CHAPTER 2
LITERATURE SURVEY

2.1 Introduction

Audio features play a very important role in computer based listening system and pattern recognition, especially in describing the contents of the audio clips. This chapter reviews the work carried out in the area of feature extractions and classification methods used for audio signal classification. Literature survey is divided into three sections based on three important application areas of audio classification considered in this investigation. First, basic papers on content based audio analysis are surveyed. Second, a review of methods used for music genre classification, transcription based literature is presented. Third, papers based on speaker recognition are studied and analyzed. This work, published in various articles in international and national journals, at different conferences, in standard books/ titles, in society letters etc. has been presented with birds eye-view and motivation for the present investigation has been concluded from this literature survey.

2.2 Review of Literature related to Content Based Audio Classification

J. Saunders [1] developed a technique which is successful at discriminating speech from music on broadcast FM radio. The computational simplicity of the approach could lend itself to wide application including the ability to automatically change channels when commercials appear. The algorithm provides the capability to robustly distinguish the two classes speech and music and runs easily in real time. The strict use of time-domain
features avoids the need for an FFT processor to compute spectral features. Given that this algorithm has only been developed in an exploratory fashion, it is expected that improved performance beyond that reported is possible.

E. Scheirer et al. [2] reported on construction of a real-time computer system capable of distinguishing speech signals from music signals over a wide range of digital audio input. They have examined 13 features intended to measure conceptually distinct properties of speech and/or music signals, and combined them in several multidimensional classification frameworks. They have provided extensive data on system performance and the cross-validated training/test setup used to evaluate the system. For the datasets currently in use, the best classifier classifies with 5.8% error on a frame-by-frame basis, and 1.4% error when integrating long (2.4 second) segments of sound.

J. Foote [3] presented a method of rapidly determining the characteristics of audio samples, using a supervised tree-based vector quantizer trained to maximize mutual information (MMI). Such a measure has proved successful for talker identification, and the extension from speech to general audio, such as music, is straightforward. A classifier that distinguishes speech from music and non-vocal sounds is presented, as well as experimental results showing how perfect classification accuracy may be achieved on a small corpus using substantially less than two seconds per test audio file. The techniques presented here may be extended to other applications and domains, such as audio retrieval-by-similarity, musical genre classification, and automatic segmentation of continuous audio.
M. J. Carey et al. [4] examined the discrimination achieved by several different features using common training and test sets and the same classifier. The database assembled for these tests includes speech from thirteen languages and music from all over the world. In each case the distributions in the feature space were modeled by a Gaussian Mixture Model (GMM). Experiments were carried out on four types of feature, amplitude, cepstra, pitch and zero crossings. In each case the derivative of the feature was also used and found to improve performance. The best performance resulted from using the cepstra and delta cepstra which gave an equal error rate (EER) of 1.2%. This was closely followed by normalized amplitude and delta amplitude. This however used a much less complex model. The pitch and delta pitch gave an EER of 4% which was better than the zero crossing which produced an EER of 6%.

T.Zang et al. [5] presented a hierarchical system for audio classification and retrieval based on audio content analysis. The system consists of three stages. The first stage is called the coarse-level audio classification and segmentation where audio recordings are classified and segmented into speech, music, several types of environmental sounds, and silence, based on morphological and statistical analysis of temporal curves of short time features of audio signals. In the second stage, environmental sounds are further classified into final classes such as applause, rain, birds sound, etc. This fine-level classification is based on time-frequency analysis of audio signals and use of hidden Markov model (HMM) for classification.
K. El-Maleh, et al. [6] presented results of combining the line spectral frequencies (LSFs) and zero-crossing-based features for frame-level narrowband speech/music discrimination. The classification results for different types of music and speech show the good discriminating power of these features. The classification algorithms operate using only a frame delay of 20 ms, making them suitable for real-time multimedia applications. KNN classifier and Quadratic Gaussian Classifier (QCG) are used for further classification.

L. Lu et al. [7] presented a study of audio content analysis for classification and segmentation, in which an audio stream is segmented according to audio type or speaker identity. They proposed a robust approach that is capable of classifying and segmenting an audio stream into speech, music, environment sound, and silence. Audio classification is processed in two steps, which makes it suitable for different applications. The first step of the classification is speech and non-speech discrimination. In this step, a novel algorithm based on K-nearest-neighbor (KNN) and linear spectral pairs-vector quantization (LSP-VQ) is developed. The second step further divides non-speech class into music, environment sounds, and silence with a rule-based classification scheme. They have used a set of features such as the noise frame ratio and band periodicity, apart from the traditional features, as an input to an audio classifier.

Bugatti et al. [8] compared between two different techniques for speech/music discrimination. The first method is based on zero crossing rate and Bayesian classification. It is very simple from a computational point of view, and gives good
results in case of pure music or speech. The simulation results show that some performance degradation arises when the music segment contains also some speech superimposed on music, or strong rhythmic components. To overcome these problems, they proposed a second method that uses more features, and is based on neural networks (specifically a multi-layer Perceptron). In this case they obtained better performance, at the expense of a limited growth in the computational complexity.

Costas Panagiotakis et al. [9] dealt with the characterization of an audio signal which may be a part of a larger audiovisual system. They developed a system for segmentation of audio signal and then classification into one of the two main categories: speech or music. Amongst the systems requirements are its processing speed and ability to function in a real time environment with small responding delays. Because of restriction to two classes the characteristics that are extracted are considerably reduced and moreover required computations are straightforward. Segmentation is based on mean signal amplitude distribution where as classification utilizes additional characteristic related to frequency. The resulted reported shows correct classification accuracy about 95%

W. H. Abdulla et al. [10] separated the speaker datasets based on the gender to build gender dependent hidden Markov model for each word. Average pitch frequency of the speaker was used as the gender separation criterion. Speaker attributed variability are undesirable in speaker independent speech recognition systems. The gender of the speaker is one of the influential sources of this variability. Experimental evaluation
shows significant improvement in word recognition accuracy over the gender independent method with a slight increase in the processing computation.

_Dongge Li et al._ [11] addressed the problem of continuous general audio data for content based retrieval, and described a scheme to classify audio segments into seven categories. They have used Mel Frequency Cepstral Coefficient (MFCC) and Linear Predictive Coefficients (LPC) as feature vectors. The system provides 90% accuracy and is much faster than the playing rate.

_Michael Cowling et al._ [12] presented a comprehensive comparative study of artificial neural network, learning vector quantization and dynamic time warping classification techniques combined with stationary/non-stationary feature extraction for environmental sound recognition. Results shows 70% recognition using MFCC or continuous wavelet transform with dynamic time warping.

_C. J. C. Burges_ [13] explained the maximal margin separating hyper plane between two classes of data. It gives details of the fundamentals of Support Vector Machine like the kernel function which maps the input points to a high dimensional space. Multiclass SVM is also explained. There are two approaches to achieve this, “one versus all” and “one versus one”. The tutorial starts with an overview of the concepts of VC dimension and structural risk minimization. Linear SVMs for separable and non-separable data, working through a non-trivial example in detail are nicely described. They have also described how support vector training can be practically implemented.
Vladimir N. Vapnik [14] presented an overview of statistical learning theory including both theoretical and algorithmic aspects. The goal of this overview was to demonstrate how the abstract learning theory established conditions for generalization which are more general than those discussed in classical statistical paradigms and how the understanding of these conditions inspired new algorithmic approaches to function estimation problems. A more detailed overview of the theory (without proofs) can be found in Vapnik (1995). In Vapnik (1998) one can find detailed description of the theory (including proofs).

S. Sathiya Keerthi et al. [15] enlisted the reasons of the popularity of the Gaussian RBF kernel in SVM classification. Model selection in this class of SVMs involves two hyper parameters: the penalty parameter $C$ and the kernel width gamma. They have analyzed the behavior of the SVM classifier when the hyper-parameters take very small or very large values and provided a thorough graphical explanation of the over-fitting and under-fitting problem in SVM classifiers. The results help in understanding the hyper-parameter space that leads to an efficient heuristic method of searching for hyper parameter values with small generalization errors. The analysis also indicates that if complete model selection using the Gaussian kernel has been conducted; there is no need to consider linear SVM.

Chih-Wei Hsu et al. [16] proposed a comprehensive literature on support vector machine. The concept of cross-validation has been explained in brief, wherein, SVM datasets are partitioned into equal parts and the model so created is validated with each
data part. The use of the python program ‘grid.py’ in finding out efficient model parameters has elaborated in detail with examples. The decision of usage of linear kernel or RBF kernel in model creation has been explained considering different instances. SVM is a popular technique for classification. However, beginners who are not familiar with SVM often get unsatisfactory results since they miss some easy but significant steps. In this guide, they proposed a simple procedure, which usually gives reasonable results.

Guodong Guo, [17] used Support vector Machines with binary tree recognition strategy to tackle the audio classification problem. They have compared the SVMs based classification with other popular approaches. The potential of SVMs on a multiclass audio database is illustrated. For audio retrieval, they have proposed a new metric called distance-from-boundary. When a query audio is given, the system first finds a boundary inside which the query pattern is located. Then, all the audio patterns in the database are sorted by their distances to this boundary. All boundaries are learned by the SVMs and stored together with the audio database. Experimental comparisons for audio retrieval are presented to show the superiority of this novel metric to other similarity measures.

Vikramjit Mitra et al. [18] proposes a set of audio content features and parallel Neural Network architecture that addresses the task of automated content based audio classification. Feature set based on signal periodicity, beat information, sub-band energy, MFCC and wavelet transforms are proposed and each of the feature sets are individually analyzed for their pertinence in the proposed task.
\textit{J shirazi et al.} [19] presented a new feature set based on sinusoidal modeling of audio signal. Duration of the longest frequency track is used as a measure of harmony. The performance of this sinusoidal model feature is evaluated through classification of audio into speech and music using GMM and SVM classifiers. They have achieved a classification accuracy of 94.32\% using feature extracted from one second segment of signal.

\textit{Zhong –Hua Fu et al.} [20] presented two noise robust features namely Average Pitch Density (APD) and Relative Tonal Power Density (RTPD) for speech/music discrimination in real time telecommunication. APD refers to the differences in tone characteristics of music and speech signals and RTPD focuses on the distinct properties of percussion instruments. These features are compared with several state of the art features and found to achieve significant robustness.

\textbf{2.3 Survey of Literature related to Music Classification/Transcription}

\textit{Siddharth S Malu et al.} [21] discussed about the acoustics of the Indian drum and its harmonic nature. It clearly explains the difference between the Indian Tabla and Western drums. A musical sound has its own definite wave patterns, so that any arbitrary sound cannot be called music. Music has the pitch and timber as two important attributes. The notes produced by musical instruments consist of sine waves of one fundamental frequency and of higher frequencies called overtones. For the resultant waveform to be periodic, the overtones have to be integral multiples of the fundamental frequency.
Overtones that are integral multiples of the fundamental frequency are called harmonics. Unless a majority of overtones in a particular sound are harmonic, the waveform will not be periodic and will not have a discernible pitch. Thus for a note to sound musical, a majority of the overtones must be harmonic. They have shown that Indian Tabla is harmonic where as its counterpart western drum is aharmonic.

*E Schierer* [22] described real time beat tracking system for audio signal with music. A method is presented for using a small number of band-pass filters and banks of parallel comb filters to analyze the tempo of, and extract the beat from, musical signals of arbitrary polyphonic complexity and containing arbitrary timbres. This analysis is performed causally, and can be used predictively to guess when beats will occur in the future. Results in a short validation experiment demonstrate that the performance of the algorithm is similar to the performance of human listeners in a variety of musical situations. Aspects of the algorithm are discussed in relation to previous high-level cognitive models of beat tracking.

*G. Tzanetakis et al.* [23] described a beat histogram calculation method to find the beat strength which can be used to find tempo of music clip. They explored the automatic classification of audio signals into a hierarchy of musical genres. They proposed three feature sets for representing timbral texture, rhythmic content and pitch content. The performance and relative importance of the proposed features is investigated by training statistical pattern recognition classifiers using real-world audio collections. Using the
proposed feature sets, classification of 61% for ten musical genres is achieved. This result is comparable to results reported for human musical genre classification.

**J Paulas et al.** [24] explored a dynamic approach for measuring the similarity of rhythmic patterns. The patterns were represented as acoustic signals, and were not assumed to have been performed with similar sound sets. A probabilistic musical meter estimation process was described, which segments a continuous musical signal into patterns. As a side-product, the method outputs beat, and measure lengths. A subsequent process performs the actual similarity measurements. Acoustic features are extracted which model the fluctuation of loudness and brightness within the pattern, and dynamic time warping is then applied to align the patterns to be compared. In simulations, the system behaved consistently by assigning high similarity measures to similar musical rhythms, even when performed using different sound sets.

**J Jensen et al.** [25] introduced recently a representation for rhythmic patterns that is insensitive to minor tempo deviations and that has well defined behavior for larger changes in tempo. They have combined the representation with a Euclidean distance measure and compared it to other systems in a classification task of ballroom music.

**Klapuri A.** [26] described sound onset detection by applying psychoacoustic knowledge which propounds the difference of the log spectral power in bands as a more psychoacoustically relevant feature related to the discrimination of intensity. He originally proposed an onset detection model combining detection in multiple bands where the
salience of onsets is rated by a loudness summation based on the Moore, Glasberg and Baer loudness model. His most recent onset detection scheme generalizes the logarithmic compression, using the same analysis frontend as a recent beat induction model.

Stephen Hainsworth [27] discussed onset detection in musical audio signals and presented an equivalent formulation in the context of spotting harmonic content change, using a 4096 point FFT with a restriction of contributing bands to those in the range 30Hz-5 KHz. The detection function may be formed from the direct output of a loudness model, or a first order difference of one to enhance change detection.

Ajay Kapur et al. [28] described an automatic system for detecting and transcribing low and medium-high frequency drum events from audio signals. Content-based analysis of music can help manage the increasing amounts of music information available digitally and is becoming an important part of multimedia research. The use of drums and percussive sounds is pervasive to popular and world music. Two different sub-band front-ends are utilized. The first is based on band-pass filters and the second is based on wavelet analysis. Experimental results utilizing music, drum loops and Indian tabla thekas as signals are provided. The proposed system can be used as a preprocessing step for rhythm-based music classification and retrieval. In addition it can be used for pedagogical purposes.

O. Gillet et al. [29] proposed several methods for drum loops transcription where the drums signals dataset reflects the variability encountered in modern audio recordings
(real and natural drum kits, audio effects). The approaches described were based on Hidden Markov Models (HMM) and Support Vector Machines (SVM). Promising results were obtained with an 83.9% correct recognition rate. Recent efforts in audio indexing and retrieval in music databases mostly focus on melody. For polyphonic music signals, specific approaches are needed for systems dealing with percussive audio signals such as those produced by drums or tabla. Most studies of drum signals transcription focus on sounds taken in isolation.

K. Yoshii, et al. [30] described a system for the automatic description of drum sounds for real world musical audio signals. They proposed template adaption and template matching methods for identification of drum sound variation problem. Experimental results showed that the accuracy of identifying bass and snare drums in popular music was about 90%.

N.J. Hunt, et al. [31] described an exploratory implementation of a syllable based recognizer. Continuous speech was first divided into syllabic unite, and the units are then matched against syllable templates using a dynamic programming algorithm. A hierarchical transition network is used to limit the syllable search to possible continuations of the current partial sentence hypotheses. Competing hypotheses are pruned by a beam search. Experiments are reported on automatic recognition of English sentences with a 70 word vocabulary and restricted syntax produced by one male speaker. 85% of the sentences were correctly recognized. A method of scaling the distance measures used in the syllable matching is described. This scaling takes into account
variability in syllable production, both as a function of position within the syllable and as a function of the various spectral parameters being used.

2.4 Review of Literature related on Speaker Recognition

*Joseph P Campbell, Jr.* [32] a resourceful tutorial on Speaker Recognition explains the basic steps involved in speech processing and gives a detailed explanation about automatic speaker recognition. Automatic speaker recognition is the use of a machine to recognize a person from a spoken phrase. These systems can operate in two modes: to identify a particular person or to verify a person’s claimed identity. Speech processing and the basic components of automatic speaker recognition systems are shown and design tradeoffs are discussed. A speaker recognition system that uses an information-theoretic shape measure and Linear Spectral Pairs frequency features to discriminate between speakers is presented. This system yielded 98% accuracy for closed-set speaker identification.

*Y. Linde, et al.* [33] presented an efficient and intuitive algorithm for the design of vector quantizers based either on a known probabilistic model or on a long training sequence of data. The basic properties of the algorithm are discussed and demonstrated by examples. Quite general distortion measures and long block-lengths are allowed, as exemplified by the design of parameter vector quantizers of ten dimensional vectors arising in Linear Predictive Coded speech compression with a complicated distortion measure arising in LPC analysis that does not depend only on the error vector.
Md. Zulfiquar et al. [34] used the well-known Mel-frequency Cepstrum Coefficient (MFCC) as a feature for designing a Bengali text dependent speaker identification system. The extracted speech features (MFCC’s) of a speaker are quantized to a number of centroids using LBG algorithm. These centroids constitute the codebook of that speaker. Speakers uttered different Bengali words once in a training session and once in a testing session later. The quantization distance between the MFCC’s of each speaker in training phase to the centroids of individual speaker in testing phase is measured and the speaker is identified according to the minimum quantization distance. The resulting identification accuracy lies between 70-85%.

Wei Han et al. [35] introduced a new algorithm of extracting MFCC for speech recognition. This new algorithm differed from the conventional approach mainly in the windowing and overlap stages. This algorithm was computationally efficient and was suitable for hardware implementation. The new algorithm reduces the computation power by 53% compared to the conventional algorithm. Simulation results indicate the new algorithm has a recognition accuracy of 93 %. There is only a 1.5% reduction in recognition accuracy compared to the conventional MFCC extraction algorithm, which has an accuracy of 94.5%. However, the number of logic gates required to implement the new algorithm is about half of the MFCC algorithm, which makes the algorithm very efficient for hardware implementation.

Sangeeta Biswas et al. [36] presented, a speaker identification system using cepstral based speech features with Discrete Hidden Markov Model (DHMM). The features
represented by the speech signal are potentially characterized by the cepstral coefficients. The commonly used cepstral based features; mel- frequency cepstral coefficient (MFCC), linear predictive cepstral coefficient (LPCC) and real cepstral coefficient (RCC) are employed with DHMM in the speaker identification system. The performances of the proposed method are compared with respect to each of the three feature spaces. The experimental results show that the identification accuracy with MFCC is superior to both of LPCC and RCC. The highest accuracy achieved was 87.5% using MFCC in the case of text-dependent speaker recognition system.

*Makhoul John* [37] presented the theory of linear prediction in the analysis of discrete signal. The signal is modeled as a linear combination of past values and present and past value of hypothetical input to the output whose output is given signal. In the frequency domain this is equivalent to modeling the signal spectrum by a pole zero spectrum. The major part of this paper is devoted to all-pole models. The model parameters are obtained by a least square analysis in time domain. The advantages and disadvantages of least square error criteria are also discussed.

*Talal Amin et al.* [38] Speech Recognition is a technology enabling human interaction with machines. This paper describes an isolated word, speaker dependent speech recognition system capable of recognizing spoken words at sufficiently high accuracy. This technology can serve as a means of data inter operability and distribution by allowing a mobile user to retrieve information from the data networks using client server architecture. The satellite system can be used as a wireless medium for accessing the data
network. The system has been tested and verified on MATLAB as well as the TMS320 C6713 DSK with an overall accuracy exceeding 90%.

Debadatta P et al. [39] demonstrated the feasibility of excitation source information obtained by non parametric vector quantization for speaker recognition task. Linear Prediction residual is used as representation of excitation source information. Combination of LPC and MFCC is used as feature vectors for the proposed Speaker recognition system. The overall accuracy of the system for 30 speakers of TIMIT data base is 98.33%

Yegnanarayana B and Sharat Reddy, et al. [40] presented the effectiveness of the features extracted from the source and the system components of speech production process for the purpose of speaker recognition. The source and system components are derived using linear prediction analysis of short segments of speech. The source component is LP residual derived from the signal, and system component is a set of weighted linear predictive coefficients. The features are captured implicitly by feed forward auto associative neural network (AANN). Two separate speaker models are derived by training two AANN models using feature vectors corresponding to source and system components. A speaker recognition system of 20 speakers is built and tested using both the models to evaluate the performance of the source and system features. The study demonstrated the complementary nature of the two components.
Yegnarayana B and Prasanna S R M, et al. [41] proposed a text dependent speaker verification system which uses different type of information for making decision regarding the identity of claimed speaker. The base line system used is Dynamic Time Warping (DTW) technique for matching. Detection for the location of the end point is crucial for the performance of DTW based template matching. A method based on vowel onset point (VOP) is proposed locating the end point of an utterance. The proposed method for speaker verification used the suprasegmental and source features, besides the spectral features. The suprasegmental features such as pitch and duration are extracted from the warping path information of the DTW algorithm. Features of the excitation source, extracted using neural network models are also used for text-dependent speaker verification system.

2.5 Summary of Literature Survey

1. **Content Based Audio Classification:**

   As apparent from literature survey [1-20], intensive studies have been conducted on audio classification and segmentation by employing different features and classifiers. One of the more common audio signal classification problems tackled recently is that of speech-music classification. Due to a large variety in multimedia applications, multiclass audio classification has become necessary for content based storage and retrieval of audio signals. Recently Support Vector Machines have been popularly used as a reliable learning algorithm for pattern classification. Their appeal lies in their strong connection to the underlying statistical learning theory, in particular the theory of Structural Risk Minimization. In many problems across different applications, SVMs have been shown to
perform much better than other non-linear classifiers such as artificial neural networks. Content based audio classification and retrieval is a challenging research topic. The topic if researched and developed properly will help in new and exciting full-fledged applications which will be beneficial to the technology development, in general.

2. **Music Classification and Transcription:**

From the references [21-31] it is clear that most of the work done in music genre classification is for western music and instruments. The rhythmic features like tempo and beat histogram are analyzed for western drums. To the best of my knowledge, there has been little research in feature extraction and classification with explicit goal of classifying *Tabla-Taals* and *bols* in a widely used accompanying instrument in most of the Indian classical music recitals.

3. **Speaker Recognition:**

From the references cited in literature survey [32-41], MFCC as a feature vector and VQ codebook as a speaker database is widely used for speaker verification and identification systems. The computation time is a function of number of speakers. There is a scope for improvement in computational time by optimum selection of type and dimensions of feature vectors and the codebook size. Also the performance of the system needs to be improved in noisy conditions.