Indexing and Retrieval of Music using Gaussian Mixture Model Techniques

R. Thiruvengatanadhan
Assistant Professor
Department of CSE
Annamalai University

P. Dhanalakshmi
Associate Professor
Department of CSE
Annamalai University

ABSTRACT

Audio processing systems have taken gigantic leaps in everyday life of most people in developed countries. The technologies are getting entrenched in providing entertainment to consumers. Digital audio techniques have now achieved domination in audio delivery with CD players, internet radio, mp3 players and iPods being the systems of choice in many cases. With the huge growth of the digital music databases people begin to realize the importance of effectively managing music databases relying on music content analysis. The goal of music indexing and retrieval system is to provide the user with capabilities to index and retrieve the music data in an efficient manner. For efficient music retrieval, some sort of music similarity measure is desirable. In this paper, we propose a method for indexing and retrieval of the music using Gaussian mixture models. Acoustic features namely MFCC, chromagram, tempogram and MPEG-7 features are used to create the index. Retrieval is based on the highest probability density function and the experimental analysis shows that the rate of average number of clips retrieved for each query is 5 clips.

Keywords

Acoustic Feature Extraction, Indexing and Retrieval, Gaussian Mixture Model and Probability density function

1. INTRODUCTION

Audio content analysis has been an extremely active research topic. Throughout the 1990's, with the appearance of digital video and audio, analysis on the audio and video retrieval has become equally necessary. Retrieval of multimedia system information is totally different from retrieval of structured information. Music audio retrieval is normally done by annotating the media with text and uses text based retrieval systems to perform music retrieval [1]. But when the music information is voluminous, text based annotation becomes a tedious job. However, audio data have enormous information and expressing such information using text may not be appropriate [2].

2. MUSIC INFORMATION RETRIEVAL

There has been a rapid growth of Music Information Retrieval (MIR) in the field of audio processing. Search engines like Google facilitate in audio retrieval for music fans and the reputations of music have also shown exponential growth [3]. In World Wide Web, some websites were professionally maintained to provide or satisfy the requirement of music fans [4]. However, these search engines were extremely simple and they were developed on the basis of text query. Moreover, these systems realize difficulties in search of similar music or songs [5]. Inspite of explosive growth of multimedia system information, most of the traditional approaches follow the textual information (file name, title, composer or subject classification) for retrieving music data. Due to their incompleteness in textual information, there are several difficulties in satisfying specific needs of applications.

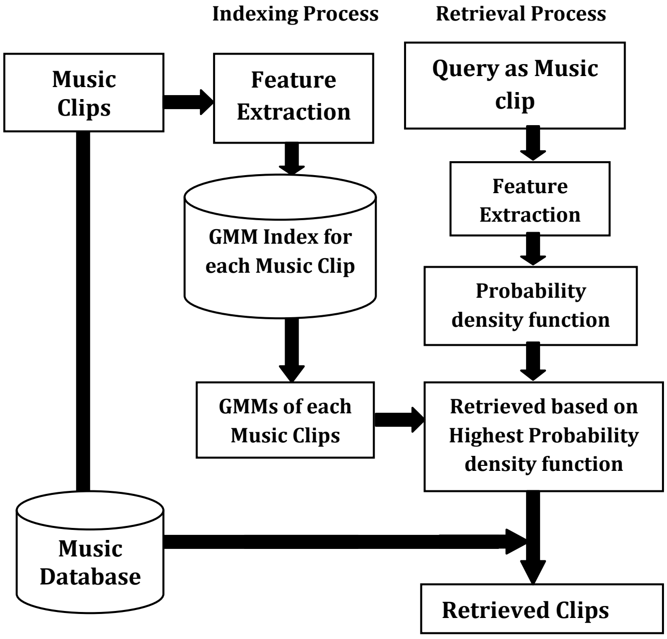

Fig 1: Proposed Methods for Music Indexing and Retrieval System.

Figure 1 shows the proposed method for music indexing and retrieval system. Recent research interest includes Relevance Feedback (RF) method for reforming the query system. RF with evolutionary based genetic algorithm is used to retrieve the musical information. In music retrieval field, these systems are still at initial stages [6]. Music audio retrieval using query-by-example takes recorded audio as input, and automatically retrieves documents from a given music collection containing the music as a part or similar to it [7].

3. ACOUSTIC FEATURE EXTRACTION

Music follows certain rules which are highly structured and it provides strong regularities, whereas music signals are random and chaotic [8]. In these work, acoustic features namely MFCC, tempogram, chromagram and MPEG-7 features are extracted for the indexing and retrieval.

3.1 Mel-Frequency Cepstral Coefficients

Mel-Frequency Cepstral Coefficients (MFCCs) are short-term spectral features which are used in various areas of music processing [9]. The effectiveness of mel-frequency cepstrum in music structures and modelling the subjective pitch and content of frequency has proved to be very high. MFCCs are highly effective in modelling the information of pitch in the field of audio processing as a short-term spectral feature [10].

The Mel-frequency cepstrum has proven to be highly effective in recognizing the structure of music signals and in modeling the subjective pitch and frequency content of audio signals [11]. The MFCCs have been applied in a range of audio mining tasks, and have shown good performance compared to other features. MFCCs are computed by various authors in different methods. It computes the cepstral coefficients along with delta cepstral energy and power spectrum deviation which results in 26 dimensional features. The low order MFCCs contains information of the slowly changing spectral envelope while the higher order MFCCs explains the fast variations of the envelope [12].

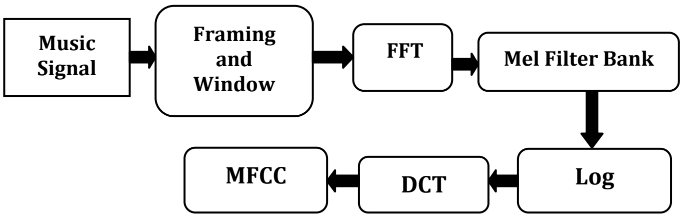

Fig 2: Extraction of MFCC from Audio Signal.

MFCCs are based on the known variation of the human ears critical bandwidths with frequency. The filters are spaced linearly at low frequencies and logarithmically at high frequencies to capture the phonetically important characteristics of speech and audio [13]. To obtain MFCCs, the audio signals are segmented and windowed into short frames of 20 milliseconds. Figure 2 describes the procedure for extracting the MFCC features.

Magnitude spectrum is computed for each of these frames using Fast Fourier Transform (FFT) and converted into a set of mel scale filter bank outputs. Logarithm is then applied to the filter bank outputs. Discrete Cosine Transformation (DCT) is applied to obtain the MFCCs. In this work, the number of MFCC parameters is chosen as 13.

3.2 Chromagram

Pitch is a property related to perception and sound is ordered on a scale related to frequency [14]. The audio signal is decomposed into bands of varying frequency [15]. Chroma feature representation is an effective and powerful method to describe harmonic information in music structure analysis [16]. Pitch class is a collection of pitches that share the same chroma. Two dimensions characterize music, tone height and chroma [17]. The dimension of tone height is partitioned into the musical octaves. The range of chroma is usually divided into 12 pitch classes, where each pitch class corresponds to one note of the twelve tone equal temperament. The spectral energy of each of the 12 pitch classes is represented by chromogram. It is based on a logarithmized short time Fourier spectrum. The chromagram represents an octave-invariant (compressed)spectrogram that takes properties of musical perception into account [18].

The chromagram is extracted as follows: In the pre-processing stage, short segmented frames are extracted and windowed from music signal using framing and windowing Following that, shifting the centre frequencies of the subband filters of the multi rate bank is necessary for the global tuning of a recording as shown in Figure 3. An average spectrum vector is calculated and the derivation of an estimate for tuning derivation is done by stimulating the filter bank shifts using weighted binning techniques [1].

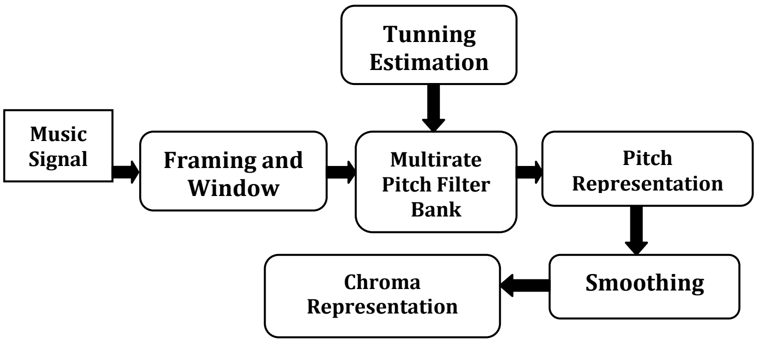

Fig 3: The Chromagram Computation.

The pitch representation is performed by the decomposition of a given music signal on 88 frequency bands with centre frequencies corresponding to the pitches A0 to C8 (MIDI pitches p=21 to p=108) in order to extract the chroma features. The mapping of DFT values to a semitone spectrum of a pitch is done using the following mapping function:

$$p_s(f_k) = 12log_2\left(\frac{f_k}{440}\right) + 69, p_s \in \Re^+ \qquad (1)$$

where, fk denotes the frequencies of the Fourier transform and p_s represents the scale values of semitone pitches. For obtaining a proper spectral resolution for the lower frequencies, there is a necessity for knowing a low sampling rate or a larger temporal window.

The smoothening of semitone pitch spectrum is accomplished by the use of median filtering that reduces extraneous contents from the signal. Once smoothing is done, mapping of the semitone pitches p_s to the corresponding pitch classes c_p is performed by the use of following mapping function:

$$c_p(p_s) = \mathrm{mod}(p_s, 12) \qquad (2)$$

Hence 12 dimensional chromagram feature vectors are used [32].

3.3 Tempogram

An element which gives shape to the music in temporal dimension is the rhythm. Rhythmic feature arranges sounds and silences in time. A predominated pulse called beat which serves as basis for temporal structure of music is induced [19]. Tempogram captures the local tempo and beat characteristics of music signals. The Fourier tempograms are used in the research work.

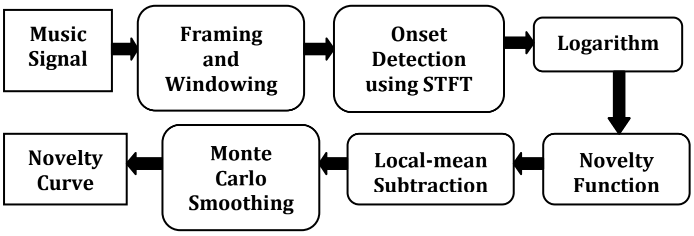

Fig 4: Novelty Curve Computations.

Human perceives rhythm as a regular pattern of pulses as a result of moments of musical stress. Abrupt changes in loudness, timbre and harmonic causes the occurrences of musical accents [20]. In instruments like piano, percussion instruments and guitar, occurs a sudden change in signal energy accompanied by very sharp attacks. A novelty curve is based on this observation and is computed for extracting meaningful information regarding note onset e.g. pieces of songs which are dominated by instruments [21]. In the pre-processing, stage short segmented frames have been extracted and windowed.

Novelty curve computation is shown in Figure 4. Onset changes are calculated from spectrogram using short-time Fourier transform. Logarithm is applied to calculate the intensity of sound to the spectrogram of magnitude. Find Z's discrete temporal derivative. Positive intensity changes are added up for getting the novelty function. Subtract the mean and take positive values. Novelty curve is the form of irregular impulse spikes in onset properties, to remove the locally periodical information for the limitation of human perception [22]. The unobserved states in a non-linear model estimates provide the Monte Carlo smoothing filtering [23].

The novelty curve computed, as described above, indicates peaks which represent note onset values. A hamming window function is applied to avoid boundary problems as smoothing. The Fourier tempogram is calculated as follows:

The complex Fourier coefficients $\mathcal{F}(\mathrm{t},\omega)$

$$\mathcal{F}(\mathrm{t},\omega)=\sum_{\mathrm{n}\epsilon\mathbb{Z}}\Delta(\mathrm{n}).\mathrm{W}(\mathrm{n}-\mathrm{t}).\mathrm{e}^{-2\pi\mathrm{i}\omega\mathrm{n}} \qquad (3)$$

Tempo related to musical context is a measure of beats per minute. This work uses a tempo parameter τ =60.ω. Four tempo octaves $\tau = 30$ to $\tau = 480$ are used in the implementation. This interval is sampled to cover each tempo octave. The Discrete Fourier Transformation (DFT) is given as follows:

$$\mathrm{T}^{\mathrm{F}}(\mathrm{t},\tau)=\left|\mathcal{F}\left(\mathrm{t},{}^{\tau}/_{60}\right)\right|$$
(4)

Finally, the histogram is computed for each frame resulting in 12 dimensional feature vectors.

3.4 MPEG-7 Audio Descriptors

The Moving Picture Experts Group (MPEG) comprises ISO/IEC standards for digitally coded illustration of audio and video [7]. The MPEG-7 standard is represented by two level descriptors. The first one is low-level audio descriptor and the other is high-level descriptor to obtain the feature extraction of the audio content. The MPEG-7 content describes the storage information and structural information on temporal components [24]. The high-level descriptors are used to reduce the dimensions in the features and low-level descriptors describe the content of the audio or music signals. This work aims to extract features of music signal using spectral properties as low-level descriptors [2]. The spectral properties are used to describe the music over frequency variations or time variations. Low-level descriptors like audio spectrum centroid (ASC), audio spectrum spread (ASS), audio spectrum flatness (ASF) and zero crossing rate (ZCR) are extracted in this work [25].

3.4.1 Audio Spectrum Centroid

Audio Spectrum Centroid (ASC) describes the center gravity of a power spectrum's log frequency. i.e., the location of the center-of-mass of the spectrum is indicated by this brightness of a sound.

$$\mathrm{ASC}=\frac{\sum_{\mathrm{n}=0}^{\mathrm{N}-1}\mathrm{f}(\mathrm{n})\mathrm{x}(\mathrm{n})}{\sum_{\mathrm{n}=0}^{\mathrm{N}-1}\mathrm{x}(\mathrm{n})} \qquad (5)$$

where, x(n) denotes weighted frequency value of n, and f(n) is the center frequency.

3.4.2 Audio Spectrum Spread

Audio spectral spread (ASS) is the bandwidth of spectral shape measure. The features are extracted using RMS of the spectrum deviations on audio centroid spectrum.

$$\mathrm{ASS}=\sqrt{\frac{\sum_{\mathrm{k}'=0}^{\left(\frac{\mathrm{N}_{\mathrm{FT}}}{2}\right)-\mathrm{k}_{\mathrm{low}}}\left[\log_2\left(\frac{\mathrm{f}'(\mathrm{k}')}{1000}\right)-\mathrm{ASC}\right]^2\mathrm{P}'(\mathrm{k}')}{\sum_{\mathrm{k}'=0}^{\left(\frac{\mathrm{N}_{\mathrm{FT}}}{2}\right)-\mathrm{k}_{\mathrm{low}}}\mathrm{P}'(\mathrm{k}')}} \qquad (6)$$

here, klow is the power coefficient index of discrete frequency bin scale, NFT is the frequency interval between two FFT bins.

Compute f'(k') frequencies corresponding to the new bins k' and (P'(k')) power spectrum modified coefficients corresponding to the new bins k' similar to that of the ASC descriptor. In ASS, centroid indicates the spectrum distributions [2].

3.4.3 Audio Spectrum Flatness

Power spectrum has a property called Audio spectrum flatness (ASF) to reflect the flatness. ASF is computed by dividing arithmetic and geometric mean of bank b, as shown in Equation (7) where P(k) is the power coefficient.

$$\mathrm{ASF}=\frac{\sqrt[\mathrm{hi\,k'}_{\mathrm{b}}-\mathrm{lo\,k'}_{\mathrm{b}}+1]{\prod_{\mathrm{k}'=\mathrm{lo\,k'}_{\mathrm{b}}}^{\mathrm{hi\,k'}_{\mathrm{b}}}\mathrm{P'}_{\mathrm{g}}(\mathrm{k}')}}{\frac{1}{\mathrm{hi\,k'}_{\mathrm{b}}-\mathrm{lo\,k'}_{\mathrm{b}}+1}\sum_{\mathrm{k}'=\mathrm{lo\,k'}_{\mathrm{b}}}^{\mathrm{hi\,k'}_{\mathrm{b}}}\mathrm{P'}_{\mathrm{g}}(\mathrm{k}')}\,(1\leq\mathrm{b}\leq\mathrm{b}) \qquad (7)$$

For each b and, $\mathrm{P'}_{\mathrm{g}}(\mathrm{k}') = \mathrm{P}(\mathrm{k}')$ between $\mathrm{k}' = \mathrm{lok'}_{\mathrm{b}} = \mathrm{lok}_{\mathrm{b}}$ and $\mathrm{k}' = \mathrm{hik'}_{\mathrm{b}} = \mathrm{hik}_{\mathrm{b}}$. Equation (6) is used for bands that require power above 1 kHz [2].

3.4.4 Zero Crossing Rate

The zero crossing rate (ZCR) is computed as follows (Ausgef¨uhrt, 2006):

$$\mathrm{Z}_{\mathrm{m}}=\sum_{\mathrm{n}}|\mathrm{sgn}[\mathrm{x}(\mathrm{n})]-\mathrm{sgn}[\mathrm{x}(\mathrm{n}-1)]|\mathrm{w}(\mathrm{m}-\mathrm{n}) \qquad (8)$$

where the sgn function is $\mathrm{sgn}[\mathrm{x}(\mathrm{m})]=\begin{cases}1, \mathrm{x}(\mathrm{m})\geq 0\\ -1, \mathrm{x}(\mathrm{m})<0\end{cases}$ and x(m) is the time domain signal for frame m.

4. TECHNIQUES FOR MUSIC INDEXING AND RETRIEVAL

Acoustic features namely MFCC, chromagram, tempogram and MPEG-7 are extracted from music audio clips. Index is created for the feature vectors using GMMs. Retrieval is made depending on the maximum probability density function.

4.1 Gaussian Mixture Model

Parametric or non-parametric methods are used to model the distribution of feature vectors. Parametric models are based on the shape of probability density function [26]. In non-parametric modeling only minimal or no assumption regarding the probability density function of feature vector is made [27], [28]. The basis for using GMM is that the distribution of feature vectors extracted from a class can be modeled by a mixture of Gaussian densities.

GMM's represent the feature vectors using Gaussian components and are characterized by the mean vector and the co-variance matrix [29]. Even in the absence of other information,

GMM models have the capability to form an arbitrarily shaped observation density [30]. For a D dimensional feature vector x, the mixture density function for category s is defined as

$$p(\frac{x}{\lambda_s}) = \sum_{i=1}^{M} \alpha_i^s f_i^s(x) \qquad (9)$$

The mixture density function is a weighted linear combination of M component uni-modal Gaussian densities $f_i^s(.)$.Every Gaussian density function $f_i^s(.)$ is categorized by the mean vector μ_i^s and covariance matrix $\sum_i^s$ using

$$f_i^s(x) = \frac{1}{\sqrt{(2\pi)^d |\sum_i^s|}} \exp(-\frac{1}{2}(x - \mu_i^s)^T (\sum_i^s)^{-1} (x - \mu_i^s)) \qquad (10)$$

where, $(\sum_i^s)^{-1}$ and $|\sum_i^s|$ denote the inverse and determinant of the covariance matrix $\sum_i^s$,respectively. The iterative Expectation Maximization (EM) algorithm is used to estimate the parameters of GMM.

EM algorithm is one of the most popular clustering algorithms used to estimate the probabilistic models for each Gaussian component. The Expectation step (E-step) and Maximization step (M-step) are iterated till the convergence of the parameter [31]. EM algorithm finds out maximum likelihood estimation of parameters.

5. PROPOSED METHOD FOR INDEXING AND RETRIEVAL OF MUSIC CLIPS

The algorithm for indexing and retrieval of Music clips is described below:

5.1 Algorithm for Indexing of Music

Step 1: Collect 100 music clips $m_{1,} m_2, \ldots, m_{100}$ each of 50 seconds duration from TV broadcast music channels.

Step 2: 13-dimensional MFCC features are extracted from all 100 music clips to form the music index.

Step 3: A GMM is fit for all 100 music clips using the MFCC features.

Step 4: Repeat step 2 and step 3 for chromagram, tempogram and MPEG-7 features.

5.2 Algorithm for Retrieval of Music using Index

Step 1: A music query clip of 10 seconds duration is extracted from the music wave file.

Step 2: MFCC features are extracted from the music query.

Step 3: The probability of the query feature vectors belonging to the 100 GMMs is computed.

Step 4: The maximum probability density function corresponds to the music query audio.

Step 5: The music clips which have the maximum probability density function are retrieved.

Step 6: Repeat steps 2 to 4 for chromagram, tempogram and MPEG-7 features.

5.3 Performance Measures

The accuracy of retrieval and average number of clips retrieved for each query are the measures used to analyse the performance of the music audio indexing system.

5.3.1 Accuracy of retrieval

Accuracy of retrieval is a performance measure for music audio indexing system. It is measured using Equation (11).

$$R_m = \frac{K_m}{Q_m} \times 100 \qquad (11)$$

where R_m is the accuracy of retrieval, K_m is the number of music clips retrieved in the top 'n' ranked list and Q_m is the total number of queries.

5.3.2 Average number of clips retrieved for each query based on a threshold

The music retrieval performance of the system is measured with T as the number of clips retrieved on an average for each query based on a predefined threshold as shown in the Equation (12)

$$T = \frac{\sum_i^N \text{Retrieved}_i}{N} \qquad (12)$$

where T is the average number of clips retrieved based on threshold and Retrieved_i is the number of clips retrieved for a given query audio clip.

6. EXPERIMENTAL RESULTS

For music indexing and retrieval, experiments are conducted to study the performance of the music retrieval algorithms in terms of performance measures.

6.1 Database for Music Indexing

Experiments are conducted for indexing music audio using the television broadcast. The music data is collected from Tamil music channels using a TV tuner card. A total dataset of 100 different songs is recorded, which is sampled at 22 kHz and encoded by 16-bit. Fixed duration music clips of first 50 seconds are used for creating a music database and last 10 seconds duration of 1 minute music clip is used for query.

6.2 Acoustic Feature Extraction

Each music clip with 50 seconds of duration for MFCC, chromagram, tempogram and MPEG-7 features are extracted. A frame size of 20 ms and a frame shift of 10 ms are used. Thereby 13 MFCC features are extracted for each music audio clip of 50 seconds. Hence 5000 × 13 feature vectors are arrived at for each of the 50 second clip and this procedure is repeated for all 100 clips. Similarly, experiments are conducted to extract chromagram features of 5000 × 12, tempogram features of 5000 × 12 and MPEG-7 features of 5000 × 4 dimensions respectively. Same procedure is repeated for all the 100 clips.

6.3 Creation of Index

GMMs are constructed for 100 music clips using MFCC features which form the index. Experiments are also conducted with GMMs using chromagram, tempogram and MPEG-7 features to create index.

6.4 Retrieval of a Music using Index

For retrieval, the last 10 seconds in a music clip is used as query. For every frame in the query the probability density function that the query feature vector belongs to the first Gaussian is computed. The same process is repeated for all the feature vectors.

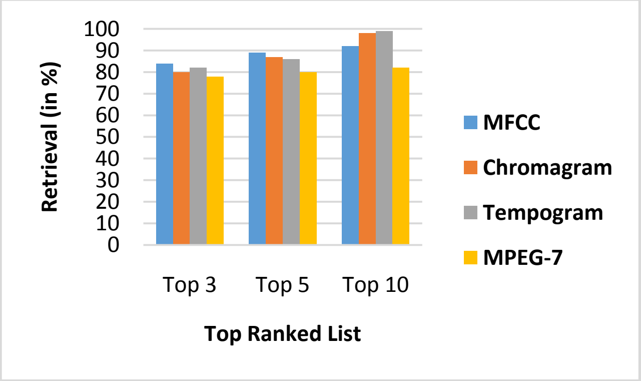

Fig 5: Accuracy of Retrieval of Music Clips in the Top n Ranked List

The average probability density function is computed for the first Gaussian. Similarly the average probability density function for all the Gaussians is computed. Retrieval is based on the maximum probability density function. Figure 5 shows the percentage of retrieval in top ranked list.

The GMM is used to capture the distribution of features namely MFCCs, chromagram, tempogram and MPEG-7 features. The performance of GMM for different mixtures is shown in Figure 6. Various experiments for different mixtures are carried out and the retrieval is based on the highest probability density function. With GMM, the performance is found to increase as the mixture increased from 2 to 5 and the optimal performance is achieved with 5 mixtures. When mixtures increased from 5 to 10, the performance remained stable. After 10 mixtures, the performance deteriorated.

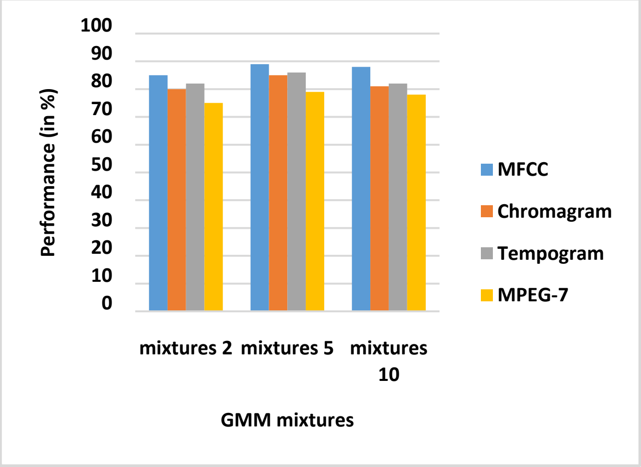

Fig 6: Performance of GMM for Different Mixtures.

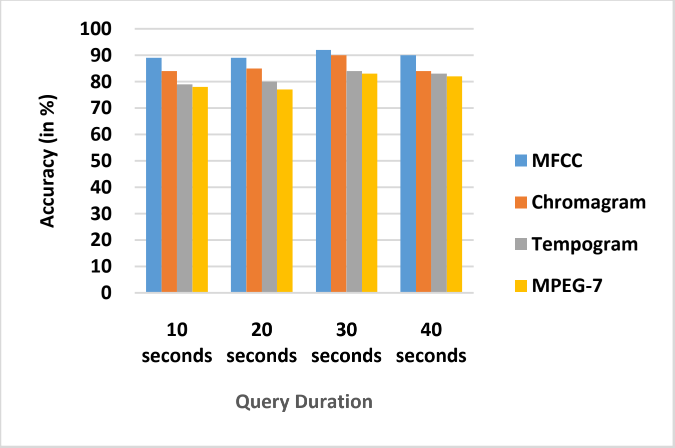

Fig.7. Performance of Music Retrieval for Different Durations of Query Clips.

For studying the performance, several experiments have been performed on music indexing for different durations of query music clips at 10, 20, 30 and 40 seconds respectively. Accuracy of retrieval is shown in Figure 7.

Table 1. Average Number of Clips Retrieved.

No. of clips used for index creation	MFCC	Chromagram	Tempogram	MPEG-7
50	5.3	5.0	7.1	4.9
100	6.7	5.9	6.5	5.8

Table 1 shows average number of clips retrieved for a given query.

Figure 8 shows the average number of clips retrieved for a query based on threshold = 0.53. 100 music clips are used to create the index.

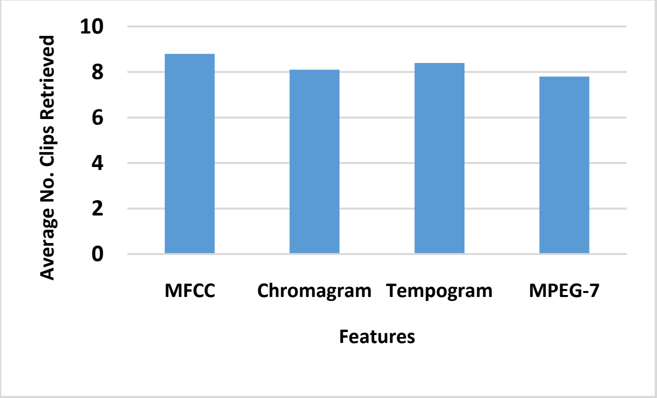

Fig.8. Average Number of Clips Retrieved for a Query Based on Threshold.

Table 2 shows performance of indexing and retrieval of query music clips using various feature sets.

Table 2. Performance of Music Indexing system using various features.

Features	Accuracy of retrieval (in %)
MFCC	97.3
Chromagram	75.4
Tempogram	81.1
MPEG-7	62.4

7. CONCLUSION

In this paper, a new method is proposed for music indexing and retrieval. MFCC, chromagram, tempogram and MPEG-7 features are extracted. For the music clips, GMMs are used to create an index based on the features extracted. For retrieval, the probability of the query feature vector belonging to each of the Gaussian is computed. Average probability density function is computed for all the Gaussians and the retrieval is based on the highest probability density function. The retrieval performance is studied for different features. Efficiency of music audio retrieval system is evaluated for 100 clips and the method achieves about 89.0% accuracy using MFCC features. This method is sensitive to the duration of the query clip.

8. REFERENCES

[1] Gal Chechik, Eugene Ie, Martin Rehn, Samy Bengio and Dick Lyon, 2008, Large-Scale Content-Based Audio Retrieval from Text Queries, International Conference on Multimedia Information Retrieval, pp. 105-112.

[2] Hyoung-Gook Kim, Nicolas Moreau and Thomas Sikora, 2004, MPEG-7 Audio and Beyond Audio Content Indexing and Retrieval, John Wiley and sons Ltd.,.

[3] Masataka Goto, 2004, A Real-Time Music-Scene Description System: Predominant-*F0* Estimation for Detecting Melody and Bass Lines in Real-world Audio Signals, Speech Communication, no. 43, pp. 311-329.

[4] Cheong Hee Park, 2015, Query by Humming Based on Multiple Spectral Hashing and Scaled Open-End Dynamic Time Warping, Signal Processing, no.108, pp. 220-225.

[5] BalajiThoshkahna and K.R.Ramakrishnan, 2005, Projektquebex: A Query by Example System for Audio Retrieval, IEEE International Conference on Multimedia and Expo, pp. 265-268.

[6] Seungmin Rho, Byeongjun Han, Eenjun Hwang, and Minkoo Kim, 2007, Musemble: A music Retrieval System Based on Learning Environment, IEEE International Conference on Multimedia and Expo, pp. 1463-1466.

[7] Ausgef¨uhrt, 2006, Evaluation of New Audio Features and Their Utilization in Novel Music Retrieval Applications, Master's thesis, Vienna University of Technology.

[8] Peter M. Grosche, 2012, Signal Processing Methods for Beat Tracking, Music Segmentation and Audio Retrieval, Thesis, Universit¨at des Saarlandes.

[9] Yin-Fu Huang, Sheng-Min Lin, Huan-Yu Wu and Yu-Siou Li, 2014, Music Genre Classification Based on Local Feature Selection using a Self-Adaptive Harmony Search Algorithm, Data & Knowledge Engineering, no.92, pp. 60–76.

[10] Jesper Højvang Jensen, 2009, Feature Extraction for Music Information Retrieval, Thesis, Aalborg University.

[11] Ahmad R. Abu-El-Quran, Rafik A. Goubran, and Adrian D. C. Chan, 2006, Security Monitoring using Microphone Arrays and Audio Classification, IEEE Transaction on Instrumentation and Measurement, vol. 55, no. 4, pp. 1025-1032.

[12] Meng .A and J. Shawe-Taylor, 2005, An Investigation of Feature Models for Music Genre Classification using the Support Vector Classifier, International Conference on Music Information Retrieval, Queen Mary, University of London, UK, pp. 604-609.

[13] Francois Pachet and Pierre Roy, 2009, Analytical Features: A Knowledge-Based Approach to Audio Feature Generation, Journal on Applied Signal Processing.

[14] FitzGerald. D and J. Paulus, 2006, Unpitched Percussion Transcription, in Signal Processing Methods for Music Transcription, Springer, pp. 131-162.

[15] Zhe Zuo, 2011, Towards Automated Segmentation of Repetitive Music Recordings, Master's Thesis, Saarland University.

[16] Müller. M, 2007, Information Retrieval for Music and Motion, Database Management & Information Retrieval Springer.

[17] Shepard, 1964, Circularity in Judgements of Relative Pitch, The Journal of the Acoustical Society of America, vol. 36, pp. 2346-2353.

[18] Schroder M. R., B. S. Atal, and J. L. Hall, 1979, Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear, Journal of the Acoustical Society of America, vol. 66, pp. 1647-1652.

[19] Philipp von Styp-Rekowsky, 2011, Towards Time-Adaptive Feature Design in Music Signal Processing, Master's Thesis, Saarland University.

[20] Eronen, A. and Klapuri, A , 2010 “Music Tempo Estimation with *k*-NN regression,” IEEE Transactions on Audio, Speech and Language Processing, vol. 18, no. 1, pp. 50-57.

[21] Venkatesh Kulkarni, 2014, Towards Automatic Audio Segmentation of Indian Carnatic Music, Master Thesis, Friedrich Alexander University.

[22] Duifhuis H., Willems L. and Sluyter R., 1982, Measurement of Pitch in Speech: An Implementation of Goldstein’s Theory of Pitch Perception, Journal Acoustic Society of America, vol. 71, no.6, pp. 1568-1580.

[23] William Fong and Simon J. Godsill, 2002, Monte Carlo Smoothing with Application to Audio Signal Enhancement, IEEE Transactions on signal processing, vol. 50, Issue 2, pp. 438-449.

[24] Eberhard Zwicker and Hugo Fastl, 1999, Psychoacoustics-Facts and Models, Springer Series of Information Sciences, Berlin.

[25] PetrMotlcek, 2003, Modeling of Spectra and Temporal Trajectories in Speech Processing, PhD thesis, Brno University of Technology.

[26] Menaka Rajapakse and Lonce Wyse, 2005, Generic Audio Classification using a Hybrid Model Based on GMMs and HMMs, IEEE International Multimedia Modelling Conference, pp. 53-58.

[27] Zanoni, M., Ciminieri, D., Sarti, A. and Tubaro, S, 2012, Searching for Dominant High-Level Features for Music Information Retrieval, European Signal Processing Conference, pp. 2025-2029.

[28] Chunhui Wang, Qianqian Zhu, Zhenyu Shan, Yingjie Xia and Yuncai Liu, 2014, Fusing Heterogeneous Traffic Data by Kalman Filters and Gaussian Mixture Models, IEEE International Conference on Intelligent Transportation Systems, pp. 276-281.

[29] Rafael Iriya and Miguel Arjona Ramírez, 2014, Gaussian Mixture Models with Class-Dependent Features for Speech Emotion Recognition, IEEE Workshop on Statistical Signal Processing, pp. 480-483.

[30] Tang, H., Chu, S. M., Hasegawa-Johnson, M. and Huang, T. S., 2012, Partially Supervised Speaker Clustering, IEEE transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 5, pp. 959-971.

[31] Chien-Lin Huang, Chiori Hori and Hideki Kashioka, 2013, Semantic Inference Based on Neural Probabilistic Language Modeling for Speech Indexing, IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8480-8484.

[32] Papadopoulos and Geoffroy Peeters, 2007, Large-Scale Study of Chord Estimation Algorithms Based on Chroma Representation and HMM, International workshop on Content-Based Multimedia Indexing, pp. 53-60.