

Content-Based Audio Indexing and Retrieval: an overview

Hafiz Malik

Abstract:

This manuscript provides a brief overview of trends and developments in the area of Content-Based Audio Indexing and Retrieval (CBAIR), during the past few years. Here we explored some limitations and constrains of existing Query By Example (QBE) and Query By Humming (QBH) CBAIR systems. We talked about different methods to represent musical objects, like feature-based representation, musical parameter-based representation; similarly retrieval strategies, like feature based retrieval as well as melody or theme based retrieval of musical objects, in this paper. Moreover, some important issues regarding to indexing and retrieval performance i.e. efficient indexing and retrieval complexity, in this area are discussed thoroughly. Finally, hypothetical solutions for the existing problems in this area are proposed to improve the performance.

Introduction:

During the past few years, the demand for and growth of multimedia data around the world has been increasing rapidly and dramatically with the availability of high performance but low cost personal computers, smart wireless hand-held devices (mobile sets, notepads, palmtops etc.) with more computational power, and high speed integrated digital networks (ATM, ISDN etc.) with even more bandwidth. In this modern age, we are living in a unified society called “Global Village”, where importance of multimedia and its related fields became more vital than before. In this Village where any one around the globe can access multimedia database system to his/her desired document; to meet these requirements, we have to have very large multimedia databases that should contain music, images, videos etc. of different categories and cultures; so that fast and accurate retrieval of desired multimedia document is possible. Now state of the art technologies enable people to share more multimedia data for entertainment, education, business and other purposes, this requires new retrieval methods to cope with the future challenges in this active area of research.

Currently available search engines (like Yahoo, Google, AltaVista etc.) do similarity search based on Key-word, but for many users it is difficult to describe query in terms of key words especially when they are looking for some multimedia document from a multimedia database, for example, its hard to formulate a key-word based query for “red apples hanging on a tree with raining landscape” or “a slow tempo song of unknown lyrics and singer”. Content-Based retrieval is considered as an effective solution to this problem; in recent years many retrieval systems [21-46] implemented this idea (i.e. content-based retrieval) quite successfully.

Content-Based retrieval systems accept data type queries i.e. drawing sketch for an image retrieval; hummed, sung or original clip of a song for a song retrieval; and a video clip or set of images from some video short for a video retrieval. Content-Based retrieval allows users to describe the query as what they want, so it makes query formulation more comprehensive and easier than key word based retrieval. Moreover, content-based retrieval system permits more tolerance towards erroneous queries, as in these systems queries contain more errors; so for such search keys similarity search based on approximate matching produce better results compare to exact matching.

In this paper, CBAIR systems [21-46] developed during recent years are discussed in detail; these systems are classified based on audio object representation, indexing structure and retrieval technique used. Throughout this paper, by audio we mean both natural sounds (excluding speech) as well as music only. We looked into limitations and constrains of existing CBAIR systems; common limitations of existing systems are, audio object representation, retrieval complexity, memory requirement for the database, retrieval time etc. then based on these limitations and constrains a prototype system is purposed, that might be helpful for the development of a retrieval time efficient as well as more compact musical object representation CBAIR system for very large database, in future. For such systems, compact and more comprehensive music representation along with more efficient indexing structures and retrieval strategies would be main consideration.

Generally music is represented by, written score, recorded performance and MIDI (Musical Instrument Digital Interface) format. Currently almost all existing CBAIR systems used songs in MIDI format for their databases. Now a days hundreds of thousands songs in MIDI format of almost all kinds are easily available on the web, and this might be one of the reasons for using songs in MIDI format for the database. According to Dawling [1] a listener can remember melody longer than all other perceived parameters and this is one of perceived parameters that is easy to reproduce. This is the basic principle of QBH CBAIR systems, where user hums, or sings the tune of desired song. There are number of melody extraction algorithms from MIDI format developed by [29-38], but the main idea is same every where, that is, convert polyphonic tunes into monophonic tune on the basis of highest energy or intensity for each note and chord deletion.

Rest of the paper is organized as; in section II audio object modeling is discussed, Section III is about retrieval technique, comparison of existing CBAIR systems is viewed in section IV, prototype system is purposed in section V and concluding remarks are in section VI.

Section 2

Audio Object Modeling:

In this paper, each single entity in an audio database or a query for CBAIR system is considered as an object and each object has its specific attributes that characterize it. Our definition for audio in this paper covers music and natural sounds only. Moreover for the rest of the paper, by an object we mean both music as well as natural sounds where as a musical object will cover music only and an audio object will stand for natural sounds only. Attributes of an object can be divided into tow groups based on their extraction technique, that is, frame level attributes like zero-crossing rate, energy, pitch frequency, loudness, intensity, transformed domain coefficients etc. and global attributes like timber, loudness contour, first order and second order statistic of transformed domain coefficients, pitch contour or melodic contour, tempo, rhythm etc. Frame level attributes reflect the non-stationary behavior of an object where as global features give the overall characteristic of an object; collectively these of features characterize a specific object in the database and this specific character of an object is actually used for retrieval. Frame level feature extraction commonly employs signal-processing techniques [5, 6, and 7] and for global feature extraction generally statistical analysis [19] is applied to the frame level features. We can model an audio object based on its signal parameters where as a musical object can be modeled based on signal parameters as well as based on its musical parameters, so we can say that an object can be modeled as,

1. Signal Parameter Based Modeling:

This type of modeling is applicable to musical objects as well as audio objects, because both can be characterized by signal or acoustical parameters. To model an object we need both frame level as well as global parameters. General attributes required for this type of modeling are, zero-crossing rate, energy, pitch frequency, timber, energy contour, and loudness contour etc. figure1 illustrate this model.

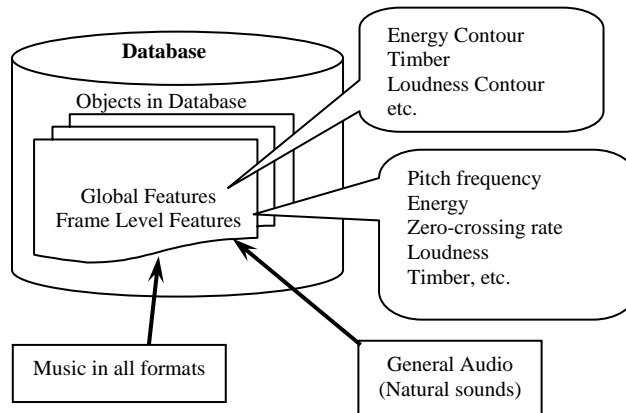


Figure 1: signal parameter based object model

CBAIR systems where databases contain objects modeled based on signal parameter, can support only QBE only because hummed, sung, or whistled query would have entirely different parameters than actual object.

2. Musical Parameter Based Modeling

This type of modeling is applicable to musical objects only, because this type modeling may not work for audio objects; reason is very simple, that is, according to our definition of audio objects they cover only natural sounds like thunder, sound of raining etc., and musical parameters for this kind of audio is meaningless. Musical object model using musical parameters is given in figure 2.

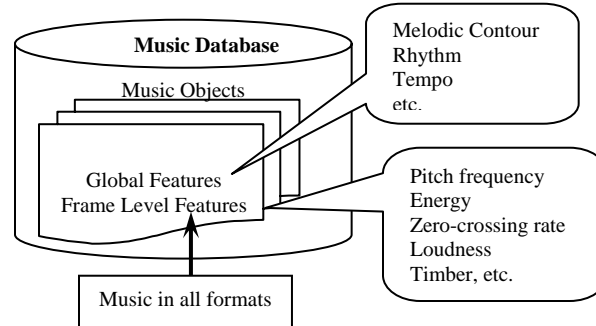


Figure 2: Musical parameter based model

This type of model requires melodic contour, rhythm, tempo, timber etc. CBAIR systems where databases are based on this type of modeling can support both QBH as well as QBE.

Section 3

Retrieval Techniques

Retrieval strategies assign a measure of similarity between a query and a document in database. These strategies are based on simple notion that more often terms found in both the document (object in case of CBAIR system) and the query, the more relevant the document is seem to be the query.

Mathematically, a retrieval strategy is an algorithm that takes query Q and set of documents $D_1, D_2 \dots D_n$, and evaluates the similarity coefficient (SC) for each document in the database. Detailed discussion on retrieval strategies currently employed in the Information Retrieval (IR) area can be found [8].

Retrieval strategies employed in CBAIR systems in the past are---

- **Vector Space Bases Retrieval or Vector Based Model**

Both the query and each object are represented as vectors in terms of n-dimensional space. A measure of similarity between the query and each object in the database is computed. Figure 3 illustrate the basic notion of the vector space model.

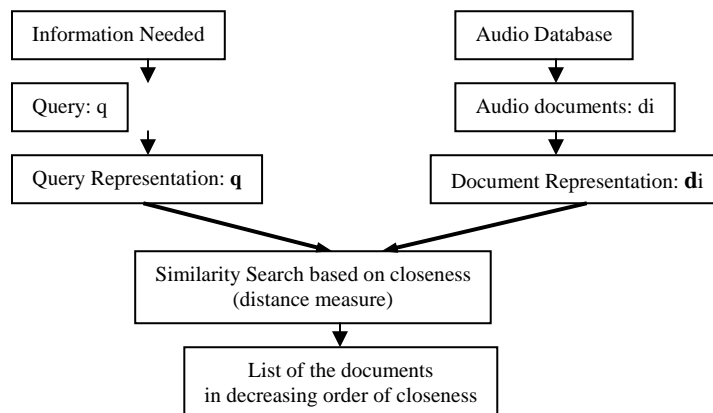


Figure 3 the vector space retrieval model

This model shows that query is transformed into a feature vector \mathbf{q} , now search engine finds the most relevant document d for the entire database that contains documents (objects) as features vectors $\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_n$. This model involves the construction of vectors representing the salient features of objects, so similarity between two objects is determined by measuring the closeness of two vectors in space; in existing CBAIR systems [19-26], this is done by computing distance between two vectors. As mentioned in section 2 that we can represent musical objects as well as audio objects in the database by n -dimensional feature vector in [8]. Now relevance between two objects is just a matter of, how closely they are located in n -dimensional space, or in other words by finding distance between them. CBAIR systems [19-26] are using distance measure for IR commonly employs either Euclidean distance or Cosine distance or both.

There is one limitation of this technique that is, its retrieval complexity is linear $O(n)$ where n is the number of objects in the database; linear complexity works efficiently for small databases but for very large databases linear complexity is not good. To get rid of this problem k -mean clustering [20] or other kind of advance clustering techniques [20] can be used for retrieval, complexity of this retrieval method is still linear but with better constant that is, its $O(n/k)$, where k is number of clusters in the database.

- **Pattern Matching Based Retrieval or String Matching Bases Retrieval**

In pattern matching based retrieval, both the query and the document are represented by a sequence of characters, integers, words etc., and similarity between them is computed based on, how similar two sequences are? Similarity matching between them can be determined either by exact sequence matching or by approximate sequence matching.

The approximate pattern matching or approximate string matching problems has been extensively studied over the last few years. Efficient algorithms for computing the approximate repetition have applications in the area of molecular biology [13], especially in DNA sequencing. Many researchers [9-12] also applied approximate string matching problem in the area of music indexing and retrieval. It is obvious that exact matching can not give satisfactory results for QBH music retrieval system, where as approximate matching allows some mismatches between query and the documents; which is referred as pattern/string matching with k -mismatches. For melodic contour based representation, where a musical object is represented by a string of U, D, and R (or S) characters. So similarity between two objects is determined their closeness, i.e. how similar two strings are? Hence similarity search problem is transformed into a string-matching problem, where similarity between two strings is evaluated using Edit or hamming distance between them. Edit distance actually tells the number of mismatches between source string and target string. There are number of efficient algorithms for exact string matching [9,10] and approximate string matching with k -mismatches [11,12,13,15], Dynamic Programming (DP)[14] is commonly used for efficient implementation, which is discussed next.

Sequence-Comparison using Dynamic Programming (DP):

There are number of methods for sequence comparison, but sequence matching using DP is quite popular than all other due to its space efficient implementation and with lower complexity [19]. Sequence matching based on DP uses the concept of *edit distance*; edit distance is the cost of changing the *source sequence* (*source string*) into *target sequence* (*target string*). The cost of transposing source string into target string is calculated in terms of edit operators, common edit operators used in DP are substitution (replacement), deletion and insertion. For example to transpose a string ABCDE into ABDEF requires leaving A, B, D and E unchanged (i.e. replacing by themselves), deleting C and inserting F. Each of these operations incurs a cost or weight. Usually the cost of replacing a letter is the alphabetic distance (i.e. C-A=2, C-C = 0), and cost of insertion and deletion is 1, so following the above rules, cost of changing ABCDE into ABDEF is 2, or the distance between ABCDE and ABDEF is 2.

Section 4

Classification of CBAIR Systems:

We can classify the existing groups working on CBAIR systems into two main communities; audio or acoustic parameter based CBAIR group which basically uses purely signal processing and acoustic similarity for content based audio or music retrieval but limited to only QBE, where as the second group uses music parameters and music psychology principles and theories for content based audio or music retrieval that is applicable to both QBE as well as QBH.

1. Audio or Acoustic Parameter Based CBAIR

In this domain an audio object is characterized by its salient parameters. These parameters are extracted using some data reduction techniques, for example data transformation techniques like KL transform, Wavelet transform Cosine transform and Fourier transform etc., or some frame level parameters of an audio object like zero-crossing rate, pitch frequency and energy. For feature selection, an audio object is processed at frame level and based on the predefined parameters, transformed into feature vector. For indexing multidimensional space, with or without clustering, is generally used and for retrieval, distance measure is applied.

CBAIR systems based on frame level parameters work quite effectively for general audio (music as well as speech) retrieval, but these systems have some intrinsic limitations, like their performance deteriorate drastically in case of QBH, because feature vector of the hummed query would be entirely different from the feature vector of the target song (accompanying with musical instruments), this way.

I think there are few algorithms that can speed up retrieval process in this domain, first by using clustering technique for indexing and retrieval and second using some modified version of SVM for retrieval. (?????)

- **CBAIR Systems Based on Audio Parameters**

Audio parameters based systems[21-27] have been extensively used for speak recognition and speaker identification systems for more than two decades and these systems are still popular in this area, but unfortunately, for CBAIR systems based especially for music retrieval systems; audio parameter based systems could not gain same popularity. Main reason might be, that these systems did not support QBH, which is very popular in the area of music retrieval now a day. During past few years many researchers developed systems for CBAIR based on audio parameters that support QHE only.

J. T Foote in [22] presented an idea for the representation of an audio object by a template that characterizes the object in his purposed system. For construction of a template; an audio signal is first divided into overlapping frames of constant length then using simple signal processing techniques, for each frame a13-dimensional feature vector is extracted (12 Mel-Frequency Cepstral Coefficients plus Energy) at a 500Hz, and then these feature vectors are used to generate templates using tree-based Vector Quantizer trained to maximize mutual information (MMI). For retrieval, query is first converted in to template in the same way described earlier then for its similarity search template matching is applied which uses distance measure, and finally a ranked list is generated based on minimum distance. In this system performance of the system with Euclidean distance as well as Cosine distance, is also compared, and experimental results show that cosine distance performs slightly batter than Euclidean distance.

This system may fail for music retrieval if either query is corrupted with noise or bad quality recorded.

Muscle fish group [25], in this system an audio object is characterized by its frame level and global acoustical and perceptual parameters. These features are extracted at frame level using signal processing techniques and globally using statistical analysis based on frame level features and musical features (for music signals only) using musical analysis. Frame level features consist of loudness, pitch, tone (brightness and bandwidth), MFCCs and derivative. Global features are determined by applying statistical modeling techniques on the frame level features that is, using Gaussian and Histogram Modeling techniques to analyze audio objects. For musical objects, musical features (i.e. rhythm, events and distance (interval)) are extracted using simple signal processing techniques like pitch tracking, voiced and unvoiced segmentation and note formation.

For indexing, multidimensional features space is used. For retrieval, distance measure is used and to improve the performance, a modified version of query-point-expansion [16, 17] technique is used, but here expansion for the refinement of the concept if achieved by standard deviation and mean of the objects in the expected region.

This system again bounded by its inherited limitation, and works for QBE only.

G. Li and A. Khokhar [26] represented an audio object by 3-dimensional feature vector for their system. Feature vectors are extracted at frame level using Discrete Wavelet Transform (DWT). They applied 4-level DWT decomposition to audio signal, then in transformed domain variance, zero-crossing rate and mean of wavelet coefficients are determined to form feature vectors. For indexing structure, B-tree structure [18] is used, which is constructed using clustering technique along with multiresolution property of the wavelet transform. Similarity search is applied using weighted Euclidean distance, and based on minimum distance a ranked target list is retrieved for the desired query.

S. R. Subramenya and A. Youssef [27] presented a signal processing based approach using Discrete Wavelet Transform(DWT) for feature vector of an audio object. First of all an audio signals is decomposed using 4-level DWT and then from wavelet coefficients, feature vector is formed using all approximate coefficients and 20%-30% of detail coefficients of all levels obtained during wavelet decomposition. For query processing same procedure is applied. They did not specified indexing technique but for similarity matching they used Euclidean distance measure.

2. *Melodic Parameter Based CBAIR*

These systems are mainly based on the music psychology principles and theories, like melody perception, music memory, melodic organization, melodic similarity, etc. Melody or theme is a perceptual feature of music. Human perception can easily remember melody due to its overall shape or some time called “melodic contour”, rhythm and dynamic behavior. There has been extensive research in past few decades on how human perceive and remember music. Dowling[1] discovered that melodic contour is easy to remember than exact melodies. Contour refers to the shape of the melody, indicating the relative pitch difference of each note from the previous note. According to him, human memory store melodic contour based on relative pitch difference rather than absolute pitch difference; which might be one of the reasons that in this domain almost every system spouts Query By Humming retrieval along with Query By Example, because its easy for user to hum or sing a query with relative melodic contour similar to the target song.

So an audio object is usually first converted into melodic contour which is mapped onto string like data types, tree like data structures are usually used to index audio objects and string matching techniques (using dynamic programming) are used for retrieval in this domain.

Only limitation for these systems according to us, is their limited database size; because systems based on the strategies mentioned above can perform well for small music databases, but for very large music databases (especially with diverse kind of music i.e. pop, rock, folk etc. and music from different cultures like western, Chinese, Arabic, etc.) they may not be equally effective; and retrieval time may be too high if only string matching technique is used for retrieval.

So there is a room for a more compact music representation along with new indexing structures and more effective retrieval strategies.

- **CBAIR systems based on Melodic parameters:**

In past few years number of groups have been developing Content Based Music Indexing and Retrieval (CBMIR) systems, like [28-46], but with the rapid growth of multimedia documents and increasing popularity of the Internet; now a days researchers are working to develop CBMIR system for very large music databases supported on the Internet, with modified indexing and retrieval techniques to improve the performance of their systems as well as to over come the inherited limitations of this domain.

Asif Gais et al.[28], their system is considered to be the first complete CBMIR database system based on QBH, it was consist of 183 songs in MIDI file format. To represent music object, songs in MIDI format are first converted into melodic contours and then each contour is transformed into a string of characters U(up), D(down) and S(same). To generate query, pitch is extracted from recorded hummed query then it is convert into melodic contour which is converted into a string of same three characters (U, D, and S). For

retrieval, similarity for input query is evaluated using approximate string matching algorithm then a rank list of similar melodies(songs) based on the minimum *edit distance* is generated.

Performance of this system was quite satisfactory, and one of the reasons for good performance is its very small size music database.

Lie Lu et. al. [30], their proposed QBH system, uses more musical parameters (like melodic contour, pitch interval and rhythm.) for similarity matching than Asif Gais et. al.[28]. Melodic information are represented by pitch contour, pitch interval and pitch duration in this system, where as melodic contour is represented by UD string only. Music database consist of MIDI files. For indexing, multi-track MIDI files are first converted into melodic contours and then transformed into melodic triplet, i.e.(pitch contour, pitch interval, pitch duration). For query processing, silence and noise are removed (using zero-crossing rate and energy contour) from hummed query, and then using pitch detection and adaptive note segmentation query is transformed into melodic triplet. For similarity matching, a rank list is generated based on the melodic contour similarity (using string matching), pitch interval matching and pitch duration matching (from matched contour, using Euclidian distance).

McNab et. al. [32,33,34] have constructed a working retrieval system called MELDEX, with 9400 melodies in MIDI format This system uses melodic contour, musical interval and rhythm as a search criterion. They used UDR string for melodic contour representation. For query processing, simple signal processing technique pitch is extracted from hummed query, for pitch representation they used western music transcription, (i.e. each note is identified to its nearest semitone) which is then transformed into melodic representation using UDR string. They used approximate string matching using DP for similarity matching; they also compared retrieval results under various matching regimes, The dimensions of matching include whether interval or contour is used as the basic pitch metric, whether or not account is taken of rhythm, and whether matching is exact or approximate. Based on these dimensions, they examined exact matching of: 1)interval and rhythm, 2)interval regardless of rhythm, 3)contour and rhythm and 4)contour regardless of rhythm; and for approximate matching (using dynamic programming) of: 1)interval and rhythm and 2)contour and rhythm. Their search results show that approximate contour matching produces the best result.

Kjell and Pauli [35] proposed a new representation for music data and called it *inner representation*, established from MIDI-files; for their prototype MIR system. This inner representation consists of 12 components like pitch of the note, duration of the note, start time of the note; inner onset time of the note, interval from the previous note, and the type of the chord the note is in etc., and they claimed that this representation will improve the retrieval efficiency. For more precise melodic contour representation, they proposed seven pitch intervals (small, medium and large, Up- or Down, and Same) instead of three pitch intervals (Up, Down and Same). They also proposed new encoding scheme called *2-dimensional relative code*, (a modified version of relative pitch interval), extracted from their inner representation. They used well-known string matching data structure, *suffix-trie*, for indexing structure, and approximate string matching for similarity matching. For this system input query (search key) can be given by singing, humming, whistling, playing a MIDI instrument or musical notation. So depending on the type of search key, search key is divided in three categories i.e. exact (by musical notation), semi-exact (by MIDI instrument) and noisy (by humming etc.). This system requires large storage space due to its very detailed music data representation.

Chou, Chen and Liu [36] used chordal reduction based on the melodic line to represent music data. They claimed that chord representation model, the input fault ability is equipped and requires very small storage space; but their chord representation is not according to the music theory. For indexing structure they used PAT-tree and for similarity search, approximate string matching using DP is used. Due to series of insertion and deletion cause the PAT-tree unbalance, to overcome this issue they used B+ tree data-structure. Query can be posed by number of ways, like using MIDI keyboard, by singing or by musical notation drawing on the staff provided with the GUI; no matter which way user posed the query, it is first transformed into corresponding melody and rhythm and then using proposed ***Chord Decision Algorithm***, converted into a string of chords which is used to traverse the PAT-tree for similarity matching.

In his master thesis [37], *Wei Chai* implemented a melody retrieval system that supports web based communication, his client-server architecture mainly consists of, 1) Music Database where target data in the system are all in MIDI format; 2) Melody Description Objects that capsule the melody information for efficient indexing; 3)QBH Server that receives that query from QBH Client, matches it with the melodies in the melody description objects and retrieves the target songs in a rank list; 4) QBH Client that extracts melodic contour and beat information from hummed query and sends it to the QBH Server; and 5)Melody extraction tool that extracts melody information and transforms them into melody description objects. For melody representation he used a < time signature, pitch contour vector, absolute beat number vector> triplet; and approximate string matching using DP is used for similarity matching. For query processing, simple signal processing techniques are used. He also gave a statistical overview about pitch, interval, query length (number of notes per query and seconds per query), note length and tempo of the hummed query.

According to us this system has one main limitation, that is, whenever QBH Client sends a query to the QBH Server, for matching QBH Server loads all melody description objects into its main memory for similarity matching; for small music databases this can work efficiently but for very large and diverse kind of music databases this method may fail.

Naoko Kosugi et. al.[38-40] (working at NTT Labs Japan) came up with a new QBH system that retrieve target songs based on beat information instead of note information; commonly used by earlier QBH systems. Their database consists of more than 10,000 songs in MIDI format and retrieval time is less than one second. They first they performed musical analysis over the whole database, and then based on statistical results, like note distribution, repetition structure of music, tempo distribution and interval distribution; they design indexing and retrieval method for their system. This might be one of the reasons of reasonably good performance of their system. For database construction, musical data in MIDI format is converted in to melody because most users remember a song by its melody. The melody data is then chopped up into melody pieces of constant length by sliding window method; they called these pieces as “sub-data”. Then this sub-data is converted into feature vectors. In this system each musical object is transformed into a set of feature vectors, that is, tone transition feature vector, partial tone transition feature vector and tone distribution feature vector, parameters of these vectors depends on the statistical results obtained during music analysis. So in this system each musical object is represented in multidimensional space. For indexing an improved version of VAM Split R-tree is used. For query processing, hummed query is converted into MIDI format, after double pitch error correction hummed MIDI is chopped up into hummed pieces of same window size and overlapping that of sub-data, thus multiple search keys are generated that are further used by Or-ed retrieval strategy for retrieval. Similarity retrieval is found using minimum Euclidean distance between search key feature vectors and indexed feature vectors, and a final rank list of target songs is generated.

This system may work well for QBH music retrieval but for general audio retrieval like QBE, especially when query is a characterized sound like sound of a car engine or sound of a thunder, instead of a song; this system may fail, because in this system retrieves the target object based on musical parameters instead of acoustical parameters.

Section 6

Conclusion:

Most of the previous work on melody based indexing and retrieval has been done on songs in MIDI format. This format is a descriptive format of the song. MIDI format is a method of representing musical performance information, rather than the digital recordings of real sounds. A MIDI file contains instructions to perform particular commands, such as note on and off, preset changes, events and timing information.

The advantage of using MIDI format compare over other audio file formats(.wav, .mp3, .ra, .au etc.) is that, melody extraction from MIDI format is comparatively easier than other audio formats. Melody extraction from MIDI format requires simple algorithms based on fundamental principles of melody, where as melody extraction from other audio format needs signal processing based pitch extraction, transformation of pitch contour into notes, which is then converted into melodic contour or melody line.

But MIDI format representation has its own limitations like in MIDI format tunes usually consist of multiple channels (like there are 17 channels, for standard MIDI format), where each channel usually

corresponds to a different instrument. Furthermore, there is usually polyphony (several notes played simultaneously) within a single channel. Systems using MIDI format, for melody extraction often consider each channel as a tune and polyphonic channels are made monophonic by using some algorithm that transforms polyphonic pitches into monophonic pitched based on highest pitch, chord deletion and highest energy.

Secondly MIDI format does not contain the original piece of music as would be heard in an audio representation of the piece, and it is not usually of high enough quality for casual listening.

Finally, search key extraction from the query requires more computation if melodies of target songs are extracted from MIDI format than other audio format, because after frame level processing, query would be converted into MIDI format first and then using melody extraction algorithm is transformed into melody for similarity matching.

So based on the above discussion it is clear that MIDI based indexing and retrieval systems have some inherited drawbacks, hence we have to look an alternative method to extract melody other than MIDI format from musical signal as well as more efficient indexing and retrieval strategies to meet the future challenges in this area of research. A prototype system is purposed in the next section.

References:

Psychoacoustics

- [1] W. J. Dowling, "Scale and contour: two components of a theory of memory for melodies." *Psychological Review* 85: 341-354., 1978.
- [2] R. J. Ristama., "Frequencies dominant in the perception of the pitch of complex sounds" *J. Acoustic Soc. America*, 24(1): 191-198, 1967.
- [3] R. Plomp, "pitch of complex tones" *Acoustic Soc. America*, 41(6): 1526-1533, 1967
- [4] R. Plomp, "Aspects of Tone Sensation: A Psychophysical Study." Academic Press, London, 1976.

Pitch Extraction:

- [5] D. J. Liu, C. T. Lin, "Fundamental frequency estimation based on the joint time-frequency analysis of harmonic spectral structure" PP. 609-621 *IEEE Trans.* 2001.
- [6] J. J. Dubnowski, R. W. Schafer, and L. R. Rabiner. "Real-Time Digital Hardware Pitch Detector." *IEEE Trans. On Acoustics, Speech, and Signal Processing*, ASSP-24(1):2-8, February 1976.
- [7] L. R. Rabiner and R. W. Schafer "Digital processing of speech signals" *Prentice-Hall*, Signal processing Series, 1978.

Algorithms:

- [8] D. A. Grossman and O. Frieder, "Information Retrieval: Algorithms and Heuristics", Kluwer, August 1998
- [9] A. V. Aho and M. J. Corasick, "Efficient String Matching" *Communication of the ACM* 18(6) 333-340, 1975
- [10] R. Boyer and S. Moore, "A Fast String Matching Algorithm" *Communication of the ACM* 20(10) 762-772, 1977
- [11] Baeza-Yates, and Gonet, "Fast Text Searching Allowing Errors" *Communication of the ACM* 35(10) 74-82, 1992
- [12] S. Wu and U. Manber., "Fast Text Searching Allowing Errors." *Communications of the ACM*, 35(10):83-91, 1996.
- [13] O. Gotoh, "An Improved Algorithm for Matching Biological Sequences", *J. Mol. Biol.*, 162:705-708, 1982.
- [14] R. Bellman, "Dynamic Programming" Princeton University Press, 1957.
- [15] M. J. Atallah, F. Chyzak, P. Dumas, "A Randomized Algorithm for Approximate String Matching" *Algorithmica* 29: 468-486 (2000)
- [16] K. Chakrabarti and S. Mehrotra, "The Hybrid Tree: an Indexing Structure for High Dimensional Feature Space" In *Int'l. Conf. on Data Eng. Match*, 1999.
- [17] M. Charikar, C. Chekuri, T. Feder and R. Motwani, "Incremental clustering and dynamic information retrieval" *Proc. of ACM Symposium on Theory of Computing*, 1997.
- [18] D. Comer. "The Ubiquitous B-Tree". *ACM Computing Surveys*, 11(2):121--128, June 1979
- [19] S. Skiena, "The Algorithms Design Manual" Telos/Springer-Verlag, ISBN 0-387-94860-0 (1997).
- [20] A. K. Jain and R. C. Dubes, "Algorithms for Clustering Data" Prentice-Hall, 1988.

Signal Parameter Based Audio/Music Retrieval:

- [21] Brown, M., Foote, J., Jones, G., Spärck-Jones, K., Young, S., "Open-Vocabulary Speech Indexing for Voice and Video Mail Retrieval," *Proceedings ACM Multimedia 96*, Boston, November 1996.

- [22] J. T. Foote., "Content-Based Retrieval of Music and Audio." In *Proc. SPIE, vol3229*, pp. 138-147, 1997.
- [23] J. Foote., "An overview of audio information retrieval." In *Multimedia Systems 7*, pp 2-10. ACM, January 1999.
- [24] J. Foote., "Visualizing Music and Audio using Self-Similarity." In *Proc. ACM Multimedia 99*, pp 77-80,1999.
- [25] Muscle Fish LLC. <http://www.musclefish.com/>.
- [26] A. Khokhar, G. Li "Content-based Indexing and Retrieval of Audio Data using Wavelet" ICME 2000
- [27] S. R. Subramanya and A. Youseef, "Wavelet-based Indexing of Audio Data in Audio/ Multimedia Databases" 4th IEEE Int'l. Workshop on Multimedia DBMS, Dayton, Ohio, August 1998.

Melodic Based Music Retrieval:

- [28] A. Ghias, J. Logan, and D. Chamberlin. B. C. Smith, "Query By Humming." In *Proc. ACM Multimedia 95*, pp. 231-236, 1995.
- [29] TC, ALP Chen, and CC Liu, "Music Databases: Indexing Techniques and Implementation," in Proc. of IEEE Intl. Workshop Multimedia Data Base Management System, 1996.
- [30] L. Lu, H. You, H. J. Zhang, "A New Approach to Query by Humming in Music Retrieval" in ICME2001, Tokyo, August 2001
- [31] S. Blackburn and D. DeRoure. "A Tool for Content Based Navigation of Music." In *Proc. ACM Multimedia 98*.
- [32] McNab, R.J., Smith, L.A., Witten, I.H., Henderson, C.L., and Cunningham, S.J, "Towards the digital music library: tune retrieval from acoustic input." In Proceedings of ACM Digital Libraries '96, 1118.
- [33] R. J. McNab, L. A. Smith, D. Bainbridge, and I. H. Witten., "The New Zealand Digital Library MELody inDEX." <http://www.dlib.org/dlib/may97/meldex/05written.html>, May 1997.
- [34] Bainbridge, D. (1997) Extensible optical music recognition. PhD thesis, Department of Computer Science, University of Canterbury, New Zealand., 1997
- [35] L. Kjell and L. Pauli, "Musical Information Retrieval using musical Parameters" International Computer Music Conference, Ann Arbour, 1998.
- [36] A. LP Chen, M. Chang and J. Chen. "Query by Music Segments: An Efficient Approach for Song Retrieval". In the Proc. of ICME 2000.
- [37] W. Chai. "Melody Retrieval on The Web", Master thesis Media Art and Science at MIT, 2001.
- [38] N. Kosugi, Y. Nishihara, T. Sakata, M. Yamamuro, and K. Kushima., "A Practical Query-By-Humming System for a Large Muac Database." In *Proc. of the 8th ACM International Conference on Multimedia*, 2000.
- [39] N. Kosugi, Y. Nishihara, S. Kon'ya, M. Yamamuro, and K. Kushima., "Let's Search for Songs by Humming!" In *Proc. ACM Multimedia 99 (Part 2)*, pp 194, 1999.
- [40] N. Kosugi, Y. Nishihara, S. Kon'ya, M. Yamamuro, and K. Kushima., "Music Retrieval by Humming." In *Proc. Of PACRIM'99*, pp. 404-407. IEEE, 1999.
- [41] P. Y. Rolland, G. Raskinis, and J. G. Ganascia., "Musical Content-Based Retrieval: an Overview of the Melodiscov Approach and System." In *Proc. ACM Multimedia 99*, pp. 81-84, 1999.
- [42] A. Uitdenbogerd and J. Zobel., "Melodic Matching Techniques for Large Music Database." In *Proc. ACM Multimedia 99*, pp. 57-66, 1999.
- [43] A. Yoshitaka and T. Ichikawa., "A Survey on Content-Based Retrieval for Multimedia Databases." *IEEE Trans. Knowledge and Data Engineering*, 11(1):81-93, 1999.
- [44] T. Kageyama, Y. Takashima, "A Melody Retrieval Method with Hummed Melody", Journal of The Institute of Electronics Information and Communication Engineers, pp.1543-1551,1994
- [45] Keislar, D., Blum, T., Wheaton, J., and Wold, E., "Audio Databases with Content-Based Retrieval," Proc. of the International Computer Music Conference 1995, pp. 199-202. 1995.
- [46] Wold, E., Blum, T., Keislar, D., and Wheaton, J., "Content-Based Classification, Search and Retrieval of Audio," *IEEE Multimedia*, 3(3), 27-36, Fall 1996.
- [47] Blum, T., Keislar, D., Wheaton, J., and Wold, E., "Audio Databases with Content-Based Retrieval," in *Intelligent Multimedia Information Retrieval*, AAAI Press, Menlo Park, California, 113-135, 1997.