the correction, bar i has two chord transitions and a chord in bar i+1.


Figure 4-15: Correction of chord transition

## 5 Music Region and Music Similarity detection

This chapter details the procedures for extracting the information in layers 3 and 4 of the music structure pyramid. Section 5.1 discusses different features which can be used for music region detection via statistical learning. We incorporate music knowledge to design better features with to characterize the signal sections in the music regions.

Section 5.2 is focused on detecting two types of similarity regions in the music, namely, melody-based similarity regions and content-based similarity regions. These similarity regions help us to formulate a procedure to detect the components of popular song structures (i.e. Intro, Verse, Chorus, Bridge, Middle eighth, INST, and Outro). This procedure is explained in section 5.3.

### 5.1 Music region detection

Based on the music source mixtures (accompaniment), an acoustical music signal can be divided into different regions. Figure $5-1$ shows the possible regions in the music, Pure vocals (PV), Instrumental Mixed Vocals (IMV), Pure Instrumental (PI), and Silence (S). Silence regions are identified in the pre-processing detailed in chapter 4.2.


Figure 5-1: Regions in the music

Music region detection can simplify the process of music content analysis because it helps to identify the signals sections, which have different music sources. The points discussed below highlight the necessity of accurate region detection in achieving better performance in music applications.

- The information in the vocal line (PV and IMV regions) is important for building a content-based music retrieval system and an automatic lyrics generator.
- PI and IMV should be carefully analyzed in order to extract the music note information in a music transcription system. This information can be about the type of the instrument, the name of the instrument which generates the note, the position of the note in the music score and the note characteristics i.e. attack, sustain, decay times.
- For music summarization, repetitive portions in a song should be accurately identified in order to avoid their repetition in the summary. The similarity analysis between the music regions assists in identifying semantic clusters in the music. For example, choruses in a song usually have similar vocal and instrumental regions (Maddage et al 2004 [75]).
- For singer identification, the first step is to detect the vocal regions in the music.
- Signal complexity varies between different music regions. Therefore, customized signal restoration techniques can be applied to these music regions in order to conceal errors in music streaming (Maddage et al 2004 [75]).
- Compared to the properties of singing voice, properties of instrumental music are spread across a wide spectrum. Music region detection is helpful in designing a robust music watermarking scheme. Watermarks can be embedded in different ways for vocal and instrumental signal sections so that the signal properties of these regions are not distorted.
- Removal of the vocal line from music is useful for applications such as Karaoke music generation.

The above application scenarios motivate the need for identifying music regions. These applications are discussed in chapter 7. It is found in our survey (chapter 2.5) that on an average, singing voice occupies $65 \%$ of the duration of a popular song. These vocals can be pure vocals or instrumental mixed vocals. Both IMV and PI are the most frequently occurring regions in popular music. Earlier research efforts reveal that vocal region detection is a difficult task in comparison with the detection of other regions (Gao et al 2003 [41], Maddage et al 2003 [72]). The vocal cord is the oldest music instrument. Both the human auditory physiology and the human perceptual apparatus have evolved to develop a high level of sensitivity to the human voice. After more than three decades of extensive research speech recognition technology has matured to the level of practical applications. However, speech processing techniques have limitations when they are applied to singing voice analysis because
speech and singing voice differ significantly in terms of their production as well as their perception by humans (Sundberg 1987 [118], Miller 1986 [84]). Singing voice has more dynamic characteristics than speech (Saitou et al 2002 [101]). The dynamic range of the fundamental frequency (F0) contours in singing voice is wider than that of speech. The F0 fluctuation of singing voice is larger and more rapid than that of in speech (Everest 2001 [37], Sundberg 1987 [118]).

We believe that a combination of bottom-up and top-down approaches, which combine the strength of low-level features and high-level music knowledge, can provide us a powerful tool towards improving our music analysis. In our recent work (Maddage et al 2004. [73], [74], [75]), we have incorporated music knowledge for both segmentation and feature extraction for music content characterization. We investigated the performances of speech processing techniques on the task for music region detection, and found that it does not work well. Therefore, we have developed a novel approach, one that considers both signal processing and music knowledge, to detect music regions.

The block diagram of the proposed method is shown in Figure 5-2. After the music piece is segmented into the smallest note length frames, the silent frames are then identified. Music segmentation and silence region detection have been discussed in chapter 4.1 and 4.2, respectively.


Figure 5-2: The steps for vocal instrumental region detection

### 5.1.1 Applying music knowledge for feature extraction

Feature selection is very important for music content analysis. In this section, we analyze both time and frequency domain features, i.e. the cepstral coefficient feature and the linear prediction coefficient feature. These features are traditionally used in speech processing because of their excellent ability to model speech in terms of both production and perception. Though these features have been tested for speech signals, there has not been sufficient analysis to determine to what degree they can be applied to music content analysis.

### 5.1.1.1 Cepstral Coefficients

The cepstral coefficients have been widely used for detecting the pitch and formant structures of voice/speech (John et al 1999 [58]). Equation (5-1) describes the
calculation of the $n^{\text {th }}$ cepstral coefficient $\{C(n)\}$, where $S(\omega)$ is the strength of the spectrum of the signal at angular frequency $\omega$.

$$
\begin{equation*}
C(n)=\frac{1}{2 \pi} \int_{-\pi}^{\pi} \log |S(\omega)| e^{j \omega n} d \omega \tag{5-1}
\end{equation*}
$$

The basic steps for calculating the cepstral coefficients are shown in Figure 5-3. The signal is first segmented using a sliding window (L points). For music content analysis, we assume a non-overlapping rectangular window whose size (L points) is equal to the beat space segment (chapter 4.2). The signal properties are assumed to be stationary inside the window and zero signal strength is assumed to be outside the window (Rabiner at al [94], John and Dimitris 1995 [57]).


Figure 5-3: Steps for calculating cepstral coefficients

Second, the linearly scaled spectrum is computed using a Discrete Fourier transform (DFT). In order to highlight the low varying frequency characteristics of the signal, the strengths of the frequencies are scaled using their log magnitudes. Cepstral coefficients are sensitive to the sources, which are multiplied together to generate a signal. The Log magnitude of the DFT of a signal reflects the sum of the transformation of the individual signals. Cepstral coefficients can identify these individual signal characteristics if they are highlighted at different places of the spectrum (Moorer (1975) [85]). Therefore, a frequency scaling block is used for emphasising the important areas of the spectrum. Mel-scale, which arrange the frequency spectrum of the signal according to the perception mechanism of the human ear, has popularly been used in speech processing (see next sub-section a). We
propose an octave scale for music processing because, as discussed in chapter 2 and chapter 3.2, music properties fluctuate in an octave scale.

Finally, we take the inverse DFT (IDFT) of the spectrum to calculate the cepstral coefficients. The DFT and IDFT are computed according to equation (5-2) and (5-3), where $N$ is the frequency resolution. To avoid time domain aliasing in the IDFT, the value of $N$ is always chosen to be $N \geq L$ ( $L$ is the window size). The value $\mathrm{x}(n)$ is the $\mathrm{n}^{\text {th }}$ sample value of the signal, which is within the window.

$$
\begin{align*}
& X(k)=\sum_{n=0}^{N-1} x(n) e^{-\frac{j 2 \pi k n}{N}} \quad k=0,1,2 \ldots \ldots . . N-1  \tag{5-2}\\
& x(n)=\frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{\frac{j 2 \pi k n}{N}} \quad n=0,1,2 \ldots \ldots . N-1 \tag{5-3}
\end{align*}
$$

In the following two sub-sections, we discuss two frequency scaling techniques, i.e. Mel scale and Octave scale, for deriving the cepstral coefficients. Mel scale has popularly been used among the speech community because it measures the perceived pitches and their harmonics in the spectrum. The mechanism behind Mel scale based cepstral coefficients is detailed in section (a). Then we propose Octave scale based cepstrum coefficients in section (b). Brown (1991) [15] has also acknowledged that octave scaling has advantages for highlighting music properties in the frequency domain.

## (a) Mel-Frequency Cepstral Coefficients (MFCCs)

The Mel-frequency cepstral coefficient has been found to be highly effective in recognizing speech signals and modelling the subjective pitch and harmonic content of audio signals (John et al 1999 [58]). Figure 5-4 illustrates the relationship between
the Mel scale and the linear frequency scale. The positions of the frequencies in the linear frequency scale are transformed into the Mel scale using the equation (5-4), where $C$ is a scaling factor which controls the slope of the curve (John et al 1999 [58]).

$$
\begin{equation*}
F_{\text {mel }}=\frac{C \log _{10}\left(1+\frac{F_{\text {linear }}}{C}\right)}{\log _{10}(2)} \tag{5-4}
\end{equation*}
$$

When C is $100,300,500$ and 700, Figure $5-4$ shows the respective values in the Mel scale for $f_{\text {linear }}=3000$. The non-linear filter positions in the linear frequency scale ( $f_{\text {linear }}$ ) are calculated by transforming the equally spaced filter positions in the Mel scale using equation (5-4). It can be visualized that when C is increased, the number of filters placed in the lower frequencies, 200 Hz to 1200 Hz of the linear frequency scale $\left(f_{\text {linear }}\right)$ is lower (compare $\mathrm{C}=700$ with $\mathrm{C}=100$ ). Thus, tuning C in equation (5-4) is essential to positioning triangular filters accurately on the linear frequency axis to extract dominant pitches and their harmonics in the spectrum. The value of C in the range of 250 to 350 has been empirically found to be efficient for detecting vocal and instrumental signals.

After filter positions in the linear frequency scale are computed, the full band spectrum is represented piece-wise according to the triangular filter regions. As shown in Figure 5-4 (bottom), $Y(i)$ represents the strength of the $i^{\text {th }}$ triangular filter region and is calculated using equation (5-5). The values $S(j)$ and $H_{i}(j)$ are magnitude of the $j^{\text {th }}$ frequency component of the spectrum and the $j^{t h}$ output of the $i^{\text {th }}$ triangular filter, respectively. The $i^{\text {th }}$ filter boundaries in the linear frequency scale and Mel scale are $\left\{m_{i}, n_{i}\right\}$ and $\left\{\underline{m}_{i}, \underline{n}_{i}\right\}$, respectively. The centre frequency of the triangular filter
region is considered to be the frequency representative for the particular region. The value $f_{\text {linear }}=k_{i}$ is the centre frequency of the $i^{\text {th }}$ filter region.

$$
\begin{align*}
& Y(i)=\sum_{j=m_{i}}^{n_{i}} \log _{10}[S(j)] H_{i}(j)  \tag{5-5}\\
& C(n)=\frac{2}{N} \sum_{i=1}^{N_{c b}} Y(i) \cos \left(k_{i} \frac{2 \pi}{N} n\right) \tag{5-6}
\end{align*}
$$

The Mel scale cepstral coefficients are calculated according to equation (5-6), where $\mathrm{N}, \mathrm{N}_{\mathrm{cb}}$, and n are the number of frequency sample points, number of band-pass filters, and number of cepstral coefficients, respectively. In contrast to equation (5-3), equation (5-6) omits the phase information in the spectral transformation.


Figure 5-4: The filter distribution in both Mel scale and linear scale.

## (b) Octave Scale Cepstral Coefficients (OSCC)

The locations of the fundamental frequencies (F0s) and the harmonics of the music notes are closely related with the octave scale. The pitch contour of the professional vocal line is kept as close as possible to the instrumental line. Therefore, vocal F0s and harmonics also fluctuate in octave. Figure 5-5 shows the frequency domain representation of quarter note length music signals and fixed length speech signal. Figure 5-5 (a) and (b) depict spectral envelopes scaled in octaves. However, such octave scale spectral envelopes cannot be noticed in the spectrum of the speech signal. This is because the vocal code, which generates the speech, is not tuned for the octave scale. Ideal octave scale spectral envelopes which encapsulate the music signal spectrum are drawn in Figure 5-5 (d). This octal behaviour of music signals is experimentally discussed in chapter 6 .


Figure 5-5: Music and speech signal characteristics in frequency domain. (a) Quarter note length (662ms) instrumental (Guitar) mixed vocal (male) music, (b) Quarter note length ( 662 ms ) instrumental (Mouth organ) music, (c) - Fixed length (600ms) speech signal, (d) - Ideal octave scale spectral envelopes.

Because of these noticeable octave scale spectral envelopes, we can assume that the fluctuation of the spectral properties of the music signal is in octave scale. Therefore,
we propose using an "Octave Scale" instead of the Mel scale. In octave scaling, filters are positioned on octaves so that they can capture the spectral properties, which fluctuate on octaves. The cepstral coefficient calculated on the octave scale is called the Octave Scale Cepstral Coefficient (OSCC).

In our approach, filters are positioned on all of the octaves in the 0 to 16.4 kHz frequency range where 44.1 kHz audio sampling frequency is assumed. The frequency range of 0 to 128 Hz , which comprises the first two octaves $\mathrm{C} 0 \sim \mathrm{~B} 0$ and $\mathrm{C} 1 \sim \mathrm{~B} 1$, is mostly dominated by percussion instruments (e.g. Bass drums). The spectral properties of the percussion instruments can be considered unstructured ${ }^{8}$, and spectral behaviours are not closely related to the octave scale (Rossing 2001[98]). Figure 5-6 illustrates the log magnitude spectrum of the bass drum and the side drum. It can be seen that these instruments concentrate a higher energy in the spectrum at the $0 \sim 128 \mathrm{~Hz}$ frequency range, and their spectrums are weak in harmonic structures. Thus, twelve $50 \%$ overlapping filters are linearly placed on the first two octaves i.e. C0~B0 to C1~B1 (0~128 Hz).


Figure 5-6: Log magnitude spectrums of bass drum and side drum
The octave scaling is used to find the filter positions within the rest of the octaves (C2~B2, to C9~B9 - eight octaves). The equation (5-7) describes how to transform

[^6]linear frequencies $f_{\text {linear }}$ to the octave scale $f_{\text {octave }}$, where $F s$ and $N$ are the sampling frequency and the number of FFT points, respectively. The variable $F_{\text {ref }}$ refers to the starting point of the linear to octave frequency mapping and C is the number of equally spaced points in an octave.
\[

$$
\begin{equation*}
f_{\text {octvae }}=\left[C * \log _{2}\left(\frac{F_{s} * f_{\text {linear }}}{N * F_{\text {ref }}}\right) \bmod C\right] \tag{5-7}
\end{equation*}
$$

\]

Compared with Mel-scale, which is approximately in $\log _{10}($.$) (See equation (5-4)),$ Octave Scale can be approximated in $\log _{2}($.$) . In the equation (5-7), F_{\text {ref }}$ is set to 128 Hz , because the filter positions in C2~B2 and higher octaves (i.e. $128 \sim 22050 \mathrm{~Hz}$ ) are calculated using the octave scale.


Figure 5-7: The filter band distribution in Octave scale for calculating cepstral coefficients

In western music, twelve pitches are considered to be within an octave (see chapter 2).
These pitches are noted in Table 2-1. To capture the effects of the pitches and their harmonics in the spectrum, 12 filters are placed in each octave. By setting $\mathrm{C}=12$ in equation (5-7), the upper limit of $f_{\text {octave }}$ is set to 12 and Figure 5-7 depicts the linear positioning of 12 filters in the octave scale and their positions' transformation to
linear frequency scale $f_{\text {linear }}$ in the $\mathrm{C} 4 \sim \mathrm{~B} 4$ octave. A Total of 96 filters accommodate the full spectrum $(0 \sim 16.4 \mathrm{kHz})$ and they are detailed in Table 5-1.

Table 5-1: Filter distribution for computing Octave Scale Cesptral Coefficients

| Sub-band No | S 1 | S 2 | S 3 | S 4 | S 5 | S 6 | S 7 | S 8 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Frequency <br> range (Hz) | $0 \sim 128$ | $128 \sim 256$ | $256 \sim 512$ | $512 \sim 1024$ | $1024 \sim 2048$ | $2048 \sim 4096$ | $4096 \sim 8192$ | $8192 \sim 16384$ |
| Octaves | $\mathrm{C} 0 \sim \mathrm{~B} 0$ | $\mathrm{C} 1 \sim \mathrm{~B} 1$ | $\mathrm{C} 2 \sim \mathrm{~B} 2$ | $\mathrm{C} 3 \sim \mathrm{~B} 3$ | $\mathrm{C} 4 \sim \mathrm{~B} 4$ | $\mathrm{C} 5 \sim \mathrm{~B} 5$ | $\mathrm{C} 6 \sim \mathrm{~B} 6$ | $\mathrm{C} 7 \sim \mathrm{~B} 7$ |
| Number of <br> filters | (linearly spaced with $50 \%$ <br> overlap) | 12 | 12 | 12 | 12 | $\mathrm{C} 8 \sim \mathrm{~B} 8$ |  |  |

Singular value decomposition (SVD) can be used to measure the level of uncorrelation among the coefficients derived from the Octave Scale and the Mel Scale. Appendix-A discusses the relationship between principle component analysis and SVD. Singular values in the diagonal matrix reflect both the noise and uncorrelated coefficients when diagonal values are low and high respectively. Figure 5-8 shows the normalized singular value variation of 20 OSCCs and 20 MFCCs extracted from both vocal and instrumental regions.


Figure 5-8: Plot of the 20 Singular values, which are computed from OSCCs and MFCCs for vocal and instrumental music frame.

The frame size is equal to the duration of the beat space segment. It can be seen that the normalized singular values of OSCCs are higher than that of MFCCs for both vocal and instrumental frame. Figure 5-9 describes the different averages of singular values for OSCCs and MFCCs. The average normalized singular values per OSCC for both vocal and instrumental frames are 0.1294 and 0.1325 . However, for MFCC, they are much lower, at 0.1181 and 0.1093 , respectively. As shown in the Figure $5-9$, the singular values are in descending order with respect to the ascending coefficient numbers. Lower singular values at the higher coefficient numbers describe the higher noise level or the lower uncorrelation level. The average of the last 10 singular values of OSCCs is nearly $10 \%$ higher than they are for MFCCs, which means that the last 10 OSCCs are more uncorrelated than the last 10 coefficients of MFCCs. Therefore, we can conclude that OSCCs are more uncorrelated than MFCCs.


Figure 5-9: Average of singular values

### 5.1.1.2 Linear Prediction Coefficients (LPCs)

The Linear Prediction Coefficients (LPCs), which approximate the speech parameters such as vocal tract area functions, pitch, and formants, have been widely used for speech content representation in low bandwidth. The details of calculating LPCs are well documented in the literature( John et al (1999) [58], Rabiner and Juang (1993) [94]). Since music is created by mixing different sets of sources/instruments with
respect to the time, music signals can be considered more complex than speech signals. Thus, finding out how well the LPCs can represent different layers of music information is still considered open research. In this section, we briefly discuss the frequency domain interpretation of linear prediction analysis and propose selective band linear predictive analysis (Makhoul 1975 [78]) to approximate the wide band music signal. The cepstral coefficients, which are derived from LPCs, may also be useful for music information representation and classification. The next paragraphs explain briefly the LPCs' interpretation in both frequency and time domains.

$$
\begin{equation*}
H(z)=\frac{S(Z)}{U(Z)}=G \times \frac{1}{1-\sum_{k=1}^{p} a_{k} Z^{-k}} \tag{5-8}
\end{equation*}
$$

Equation (5-8) describes a time varying all pole (the $p^{t h}$ order) modelled system $\{H(z)\}$ which approximates the characteristics of speech signals. For voice and unvoiced output signals, the system $\{\mathrm{S}(\mathrm{z})\}$ is excited with an input $\{\mathrm{U}(\mathrm{z})\}$ which consists of an impulse train and random noise respectively. The value $G$ and $\left\{a_{k}\right\}$ are the gain parameter and the digital filter coefficients, respectively, and vary slowly with time. With the above speech model, we write the $n$-th speech sample $\mathbf{s}(n)$ related with the excitation $u(n)$ in the equation (5-9).

$$
\begin{equation*}
S(n)=\sum_{k=1}^{p} a_{k} S(n-k)+G u(n) \tag{5-9}
\end{equation*}
$$

The LPCs are the approximate filter coefficients $\left\{b_{k} k=1 \ldots p\right\}$ of the original speech model coefficients $\left\{a_{k} k=1 \ldots . p\right\}$, and these coefficients are derived from the given speech samples. Then, the predicted $n^{\text {th }}$ speech sample $\widetilde{S}(n)$ from previous samples can be written with the calculated filter coefficients $\left\{b_{k}\right\}$ as according to equation (5-10). Equation (5-11) describes the $n^{t h}$ prediction error $e(n)$ between $S(n)$ and $\widetilde{S}(n)$.

It can be seen from Equation (5-11) that the error sequence is the output of a system whose transfer function $A(z)$ is noted in equation (5-12). Thus, $\mathrm{A}(\mathrm{z})$ can be considered the transfer function of the prediction error filter. If $a_{k}=b_{k}, A(z)$ is the inverse filter of the speech model $H(z)$.

$$
\begin{align*}
& \widetilde{S}(n)=\sum_{k=1}^{p} b_{k} S(n-k)  \tag{5-10}\\
& e(n)=S(n)-\widetilde{S}(n)=S(n)-\sum_{k=1}^{p} b_{k} S(n-k)  \tag{5-11}\\
& A(z)=1-\sum_{k=1}^{p} b_{k} Z^{-k} \quad \text { if } b_{k}=a_{k} \text { then } A(z)=\frac{G}{H(z)}(\text { ref. Eqn } 5-8) \tag{5-12}
\end{align*}
$$

The time domain mean-squared signal prediction error (obtain using autocorrelation method) of the $m^{\text {th }}$ speech segment is represented in equation (5-13). By applying Parseval's theorem, it can be represented in the frequency domain according to equation (5-14). $S_{m}\left(e^{j \omega}\right)$ is the Fourier transformation of the $m^{t h}$ speech segment $S_{m}($. and $A\left(e^{j \omega}\right)$ is the frequency response of the prediction error filter.

$$
\begin{align*}
& E_{m}=\sum_{n} e_{m}^{2}(n)  \tag{5-13}\\
& E_{m}=\frac{1}{2 \pi} \int_{\pi}^{\pi}\left|S_{m}\left(e^{j \omega}\right)\right|^{2}\left|A\left(e^{j \omega}\right)\right|^{2} d \omega \tag{5-14}
\end{align*}
$$

With a good estimation LPC (i.e. $b_{k} \approx a_{k}$ ) and connecting equation (5-12) with equation (5-14), we can represent the time domain mean-squared signal prediction error using the equation (5-15). Thus, by minimizing the integral of the ratio of the energy spectrum of the speech segment over the magnitude squared of the frequency response of the speech model, we minimize $E_{m}$.

$$
\begin{equation*}
E_{m}=\frac{G^{2}}{2 \pi} \int_{-\pi}^{\pi} \frac{\left|S_{m}\left(e^{j \omega}\right)\right|^{2}}{\left|H\left(e^{j \omega}\right)\right|^{2}} d \omega \tag{5-15}
\end{equation*}
$$

When the order of the filter in the speech system is large, (i.e. P is large) the expression given in equation (5-16) is valid. This implies that the power spectrum of the speech signal segment can be approximated using the all-pole model $H(z)$. As discussed in Rabiner and Juang 1993 [94], it has been suggested that the order $P$ of the linear prediction coefficients would control the degree of smoothness of the resulting spectrum.

$$
\begin{equation*}
\lim _{p \rightarrow \infty}\left|H\left(e^{j \omega}\right)\right|^{2}=\left|S_{m}\left(e^{j \omega}\right)\right|^{2} \tag{5-16}
\end{equation*}
$$

## Octave influence in LPC:

Compared to speech, music is a wide band signal (Rossing et al 2001 [99], Sundberg 1987 [118] ) spanning over a 10 kHz frequency range. The useful range of fundamental frequencies and the harmonics of tones produced by music instruments is considerably smaller than the audible frequency range. Therefore, it is important to conduct piecewise spectral analysis for the full spectrum ( $0 \sim 20$ ) kHz instead of taking full spectrum as one. Thus, we apply selective linear prediction (Makhoul 1975 [78]) to model regions of the spectrum with different order $(P)$ of the prediction.


Figure 5-10: Computation of selective band linear predictive coefficients (LPCs)

As shown in Figure 5-10, we first filter the signal according to the octaves and then the segment sub-band signals into beat space segments. The frequency ranges of the sub-bands are detailed in Table 5-2. Finally, we calculate LPCs with different orders for each sub-band beat spaced signal segment. Since these LPCs are calculated in octaves, they are called Octave Scale Linear Prediction Coefficients (OSLPCs). Table 5-2 details the design parameters of the elliptic filter bank (IIR type) which is used for decomposing the signal in octave scale. It is found that the elliptic filter characteristics give sharp frequency cutoffs with lower filter orders.

Table 5-2: Parameters of the Elliptic filter bank used for sub-band signal decomposition in octave scale

| Elliptic filter (IIR) specifications for octave scale sub-band decomposition |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Sub-band no | 01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 |
| Frequency range (Hz) | $0 \sim 128$ | $128 \sim 256$ | $256 \sim 512$ | $512 \sim 1024$ | $1024 \sim 2048$ | $2048 \sim 4096$ | $4096 \sim 8192$ | $8192 \sim 22050$ |
| Octave | C0-B0 \& C1-B1 | C2-B2 | C3-B3 | C4-B4 | C5-B5 | C6-B6 | C7-B7 | C8-B8 |
| Filter type | Low pass | Band pass | Band pass | Band pass | Band pass | Band pass | Band pass | High pass |
| Filter order | 6 | 3 | 4 | 5 | 6 | 7 | 8 | 18 |
| Pass band ripples (dB) | 0.5 | 0.5 | 0.5 | 0.5 | 0.6 | 0.6 | 0.5 | 0.5 |
| Stopband attenuation (dB) | 150 | 100 | 80 | 100 | 100 | 100 | 100 | 100 |
| Cutoff frequency (Hz) | 128 | $\{128,256\}$ | $\{256,512\}$ | $\{512,1024\}$ | $\{1024,2048\}$ | $\{2048,4096\}$ | $\{4096,8192\}$ | $\{8192\}$ |
| Stopband corner frequency <br> limits(Hz) | 145.3 | $\{107.7,290.7\}$ | $\{242.2,532.9\}$ | $\{430.7,1195\}$ | $\{942.1,2218\}$ | $\{1960,4264\}$ | $\{3994,8360\}$ | $\{8156\}$ |

Figure 5-11 depicts an example of the selective-band power spectrum approximation using the all pole model $H(z)$. For this analysis, we used a quarter note length instrumental frame of the song " 25 Minutes" by MLTR. First, we passed the song through an octave scale filter-bank. Filter specifications are discussed in Table 5-2. Then, we estimate OSLPCs using the same quarter note length instrumental frame of each sub-band signal. For sub-band 01 and sub-band 02 , we estimated $20^{\text {th }}$ order all pole models. For all the other sub-bands, $60^{\text {th }}$ order all pole models are estimated. These two orders for the modes are randomly selected to show their modelling effectiveness of sub-band spectrum against the full band spectrum. Red coloured dash lines show the approximation of the power spectrums from sub-band signal.

For a comparison, we plotted the $60^{\text {th }}$ order all pole model approximation of the full band power spectrum (Figure 5-11 bottom). It can be seen that the all-pole model doesn't effectively approximate the full band power spectrum when compared with the all-pole model's sub-band power spectrum approximation. Therefore, OSLPCs, which emphasize the octave scale sub-band signal information, can better represent the wide band music signals.


Figure 5-11: Selective-band power spectrum approximation using all pole speech model $H(z)$.

### 5.1.1.3 Linear Predictive Cepstral Coefficients (LPCC)

Linear predictive cepstral coefficients are derived from the filter coefficients of the all pole speech production system $H(z)$ assuming an impulse response to the system.

From equation (5-16) we can approximate the frequency response of the system $H(z)$ to the spectrum of the signal $S(\omega)$, and the calculation of cepstral coefficients is similar to equation (5-1). The impulse response of the all-pole system $H(z)$ described in equation (5-8) implies the output of a voiced signal, which has a minimum phase. Thus, we can neglect phase information while calculating the cepstral coefficients. Equation (5-17) describes the computational process where $C(n), b_{n}$ and $p$ are the $n^{\text {th }}$ cepstral coefficient, $n^{\text {th }}$ LPC and order of the all-pole system, respectively. The impulse response, $h(n)$, of the all-pole system $H(z)$ can be solved recursively by using LPCs as in equation (5-18), where $\delta(n)$ is the $n^{t h}$ impulse to the system $\mathrm{H}(\mathrm{z})$.

$$
\begin{array}{ll}
C(n)=b_{n}+\sum_{k=1}^{p-1}\left(\frac{k}{n}\right) C(k) b_{n-k} & n \geq 1 \\
h(n)=\sum_{k=1}^{p} b_{k} h(n-k)+G \delta(n) & 0 \leq n \tag{5-18}
\end{array}
$$

Compared with the LPC, the LPCC is more robust against sudden signal changes or noise because LPCCs are derived from the filter coefficients of the speech system with the impulse response of the system $H(z)$. Previous experiments have shown that the LPCCs perform better in detecting vocal and instrumental regions in music (Gao et al 2003 [41], Maddage et al 2003 [72], Xu et al 2005 [140]).

## Octave influence in LPCC:

For a better characterization of wide band music signals, we calculate the selective band LPCC of the signal, following the steps discussed in Figure 5-10, where earlier computed sub-band LPCs are used for calculating LPCCs in the respective sub-band. Similar to OSLPCs, these coefficients can be called Octave Scale Linear Predictive Cepstral Coefficients (OSLPCCs).

Since OSCCs do approximate the spectrum, in chapter 6.3, we compare the performances of OSCC, MFCC, and OSLPCC to determine which features are better for the vocal/instrumental region detection problem

### 5.1.1.4 Harmonic Spacing measurement using Twice-Iterated Composite Fourier Transform Coefficients (TICFTC)

Singing voice and music instruments generate harmonics, which can be visualized as spikes in the frequency domain. Figure 5-12 shows the harmonic structures extracted from beat space signal segments of male vocal, female vocal, guitar, mouth organ, guitar mixed female vocal and guitar mixed male vocal. Compared with the vocal harmonic structures, instrumental harmonic structures are widely spread over in the entire frequency spectrum (Sundberg 1987 [118]). Musical instruments generate different harmonic structures because their sound production mechanisms differ from each other.

The careful analysis of harmonic and sub-harmonic spacing of singing voice reveals that they are narrowly spaced over the 512 to 2048 Hz frequency range. However, these spikes are widely spread for music instruments for the same frequency range. These frequency spacings are shown in Figure 5-14 (upper). The differences between harmonic and sub-harmonic spacing in singing voice and instrumental music can be used as a potential measure to identify vocal and instrumental music. Thus, we propose a mathematical model whose coefficients are sensitive to the frequency spacing of acoustic signals. These coefficients are known as Twice-Iterated Composite Fourier Transform Coefficients (TICFTCs).


Figure 5-12: Harmonic structures of vocal and instrumental signal segments The procedure to calculate TICFTCs is described as follow. First, beat spaced music segments are transformed into the frequency domain by applying a FFT algorithm with a high frequency resolution (nearly 1 Hz ). The phase information of the signal frame is discarded and only frequency magnitudes are considered for further analysis.

The magnitude spectrum (strengths of harmonics and sub-harmonics) can be modelled as a frequency pulse train $X(f)$ with periodicity $F_{p}$, pulse width $\tau$ and amplitude $A$. Figure 5-13 (middle) shows the $1^{\text {st }} \mathrm{FFT}$ of the $i^{\text {th }}$ beat space segment, where the spectrum is approximated by a pulse train. A second FFT now operates to produce a sine cardinal envelope as shown in Figure 5-13 (lower), with Sinc magnitude $Y_{i}$ as:

$$
\begin{equation*}
Y_{i}(k)=\left|\frac{A \tau}{F_{p}} \frac{\operatorname{Sin}\left(\pi F_{0} \tau k\right)}{\pi F_{0} \tau k}\right| \quad k=1 . .2 . .3 \ldots \ldots . \tag{5-19}
\end{equation*}
$$

Where $\alpha$ and $\beta$ are the height and the span of the first semi globe of $Y_{i}(k)$ respectively. $Y_{i}(k)$ is inversely proportional to $F_{p}$ where $F_{p}$ is the spacing between adjacent spectral lines. Thus, it can be seen that $F_{p}$ controls the intensity distribution of the sine cardinal envelope.


Figure 5-13: Twice - iterated composite Fourier transform of $i^{\text {th }}$ signal frame Samples of log magnitude spectrums (512 ~ 2048) Hz and Sinc envelopes on real music signals are shown in Figure 5-14.


Figure 5-14: The $1^{\text {st }} \& 2^{\text {nd }}$ FFT of instrumental and vocal frames. Frame size is a quarter note length ( 735 ms )

We can immediately see the structural difference between instruments and the vocal tract in that the vocal $F_{p}$ is much shorter than that of instruments, resulting in a bigger amplitude $\alpha$ in the TICFT domain. The Sinc envelope in Figure 5-14 (lower) also reveals that the energies of the frequency components are compressed in the lower values $k$ of $Y_{i}$.

The net effect of TICFT is to compress the original energy into the lower coefficients of the $2^{\text {nd }}$ FFT. The cumulative energy within these lower regions can be used to efficiently separate the vocal and instrumental frames. According to equation (5-20), by summing $Y_{i}(k)$ along the first $\gamma$ coefficients, we calculate the first energy bin of the $i^{\text {th }}$ frame bin $B_{l}(i)$ (see Figure 5-13).

$$
\begin{equation*}
B_{1}(i)=\sum_{k=1}^{\gamma} Y_{i}(k) \quad, \quad B_{2}(i)=\sum_{k=\gamma}^{\psi} Y_{i}(k) \quad, \quad B_{j}(i)=\sum_{k=\omega}^{\sigma} Y_{i}(k) \tag{5-20}
\end{equation*}
$$

Similarly, the $2^{\text {nd }}, 3^{\text {rd }}$ and $j^{\text {th }}$ energy bins, $B_{2}(i), \ldots B_{j}(i)$ are calculated. The values $\gamma$, $\psi, \omega$, and $\sigma$ are set manually based on experiments. Since $F_{p}$ controls the intensity distribution of the sine cardinal envelope, we take the ratios of the energy bins according to equation (5-21). The value $C_{j}$ where $\mathrm{j}=1 \ldots \mathrm{n}$ measures the harmonic and sub-harmonic spacing in the frequency spectrum. These coefficients $\mathrm{C}_{\mathrm{j}=1 . . \mathrm{n}}$ are called Twice-Iterated Composite Fourier Transform Coefficients (TICFTCs).

$$
\begin{equation*}
C_{j}(i)=\frac{B_{j}(i)}{B_{j+1}(i)} \quad \text { Eg. } \quad C_{1}(i)=\frac{B_{1}(i)}{B_{2}(i)} \tag{5-21}
\end{equation*}
$$

We conducted an experiment to evaluate the ability of the first energy bin $B_{l(.)}$ in discriminating vocal and instrumental frames. In this test we consider harmonics in the $(512 \sim 2048) \mathrm{Hz}$, i.e. sub-band 04, frequency range in Table 5-1. Then, we took the $2^{\text {nd }}$ FFT over the $\log$ magnitude spectrums in the $(512 \sim 2048) \mathrm{Hz}$ range and computed the energy first bin $B_{1}($.$) . We found empirically that \gamma=50$ yielded a good performance. The value $B_{l}($.$) is normalized by removing the mean of B_{l}($.$) over all n$ frames using equation (5-22), to yield $B_{I}{ }^{M R}$ (i) for the $i^{\text {th }}$ frame.

$$
\begin{equation*}
B_{1}^{M R}(i)=B_{1}(i)-\frac{1}{n} \sum_{i=1}^{n} B_{1}(i) \tag{5-22}
\end{equation*}
$$

In our experiments, we found that $B_{1}{ }^{M R}($.$) is positive for most of the vocal frames and$ negative for instrumental frames. Figure 5-15 shows the $B_{l}{ }^{M R}$ (.) plot for a full song "Sleeping Child" by MLTR. The regions covered with grey lines show the actual instrumental and vocal mixed instrumental regions. Mean removed first energy bin $B_{I}{ }^{M R}$ (.) can correctly classify the vocal/instrumental frames of the song with $83.27 \%$ accuracy.


Figure 5-15: The mean-removed bin $B_{1}($.$) with beat space ( 662 \mathrm{~ms}$ ) frames of "Sleeping child" by MLTR

Octave influence for calculating TICFTCs:
To accommodate harmonics and sub-harmonics in the octaves, we divide full frequency range into 8 sub-bands whose frequency ranges are noted in Table 2-1.

Figure 5-16 shows the procedure to calculate twice -iterated composite Fourier transform coefficients (TICFTCs) for each sub-band.


Figure 5-16: Twice -iterated composite Fourier transform coefficients
The optimum number of TICFTCs can represent the characteristics of the sub-band frequency spectrum (i.e. spacing between harmonics and sub-harmonics of the subband spectrum). Equation (5-23) describes the vector representation of the $i^{\text {th }}$ frame using TICFTCs, where $j$ and $s b$ are the number of coefficients calculated from each sub-band and the sub-band number, respectively.

$$
\begin{equation*}
V_{i}=\left\{C_{j}^{s b}\right\} \quad j=1,2, \ldots . n \text { and } s b=1,2 \ldots 8 \tag{5-23}
\end{equation*}
$$

### 5.1.2 Statistical learning for vocal I instrumental region detection

Since the number of pure vocal regions (PV) in popular music is few, we merge both PV and IMV together and call the merged parts vocal regions. As shown in Figure 5-17, it is a two class classification problem. Three statistical models, i.e. HMM, SVM, and GMM, are examined for vocal/instrumental region detection. These classifiers were explained in chapter 4.4.2.


Figure 5-17: Classification

### 5.2 Music similarity analysis

Music similarity analysis is important for the semantic level understanding of music structure and useful for many music applications such as automatic music summarisation, music streaming, and music retrieval (see chapter 7). Earlier researches as surveyed in chapter 3.4 mainly focused on feature based similarity analysis. These methods are not accurate in both detecting and interpreting the level of similarities we find in the music content, because music knowledge has not been effectively exploited. For example, feature based similarity detection methods are not accurate in detecting the boundaries of the similarity regions. Different types of similarities that can be seen in the music. They are:
$>$ Beat cycle - beat pattern that repeats in every bar
> Melody based similarity - repetition of the chord patterns
$>$ Vocal similarity - similar lyrics appearing in the song i.e. chorus
> Semantic level similarity - music pieces or extracts that create similar auditory scenes or sensation

Based on these different similarities we can index the music into different similarity regions. Thus, we can call melody-based similarity regions MBSRs. If the vocals within some of the MBSRs are similar, then those MBSRs are called content-based similarity regions (CBSRs). CBSRs, which have both similar chord pattern and similar vocals, are a subset of MBSRs. These two regions usually appear in popular music and Figure 5-18 shows them in the conceptual music pyramid.


Figure 5-18: Similarity regions in the music

Thus, detection of these two types of similarity regions is necessary for the popular song structure formulation discussed in the next section. The choruses in popular music have similar melodies and similar vocals, and are considered CBSRs. Though
the melody is similar in verse, the verses do not carry similar vocals. Thus, verses are the MBSRs. The following sub-sections describe the detection of these regions.

### 5.2.1 Melody-based similarity region detection

Melody-based similarity regions have the same chord patterns. In Figure 5-19, the regions $R_{2}, R_{3}, . . R_{i}$,., $R_{j}$ have the same chord pattern as $R_{1}$ (denoted as Destination Region). Therefore, $\mathrm{R}_{1}, \mathrm{R}_{2}, \mathrm{R}_{\mathrm{i}}, \ldots ., \mathrm{R}_{\mathrm{j}}$ are melody similarity regions.


Figure 5-19: Melody based similarity region detection by matching chord patterns Since we cannot detect all the chords without error, the region detection algorithm should have tolerance to errors. We employ Dynamic Programming for approximate string matching [86] as our melody-based similarity region detection algorithm.

Figure 5-20 illustrates the matching results of both 8 and 16 bar length chord patterns extracted from the beginning of Verse 1 in the song "Cloud No 9 " by Bryan Adams. The Y -axis denotes the normalized cost of matching the pattern and the X - axis represents the frame number. We set the threshold $\mathrm{TH}_{\text {cost }}$ and analyze the matching cost below the threshold to find the pattern matching points in the song. The 8 -bar length regions R2 to R8 have the same chord pattern as the first 8-bar chord pattern (R1-Destination Region) in Verse 1. When we extend the Destination Region to 16
bars, only the $r_{2}$ region has the same pattern as $r_{1}$, where $r_{2}$ is the first 16 bars from the beginning of Verse 2 in the song.


Figure 5-20: 8 and 16 bar length chord pattern matching results

### 5.2.2 Content-based similarity region detection

Content-based similarity regions are regions, with similar lyrics. More precisely, they are the chorus regions in the song. As shown in Figure 5-19, the melody-based similarity regions $R_{i}$ and $R_{j}$ can be analyzed further to detect whether they are contentbased similarity regions through the following steps.


Figure 5-21: Vocal similarity matching in the $i^{\text {th }}$ and $j^{\text {th }}$ MBSRs

Step1: The beat space segmented vocal frames of the two regions are first subsegmented into 30 ms with $50 \%$ overlapping sub-frames. Although the two choruses have both similar vocal content (lyrics) and melody, the vocal content may be mixed with a different set of instrumental setup. Therefore, to find the vocal similarity, it is important that the extracted features from the vocal content of the regions should be sensitive only to the lyrics and not to the instrumental line mixed with the lyrics.

Typically, when the coefficient order increases in the features, then the correlation between the coefficients increases or the coefficient sensitivity to information decreases. Figure 5-22 illustrates the variation of the $9^{\text {th }}$ coefficient of the OSCC, MFCC and LPC features for the three words 'clue number one', which are mixed with notes from the rhythm guitar. It can be seen that OSCC is more sensitive to the syllables in the lyrics than MFCC and LPC. Thus, we extract 20 coefficients of the OSCC feature per sub-frame to characterize the lyrics in the regions $R_{i}$ and $R_{j}$.


Figure 5-22: The response of the $9^{\text {th }}$ OSCC, MFCC and LPC to the Syllables of the three words 'clue number one'. The number of filters used in OSCC and MFCC are 64 each. The total number of coefficients calculated from each feature is 20 .

Figure 5-23 illustrates SVD analysis of the OSCCs and MFCCs extracted from both the solo male track and the guitar mixed male vocals of a Sri Lankan song "Ma Bala Kale (®) ఎ® ゅ๐゚)". The quarter note length is 662 ms and the sub-frame size is 30 ms with $50 \%$ overlap. Singular value variation of 20 OSCCs and 20 MFCCs for both pure vocals and vocal mixed with guitar are shown in Figure 5-23 (a), (b), (d), and (e) respectively.


Figure 5-23: Vocal sensitivity analysis of OSCCs and MFCCs using SVD.
Singular values indicate the variance of the corresponding structure. Comparatively high singular values describe the number of dimension, which can be represented orthogonally, while smaller singular values indicate the correlated information in the structure and are considered noise. The percentage in variation of the singular values of each OSCC and MFCC when guitar music is mixed with solo vocals are shown in Figure 5-23 (c) and (f) respectively. When all of the 20 coefficients are considered, the average singular value variation for OSCC and MFCC are $17.18 \%$ and $34.35 \%$, respectively. When the first 10 coefficients are considered, they are $18.16 \%$ and
$34.25 \%$, respectively. It can be concluded that even when guitar music is mixed with vocals, the variation of OSCCs is much lower than the variation of MFCCs. Thus compared with MFCCs, OSCCs are more sensitive to vocal line than to the instrumental music.

Step 2: The distances between feature vectors of $R_{i}$ and $R_{j}$ are computed. The equation (5-24) explains how the $\mathrm{k}^{\text {th }}$ distance $\operatorname{dist}(k)$ is computed between the $\mathrm{k}^{\text {th }}$ feature vectors $\mathrm{V}_{\mathrm{i}}$ and $\mathrm{V}_{\mathrm{j}}$ in the regions $\mathrm{R}_{\mathrm{i}}$ and $\mathrm{R}_{\mathrm{j}}$, respectively. The ' $n$ ' distances calculated from the region pair $\mathrm{R}_{\mathrm{i}}$ and $\mathrm{R}_{\mathrm{j}}$ are summed up and divided by ' $n$ ' to calculate the " dissimilarity $\left(R_{i} R_{j}\right)$ ", which gives a lower value for the content-based similarity region pairs as shown in equation (5-25).

$$
\begin{align*}
& \operatorname{dist}_{R_{i} R_{j}}(k)=\frac{\left|V_{i}(k)-V_{j}(k)\right|}{\left|V_{i}(k)\right| \cdot\left|V_{j}(k)\right|} \quad i \neq j  \tag{5-24}\\
& \operatorname{dissimilarity}\left(R_{i}, R_{j}\right)=\sum_{k=1}^{n} \frac{\operatorname{dist}_{R_{i} R_{j}}(k)}{n} \tag{5-25}
\end{align*}
$$

Step 3: To overcome the pattern matching errors due to detected error chords, we shift the regions back and forth in one bar step with the maximum size of the shift being 4 bars. Then, Steps $1 \& 2$ are repeated to find the positions of the regions, which give the minimum value for "dissimilarity ( $R_{i} R_{j}$ )" in equation (5-25).

Step 4: We compute "dissimilarity $\left(R_{i} R_{j}\right)$ " in all region pairs and normalize them. We set a threshold $\left(\mathrm{TH}_{\text {smlr }}\right)$ such that the region pairs below the $\mathrm{TH}_{\text {smlr }}$ are detected as content-based similarity regions, which also implies that they belong to chorus regions. Based on our experimental results $\mathrm{TH}_{\text {smlr }}=0.389$ gives a good performance. Figure 5-24 illustrates the calculated content-based similarity regions between
melody-based similarity region pairs, which are found in Figure 5-20 for the song "Cloud No 9" by Bryan Adams. It is obvious that the dissimilarity is very high between $R_{1}$, which is the first 8-bar length of Verse 1 , and other regions. Therefore, if $R_{1}$ is the first 8-bar region of Verse 1 , the similarity between $R_{1}$ and other regions is not compared in our algorithm.


Figure 5-24: The normalized content-based similarity measure between regions $\mathrm{R}_{1}$ through $\mathrm{R}_{8}$ computed from melody-based similarity regions of the song as shown in Figure 5-20 (Red dash line)

### 5.3 Song structure formulation with heuristic rules

The structure of a song is detected by applying heuristics which agree with most of the songs. Popular song structure follows the verse-chorus pattern repetition [120], as shown below.
(a) Intro, Verse-1, Chorus-1, Verse-2, Chorus-2, \{Verse, Chorus, Middle Eighth\}, Outro.
(b) Intro, Verse-1, Verse-2, Chorus-1, Verse-3,\{Verse, Chorus, Middle Eighth\}, Outro.

According to our survey of 220 songs (see chapter 2.5), most of the songs follow either structure (a) or (b) and they are followed by the parts in the brackets i.e. \{Verse, Chorus, Middle Eighth\}. For example, some songs may have five choruses with eight verses and middle eighth component may also appear in the song. Following constraints are considered for song structure analysis:

- The minimal number of verses that appears in a song is 2 .
- Verse and chorus are 8 or 16 bars long.
- All the verses in a song share a similar melody and all the choruses also share a similar melody. In some songs, the melody of the chorus may be partially or fully identical to the melody of the verse.
- In a song, the lyrics of all verses are quite different, but the lyrics of all the choruses are similar.
- The length of the middle eighth is 8 or 16 bars


### 5.3.1 Intro detection

Since Verse 1 starts at either the beginning of the bar or the second half note in the bar, we extract the instrumental section until the $1^{\text {st }}$ vocal frame of Verse 1 and designate that section as the Intro. If silent frames are present at the beginning of the song, they are not considered to be part of the Intro because they do not carry a melody.

### 5.3.2 Verses and Chorus detection

The end of the Intro is the beginning of Verse 1 . Thus, we can detect Verse 1 if we know whether it is of length 8 or 16 bars and then detect all the melody-based
similarity regions. Since the minimum length of the verse is 8 bars, we find the melody-based similarity regions (MBSR) based on the first 8-bar chord pattern of Verse 1 according to the method specified in section 5.2.1. We assume the 8 -bar MBSRs are $R_{1}, R_{2}, R_{3} \ldots . \mathrm{R}_{\mathrm{n}}$ in a song where $n$ is the number of MBSRs. Case $1 \& 2$ describe how to detect the boundaries of the verses and the choruses when the number of MBSRs ( n ) is smaller and equal to three and greater than three.

Case 1: $\mathrm{n} \leq 3$

The melodies of the verse and chorus are different in this case.

## Verse boundary detection:

To decide whether the length of the verse is 8 or 16 bars, we further detect the MBSRs based on the first 16-bar chord pattern extracted from the beginning of Verse 1. If the detected number of 16 -bar MBSRs is same as the earlier detected 8 -bar MBSRs (i.e. $n$ ), then the verse is 16 bars long. Otherwise, it is 8 -bars long.

## Chorus boundary detection:

Once the verse boundaries are detected, we check the gap between the last two verses. If the gap is more than 16 bars, the length of the chorus is 16 bars otherwise 8 bars. Since the chorus length is computed, we find the chorus regions in the song according to section 5.2.1.

The verse-chorus repetition patterns described in song structures (a) and (b) imply that Chorus 1 appears right before and right after Verse 2, respectively. A bridge may appear between Chorus 1 and the verse before it. Thus, we assume that Chorus 1 ends at the beginning of the Verse (i.e. in (a) this is Verse-2 and in (b) this is Verse-3) and
then the MBSRs are found based on the chord pattern of the approximated Chorus 1. In order to find the exact boundaries of the choruses, we use a content-based similarity measure (see section 5.2.2) between the detected chorus regions.

- We compute the dissimilarity of Chorus 1 and other estimated chorus regions based on step 1, 2, and 3 in section 5.2.2. We sum all the dissimilarities as Sum_dissm (0) where 0 is the zero shift.
- We shift the chorus backward by one bar and re-compute Sum_dissm (-1B), where-1B is 1-bar backward shift.
- We repeat the process of shifting and computing of Sum_dissm () until Chorus 1 comes to the end of the $2^{\text {nd }}$ to last verse.
- The position of Chorus 1 which gives the minimum value for Sum_dissm () defines the exact chorus boundaries.

Case 2: $n>3$,
The melodies of the chorus and verse are partially or fully identical in this case. It can be seen from Figure 5-20 that there are 8 MBSRs detected with the 8 -bar length verse chord pattern.

- First, we compare content-based similarities among all the regions except $\mathrm{R}_{1}$ based on step 1, 2, 3 and 4 in section 5.2.2. The region pairs, which have dissimilarities lower than $\mathrm{TH}_{\text {smlr }}$, are the 8-bar length chorus sections (see equation (5-25)).
- If the gap between $R_{1}$ and $R_{2}$ is more than 8 bars, the verse is 16 bars. Then based on the 16 -bar long chord pattern of Verse 1 we find the other verse regions.
- If a found verse region overlaps with an earlier detected 8 -bar chorus region, the verse region is not considered to be a verse.
- Once the verse regions are found, we can detect the chorus boundaries in the same way as was described in Case 1.


### 5.3.3 Instrumental sections (INST) detection

The Instrumental section may have a melody similar to the chorus or verse. Therefore, the melody-based similarity regions, which have only instrumental music, are detected as INSTs. However, some INSTs contain a different melody. In that case, we run a window of 4 bars to find regions which have INSTs.

### 5.3.4 Middle eighth and Bridge detection

The middle eighth is 8 or 16 bars long and its key is different key from the song's main key. Chapter 4.4.2 discusses the key determination of the song where a 16 bar length window is run over the detected chord to find the key. If a key different from the main key of the song is detected at any point, we further check to see whether the changed key area is of 16 bar length or 8 bar length.

Once the boundaries of verses, choruses, INSTs, and middle eighths are defined, the appearance of a bridge can be found by checking the gaps between these regions.

### 5.3.5 Outro detection

Before the Outro, there is usually a chorus in the song. Thus, we detect the Outro based on the length between the end of the final chorus and the end of the song.

## 6 <br> Experimental Results

This chapter presents the results of our analysis and the system level performance of the techniques used for information extraction in the layers of the music structure pyramid (Figure 2-1). All experimental data are limited to songs with $4 / 4$ meter (see chapter 4.1), which is the commonly used meter of popular music (Goto 2001[48]). Therefore, time information extraction (the bottom layer of the pyramid) only focused on identifying the smallest note length of a song. Section 6.1 discusses the performance of the proposed system for time information extraction. The analysis results of the proposed musically and perceptually modified features, which characterise the information in layer 2 and layer 3 of the pyramid, (i.e. information about harmony/melody contours and vocal/instrumental regions), are discussed in section 6.2 and 6.3, respectively. The effectiveness of song composition level heuristic rules for finding semantic clusters (i.e. Intro, Verse, Chorus, Bridge and Outro) in a song is discussed in section 6.4.

All the experimental data are sampled at 16 bits 44.1 kHz stereo format and are converted to mono to reduce the high computational power.

### 6.1 Smallest note length calculation and silent segment detection

The steps for calculating the smallest note length have been discussed in chapter 4.1. We use 120 songs, which consist of 10 songs by each artist. They are detailed in

Table 6-2. To compute the average length of the smallest note, is seen in a song, we test the first 30,60 , and 120 seconds of the song and then compute the average length of the smallest note seen in a song. The smallest note lengths of 106 songs were correctly detected with $\pm 30 \mathrm{~ms}$ error margin, i.e. 89.16 \% accuracy. A 30 ms error margin is set because the window size (frame size) in our rhythm tracking system is 60 ms with $50 \%$ overlap. Figure $6-1$ shows the actual and computed $16^{\text {th }}$ note lengths. Listening tests with the help of music scores sheets are carried out to compute the actual $16^{\text {th }}$ note lengths of the songs.


Figure 6-1: Actual and computed $16^{\text {th }}$ note lengths of songs
We set the frame size to be equal to the smallest note length and then segment the music. As discussed in chapter 4.1, this segmentation is called Beat Space Segmentation (BSS). The frames with normalized short time energies below a threshold ( $\mathrm{THs}=0.18$ ) are detected as silence frames

### 6.2 Chord detection for creating harmony contour

As discussed in chapter 4.4, we use statistical learning techniques for modelling 48 music chords using the training data. The first step is to optimize the parameters of both the polyphonic pitch representation features (pitch class profile PCP feature and
psycho acoustic profile PAP feature) and the statistical models (SVM, HMM, and GMM). All the parameters are optimized in a synthetic environment because it is difficult to collect large annotated datasets from the original soundtracks. Synthetic data is comparatively less noisy than the original soundtracks. In the synthetic environment, we generated music chords by mixing the related notes. The performances of the optimized features and the statistical models are evaluated with the original soundtracks in a real music environment. The real music environment is the orchestra, where music notes of different instruments are played together as time progresses. This continuous chord progression is depicted in Figure 4-8.

### 6.2.1 Feature and statistical model parameter optimization in synthetic environment

We optimize the parameters of the polyphonic music pitch representation features i.e. Pitch Class Profile (PCP) feature and Psycho Acoustic Profile (PAP) feature. Using optimized features, we tune the parameters of the 3 statistical models (i.e. HMM, GMM, and SVM) which are used for modelling 48 chords.

The synthetic dataset is used for parameter optimization of the features and the classifiers. We create music chords by mixing music notes according to Table 2-2 and the music notes are generated at 100 BPM (beats /quarter notes per minutes $-4 / 4$ meter). The note mixing procedure for creating a chord is shown in Figure 6-2. By setting different delays while mixing the notes, we can generate the same chord differently. As shown in this figure, we set the delay x to be $\{0, \mathrm{~T} / 4, \mathrm{~T} / 8, \mathrm{~T} / 16, \mathrm{~T} / 32\}$ and mix the notes to generate 5 different samples of each chord. These time delays are set to make synthetically generated chords as close as possible to the real chords
generated in the orchestra. For example, the strumming delay of strings to generate a chord is proportional to the tempo of the music. We circulate the note mixing such that note 3 will take the position of note 1 , note 1 takes the note 2 position, and note 2 takes the note 3 position. Therefore, in total we can generate 13 samples of the same chord.


Figure 6-2: Note mixing procedure for creating a synthetic chord
Chord data is generated from both natural instruments such as piano, bass guitar, rhythm guitar, and synthetic instruments like acoustic grand piano, acoustic nylon guitar, electric bass fingered, and fretless bass using a Roland RS- 70 synthesizer and the cakewalk software. Notes generated from natural music instruments are tuned to the ISO 16 standard, which specifies that A4 is 440 Hz , the concert pitch (see Table 2-1). Synthesizers and MIDI tone generating softwares generally follow the same standards. All generated music notes cover the range of C3 to B7. In total, we have over 6 minutes of each chord sample in the synthetic data set.

## (a) Parameter optimisation of the polyphonic pitch representation features

In order to optimize the parameters of the PCP and PAP features, we use a 5 state (including entry and exit) HMM with 2 Gaussian mixtures for each hidden state as a
test model to evaluate modelling accuracy when the feature parameters are changed. We use the entire synthetic dataset with cross-validation where, half of the data of each chord is used for training. The HTK toolbox [142] was used to model 48 chords with the HMM. We segment the music chords into 600 ms beat space segments because the tempo of music notes which comprise the chord is 100BPM. It is found in our initial experiments that the chord detection accuracies of both the features (PCP and PAP) are higher when the frequency range for feature calculation is set to (128 ~ 8192) Hz.

The chapter 4.4.1.1 discusses technical details about the feature calculation. The optimized number of coefficients for PCP is 60 and they are calculated by setting value of $\mathrm{C}=1200$ in equation (4-4) and following the other equations (4-5) and (4-6). This PCP feature parameter setting can improve the average chord detection accuracy up to $8.2 \%$, compared to the value of $\mathrm{C}=12$ (total 12 PCP coefficients) in the equation (4-5). When the value of $\mathrm{C}=12$, then the number of coefficients in the PCP feature vector is 12 . Since we approximate the harmonic and sub-harmonic to the nearest music tone in the corresponding octave, we have 12 coefficients for each octave in the PAP feature vector. Therefore in the $128 \sim 8192 \mathrm{~Hz}$ frequency range we have 72 optimized coefficients. With the optimized parameters, $81.59 \%$ and $86.28 \%$ accuracies are reported by both PCP and PAP features respectively.

## (b) Parameter optimisation of the statistical models

We use optimized the PCP and PAP features to tune the parameters of the statistical models, i.e. HMM, GMM, and SVM. The technical details of the SVM, GMM, and

HMM classifiers are discussed in chapter 4.4.2. Experimentally discovered optimized parameters for these classifiers are listed below.
> SVM: Radial based kernel function (RBF) with the $\mathrm{C}=0.145$ in Equation (4-10)
> GMM: 13 Gaussian mixtures (GMs) in the GMM for each chord
> HMM: 7 states HMM including entry and exit states, two GMs in each hidden states.

Figure 6-3 shows the chord classification performance of the HMM, SVM, and GMM. It can be seen that the HMM outperforms the SVM and GMM in modelling chords. All the statistical models are able to model PAP features better than PCP features. With HMMs, the PAP feature can achieve around $4 \%$ higher average accuracy for chord detection than the PCP feature. We then use these optimized parameters of both the HMM and the PAP feature for real music environment experiments.


Figure 6-3: Average chord classification accuracy of the statistical models

### 6.2.2 Performance of the features and the statistical models in the real music environment

We use 50 songs by cross validation, where 30 songs are used for training and 20 songs are used for testing during each turn. In addition to the song training chords, we
use a synthetic dataset to train the 48 HMM models. For the ground truth we manually annotate the songs. Figure 6-4 shows a sample annotation of a song section.


Figure 6-4: Manually annotated the intro and the verse 1 of the song "Cloud No 9 by Bryan Adams"

The reported average frame-based accuracy of chord detection is $77.48 \%$. We managed to determine the correct Key of all the songs. Therefore, frame-based accuracy of $83.25 \%$ is achieved after error correction with the Key information.

## Results comparison with a previous method

Using both our data set and the procedures described in section 6.2.2 for data training, and testing, we compare the performances of our method with the method proposed by Sheh and Ellis (2003) [108]. The technical details of both the methods are highlighted in Table 6-1.

Table 6-1: Technical details of our method and the other method

|  | Sampling frequency <br> (fs) of the data | Frequency resolution used <br> for feature calculation | Feature | Segmentation step size | Statistical model |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Other method | 11025 Hz | 2.7 Hz | PCP - $24 \operatorname{dim}$ | Fixed length (100 ms), non <br> overlapping frames | HMM (3-hidden states, <br> 32 mixtures per state) |
| Our method | 44100 Hz | 2 Hz | PAP - 72 dim | Beat space segment | HMM (5-hidden states, <br> 3 mixtures per state) |

We can achieve up to a 67.12 \% frame-based accuracy with Sheh and Ellis's method, while a $77.48 \%$ frame based accuracy has been reported by our method. Our experiments found that the inclusion of information from harmonics, sub-harmonics
information, and fundamental frequencies (F0s), can more effectively characterize polyphonic pitches (PAP feature) than the usage of only the F0s of the pitches (PCP). As shown in Table 6-1, Sheh and Ellis's method considered only the 0 to 5 kHz frequency range. Most of the F0s of the music pitches are within this frequency range. Sheh and Ellis made the correct choice in using the 0 to 5 kHz frequency range for calculating the PCP feature because this feature only considers the F0s of the music pitches. However, the harmonics and the sub harmonics of the music pitches are spread wider than the $(0 \sim 5) \mathrm{kHz}$ range. Therefore, we used a broader frequency range, i.e. $(0 \sim 20) \mathrm{kHz}$, to account for most of the harmonics and sub-harmonics of the music pitches.

### 6.3 Vocal/instrumental region detection

Experiments have been conducted to find how robustly the musically modified features discussed in chapter 5.1.1 can characterize these vocal and instrumental regions and how accurately existing statistical learning techniques (SVM, GMM and HMM) model these features to classify the regions. Thus, the experimental procedures consist of the parameter optimization of both individual features and classifiers. The Vocal/instrumental region detection problem must be language independent. Thus, features must be least sensitive to the language of singing. The following subsections describe these experimental procedures. Experiments have been conducted using 120 popular songs sung by male and female artists with each artist singing ten of these songs. All the artists are categorized according to their gender and language. They are detailed in Table 6-2. Each song is over 3.5 minutes in average length. A total of around 420 minutes of data have been used in this series of experiments.

Table 6-2: Details of the Artists

| $A_{i t i s t_{s}}$ |  | Language |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | English |  |  | Chinese |  |  |
| $\left\|\begin{array}{l} \stackrel{\rightharpoonup}{0} \\ \stackrel{\rightharpoonup}{0} \\ 0 \end{array}\right\|$ |  | Michael Learns To Rock (MLTR) | Bryan Adams | Westlife | Huang Pingyuan | Wen Zheng | A Du |
|  | \% | Shania Twain | Mariah Carey | Celine Dion | Liu Ruoying (Rene) | Leung (Jasmine) | Li Qi |

### 6.3.1 Manual labelling of experimental data for the ground truth

All the experiment results are compared with the manually annotated data. A sample of this manual annotation is shown in Figure 6-5.

| SONG :BSS No- |  | On a day like Today |  |  | $\begin{array}{r} \text { ARTIS } \\ \text { Segment } \end{array}$ | ST :- Bryan Numbers | Adan |  | 16 th note length $=182.49052$ milliseconds |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Vocal section |  |  |  |  | Instrumental section |  |  |  | Lyrics ( Music Phrases) |
|  | Time duration (Sec) |  | BSS No |  | Time duration (Sec |  | BSS No |  |  |
| 1 |  |  |  |  | 0 | 14.2343 | 1 | 78 |  |
| 2 | 14.2343 | 16.05917 | 79 | 88 | 16.0592 | 16.9716 | 89 | 93 | Free |
| 3 | 16.9716 | 18.61403 | 94 | 102 | 18.614 | 20.4389 | 103 | 112 | is all you gotta be |
| 4 | 20.4389 | 24.45373 | 113 | 134 | 24.4537 | 26.0961 | 135 | 143 | dream dreams no one else can see |
| 5 | 26.0961 | 30.29343 | 144 | 166 | 30.2934 | 31.3884 | 167 | 172 | sometimes ya wanna run away |
| 6 | 31.3884 | 36.68059 | 173 | 201 |  |  |  |  | but ya never know what might be comin' round your way |
| 7 | 36.6806 | 39.41795 | 202 | 216 | 39.418 | 39.6004 | 217 | 217 | ya ya ya |
| 8 | 39.6004 | 42.3378 | 218 | 232 | 42.3378 | 43.0678 | 233 | 236 | On a day like today |
| 9 | 43.0678 | 45.25765 | 237 | 248 | 45.2576 | 45.8051 | 249 | 251 | the whole world could change |
| 10 | 45.8051 | 48.1775 | 252 | 264 | 48.1775 | 48.725 | 265 | 267 | the sun's gonna shine |
| 11 | 48.725 | 50.91486 | 268 | 279 | 50.9149 | 51.2798 | 280 | 281 | shine thru the rain |
| 12 | 51.2798 | 54.01719 | 282 | 296 | 54.0172 | 55.8421 | 297 | 306 | on a day like today |
| 13 | 55.8421 | 61.49931 | 307 | 337 | 61.4993 | 62.7767 | 338 | 344 | ya never wanna see the sun go down |
| 14 | 62.7767 | 66.97402 | 345 | 367 | 66.974 | 71.5363 | 368 | 392 | ya never wanna see the sun go down |

Figure 6-5: This manual annotation describes the time information of the vocal and instrumental boundaries in the first few phrases of the song "On a day like today" by Bryan Adams. The frame length is equal to the $16^{\text {th }}$ note length beat space segment ( 182.49052 ms ).It is the smallest note length that can be found in the song.

The lyrics and scores of the songs have been obtained from commercially available music sheets. This information is useful for our manual annotation. Even through listening, we can determine that the length of the vocal and instrumental regions are integer multiples of the smallest notes i.e. eighth note, sixteenth note, or thirty second note, which again confirms the idea that the length of these regions are proportional to
the inter-beat interval. This manual labelling of vocal and instrumental regions in the song is labour intensive. Listening tests of over 3 hours have been carried out manually for each song, in order to find the vocal/instrumental boundaries with about 100 ms precision. We can go to such high precision to find the region boundaries by using the knowledge that regions are multiples of the smallest note that can bee seen in a song.

### 6.3.2 Feature and classifier parameter optimization

A series of initial experiments are carried out to optimize the parameters of both the features and the classifiers used for vocal / instrumental boundary detection. The construction of features, which are used for characterizing the vocal and instrumental music signal sections, are discussed in chapter 5.1.1. These features are:
> Mel-frequency cepstral coefficients (MFCCs)
> Octave scale cepstral coefficients (OSCCs)
$>$ Linear prediction coefficients (LPCs) and Octave scale linear prediction coefficients (OSLPCs)
> Linear predictive cepstral coefficients (LPCCs) and Octave scale linear predictive cepstral coefficients (OSLPCCs)
$>$ Twice-iterated composite Fourier transform coefficients (TICFTCs) and Octave scale TICFTCs (OSTICFTCs)

Exactly 44100 sample points are used for the time to frequency signal transformation in the process of calculating MFCCs, OSCCs, TICFTCs and OSTICFTCs. For the second FFT operation for calculating TICFTCs, we used 22050 points. The total number of points in each sub-band is used for the second FFT operation in calculating
the OSTICFTCs. As described in equation (5-20), we manually set the boundary limits of the bins. We found empirically that the linear boundary limits are good enough for calculating equal size energy bins for generating TICFTCs. Equation (6-1) describes the calculation of lower and upper limits of the $j^{t h}$ bin of the $i^{t h}$ frame, where $C$ is the number of points or the width of the bin.
$B_{j}(i)=\sum_{k=\omega}^{\sigma} Y_{i}(k)$ for $\mathrm{i}^{\text {th }}$ beat space segment and the $\mathrm{j}^{\text {th }}$ bin where $\omega=\{1+C *(j-1)\}$ and $\sigma=C * j$

It is noticed in our previous study (Gao 2003 [41]) that 5-state HMM, including entry and exit states and 2 mixture models for each hidden state, perform as a good classifier for vocal/instrumental boundary detection. For feature parameter optimization, the entire data set has been used with cross validation, where 5 songs for each artist were used for training. For each feature, we vary the number of extracted coefficients between 1 and 60 , and note down the vocal/instrumental classification accuracies. The set of coefficient(s) and the feature parameters (such as number of filters in MFFCs and OSCCs, value $C$ in TICFTC), which give the highest average classification accuracy are considered the optimum parameters for that feature. These optimized feature parameters are described below in Table 6-3.

Table 6-3: Optimized parameters for features

| Feature | Other parameters | Number of filters | Number of coefficients \{SB sub-band\} |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | SB-1 | SB-2 | SB-3 | SB-4 | SB-5 | SB-6 | SB-7 | SB-8 |
| OSLPC | --- | --- | 10 | 5 | 5 | 5 | 5 | 10 | 5 | 10 |
| OSLPCC | --- | --- | 5 | 5 | 5 | 10 | 10 | 5 | 5 | 5 |
| OSTICFTC | $\mathrm{C}=50$ in each sub-band | --- | 2 | 3 | 4 | 4 | 4 | 4 | 3 | 4 |
| LPC | --- | --- | 25 |  |  |  |  |  |  |  |
| LPCC | --- | --- | 15 |  |  |  |  |  |  |  |
| TICFTC | $\mathrm{C}=100$ | --- | 15 |  |  |  |  |  |  |  |
| MFCC | --- | 20 | 35 |  |  |  |  |  |  |  |
| OSCC | --- | 96 | 20 |  |  |  |  |  |  |  |

The total number of OSLPCs, OSLPCCs, and OSTICFTCs used for characterizing the signal frames are the sum of the coefficients in the individual sub-band (i.e. OSLPC55, OSLPCC -50 and OSTICFTC-23). We then tune the classifiers in order to achieve their best performance in vocal and instrumental region classification. Experimentally found optimized parameters for these classifiers are listed below.
$>$ SVM: Radial based kernel function (RBF) with $\mathrm{C}=0.145$ in Equation (4-10).
> GMM: 62 Gaussian mixtures (GMs) in vocal class GMM, 48 GMs in instrumental class GMM.
> HMM: 11-state HMM including entry and exit states, two GMs in each hidden states.

The robust vocal/instrumental boundary detector should be less sensitive to both the language of the song and the gender of the singer. After tuning the parameters of the features and classifiers, we conduct two experiments to study the language (Chinese/English) and gender (male/female) sensitivities of the feature.

### 6.3.3 Language sensitivity of the features

This experiment was conducted by dividing the dataset (Table 6-2) according to the language of the songs. First, we trained the HMM with Chinese songs ( 60 songs) and test with English songs. Then we repeated the test with the language reversed. The average frame classification accuracies of the vocal/instrumental regions are listed in Figure 6-6. All the features, except for the LPCC, are able to better identify vocal frames than instrumental frames. This implies that the features are more sensitive to singing voice than instrumental music. The average (Avg) performance of the OSCC is higher than that of other features. Thus, it is less sensitive to the language of the
song than the other features. The OSLPC, OSLPCC, and OSTICFTC, which account for the information in octaves, give higher classification accuracies than the normal LPC, LPCC and TICFTC, respectively. The average accuracy of the LPCC is nearly $3 \%$ higher than the OSLPC, and it is less sensitive to languages than the OSLPC is. Thus, it can be concluded that feature extraction in octaves, which divides the full band music information into octaves and highlights them as independent octave information, results in lower language sensitive for both vocal signals and instrumental signals than treating the full band music information as whole.


Vocal \% - percentage of the correct classification of vocal frames
Inst $\%$ - percentage of the correct classification of Instrumental (Inst) frames
Average \% - percentage of the average correct classification of vocal and instrumental frames
Figure 6-6: Average classification accuracies of the features in the language sensitivity test

### 6.3.4 Gender sensitivity of the features

A total of 120 songs are divided equally among the four gender-language categories.
Thus, we conduct a total of four experiments. First, an HMM is trained on the male Chinese artists' songs and then tested on the female Chinese artists' songs. Then, the experiment is ran again with the training and test sets reversed. We then repeated these two experiments with the English data sets.

The average accuracies of vocal and instrumental frames classification are shown in Figure 6-7. Pitche of the female voices span a wider range than the male voices covering the spectrum of 200 to 2000 Hz . Thus, spectral fluctuations are wider in female vocals than in male vocals. It can be seen that vocal and instrumental frame detection is higher with OSCCs than with other features. Again, the accuracy of OSLPCC, OSLPC, and OSTICFTC are higher than the normal LPCC, LPC and TICFTC, which implies that octave based music information extraction can better suppress the gender sensitive characteristics in vocal and instrumental music signals. We conclude that OSCC can effectively model both male and female spectral characteristics.

Gender sensitivity test


| Features | OSCC | MFCC | OSLPCC | LPCC | OSLPC | LPC | OSTICFTC | TICFTC |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Vocal \% | 88.013 | 79.6 | 73.89 | 69.375 | 76.63 | 75 | 69.23 | 63.79 |
| Inst \% | 71.32 | 68.1253 | 60.14 | 57 | 49.88 | 49.0625 | 61.67 | 59.39 |
| Average \% | 79.6665 | 73.86265 | 67.015 | 63.1875 | 63.255 | 62.03125 | 65.45 | 61.59 |

Figure 6-7: Average classification accuracies of the features in their gender sensitivity test

### 6.3.5 Overall performance of the features and the classifiers

Overall correct vocal/instrumental classification accuracies are shown in Figure 6-8. In the experiments, all the songs by each artist are used in cross validation, where five songs by each artist are trained at each turn. When the number of hidden states increased from 3 to 9 , the average classification accuracy of OSCC increased from
$89.17 \%$ to $91.105 \%$. This is explained by the fact that HMM can model temporal properties with higher a order of hidden states. Compared with the accuracies noted in section 6.3 and 6.4, this result presents a $10 \%$ increase in accuracy in OSCC.

In music recoding, the melody contour of the instrumental line is always kept as close as possible to vocal pitch (Miller 1986 [84], Sundberg 1987 [118]). Otherwise, the music would be off tune. In our perceptual analysis (Maddage et al [75]), we found that there are structural similarities (e.g. chords mixtures, music scales, rhythm, and instrumental setup) in the instrumental music composition of songs by same singer. In this experiment, vocal and instrumental HMMs are trained with 5 songs for the same artist. Thus, it can be concluded that OSCC can characterize vocal similarities independent for instrumental similarities of different spaces.


| Features | OSCC | MFCC | OSLPCC | LPCC | OSLPC | LPC | OSTICFTC | TICFTC |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Vocal \% | 97.44 | 95.016 | 80.26 | 72.98 | 78.85 | 74.31 | 78.57 | 72.73 |
| Inst \% | 84.77 | 69.115 | 71.32 | 65.4 | 64.93 | 57.43 | 70.76 | 67.49 |
| Average \% | 91.105 | 82.0655 | 75.79 | 69.19 | 71.89 | 65.87 | 74.665 | 70.11 |

Figure 6-8: Overall classification accuracy of features with HMM
From the feature performance results shown in Figure 6-6, Figure 6-7, and Figure 6-8, we can conclude that OSCC is the best feature and MFCC produces the second highest accuracy. We repeat the same training and testing setup with SVM and GMM classifiers in order to compare the classifiers performances. From the results shown in

Figure $6-9$, we can conclude that the HMM is better than all other classifiers for vocal/instrumental classification. With all the classifiers, the OSCC can characterize vocal and instrumental music more accurately than the MFCC.


Figure 6-9: Classifier performances in vocal / instrumental classification
Music structure analysis reveals that the length of the music region is proportional to the beat space segment size, and that musically driven signal properties such as music harmonic structure and octave spectral spacing change in beat space-time steps. This signal behaviour reveals that within the beat space segment, the signal section can be considered quasi-stationary. Figure $6-10$ shows the average HMM classification accuracy for extracted OSCC and MFFC from different frame size segments. X 1.0, X 0.5 , and X 0.25 denote that frame size is equal to the beat space size (the smallest note length that can be seen in a song), half, and quarter of the smallest note length, respectively. From the results, we notice that the features extracted from inter-beat proportional segments give better accuracy than those are extracted from fixed length segments ( 30 ms ). Thus, feature stability improves with beat space proportional frame size. It can be concluded that the signal section within the beat space segment can be considered more to be stationary than the signal section of fixed length. In addition, OSCC seems to perform better than MFCC in characterizing vocal and instrumental frames when they are extracted even in less signal stability conditions (fixed length segments).


Figure 6-10: Effect of classification accuracy with frame size

## Results comparison with a previous method

We compare our results with a previous method proposed by Berenzweig and Ellis (2001) [10]. They used the $12^{\text {th }}$ order perceptual linear predictive (PLP) cepstral coefficients together with delta and double delta coefficients extracted from 100 ms fixed length music frames to characterize the vocal and instrumental frames (Williams and Ellis 1999 [136]). Then, PLP cepstral coefficients are modelled with a HMM configuration similar to that of a speech recognizer (two state and single Gaussian mixtures per state). We employed our training and testing procedures, described in section 6.3.5 to evaluate their system. Their method achieved $68.56 \%$ of frame based accuracy with our data set, whereas we are able to achieve around a $91 \%$ frame based accuracy.

### 6.4 Detection of semantic clusters in the song

Semantic clusters which define the structure of the popular song are Intro, Verse, Chorus, INST-instrumental sections, Middle eighth, Bridge, and Outro. These clusters are detailed in chapter 2.4. A rule based technique which incorporates music composition knowledge in detecting these semantic clusters has been explained in chapter 5.3.

Our experiments are conducted using 70 popular English songs ( 10 by MLTR, 10 by Bryan Adams, 10 by Shania Twain, 10 by Mariah Carey, 10 by Celine Dion, 8 by Westlife, 6 by the Beatles and 6 by the Backstreet Boys). The original keys and chord timing of the songs are obtained from a commercially available music sheet. We evaluate the results of the detected semantic clusters in two aspects.

- How correctly are all the parts in the semantic cluster identified? For example, if $2 / 3$ of the choruses are identified in a song, the accuracy of identifying the choruses is $66.66 \%$.

> number of sections identified in the semantic
> Identification accuracy $=\underline{\text { cluster (i.e. no of choruses, no of verses etc.) }}$ of the sections Actual number of sections in the semantic cluster

- How correctly are the sections detected? In equation (6-3), the accuracy of the section detection is explained. For example, if the accuracies of detecting 3 chorus sections in a song are $80.0 \%, 89.0 \%$ and $0.0 \%$, then the average accuracy of detecting chorus section in the song is $(80+89+0) / 3=56.33 \%$.

$$
\begin{gather*}
\text { Detection accuracy }  \tag{6-3}\\
\text { of a section }(\%)
\end{gather*}=\frac{\text { Length of correctly detected section }}{\text { Correct length }} * 100
$$

In Table 6-4, the accuracy of both the identification and detection of structural parts in the song "Cloud No 9" by Bryan Adams is reported. Since the song has 3 choruses and they are all identified, $100 \%$ accuracy is achieved in the identification of chorus sections in the song. However, the average correct length detection accuracy of the chorus is $99.74 \%$.

Table 6-4: Evaluation of identified and detected parts in a song
Both identification and detection accuracy of the parts in the semantic clusters of the song "Cloud No 9" by Bryan Adams

| Semantic cluster names | I | V | C | INST | B | O |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Actual number of parts | 1 | 2 | 3 | 1 | 1 | 1 |
| Number of parts identified | 1 | 2 | 3 | 1 | 1 | 1 |
| Identification accuracy of Number of parts in a semantic cluster (\%) | 100 | 100 | 100 | 100 | 100 | 100 |
| Average detection accuracy of the parts in the semantic cluster (\%) | 100 | 100 | 99.74 | 99.26 | 98.88 | 100 |

I - Intro, V - Verse, C - Chorus, B - Bridge, INST- Instrumental section, O - Outro
Figure 6-11 illustrates our experimental results for the average detection accuracy of different sections. It can be seen that Intro (I) and the Outro (O) have been detected with very high accuracy. However, the detection accuracy of Bridge (B) section is the lowest.


Figure 6-11: The average detection accuracies of different sections

## Results comparison with a previous method

Using our test data set, we compare our method with a chorus cluster detection method described in Goto (2003) [49]. In his method, a 12 dimension PCP features are extracted from was used for characterizing 256 ms . Then, feature similarity analysis was performed to detect the chorus regions in a song. Table $6-5$ summarizes the key technical differences of our method and the other method. Using their method, we achieved $54.33 \%$ and $66.18 \%$ accuracies for both chorus identification and detection respectively. With our method, accuracies were over $75 \%$. The results
comparison (both identification and detection) reveals that our method is more accurate than the Goto's method.

Table 6-5: Technical detail comparison of other method with ours.

|  | Sampling <br> Freq - (Fs) | Window size | Frequency <br> resolution | Steps for chorus detection |  |  |  |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| Other <br> method | 44.1 kHz | Fixed length <br> 256ms frames <br> with 80ms <br> shift | Not given. <br> (assumed <br> $2 \mathrm{~Hz})$ | STEP 1: PCP <br> feature based <br> similarity regions <br> are detected with <br> time lag of 80ms <br> terns | STEP 2: Similarity <br> region selection- <br> threshold based on <br> discriminant <br> criterion | STEP 3: Key change <br> detection via <br> circular modulation <br> of the PCP vector | STEP 4: Assumptions - <br> 1. Chorus length - 7.7~40s <br> structure <br> 3. middle eighth regions are <br> considered as choruss |
| Our <br> method | 16 kHz | Beat-space <br> segment <br> (smallest note) | 2 Hz | STEP 1: Melody <br> based similarity <br> regions (MBSRs) <br> detection via chord <br> detection (PAP <br> feature +HMM) | STEP 2: Content <br> based similarity <br> region (CBSRs) <br> detection - (OSCC <br> feature + HMM) | STEP 3: Chorus <br> sections are detected <br> using the <br> information of <br> MBSRs +CBSRs + | STEP 4: Assumptions - <br> heuristic rules <br> derived from our <br> 2. Mostly on V-V-C and V-C-V- <br> C song structures <br> 3. Middle eighth regions are not <br> considered as choruses |

The reasons closely related for the lower accuracy of Goto's method is given below.

1. The feature i.e. PCP feature, is highly sensitive to the harmony line (see chapter 4.4). Verses and choruses of many of popular songs have similar harmonies. Therefore, PCP feature based similarity analysis frequently wrongly identifies and detects verse regions as chorus regions.
2. The fixed length signal segmentation ( 256 ms frame size) used in the other method simply can't compute the exact boundaries of chorus regions, since chorus regions are proportional to the tempo of the music
3. The assumption on chorus length ( $7.7 \sim 40 \mathrm{~ms}$ ) in the other method doesn't assist computation on exact chorus boundaries.

## A Failure cases of the semantic cluster detection algorithm

Rules are meant to be broken. Song writers always have full freedom in writing and composing songs according to their own imagination and creativity. We designed our heuristic rules for detecting popular song structures based on the survey discussed in chapter 2.5 . However, some songs do not completely satisfy all our heuristic rules.

Our heuristic rules assume that the Intro of the song is an instrumental signal and that verse 1 follows the Intro. However, in some songs, the Intro is not fully instrumental. It contains instrumental music and humming by the singer. Our music region detection algorithm, as described in chapter 5.1, classifies the humming sections as vocal regions. This classification is correct. As we assume that verse 1 starts right after the instrumental Intro, in this case, the verse 1 region would be wrongly identified and detected. Such a situation is explained in Figure 6-12 (a) and (b). Figure 6-12 (a) shows the correct song structure of the song "Can't let go" by Mariah Carey, and Figure 6-12 (b) shows the detected song structure. The identification accuracy and detection accuracy of the different semantic clusters are shown in Figure 6-12 (c). Even though verse 1 is wrongly detected, the harmony contour of the wrongly detected verse 1 is partially equal to the harmony contour of the correct verse 1 and the choruses. Therefore, verse 2 was detected with $86.8 \%$ accuracy and all the choruses were detected with $82.12 \%$ average accuracy.

(a)

(b)

(c)

| Semantic cluster names | I | V | C | INST | B | O |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Actual number of parts | 1 | 2 | 4 | 0 | 2 | 1 |
| Number of parts correctly identified | 1 | 1 | 4 | 0 | 1 | 1 |
| Identification accuracy of Number of parts in a semantic cluster (\%) | 100 | 50 | 100 | 0 | 50 | 100 |
| Average detection accuracy of the parts in the semantic cluster (\%) | 50 | 43.4 | 82.12 | 0 | 27.81 | 100 |

I - Intro, V - Verse, C - Chorus, B - Bridge, INST- Instrumental section, O - Outro
Figure 6-12: A failure case of our semantic clusters detection algorithm. Figure (a) shows the manually annotated positions of the components in the song structures. Figure (b) shows the detected components and their positions. Figure (c) shows the identification and detection accuracy of the components in the semantic clusters.

We examine the algorithm accuracies of semantic cluster identification and detection when the song is segmented with fixed length 80 ms segments. The song selected for this test is Cloud No 9 by Bryan Adams. The smallest note in the song is the eighth note and is of 272.977 ms duration. The song has "V1-C1-V2-C2-INST - C3 O" structure where all the verses are 8 bars and all the verses are 16 bars. The selected fixed frame size, 80 ms is close to the smallest sixteenth note in the 30 to 190 BPM tempo range (see chapter 4.1) and over 52.5\% of English songs have sixteenth note as the smallest note (see Figure 2-11).

Table 6-6 describes the semantic cluster detection accuracies when the song is beat space segmented and fixed length segmented. Even though semantic cluster detection algorithm managed to correctly identify all the parts in different clusters, it failed to correctly detect $3 / 4$ of the lengths of verses and choruses. However, the chorus identification and detection accuracies are better than the previous method discussed in Table 6-5. The close reasons for the low performances of our semantic cluster detection algorithm with fixed length signal segmentation are explained below.

Table 6-6: Accuracies of semantic cluster detection and identification of the song "Cloud No 9 by Bryan Adams" based on beat space and fixed length segmentations

|  | Semantic cluster name | I | V | C | INST | B | O |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Actual number of parts | 1 | 2 | 3 | 1 | 1 | 1 |
| BSS | Identification accuracy (\%) | 100 | 100 | 100 | 100 | 100 | 100 |
|  | Detection accuracy (\%) | 100 | 100 | 99.74 | 99.26 | 98.88 | 100 |
| Fixed length segmentation | Identification accuracy (\%) | 100 | 100 | 100 | 100 | 100 | 100 |
|  | Detection accuracy (\%) | 84.12 | 71.34 | 65.78 | 52.98 | 12.1 | 73.47 |
| Other method | Identification accuracy (\%) |  |  | 66.66 |  |  |  |
|  | Detection accuracy (\%) |  |  | 52.74 |  |  |  |

The experimental results for detecting harmony line and vocal/instrumental boundaries, discussed in section 6.2.2 and 6.3.5 respectively reveals that the signal characterization and modelling, with fixed length segmentation is less accurate than
with BSS. Therefore, errors in chord detection lead to shift the melody based similarity regions. It is also required to detect the correct starting point of the first verse to detect correct semantic regions by using our algorithm. However, vocal/instrumental boundary detection algorithm has wrongly detected instrumental frames as vocal frames within the last $16 \%$ of the Intro region in this song. Therefore starting point of the first is incorrect. From the heuristic rules in semantic cluster detection algorithm point of view, all the cluster regions are measured in terms of beat space segment. When the fixed length segmentation is used, then the minimum of $1 \%$ error is added to the region detection. However, higher error rate in chord detection and vocal/instrumental boundary detection with fixed length segmentation has led the majority of the errors in semantic cluster detection in the song.

### 6.5 Summary of the experimental results

We have conducted these experiments to evaluate our methodologies for layer-wise information extraction in the conceptual music structure pyramid (see Figure 1-1). Compared to speech information processing, which employs fixed length segmentation, the fundamental change we propose in the thesis is the Beat Space segmentation (BSS). To conduct the BSS, we first calculate the smallest note that can be seen in a song, using the algorithm discussed in chapter 4.1. We assume the meter of the song is $4 / 4$, which is widely used in popular music. We have achieved $89.16 \%$ accuracy in detecting the smallest note length in 120 songs with $\pm 30 \mathrm{~ms}$ error margin. After BSS, we compare the Pitch Class Profile (PCP) feature and the Psycho Acoustic Profile (PAP) feature for their effectiveness in modelling music chords. The PAP feature incorporates the knowledge of fundermental frequencies, harmonics and subharmonics generated by music tones. The PCP feature mainly considers the
fundamental frequencies of music notes in all the octaves and project them onto the twelve notes in one octave. It is found that the optimized PAP feature with HMM models the 48 chords with the highest accuracy. By running the error correcting algorithm which we discussed in chapter 4.4.2, we are able to correctly detect the music chords of 50 songs with a $83.25 \%$ frame based accuracy. Compared to fixed length segmentation, BSS can improve the average chord detection accuracy by $10 \%$.

Music regions are the information in the $3^{\text {rd }}$ layer of the structural pyramid. We incorporate octave scale music property fluctuations for feature design. We found that Octave Scale Cepstral Coefficients (OSCCs) are more robust than Mel Scale Cepstral Coefficients (MFCCs) in detecting vocal and instrument regions. Out of 120 songs, OSCCs with HMM are able to achieve a frame based average accuracy of $91.11 \%$.

Information in the $4^{\text {th }}$ layer includes semantic music clusters, i.e. Intro, Verse, Chorus, Middle eighth, Bridge and Outro. We have explained the results of our popular song structures survey in chapter 2.5 . Chapter 5.3 explained our rule-based semantic cluster detection algorithm. Out of 70 songs, we are able to correctly detect these semantic clusters with over 70\% accuracy.

## 7 <br> Applications

Music structure analysis is essential for music semantic understanding, and is useful for developing many applications in various domains such as musicology, psychology, and information technology. In this chapter, we outline how the layer information in the music structure can be helpful in developing these applications.

### 7.1 Lyrics identification and music transcription

Lyrics identification in a song is useful for many applications. One direct application is automatic lyrics alignment in Karaoke music. Wang et al 2004 [133] presented a method to align the text of lyrics with a song's tack. However, the key challenge, music audio to text conversion, remains unsolved. Singing voice carries more information than the music (Xu et al 2005 [140]). Thus, to understand the higher semantic structure of the music (the $4^{\text {th }}$ and higher layers in Figure 1-1), we need to extract the meaning(s) of the vocal phrases. Lyrics identification can help to decode semantic meaning and analyze the music scenes. These music scenes are useful for developing music documentaries.

Both time information extraction and vocal/instrumental boundary detection are the preliminary measures for lyrics identification and music transcription. The primary information required for formulating these tasks are summarized in Figure 7-1. Since music phrases are constructed with rhythmically spoken lyrics [100] (see chapter 2.3 Figure 2-8, Figure 2-8), time analysis and beat space segmentation can be used to
identify the word boundary in the polyphonic music signal. Singing voice enhancement by the reduction or separation of background music reduces the complexity of identifying voiced/unvoiced regions within the beat space signal segment (see Figure 5-22). This direction would simplify the lyrics identification process.

In addition to vocal/instrumental region detection, chord detection extracts the pitch/melody contour of music. Further analysis of beat space segmented music signals helps to estimate the signal source mixture. This is the turning point of music transcription.


Figure 7-1: Primary information required for lyrics identification and music transcription

### 7.2 Music Genre classification

Music genre is a high-level index of the music content. It is useful for many applications such music information retrieval MIR systems, music summarization and singer identification.

Music content analysis is necessary for developing a robust genre classification system. It is more difficult to discriminate between music genres than to discriminate music from speech or other sounds. Several research efforts focus on music genre classification of MIDI files. Shan et al. (2002) [106] investigated the classification of music style by melody from a collection of MIDI music. Chai [3] employed HMM to classify the melodies of Irish, German, and Austrian folk songs. Dannenberg et al. (1997) [24] extracted 13 features from MIDI and used different classifiers to recognize the music style. However, since MIDI data is a structured format, it is easy to extract features according to its structure. Real sounds such as wav and mp3 files are different from MIDI, meaning that MIDI style classification is not practical for real applications. Matityaho and Furst (1995) [81] discriminated between classic and pop music by using the average amplitude of Fourier transform coefficients and neural networks. Soltau et al. (1998) [113] classified music into rock, pop, techno, and classic genres using HMM and ETMNN to extract temporal structure from the sequence of cepstral coefficients. Han et al. (1998) [50] classified music into classic, jazz, and pop genres using simple spectral features and the nearest mean classifier. Pye (2000) [93] used Mel-frequency cepstral coefficients (MFCC) and Gaussian mixture model (GMM) to classify music into six types; blues, easy listening, classic, opera, dance, and rock. Jiang et al. (2002) [56] used octave-based spectral contrast features and GMM to classified music into five types. Tzanetakis and Cook (2002)
[126] trained GMM models using timbral texture features, rhythmic content features, and pitch content features in order to differentiate between 20 music genres.

All these methods have followed fixed length segmentation, feature extraction, and parameter modelling, which are similar to speech processing techniques. In our previous method, Xu et al 2003 [139], genre hierarchical classification techniques were proposed, yet the robustness of the genre classification systems was not satisfactory. Shaoxi et al 2004 [106] incorporated beat information for genre classification. We believe that by extracting layer-wise structural information in the music, we can design more robust genre classification systems, as each genre has an identical beat structure that's correlated with harmony/ melody contours and instrumental setup.

### 7.3 Music summarization

The creation of a concise and informative extraction describing the original digital content is extremely important in large-scale information organization and processing. Music content summarization has a high commercial value. Advertising summaries instead of full length songs in customer interaction web pages can be a way to have users from illegal downloading. Many music recording companies such as EMI, Sony, and PolyGram invest a lot of money into manually generating music summaries of newly released songs for advertisements on websites and music radio stations. Manual summarization is time consuming and labour intensive. Automatic music summarization is the straightforward answer to this problem.

A number of techniques have been proposed for music summary generation (Logan and Chu 2000 [69], Xu et al 2002 [138], Lu and Zhang 2003 [70], Chai and Vercoe 2003 [19]). All these approaches face the difficulties of avoiding content repetition and detecting boundaries of content-based similarity regions (i.e. chorus sections) which are assumed to be most suitable for music summarization. In regards of legal issues, a summary should be a continuous clip of the original song, and the merging of multiple clips is not allowed. Thus, this task's key challenge is the choosing of a song section which represents the semantic meaning of the entire song. This kind of music summarization can be considered "Legal". However, a legal summary would not contain all the scenes in the song. Rather, it would contain the most important music scene (or the key message). This presents a big drawback in representing a concise summary of the full music content. To represent all the important elements of the song, we need to merge these clips. This described "Technical Summary" is useful for music information retrieval (MIR) systems, since low-level coefficients of the technical summary can be considered as agent which represents the entire music content and interacts with the query request in order to reduce the query processing time. The following subsections discuss how music structural analysis can be used for legal summary making and technical summary making.

### 7.3.1 Legal summary making

Summary making is subjective. Since the chord patterns (melody and harmony contours) in verses and choruses are detected, perceptually attractive melody regions with continuous verse chorus sections can be used to generate a good summary (Maddage et al 2004 [75]). The rhythm information is useful for aligning music phrases so that the generated summary has a smooth melody. Figure 7-2 illustrates the
process of generating a music summary using music structure analysis. First, the choruses are detected using our proposed method. Since the chorus occurs more than once in a song, we can create the music summary based on just one of them. If the length of selected chorus is less than the desired length of the final music summary, we can include the music phrases anterior or posterior to the selected chorus. For the summary to be acceptable, this music phrase should be intact, which can be achieved by proposed rhythm analysis, as music phrases in a song last a fixed number of music bars (generally, 4, 6 or 8 bars for popular music).


Figure 7-2: Illustration of music summary generation using music structure analysis.

### 7.3.2 Technical summary making

Generally, music scenes embedded in music structures are correlated with the genre of the song. Further, there are two types of music. One type is pure instrumental music, and the other type is instrumental mixed or vocal music. When the song is of pure instrumental, music scenes are sensitive to the harmony/melody created by different instruments, whereas when the song is vocal mixed, many of the music scenes are described by vocal phrases.

The basic idea of technical summary generation is to merge as much important information as possible in the required summary length. In order to detect important scenes, we need to first identify the nature of the song (i.e. its genre and whether it's composed of pure instrumental music or instrumental mixed vocal music).


Figure 7-3: Technical summary making steps
Figure 7-3 illustrates the block diagram of the proposed technical summary making steps. The first step, genre identification, was described in section 7.2. After the song is identified as pure instrumental or mix music, we can define content specific features to characterize the content. For example, power related features such as power spectrum, amplitude envelopes, and cepstral coefficients can be used for pure instrumental music characterization. For mixed music characterization, vocal related features such as LPCC, OSCC spectrum flux, zero crossing, and cepstrum flux are used (Xu et al 2005 [140]).

Similar music scenes at different locations of the song are grouped with the help of clustering tools. It is useful to embed semantic meaning in the clusters so that different clusters represent different music scenes. Cluster distance measure is a good tool in that respect. When the cluster distance is high, it is more likely that those clusters contain redundant information (i.e. different music scenes). Finally, we can pick beat space segmented frames from different clusters to make the summary. By applying audio fading techniques for smooth perception, we can create a perceptually smooth technical summary. It is a difficult to task to define generic procedure to evaluate the quality of the generated summaries. A subjective listening test is one method which can be used for summary evaluation.

### 7.4 Singer identification system

Singer identification is one piece of useful information which is required in music information retrieval (MIR) systems. Singing voice is human beings' oldest musical instrument. Human auditory, physiology, and perceptual apparatuses have evolved to a high level of sensitivity to the human voice. After over three decades of extensive research on speech recognition, such technologies have matured to the level of practical applications. However, speech recognition techniques have limitations when applied to singing voice identification, because speech and singing voice differ significantly in terms of their production and perception (Sundberg 1987 [118]).

Several approaches have been proposed to identify the singer of a query song from databases. Zhang 2003 [146] trained GMMs using Linear Prediction derived Cepstral Coefficients (LPCC) calculated from manually labelled vocal sections of each singer. In this method, the beginnings of the vocal sections were detected using simple
threshold settings calculated from extracted features, i.e. energy, zero crossing rate, spectral flux, and harmonic coefficients. It was assumed that vocal sections lasted for up to 10 to 30 seconds, and these vocal sections were fed into GMMs for further singer identification. Berenzweig et. al., 2002 [11] trained a multilayer perceptron neural network with LPCCs to detect the vocal passages in the song, and the same neural network was trained with Mel-Frequency Cepstral Coefficients (MFCC) for singer identification. Kim and Brian 2002 [62] used inverse combo filter bank to analyze the harmonicity, and the vocal regions were detected by setting a fixed hatmonicity threshold. Then, GMM and SVM classifiers were trained with the warped Linear Prediction Coefficient to identify the singer. Although the above mentioned methods have achieved frame level accuracies of up to $80 \%$, their performances are inefficient due to reasons given below.
$>$ Experiments are performed on studio recorded pure vocal music, not on normal instrumental mixed vocal music.
$>\quad \mathrm{Vocal} /$ instrumental boundary detection in the music is inaccurate.
> Music knowledge has not been effectively exploited for modelling the singers in existing (mostly bottom-up) methods.

We believe that a combination of the bottom-up and top-down approaches, which combines the strengths of low-level features and high-level music knowledge, i.e. structural information of the music, can provide a powerful tool for improving system performance.

Usually, in their albums, popular singers follow similar instrumental setup and music patterns such as chord combinations and music scale changes. Therefore, the melody
contour of a song is closely correlated with the formant structures of the singer. In the proposed singer identification technique, in addition to vocal tract characteristics such as formant and harmonic structures of the singing voice, we also use the structural similarities of the instrumental music sections of the same singer, in order to identify the singer with a high confidence level (Maddage et al 2004 [74]).

### 7.4.1 Singer characteristics modelling at the music archive

Figure 7-4 illustrates the proposed steps for modelling both singer and surrounding instrumental music. New techniques for music segmentation and vocal / instrumental region detection are explained in chapter 4 and chapter 5, respectively.


Figure 7-4: Vocal and the relative instrumental section modelling of songs of same singer.

As discussed in section 2.3 in chapter 2, in rhythmic phrases, the lengths of words are proportional to the beat space signal segments. However, in order to model the singer's vocal structure, we need to break the beat spaced signal segments into smaller sub-frames ( $20 \sim 40 \mathrm{~ms}$ ) which increase the sensitivity of the signal at the phonetic level. Then, vocal structure sensitivity features (E.g. Octave scale cepstral coefficients-OSCC and linear predicted coefficients' derived cepstral coefficientsLPCC) are extracted from the sub-frames. To characterize harmony transition and instrumental dynamics, sensitive features (E.g. OSCCs) are extracted from the beat space segmented instrumental frames. Then, existing statistical/temporal modelling techniques can be employed to build a singer model, which includes both vocal and instrumental models.

### 7.4.2 Test song identification

To identify the singer of a test song, the vocal and instrumental regions of the song are identified. First the content of these regions are characterized using vocal and instrumental feature vectors, respectively. The block "Feature extraction process of singer model" in Figure 7-5 is same as the "Feature extraction" in Figure 7-4. The vocal feature vectors of the test song are fed to the vocal models of different singers in the database to find a close match. Similarly, instrumental feature vectors of the test song are matched with known instrumental models in the database. The singer model, which gives the highest match of both vocal and instrumental models with the test song, is considered the singer of the test song. The combination of vocal and instrumental model responses for singer identification decision making is discussed below.
$Y^{i}$ in Figure 7-5 is the final response of the $i^{\text {th }}$ singer model in the test song and is calculated using Equation (7-1), where the scalar weights $\alpha$ and $\beta$ are the degrees of the responses of both vocal and instrumental models respectively. The calculation of total vocal or instrumental model (i.e. $\langle\text { Model }\rangle_{\text {Vocal or Inst }}^{i}$ ) corresponding to the vocal or instrumental frames of the test song is described in Equation (7-2), where $n$ is the total number of frames of either the vocal or the instrumental model.
$Y^{i}=\alpha .\langle\text { Vocal Model }\rangle^{i}+\beta .\langle\text { Instrumental Model }\rangle^{i}$

For a test song, if $Y^{i}$ is greater than the response of the rest of the models, the $i^{\text {th }}$ singer model will be assigned to be the singer of the test song. By taking vocal and instrumental model behaviours into consideration, the scalar weights $\alpha$ and $\beta$ can experimentally be computed.


Figure 7-5: Singer identification of the test song

The combination of vocal and instrumental models in singer identification system is suitable for MIR systems because in query by humming, we are more likely to map it into harmony contours than vocal models. When the clip of the song is the query segment, we can benefit for using either or both of the vocal and instrumental models.

### 7.5 Music information retrieval (MIR)

Ever increasing music collections require efficient and intuitive methods of searching and browsing. Music information retrieval (MIR) explores how music databases can best be searched by providing input queries in music form. These queries can be in the form of text based, humming, or a clip of music. For people who are not trained or educated in music theory, humming is the most natural way to formulate music queries. MIR systems can help musicians interact with other music according to their interests, which may include composition level and background information (history).

Many commercial players read the song information which is written in the header file (in text format) of the original recoding of the song. However, when the original song is mixed, edited, or clipped, such song information can only be retrieved via analysis of the song content. For example, Figure 7-6 (upper) shows the original album "Sings the Standards" played by Cliff Richard in Windows Media Player 10. We can retrieve all song titles, singer names, genres, and online shopping information. However, when the same album is converted to wave format and in the same player, we can no longer find such information about the album (Figure 7-6 bottom). This problem illustrates the importance of music content analysis.

In most MIR by humming systems, a fundamental frequency tracking algorithm is used to parse a sung query for melody content (Ghias et al 1995 [44]). The resulting melodic information is used to search a music database using either string matching techniques (McNab et al 2000 [83]) or other models such as Hidden Markov Models (Shifrin et al 2002 [111]). However, due to the lack of the music structure information, these MIR systems can solve only the simple and synthetic "toy" problem.


Figure 7-6: Singer information retrieval comparison when original album and converted wave files are played on Windows Media Player (version 10).Test album is Sings the Standards by Cliff Richard.

Figure 7-7 illustrates the proposed architecture and the major components of a MIR system. The input query can be either microphone input such as a duration of humming, a clip of pre-recorded music, a text-based description of the song, or a combination of the above mentioned query types. The audio-based queries are then pre-processed to extract important information, which can help the query matching. Complex pitch contour is an important piece of information that can be extracted from the audio-based queries. Since unprofessional vocalists may generate humming queries, the pitch contours of the query are then fine-tuned to match the possible harmonic/melody contours. Audio clip query exhibits more accurate pitch contours than audio humming query since they are generated by the professionals in a high quality sound recording environments. In addition to pitch contour, information about the harmony /melody line, possible beat structures, and information about the
instrumental/vocal information can be gathered from the audio queries with the help of music structure analysis tools. Text based query is limited to how much information the user remembers about the song, which can include instrumental setup, genre, singer gender, language, and additional music information like Key, tempo, meter, etc. These queries are then transmitted over the channel to music data archive which is at remote location.


Figure 7-7: Architecture of music information retrieval system

Music data archive management itself is an open research problem because of the difficulty of music data indexing. Achieving of MIR real sounding recordings, requires extracted information such as instrumental setup, rhythm, harmony/melody contours, key changes, and multi-source vocal information. The blocks of the MIR system, music summarization, music transcription, singer identification, and music
genre classification, are discussed in the previous sections. To support the query decision making, we believe that music content should be grouped into a high dimensional space based on genre, singer, and transcription. Such a grouping requires different levels of understanding of the music content. This is where music structural analysis plays a big role. The technical summary of the music content, an agent that represents the entire song, can be used to match the query to the song. The agent's interaction with the query can reduce the retrieval time.

### 7.6 Music streaming

There has been great concern regarding how to stream media content in real-time over different networks, which gives a chance for distant viewers to experience real-time interaction with the same event. For example, listening to a live music concert which is held in a remote location or listening to a music station from a different country would be interesting opportunities for distant music fans to experience in real time. Figure 7-8 shows an online music station application in Yahoo. This application, known as Yahoo Music Launchcast, allows users to listen to, select, and skip songs of their choice. The concept of media streaming has drawn the public's attention for over two decades, yet the quality of service ( QoS ) at the receiver continues to hinder adoption due to various limitations in modern networks.

We are particularly concerned about how to solve or improve the problem of continuous music streaming over unreliable networks such as Internet and wireless networks. The objective of packet loss recovery schemes in audio streaming is to reconstruct the data packets so that the received audio is perceptually indistinguishable or sufficiently similar to the original audio. Some schemes, which
have been proposed for audio streaming mainly for speech, were well surveyed in Perkins et al (1998) [91] and Wah et al (2000) [130]. Different error concealment schemes are discussed in the following sub sections to give an idea as to where the music structural analysis can play a major role in improving perceptual QoS in music streaming.


Figure 7-8: Music streaming software "Yahoo Music Launchcast Radio" given in Yahoo messenger for listening to the songs played at different music stations.

### 7.6.1 Packet loss recovery techniques for audio streaming

Packet loss recovery techniques for voice and speech transmitted over networks such as the Internet have achieved high perceptual QoS based on single channel audio transmission. Yet, stereo channel and multi-channel audio transmission are commonly practised for high bandwidth music streaming. Existing error concealment methods
can be divided into 3 types, and the key challenges of these methods are mentioned below.
$>$ Receiver-based methods - They are only effective if the packet loss is infrequent, the packet size is small, or the signal behaviours are quasistationary.
> Network based methods - Must consider system latency and availability of feedback channels
$>$ Sender-Receiver based method - Must balance between redundancy and QoS.

## (a) Receiver based schemes

Receiver based error concealment schemes act independent of from the sender information of the packet. Concealment is done by analyzing the behaviours of the neighbouring packets. The possible techniques are given below.

1. Packet replacement - A lost packet is replaced with silence, white noise, or the packet before or after the lost packet.
2. Packet repairing - different interpolation or extrapolation techniques (Kauppinen, 2002 [61]) can be applied for packet repair. Time domain or frequency signal characteristics of neighbouring packets can be used for to predict the lost packet.

These techniques work well only when losses are infrequent and when packet size is small. Due to the high probability of loss in the Internet and other networks like wireless networks, these techniques are not promising.

## (b) Network-Based schemes

The activities of networks and their controlling protocols are predefined. Thus, network based error concealment techniques are narrow in scopes due to the low level of involvement of high computing in the routers and the other controlling hubs. The retransmission or duplicate transmission of packets without receiver acknowledgement is a good option within local area networks, in order to avoid packet loss.

## (c) Sender-Receiver Based schemes

At sender's end, this scheme analyse the signal and encode the information. At the receiver's end, those encoded information is decoded to recover the lost packet. These schemes are sometimes called Sender Based schemes due to the bigger role played by the sender side. However, from our point of view, the sender and the receiver hold equal roles in the error concealment task. The existing schemes are discussed below.

- Retransmission

In the sender- receiver interaction, the simplest error concealment scheme is the retransmission of the lost packet with the receiver's acknowledgement.

- Forward error correction (FEC)

Several forward error correction schemes have been proposed in the literature (Perkins et al 1998 [91]). Two schemes are shown in Figure 7-9. First, signal analysis is carried out at the sender's end. Then, information about the signal characteristics are embedded with the original audio stream. At the receiver's end, the sent signal characteristics information is utilized for lost packet recovery. In the scheme shown in Figure 7-9 (a), packet FEC carries the information about the signal characteristics
of packet $1,2,3$, and 4 . In general, this scheme can be formulated as $(n-k)$ additional packets are transmitted with $k$ original packets. This concept can be seen as a parallel FEC. In Figure 7-9 (b), a packet carries information about the next packet to be sent. This concept can be called series FEC.


Figure 7-9: Forward error correction (FEC) mechanism for packet repair

- Interleaving

Here, the packet size is divided into smaller sections called "Units". These units are re-sequenced before transmission. The formulation steps of the interleaving scheme are shown in Figure 7-10. It can be seen that the loss of single packet in interleaving scheme results in multiple small gaps in the reconstructed stream at the receiver's end. This scheme may not be suitable for music streaming since these small gaps in the reconstructed stream may distract the audience.


Figure 7-10: Interleaving mechanism for packet repair

### 7.6.2 Role of the music structure analysis for music streaming

The brief survey of existing error concealment techniques in section 7.6.1 reveals that an efficient scheme for error repairing at the receiver's end needs information from the sender. Thus, sender-receiver based schemes, especially media specific FEC schemes, are more common in music streaming over networks. Few techniques have been proposed for packet loss recovery in music streaming. Wang et al 2004 [134] discussed a multi-stage interleaving scheme which reorganizes units by taking perceptually less significant frequency components of the packets into consideration. An unequal sized packetization method was proposed for the subsequent packet transmission. The drawback of this method is that it guarantees equal treatment (frequency domain) for all music content, though the different frequency ranges in the music have significantly different impacts on the human auditory system. Another approach, the content-based unequal error protection technique (Wang et al 2003 [132] ), effectively repairs lost packets which have percussion signals. Wyse et al 2003 [137] synthesized percussion sounds in the music signal using LPCs and FFTderived power spectrum coefficients. These coefficients are used as an audio codebook for packet recovery at the receiver end. However, these methods are inefficient at repairing lost packets, which contain signals other than percussion sounds (i.e. vocal signals and string, bowing \& blowing types of instrumental signals).

We conclude that the identification of music structure is necessary for an efficient packet loss recovery scheme. For example, instrumental/vocal boundary detection simplifies signal content analysis at the sender's side. Such analysis, together with pitch information (melody contour) is helpful for better signal restoration at the receiver's side. Content-based similarity region identification can be argued to be
another music signal compression scheme. Since structure analysis helps to identify content-based similarity regions such as chorus and instrumental music sections, we can avoid re-transmitting packet from similar regions and reduce bandwidth consumption. Figure 7-11 shows the schematic representation of the sender-receiver based structural information embedded packet loss recovery scheme. While designing such an architecture, there is always a trade off between bandwidth consumption and computing complexity. These are all open issues to be taken into consideration at the design stage.


Figure 7-11: Sender-receiver based music information embedded packet loss recovery scheme

### 7.6.3 Music compression

The greatest challenge in music compression is how to deal with the trade-off between the size of the music file (storage space) and the loss signal information (perceptual quality). MPEG-1, MPEG-2, ATRAC-2, ATRAC -3 and DOLBY AC-3 are some existing compression techniques. However, these compression approaches are not robust. Figure 7-12 details the architecture of the most popular compression technique (MP3) in the world.


Figure 7-12: MP3 codec architecture

The most important block in the compression scheme is the signal characterization block, where different signal processing techniques are used to represent the signal sections. Existing schemes at the encoding stage commonly use fixed length signal segmentation and logarithmic filter-banks to extract signal indexing coefficients from
the frequency domain. The psychoacoustic model guides the process of extracting these coefficients.

Earlier research on music perception reveals the strong relationship between the intervallic structure (harmonics) of the music tones and our cognitive mechanism (Pitts and McCulloch 1947 [92], chapter 2. section 3.2). The music structure, which reasons the cognitive research findings, is useful for designing or modifying the signal characterization block. With the help of beat space signal segmentation, within which the music signal can be considered quasi-stationary, octave scaling in the filter design can minimize the loss of music information in the compression.

Compared with conventional audio compression techniques such as MP3, which produces a 5:1 compression ratio, incorporation of music structure analysis (especially with semantic similarities in the music) produces much higher compression ratios, which can reach up to10:1 or even higher.

### 7.7 Watermarking scheme for music

Existing watermarking techniques are evenly applied to the entire music content. However, these techniques may not detect the watermark on the randomly clipped song section. We can use music structure information to design a robust watermarking scheme so that there is a high probability that the watermark can be detected in any possible music extract of the song. For example, listeners can remember and recall chorus sections better than verse sections. So, there is a high probability that the song clips contain choruses rather than verses. With this
knowledge, we can design a watermark scheme which gives higher priority to watermarking chorus rather than verses.

A song is measured in terms of bars, and all music content fluctuations are synchronized according to the beat structure. Most of the editing, mixing, or clipping of the song are proportional to multiples of the beat space segments of the song. Thus, beat space level watermarking is better than evenly distributed watermarking from the point of view of testing a section of the music. Figure 7-13 describes the design level architecture of the music content specific watermarking scheme. With the help of structural information, the music is segmented at the beat level. Instrumental /vocal region detection can identify the different source mixtures in the song, which then gives useful information for content specific watermark design. Watermark testing scheme follows the reverse procedure. Structural analysis of the watermarked music content assists the scheme in identifying the watermarked location and nature of the watermark for its recovery.


Figure 7-13: Design platform for content specific watermarking scheme

### 7.8 Computer aid tools for music composers and analyzers

Not all musicians can naturally analyze music content. Therefore, it is useful to have computer music tools which can assist musicians to analyze not only others' music but also their own music. Music transcription and summarization can aid in the understanding all music. Students who are trained to become musicians mostly analyze music which not composed by themselves. They incorporate pieces of those music to bridge a gap between creating new music and making mixes out of bits and pieces. To edit such music, music tools to find beat structures (tempo, meter, and time signature) of the pieces are required to create smooth drum, harmony, melody loops, and content mixers. Being able to study one's own music is the best way to understand one's own mistakes. All the music structure analysis ideas discuss in this thesis can produce a complete set of tools which can help musicians analyze music from a completely objective point of view and a logical prospective.

Music region detection, which isolates and extracts individual vocal and instrumental parts of the music, can create whole new ways of composing music. Today, DJs (Disc Jockeys) are making mixers from manually cutting vocal phrases and music pieces from various songs. However, with the help of computer-aid music structure analysis tools, they can effectively analyze much bigger music archives in a short time and find an interesting music clips. By combining these extracted clips we can create remix version of songs which sound much more organic and naturally miraculous that the original one.

### 7.9 Music for video applications

Automatic insertion of music into video content explores efficient ways of making different audio-visual documentaries. For example, TV channels like National Geographic would like to have an audio-visual platform which can assist in the selection of different background music for visual content. Entertainment channels like MTV may be interested in making automatic music videos, and sport channels like ESPN may consider generating music sport videos.

To formulate these audio-visual applications, it is required to understand both audio and video contents. Music structure analysis can help in selecting suitable music for the visual content.

## 8 Conclusions

### 8.1 Summary of contributions

This thesis emphasizes a music structure analysis framework which integrates music theory knowledge with digital audio signal processing techniques. The conceptual music structure pyramid has been proposed to explain the important components of music structure for the popular songs genre (chapter 2). Then, we develop a framework to extract the important pieces of information in the music structure pyramid (i.e. time information, harmony, music regions, and music semantics). The contributions of this thesis include:

## 1. Music segmentation

A new segmentation technique, i.e. Beat Space Segmentation (BSS), has been proposed to replace conventional fixed length segmentation for processing the music information in the $2^{\text {nd }}, 3^{\text {rd }}$, and upper layers of the music structure pyramid. In the new technique, we segment the music into the smallest note length frames. The idea behind the BSS is that music information within the beat space signal segment can be considered to be more stationary than those within the fixed length signal segment. The experimental results have convinced that this idea, which takes into account knowledge of music composition, leads to better music information extraction procedures. Our rhythm extraction algorithm is able to detect the smallest note length with a $89.16 \%$ accuracy.

## 2. Harmony/melody line detection (information in the $2^{\text {nd }}$ layer of the pyramid)

 We have compared the psycho acoustic profile (PAP) feature with the pitch class profile (PCP) feature for their performances in characterizing polyphonic music signals. The PCP feature, which has been commonly used for music chord detection in the literature, accounts for only the effects of the fundamental frequency (F0s) of the music notes. The PAP feature extracts the F0s, harmonics, and sub-harmonics of the music note. Laden and Keefe (1989) [67] used the PAP feature to differentiate between two chord types (Major and Minor). Yet the PAP feature has not been thoroughly studied for individual chord detection. Our experiments have revealed that the PAP feature is more robust than the PCP feature for chord detection. Three statistical learning techniques, i.e. HMM, GMM, and SVM, have been experimented for modelling the chords. HMM performs better in modelling the chords than SVM and GMM. Then, we apply music knowledge of the Key to correct the chord detection errors. Our method achieved $10 \%$ more frame level chord detection accuracy than an existing method.
## 3. Music region detection (information in the $3^{\text {rd }}$ layer)

For music region detection, we apply music knowledge to formulate features which characterize signal sections that belong to different regions (vocal/instrumental). Music signals have octave varying temporal characteristics. We modify speech analysis features, i.e. Linear Prediction Coefficients (LPCs), LPC derived Cepstral Coefficients (LPCCs), and Mel Frequency Cepstral Coefficients (MFCCs), to capture octave varying temporal properties of the music signal. These modified features are called Octave Scale LPCs (OSLPCs), Octave Scale LPCCs (OSLPCCs), and Octave Scale Cepstral Coefficients (OSCCs). We propose two other features, Twice-Iterated

Composite Fourier Transform Coefficients (TICFTCs) and Octave Scale TICFTCs (OSTICFTCs), to measure the harmonic spacing of the vocal and instrumental music.

The experiments for vocal/instrumental region that the detection reveal musically modified features, OSLPCs, OSLPCCs, OSCCs, and OCTICFTCs, are more robust than LPCs, LPCCs, MFCCs, and TICFTCs, respectively. Out of all the features we used, OSCCs are able to detect vocal/instrumental regions the most accurately. HMM is found to be more capable of modelling music regions with OSCCs than SVM or GMM. The result comparison shows that our method out performed the existing method by $20 \%$.

## 4. Semantic music cluster detection (i.e. Intro, Chorus, Verse, Middle eighth Bridge and Outro, detection $-4^{\text {th }}$ layer)

The $4^{\text {th }}$ and higher layers in the music structure pyramid describe the song structure and the semantic meaning(s), respectively. Popular song structure consists of semantically similar meaning clusters, namely Intro, Verse, Chorus, INST, Bridge, Middle eighth and Outro ( $4^{\text {th }}$ layer). Previous research focused only on the detection of choruses in a song (Goto 2003 [49], Bartsch and Wakefield 2001 [7]). However, to best of our knowledge, we couldn't find an approach which describes the detection of cluster other than the chorus semantic cluster. We conducted a survey of 220 popular songs to understand their song structures (see chapter 2.5). Based on our survey, we defined heuristic rules to detect these semantic clusters. Our framework can achieve $10 \%$ more chorus detection accuracy than the previous method.

Music structure analysis is useful for many music related applications. Our ideas on how this structural information can potentially be used in applications are discussed in chapter 7.

### 8.2 Future direction

After submitting this thesis, our immediate research focus will be on converting the proposed framework for music structure analysis to a real time system.

The current framework for music structure analysis is limited to popular music. One of the future directions will be to extend our framework to other genres such as Rock, Classical, Jazz etc. Semantic meaning(s) decoding is a very difficult problem and the current framework is limited to the identification of semantic music cluster in popular songs. One of our long-term goals is to formulate a generic procedure to decode the semantic meaning(s) of the music signals. The success of long-term music research is based on how well we can integrate domain knowledge of relevant communities such as musicology, psychology, and signal processing. In the future, we will work with these communities to synergize their research finding to music applications.

For long-term research goals, we would like to develop music related multimedia applications for different types of users such as mobile users.

## References

[1] Allen, D. (1967). Octave Discriminability of Musical and Non-musical Subjects. In Journal of the Psychonomic Science, 1967, Vol. 7, pp. 421-422.
[2] Alonso, M., Badeau, R., David, B. and Richard, G. (2003). Musical Tempo Estimation using Noise Subspace Projections. In Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). October 19-22, 2003, New Paltz, New York.
[3] Allen, P. E. and Dannenberg, R. B. (1990). Tracking Musical Beats in Real Time. In Proc. of the International Computer Music Conference (ICMA), Glasgow, 1990, pp. 140-143.
[4] Attneave, F. and Olson, R. (1971). Pitch as a Medium: A New Approach to Psychophysical Scaling. In American Journal of Psychology (Am J Psychol). 1971, Vol. 84, pp. 147-166.
[5] Bachem, A. (1950). A Tone Height and Tone Chroma as Two Different Pitch Qualities. In International Journal of Psychonomics (Acta Psychological), 1950, Vol. 7, pp. 80-88.
[6] Bachem, A. (1954). Time Factors in Relative and Absolute Pitch Determination. In Journal of the Acoustical Society of America (JASA), 1954, Vol. 26, pp. 751-753.
[7] Bartsch, M. A. and Wakefield, G. H. (2001). To Catch a Chorus: Using Chroma-based Representations for Audio Thumbnailing. In Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, New York. October 21-24, 2001.
[8] Bartsch, M. A. and Wakefield, G. H. (2004). Singing Voice Identification Using Spectral Envelope Estimation. In IEEE Transaction on Speech and Audio Processing, March 2004, Vol. 12, No. 2, pp. 100-109.
[9] Bello, J. P. and Sandler, M. B. (2003). Phase-Based Note Onset Detection for Music Signals. In Proc. International conference on Acoustics, Speech, and Signal processing (ICASSP), Hong Kong, April 6-10, 2003.
[10] Berenzweig, A. L. and Ellis, D. P. W. (2001). Location singing voice segments within music signals. In Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, New York. October 21-24, 2001, pp. 119-122.
[11] Berenzweig, A. L. and Ellis, D. P. W. and Lawrence, S. (2002). Using Voice Segmentations to Improve Artist Classification of Music. In Proc. of 22nd Audio Engineering Society's (AES-22) International Conference on Virtual, Synthetic and Entertainment Audio (ICVSEA), Espoo, Finland, June 15-17, 2002.
[12] Bharucha, J. J. and Stoeckig, K. (1986). Reaction Time and Musical Expectancy: Priming of Chords. In Journal of Experimental Psychology: Human Perception and Performance, 1986, Vol. 12, pp. 403-410.
[13] Bharucha, J. J. and Stoeckig, K. (1987). Priming of Chords: Spreading Activation or Overlapping Frequency Spectra? In Journal of Perception and Psychophysics, 1987, Vol. 41, No. 6, pp.519-524.
[14] Biasutti, M. (1997). Sharp Low-and High-Frequency Limits on Musical Chord Recognition. In Journal of Hearing Research, 1997, Vol. 105, pp. 7784.
[15] Brown, J. C. (1991). Calculation of a Constant Q Spectral Transform. In Journal of the Acoustical Society of America (JASA), January, 1991, Vol. 89, pp. 425-434.
[16] Brown, J. C. and Cooke, M. (1994). Perceptual Grouping of Musical Sounds: A Computational Model. In Journal of New Music Research, 1994, Vol. 23, pp. 107-132.
[17] Brown, J. C. (1999). Computer identification of musical instruments using pattern recognition with Capstral coefficients as features. In Journal of Acoustic Society America, March, 1999, Vol. 105, No. 3, pp. 1933-1941.
[18] Cemgil, A.T., Kappen, H.J., Desain, P.W.M., and Honing, H.J. (2001). On tempo tracking: Tempogram representation and Kalman filtering. In Journal of New Music Research, 2001, Vol.29, No.4, pp.259-273.
[19] Chai, W. and Vercoe, B. (2003). Music Thumbnailing via Structural Analysis. In Proc. ACM International conference on Multimedia (ACM MM), 2003, Berkeley, CA, USA, November 2-8. pp. 223-226.
[20] Chai, W. and Vercoe, B. (2001). Folk Music Classification Using Hidden Markov Models. In Proc. of $5^{\text {th }}$ International Conference on Applied Informatics, Eger, Hungary, January 28 -February 3, 2001.
[21] Collobert, R. and Bengio, S. (2001). SVMTorch: Support Vector Machines for Large-Scale Regression Problems. In Journal of Machine Learning Research, 2001, Vol. 1, pp. 143-160.
[22] Cooper, M. and Foote, J. (2002). Automatic Music Summarization via Similarity Analysis. In Proc. $3^{\text {rd }}$ International Symposium of Music Information Retrieval (ISMIR), Paris, France, October 13-17, 2002.
[23] Cosi, P., DePoli, G. and Prandoni, P. (1994). Timbre characterization with Mel- Cepstrum and neural nets. In Proc. of International Computer Music Conference (ICMC), Aarhus, Denmark, September 12-17, 1994, pp. 42-45.
[24] Dannenberg, R. B., Thom, B. and Waston, D. (1997). A Machine Learning Approach to Musical Style Recognition. In Proc. of International Computer Music Conference (ICMC), Thessaloniki, Greece, September 25 - 30, 1997.
[25] Dannenberg, R. B. and Hu, N. (2002). Discovering Musical structure in Audio Recordings. In Proc. $2^{\text {nd }}$ International Conference of Music and Intelligence (ICMAI), Edinburgh, Scotland, UK, September 12-14, 2002, pp. 43-57.
[26] Davies, M. E. P. and Plumbley, M. D. (2004). Causal Tempo Tracking of Audio. In Proc. of $5^{\text {th }}$ International Symposium/Conference of Music Information Retrieval (ISMIR), October 10-15, 2004, Barcelona, Spain.
[27] Deutsch, D. (1973). Octave Generalization of Specific Interference Effects in Memory for Tonal Pitch. In Journal Perception and Psychophysics, 1973, Vol. 23, pp. 227-230.
[28] Deutsch, D. (1999). The Psychology of Music. Academic Press, San Diego, Series in Cognition and Perception, $2^{\text {nd }}$ edition, 1999.
[29] Dixon, S. (2001). Automatic Extraction of Tempo and Beat from Expressive Performances. In Journal of New Music Research, 2001, Vol.30, No.1, pp.39-58.
[30] Dowling, W. J. (1978). Scale and Contour: Two Components of a Theory of Memory for Melodies. In Journal of Psychological Review, 1978, Vol. 85, pp. 341-354.
[31] Dowling, W. J. and Harwood, D. L. (1986). Music Cognition. Academic press, San Diego, Series in Cognition and Perception, 1986.
[32] Dubnov, S. and Rodet, X. (1998). Timbre Recognition with Combined Stationary and Temporal Features. In Proc. International Computer Music Conference (ICMC), Michigan, USA, October 1-6, 1998.
[33] Duda, R.O., Hart, P. E. and Stork, D. G. (2000). Pattern Classification. Second edition, John Wiley \& Sons publications, 2000.
[34] Duxburg, C., Sandler, M. and Davies, M. (2002). A Hybrid Approach to Musical Note Onset Detection. In Proc. of $5^{\text {th }}$ International Conference on Digital Audio Effects (DAFx-02), Hamburg, Germany, September 26-28, 2002.
[35] Eggink, J. and Brown, G. J. (2004). Extracting Melody Lines from Complex Audio. In Proc. of $5^{\text {th }}$ International Symposium/Conference of Music Information Retrieval (ISMIR), Barcelona, Spain, October 10-15, 2004.
[36] Eronen, A. and Klapuri, A. (2000). Musical Instrument Recognition Using Cepstral Coefficients and Temporal Features. In Proc. of International Conference on Acoustic, Speech and Signal Processing (ICASSP), Istanbul, Turkey, June 05-09, 2000.
[37] Everest, F. A. (2001). Master Handbook of Acoustics. McGraw-Hill, $4^{\text {th }}$ Edition, 2001.
[38] Foote, J., Cooper, M. and Girgensohn, A. (2002). Creating Music Video using Automatic Media Analysis. In Proc. International ACM Conference on Multimedia (ACM MM), Juan-les-Pins, France, December 1-6, 2002.
[39] Fujinaga, I. (1998). Machine Recognition of Timbre Using Steady-state Tone of Acoustic Musical Instruments. In Proc. International Computer Music Conference (ICMC), Michigan, USA, October 1-6, 1998, pp. 207-210.
[40] Fujishima, T. (1999). Real-time Chord Recognition of Musical Sounds: A System using Common Lisp Music. In Proc. of International Computer Music Conference (ICMC), 1999, Beijing, pp. 464-467.
[41] Gao, S., Maddage, N. C. and Lee, C. H. (2003). A Hidden Markov Model Based approach to Musical Segmentation and Identification. In Proc. IEEE Pacific-Rim Conference on Multimedia (PCM), 2003.
[42] Gao, S. and Lee, C. H. (2004). An Adaptive Learning Approach to Music Tempo and Beat Analysis. In Proc. of International Conference on Acoustic, Speech and Signal Processing (ICASSP), Montreal, Canada, May 17-21, 2004
[43] Gao, S., Lee, C. H. and Zhu, Y. (2004). An Unsupervised Learning Approach to Music Event Detection. In Proc. of IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, June 27-30, 2004.
[44] Ghias, A., Logan, J., Chamberlin, D. and Smith, B. C. (1995). Query By Humming: Musical Information Retrieval in an Audio Database. In $3^{r d} A C M$ International conference on Multimedia (ACM MM), San Francisco, California, USA, November 5-9, 1995, pp. 231-236.

Goldstein, J. L. (1973). An Optimum Processor Theory for the Central Formation of the Pitch of Complex Tones. In Journal of the Acoustical Society of America (JASA), 1973, Vol. 54, pp. 1496-1516.
[46] Goto, M. and Muraoka, Y. (1994). A Beat Tracking System for Acoustic Signals of Music. In Proc. $2^{\text {nd }}$ ACM International Conference on Multimedia, San Francisco, California, USA, October 15-20, 1994, pp. 365372.
[47] Gota, M. (2001). A Predominant F0 Estimation Method for CD Recordings: MAP Estimation using EM Algorithm for Adaptive Tone Models. In Proc. of International conference on Acoustics, Speech, and Signal processing (ICASSP), Sault lake city, Utah, May 7-11, 2001, pp. 3365-3368.

Goto, M. (2001). An Audio-based Real-time Beat Tracking System for Music With or Without Drum-sounds. In Journal of New Music Research, June 2001, Vol.30, No.2, pp.159-171.
[49] Goto, M (2003). A Chorus-Section Detecting Method for Musical Audio Signals. In Proc. International conference on Acoustics, Speech, and Signal processing (ICASSP), Hong Kong, April 6-10, 2003.
[50] Gouyon, F. and Herrera, P. and Cano, P. (2002). Pulse-Dependent Analyses of Percussive Music. In Proc. International Conference on Virtual, Synthetic and Entertainment Audio (AES22), Espoo, Finland, June 15-17, 2002.
[51] Han, K. P., Pank, Y. S., Jeon, S. G., Lee, G. C. and Ha, Y. H. (1998).Genre Classification System on TV Sound Signals Based on a Spectrogram Analysis. In IEEE Transaction on Consumer Electronics, 1998, Vol. 55, No. 1, pp.33-42.
[52] Houtgast, T. (1976). Sub-Harmonic Pitches of a Pure Tone at Low S/N Ratio. In Journal of the Acoustical Society of America (JASA), 1976, Vol.60, No. 2, pp. 405-409.
[53] Hartmann, W. (1993). On the Origin of the Enlarged Melodic Octaves. In Journal of the Acoustical Society of America (JASA), 1993, Vol. 93, pp. 3400-3409.
[54] Haykin, S. (1999). Neural Networks: A Comprehensive Foundation. $2^{\text {nd }}$ Edition, Prentice Hall, New Jersey, USA, 1999.
[55] Jensen, K. and Andersen, T. H. (2003). Real-time beat estimation using feature extraction. In Proc. of Computer Music Modelling and Retrieval Symposium, Lecture Notes in Computer Science, Springer Verlag, 2003, Vol. 2771, pp. 13-22.
[56] Jiang, D.N., Lu, L., Zhang, H. J., Tao, J. H. and Cai, L. H. (2002). Music Type Classification by Spectral Contrast Feature. In Proc. of IEEE International Conference on Multimedia and Expo (ICME), Lausanne, Switzerland, 2002.
[57] John, G. P. and Dimitris, G. M. (1995). Digital Signal Processing, Principles, Algorithms, and Applications, $3^{\text {rd }}$ Edition, Prentice Hall 1995.
[58] John, R.D., John, H. L. and John, G. P. (1999). Discrete-Time Processing of Speech Signals. IEEE Press September 1999.
[59] Kaminskyj, I. and Materka, A. (1995). Automatic Source Identification of Monophonic Musical Instrument Sounds. In Proc. IEEE International Conference on Neural Networks, Perth, Australia, Nov 27-Dec 1, 1995, pp. 189-194.
[60] Kashino, K. and Murase, H. (1998). Music Recognition using Note Transition Context. In Proc. of International conference on Acoustics, Speech, and Signal processing (ICASSP), Seattle, Washington, USA, May 12-15, 1998.
[61] Kauppinen, I. (2002). Audio Signal Restoration with Modern Digital Signal Processing Techniques. Ph.D. dissertation, Department of Physics, University of Turku, Turku, Finland, 2002.
[62] Kim, Y. K. and Brian, W. (2002). Singer Identification in Popular Music Recordings Using Voice Coding Features. In Proc. $3^{\text {rd }}$ International Symposium of Music Information Retrieval (ISMIR), Paris, France, October 13-17, 2002.
[63] Klapuri, A. P. (2003). Multiple Fundamental Frequency Estimation Based on Harmonicity and Spectral Smoothness. In IEEE Transaction on Speech and Audio Processing, November, 2003, Vol. 11, No. 6, pp. 804-816.
[64] Klapuri, A. P. (1999). Sound Onset Detection by Applying Psychoacoustic Knowledge. In Proc. of International conference on Acoustics, Speech, and Signal processing (ICASSP), Phoenix, Arizona, USA, March 15-19, 1999.
[65] Krishnaswamy, A. (2003). Application of Pitch Tracking to South Indian Classical Music. In Proc. of International conference on Acoustics, Speech, and Signal processing (ICASSP), Hong Kong, April 6-10, 2003.
[66] Krumhansl, C. L. (1979). The Psychological Representation of Musical Pitch in a Tonal Context. In Journal of Cognitive Psychology, 1979, Vol.11, No. 3, pp. 346-374.
[67] Laden, B. and Keefe, D. H. (1989). The Representation of Pitch in a Neural Net Model of Chord Classification. In Computer Music Journal, Winter 1989, Vol. 13, No. 4, pp. 12-26.
[68] Leung, T. W. and Ngo, C. W. (2004). ICA-FX Features for Classification of Singing Voice and Instrumental Sound. In Proc. International Conference on Pattern Recognition (ICPR), Cambridge, UK, August 23-26, 2004.
[69] Logan, B. and Chu, S. (2000). Music Summarization Using Key Phrases. In Proc. International Conference on Acoustics, Speech, and Signal processing (ICASSP), Orlando, USA, 2000.
[70] Lu, L. and Zhang, H. J. (2003). Automated Extraction of Music Snippets. In Proc. ACM International Conference on Multimedia (ACM MM), Berkeley, CA, USA 2003, pp. 140-147.
[71] Maddage, N. C., Xu, C. S., Lee, C. H., Kankanhalli, M. S. and Tian, Q. (2002). Statistical Analysis of Musical Instruments. In Proc. IEEE PacificRim Conference on Multimedia (PCM), Hsinchu, Taiwan, December 16-18, 2002, pp. 581-588.
[72] Maddage, N. C., Xu, C. S. and Wang, Y. (2003). A SVM-Based Classification Approach to Musical Audio. In Proc. International Symposium of Music Information Retrieval (ISMIR), Baltimore, Maryland, USA, October 26-30, 2003.
[73] Maddage, N. C., Wan, K., Xu, C. S. and Wang, Y. (2004). Singing Voice Detection using Composite Fourier Transform. In Proc. of IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, June 27-30, 2004.
[74] Maddage, N. C., Xu, C. S. and Wang, Y. (2004). Singer Identification Based on Vocal and Instrumental Models. In Proc. International Conference on Pattern Recognition (ICPR), Cambridge, UK, August 23-26, 2004.

Maddage, N. C., Xu, C. S., Kankanhalli, M. S. and Shao, Xi. (2004). Content-based Music Structure Analysis with Applications to Music Semantic Understanding. In Proc. International ACM Conference on Multimedia (ACM MM), New York, USA, October 10-16, 2004.
[76] Maddage, N. C., Xu, C. S., Shenoy, A. and Wang, Y. (2004). Semantic Region Detection in Acoustic Music Signals. In Proc. $5^{\text {th }}$ IEEE Pacific Rim Conference on Multimedia, Tokyo, Japan, Nov 30 - Dec. 3, 2004.
[77] Maddage, N. C. (2006). Automatic Structure Detection of Popular Music. In IEEE Multimedia Magazine. January-March 2006, Vol. 13, No.1, pp. 65-77.
[78] Makhoul, J. (1975). Spectral Linear Prediction: Properties and Applications. In EEE Transactions on Acoustics, Speech and Signal Processing, June 1975, Vol. ASSP-23, No. 3, pp. 283-296.
[79] Martin, K. D. (1999). Sound-Source Recognition: A Theory and Computational Model. Ph.D. dissertation, Massachusetts Institute of Technology (MIT), Media Lab, Cambridge, USA, June, 1999.
[80] Marques, J. (1999). An Automatic Annotation System for Audio Data Containing Music. Master's Thesis, Massachusetts Institute of Technology (MIT), Media Lab, Cambridge, USA, 1999.
[81] Matityaho, B and Furst, M. (1995). Neural Network Based Model for Classification of Music Type. In Proc. of $18^{\text {th }}$ Convention of Electrical and Electronic Engineers, Israel, March 7-9, 1995, pp.1-5.
[82] McKinney, M. F. and Delgutte, B. (1999). A Possible Neurophysiologic Basis of the Octave enlargement Effect. In Journal of the Acoustical Society of America (JASA), 1999, Vol. 106, No. 5, pp. 2679-2692.
[83] McNab, R.J., Smith, L.A., Witten, I.H., and Henderson, C.L. (2000). Tune Retrieval in the Multimedia Library. In Journal of Multimedia Tools and Applications, 2000, Vol. 10, No. 2-3, pp. 113-132.
[84] Miller, R. (1986). The Structure of Singing: System and Art in Vocal Technique. Wadsworth Group/Thomson Learning, Belmont California, USA, 1986.
[85] Moorer, J. A. (1975). On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer. Ph.D. dissertation, Department of Computer Science, Stanford University, 1975.
[86] Navarro, G. (2001). A guided tour to approximate string matching. In Proc. ACM Computing Surveys. March 2001, Vol. 33, NO.1, pp.31-88.
[87] Nwe, T. L. and Wang, Y. (2004). Automatic Detection of Vocal Segments in Popular Songs. In Proc. of $5^{\text {th }}$ International Symposium/Conference of Music Information Retrieval (ISMIR), Barcelona, Spain, October 10-15, 2004.
[88] Nwe, T. L., Shenoy, A. and Wang, Y. (2004). Singing Voice Detection in Popular Music. In Proc. International ACM Conference on Multimedia (ACM MM), New York, USA, October 10-16, 2004.
[89] Ohgushi, K. (1978). On the Role of Spatial and Temporal Cues in the Perception of the Pitch of Complex Tones. In Journal of the Acoustical Society of America (JASA), 1978, Vol. 64, pp. 764-771.
[90] Ohgushi, K. (1983). The Origin of Tonality and a Possible Explanation of the Octave Enlargement Phenomenon. In Journal of the Acoustical Society of America (JASA), 1983, Vol. 73. 1694-1700.
[91] Perkins, C., Hodson, O. and Hardman, V. (1998). A Survey of Packet Loss Recovery Techniques for Streaming Audio. In IEEE Network Magazine, September/October, 1998, pp. 40-48.
[92] Pikrakis, A., Antonopoulos, I. and Theodoridis, S. (2004). Music Meter and Tempo Tracking from Raw Polyphonic Audio. In Proc. of $5^{\text {th }}$ International

Symposium/Conference of Music Information Retrieval (ISMIR), Barcelona, Spain, October 10-15, 2004.
[93] Pye, D. (2000). Content-Based Methods for the management of Digital Music. In Proc. of International conference on Acoustics, Speech, and Signal processing (ICASSP), Istanbul, Turkey, June 05-09, 2000.
[94] Rabiner, L. R. and Juang, B. H. (1993). Fundamentals of Speech Recognition. Prentice-Hall, 1993.
[95] Reynolds, D. and Rose, R. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. In IEEE Transactions on Speech and Audio Processing, 1995, vol. 3, no. 1, pp. 72-83.
[96] Ritsma, R. J. (1967). Frequency Dominant in the Perception of the Pitch of Complex Sounds. In Journal of Acoustical Society of America. 1967, Vol. 42, No. 1, pp. 191-198.
[97] Rao, R. M. and Bopardikar, A.S. (1998). Wavelet Transforms: Introduction to Theory and Applications. Addison Wesley Longman, Inc., 1998.
[98] Rossing, T.D. (2001). Science of Percussion Instruments. World Scientific, Series in Popular Science, Vol. 3, 2001.
[99] Rossing, T. D., Moore, F. R. and Wheeler, P. A. (2001). Science of Sound, Addison Wesley, $3^{\text {rd }}$ edition, 2001.
[100] Rudiments and Theory of Music. The associated board of the royal schools of music, 14 Bedford Square, London, WC1B 3JG, 1949.
[101] Saitou, T., Unoki, M. and Akagi, M. (2002). Extraction of F0 Dynamic Characteristics and Developments of F0 Control Model in Singing Voice. In Proc. of the $8^{\text {th }}$ International Conference on Auditory Display, Kyoto, Japan, July $02-05,2002$.
[102] Scheirer, E. D. (1998). Tempo and Beat Analysis of Acoustic Music Signals. In Journal of Acoustical Society of America. January 1998, Vol. 103, No. 1, pp. 588-601.
[103] Scaringella, N. and Zoia, G. (2004). A Real-Time Beat Tracker for Unrestricted Audio Signals. In Proc. of the Conference of Sound and Music Computing (JIM/CIM), Paris, France, October 20-22, 2004.
[104] Sethares, W. A. and Staley, T. W. (2001). Meter and Periodicity in Music Performance. In Journal of New Music Research. June, 2001, Vol.30, No.2.
[105] Sethares, W. A., Morris, R. D. and Sethares, J. C. (2005). Beat Tracking of Musical Performances Using Low-Level Audio Features. In IEEE Transactions on Speech and Audio Processing, March 2005, Vol. 13, No. 2. pp. 275-285.
[106] Shan, M. K., Kuo, F. F. and Chen, M. F. (2002). Music Style Mining and Classification by Melody. In Proc. of IEEE International Conference on Multimedia and Expo (ICME), EPFL, Lausanne, Switzerland, August 26-29 2002.
[107] Shao, Xi., Xu, C.S. and Kankanhalli, M. S. (2004). Unsupervised Classification of Music Genre Using Hidden Markov Model. In Proc. of IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, June 27-30, 2004.
[108] Sheh, A. and Ellis, D. P. W. (2003). Chord Segmentation and Recognition using EM-Trained Hidden Markov Models. In Proc. $4^{\text {th }}$ International Symposium of Music Information Retrieval (ISMIR), Baltimore, Maryland, USA, October 26-30, 2003.
[109] Shenoy, A., Mohapatra, R. and Wang, Y. (2004). Key Detection of Acoustic Musical Signals. In Proc. of IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, June 27-30, 2004.
[110] Shepard, R. N. (1964). Circularity in Judgments of Relative Pitch. In Journal of the Acoustical Society of America (JASA), 1964, Vol. 36, pp. 2346-2353.
[111] Shifrin, J., Pardo, B., Meek, C. and Birmingham, W.P. (2002). HMM-Based Musical Query Retrieval. In Proc. of the $2^{\text {nd }}$ Joint International Conference (ACM \& IEEE-CS) on Digital Libraries (JCDL), Portland, Origone, USA, July 14-18, 2002, pp. 295-300.
[112] Sinha. R., Christos, P. and Chris, K. (2003). Loss Concealment for MultiChannel Streaming Audio. In Proc. of the $13^{\text {th }}$ International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV), Monterey, California, USA, June 1-3, 2003.
[113] Soltau, H. Schultz, T., Westphal, M. and Waibel, A. (1998). Recognition of Music Types. In Proc. of International conference on Acoustics, Speech, and Signal processing (ICASSP), Seattle, Washington, USA, May 12-15, 1998.
[114] Stevens, S.S., Volkmann, J. and Newman, E. B. (1937). A Scale for the Measurement of the Psychological Magnitude of Pitch. In Journal of the Acoustical Society of America (JASA), 1937, Vol. 8, pp. 185-190.
[115] Stevens, S.S. and Volkmann, J. (1940). The Relation of Pitch Frequency; a Relative Scale. In Journal of the Acoustical Society of America (JASA), 1940, Vol. 53, pp. 329-353.
[116] Su, B. and Jeng, S. (2001). Multi-Timbre Chord Classification using Wavelet Transform and Self-organized Map Neural Networks. In Proc. of

International conference on Acoustics, Speech, and Signal processing (ICASSP), Sault lake city, Utah, 2001, No. V, pp. 3377-3380.
[117] Sundberg, J. and Lindqvist, J. (1973). Musical Octaves and Pitch. In Journal of the Acoustical Society of America (JASA), 1973, Vol. 54, pp. 922-929.
[118] Sundberg, J. (1987) The Science of the Singing Voice. Northern Illinois University Press, Dekalb, Illinois.
[119] Szczerba, M. and Czyżewski, A. (2002). Pitch estimation Enhancement Employing Neural Network-Based Music Prediction. In Proc. 6th IASTED International Conference on Artificial Intelligence and Soft Computing (ASC), Banff, Canada, July 17-19, 2002.
[120] Ten Minute Master No 18: Song Structure. MUSIC TECH magazine, www.musictechmag.co.uk October, 2003, pp. 62-63.
[121] Terhardt, E. (1974). Pitch, Consonance and Harmony. In Journal of the Acoustical Society of America (JASA), May 1974, Vol. 55, No. 5, pp. 10611069.
[122] Terhardt, E. (1982). Pitch of Complex Signals According to Virtual-Pitch Theory: Tests, Examples, and Predictions. In Journal of the Acoustical Society of America (JASA), March, 1982, Vol. 71, No. 3, pp. 671-678.
[123] Tsai, W. H., Wang, H. M., Rodgers, D., Cheng, S. S. and Yu, H. M. (2003). Blind Clustering of Popular Music Recordings Based on Singer Voice Characteristics. In Proc. $4^{\text {th }}$ International Symposium of Music Information Retrieval (ISMIR), Baltimore, Maryland, USA, October 26-30, 2003.
[124] Tzanetakis, G., Essl, G. and Cook, P. (2001). Automatic Musical Genre Classification of Audio Signals. In Proc. $2^{\text {nd }}$ International Symposium of Music Information Retrieval (ISMIR), Bloomington, Indiana, October, 2001.
[125] Tzanetakis, G. (2002). Manipulation, Analysis and Retrieval Systems for Audio Signals. Ph.D. dissertation, Department of Computer Science, Princeton University, Princeton, New Jersey, USA, June, 2002.
[126] Tzanetakis, G., Cook, P. (2002). Music Genre Classification of Audio Signals. In IEEE Transactions on Speech and Audio Processing, July, 2002, Vol. 10, No. 5, pp. 293-302.
[127] Tzanetakis, G. (2004). Song-Specific Bootstrapping of Singing Voice Structure. In Proc. of IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, June 27-30, 2004.
[128] Uhle, C. and Herre, J. (2003). Estimation of Tempo, MicroTime and Time Signature from Percussive Music. In Proc. of the $6^{\text {th }}$ International Conference on Digital Audio Effects (DAFX-03), London, UK, September 811, 2003.
[129] Vapnik, V. (1998). Statistical Learning Theory. Wiley, 1998.
[130] Wah, B. W., Su, X. and Lin, D. (2000). A Survey of Error-Concealment Schemes for Real-Time Audio and Video Transmission over the Internet. In IEEE International Symposium on Multimedia Software Engineering, Taipei, Taiwan, December, 2000, pp. 17-24.
[131] Wang, Y. and Vilermo, M. (2001). A Compressed Domain Beat Detection Using MP3 Audio Bitstreams. In Proc. 9th ACM International Conference on Multimedia (ACM MM), Ottawa, Ontario, Canada, Sept. 30 - Oct.5, 2001.
[132] Wang, Y., Ahmaniemi, A., Isherwood, D. and Huang, W. (2003). Content Based UEP: A New Scheme for Packet Loss Recovery in Music Streaming. In Proc. ACM International conference on Multimedia (ACM MM), Berkeley, CA, USA, November 2-8, 2003.
[133] Wang, Y., Kan, M. Y., Nwe, T. L., Shenoy, A. and Yin, J. (2004). LyricAlly: Automatic Synchronization of Acoustic Music Signals and Textual Lyrics. In Proc. International ACM Conference on Multimedia (ACM MM), New York, USA, October 10-16, 2004.
[134] Wang, Y., Huang, W. and Korhonen, J. (2004). A Framework for Robust and Scalable Audio Streaming. In Proc. International ACM Conference on Multimedia (ACM MM), New York, USA, October 10-16, 2004.
[135] Ward, W. (1954). Subjective Musical Pitch. In Journal of the Acoustical Society of America (JASA), 1954, Vol. 26, pp 369-380.
[136] Williams, G. and Ellis, D. (1999). Speech/music Discrimination based on posterior probability features. In Proc. European Conference on Speech Communication and Technology (Eurospeech), Budapest, September, 1999.
[137] Wyse, L., Wang, Y. and Zhu, X. (2003). Application of a Content-Based Percussive Sound Synthesizer to Packet Loss Recovery in Music Streaming. In Proc. ACM International conference on Multimedia (ACM MM), Berkeley, CA, USA, November 2-8, 2003.
[138] Xu. C, Zhu. Y and Tian, Q (2002), Automatic Music Summarization Based on Temporal, Spectral and Cepstral Features, In Proc. IEEE International Conference on Multimedia and Expo, Lausanne, Switzerland, August 26-29, 2002, pp. 117-120.
[139] Xu, C. S., Maddage, N. C., Shao, Xi., Cao, F. and Tian, Q. (2003). Musical Genre Classification Using Support Vector Machines. In Proc. International Conference on Acoustics, Speech, and Signal processing (ICASSP), 2003, pp.V429-V432.
[140] Xu, C. S., Maddage, N. C. and Shao, Xi. (2005). Automatic Music Classification and Summarization. In IEEE Transaction on Speech and Audio Processing, May, 2005, Vol. 13, pp 441-450.
[141] Yoshioka, T., Kitahara, T., Komatani, K., Ogata, T. and Okuna, H. G. (2004). Automatic Chord Transcription with Concurrent Recognition of Chord Symbols and Boundaries. In Proc. of $5^{\text {th }}$ International Symposium/Conference of Music Information Retrieval (ISMIR), Barcelona, Spain, October 10-15, 2004.
[142] Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. and Woodland, P. (2002). The HTK Book (for HTK Version 3.2), Engineering department, Cambridge University, December, 2002.
[143] Zhu, Y. (2004). Content-Based Music Retrieval by Acoustic Query. Ph.D. dissertation, Department of Computer Science, National University of Singapore, October, 2004.
[144] Zhu, Y., Kankanhalli, M. S. and Gao, S. (2005). Music Key Detection for Musical Audio. In Proc. $11^{\text {th }}$ International Multimedia Modelling Conference (MMM), Melbourne, Australia, January 12-14, 2005.
[145] Zhang, T. and Kuo, C. C. J. (2001). Audio Content Analysis for Online Audiovisual Data Segmentation and Classification. In IEEE Transaction on Speech and Audio Processing, May 2001, Vol. 9, No. 4. pp 441-457.
[146] Zhang, T. (2003). Automatic Singer Identification. In Proc. The IEEE International Conference on Multimedia and Expo (ICME), Baltimore, Maryland, July 6-9, 2003.

## Appendix - A

The appendix -A highlights the relationship between principle component analysis (PCA) and singular value decomposition (SVD)

## Principle component analysis (PCA)

Principle component analysis is useful for transforming original feature vector ( X ) to another space $(\mathrm{Y})$ which gives maximum variance among the components in the vector. This idea is described in Figure A- 1.


Figure A- 1: Transformation of feature vector ' X ' to another space ' Y ' to find uncorrelated elements in the vector

Now we can write this linear transformation as:

$$
\begin{equation*}
Y=A^{T} X \tag{b-1}
\end{equation*}
$$

Let $\mu_{x}, \Sigma_{x}, \mu_{y}$ and $\Sigma_{y}$ are the mean vector and covariance matrix of vector X and Y respectively and their relationships are shown below.

$$
\begin{equation*}
\mu_{y}=A^{T} \mu_{x} \quad \text { and } \quad \Sigma_{y}=A^{T} \Sigma_{x} A \tag{b-2}
\end{equation*}
$$

To remove the mutual correlation between the elements of $\mathrm{X}, \Sigma_{y}$ must be diagonal. Thus matrix A must be the similarity transform of $\Sigma_{x}$ and columns of A are eigenvectors of $\Sigma_{x}$ (Duda et al [33]). Then the diagonal elements in $\Sigma_{y}$ are eigenvalues of $\Sigma_{x}$. Reordering the diagonal values in $\Sigma_{y}$ in descending order we can find the elements which are highly uncorrelated in Y.

## Singular Value Decomposition (SVD)

Any $m$ x $n$ matrix A can be decomposed into:

$$
\begin{equation*}
A=U \Sigma V^{T} \tag{b-3}
\end{equation*}
$$

$\mathrm{U}: \mathrm{mxm}$ - columns are left singular vectors - eigenvectors of $\mathrm{AA}^{\mathrm{T}}$
$\Sigma: m x n-$ diagonal - singular values - square roots of eigenvalues of $A^{T} A$ or $A A^{T}$
$V: n x n-$ columns are right singular vectors- eigenvectors of $A^{T} A$
Assume $\mathrm{AA}^{\mathrm{T}}$ the covariance matrix of A is $\Sigma_{x}$ in PCA. Then we know $\Sigma_{y}$ (diagonal matrix) represents the eigenvalues of $\Sigma_{x}$ and $\Sigma_{y}$ has the maximum variance. Then equation (b-4) describes the relation ship between $\Sigma_{y}$ and $\Sigma$.

$$
\begin{equation*}
\Sigma_{y}=[\Sigma]^{2} \tag{b-4}
\end{equation*}
$$

Thus the singular values can be used as a measurement to assess how uncorrelated the original data (i.e. matrix A). Higher the singular values describe higher uncorrelation between elements in the matrix A. SVD operation is useful for data compression and filtering noise in the data set. Typically small singular values in matrix " $\Sigma$ " are caused by noise. Singular values are diagonally set in the matrix $\Sigma$ in descending order.


[^0]:    ${ }^{1}$ Strong beats are the high energy impulses generated from bass drums and side drums and cymbals. They are commonly played on accents (see Figure 2-8)
    ${ }^{2}$ Weak beats are generated from low energy impulses from instruments such as snares.

[^1]:    ${ }^{3}$ Dodecaphonic notes are the twelve tones (C, C $\left.{ }^{\#}, D, D \#, E, F, F^{\#}, G, G^{\#}, A, A^{\#}, B\right)$ in an octave.

[^2]:    ${ }^{4}$ Root is the note from which the chord originates [100]. For example, note C is the root note of C major chord.

[^3]:    ${ }^{5}$ One fragment $=15$ seconds

[^4]:    ${ }^{6}$ Since meter of the song is assumed to be $4 / 4$, inter- beat interval is equal to quarter note length.

[^5]:    ${ }^{7}$ Complex tone is a mixture of pitches. A music chord, which is generated playing several notes at same time, is considered as a complex music tone.

[^6]:    ${ }^{8}$ Structured signal has a F0 and its harmonics are spread all over the spectrum

