A Survey on Query by Singing/Humming

Vyankatesh Kharat
Student,
Department of IT, Sinhgad College of Engineering, Pune, India

Kalpana Thakare
Associate Professor,
Department of IT, Sinhgad College of Engineering, Pune, India

Kishor Sadafale
Assistant Professor,
Department of IT, Sinhgad College of Engineering, Pune, India

ABSTRACT
Performing comparison search in huge databases is a difficulty of particular concern in several communities, such as music, database, and data mining. A number of query by humming/singing (QBH) systems have evolved in recent years, which can search for the song without manual input. Query by humming systems will return a structured list of songs according to the similarity between humming and intent song according to the given melodies hummed by the users. Query by humming uses a content-based music information retrieval (MIR) method which is an efficient way to search the song from a large database. This paper is focused on providing a brief overview of query by singing/humming systems and methods are available in literature.

Keywords
Query by humming/singing, Music Information Retrieval, dynamic time warping, Hidden Markov Models

1. INTRODUCTION
The popularity of mobile devices (for example, tablets and smart phones) has lead to the fast increase of a variety of applications. One of the most common applications is listening to music. Consumers can now use mobile devices to play music anywhere, for example, when they are exercising or driving. Now music search or music suggestion, automatic playlist creation are associated trouble. In requirements of song searching, one can use a song's related information (for example, song title, artist, etc.), or the content of a music categorizer (for example, melody).

A user with a mobile device can simply search for a song through a voice recognition system. So to carry out a search, the user can call for the song title or artist information. But, people often cannot remind the song title or the name of the artists and only part of the song is remembered. The QBH is one of the most powerful way to find music when users do not have any metadata such as lyrics, title of song, and name of singer. The key function for query by a singing/humming (QBH) system is to carry out song searching derived from the melody sang or hummed by the user.

The basic purpose of a QBH system is searching for the song that is most similar to the query given by the user in the DB. The main condition for the QBH system is how it can find the music accurately and rapidly. The conventional QBH systems have been developed with DB created from monophonic records such as musical instrument digital interface (MIDI) files instead of the original polyphonic music files. For commercial applications, to develop the melody extraction method from polyphonic music signals for creating DB and the matching engine is desirable.

Different query by humming/singing system are broadly classified into two different groups according to the input given by the user. The input given by the user can be a hum, whistle or singing query.

2. LITERATURE REVIEW
A basic block of query by humming/singing is shown in figure 1. The acoustic query from user is normally a few notes whistled, hummed or sung by the user, is recorded with the help of microphone**.** The signal processing is preferred for extracting the melody from the melody database and from acoustic query given by the user. The pattern matching algorithm is used to achieve proper ranked list of matching melodies. There are number of query by humming systems that are present in the literature with more advanced features in the hardware and software modules in the mobile phones. Some of QBH systems are presented here

Most of the QBH systems are realize by means of frame-based algorithm. These methods are more accurate than the note-based ones. However, the note-based methods are more efficient than the frame-based ones. Jia Liu et al. [1] proposed the design of a query by humming system based on notes, which is mainly comprised of noted-based linear scaling (NLS) and noted-based recursive align (NRA). This system introduced an accurate note-based multi-stage system. The noted-based algorithms does not need to judge against the song with the humming query frame-by-frame. This method mostly uses the pitch and time information of the note and ignores the timing information of the music. The pre-processing is done during and after segmentation to remove the tracking errors and to get appropriate note features after the processing. One of the most significant advantage of this system is it uses note-based methods. The note-based methods are more efficient than the frame-based one but the frame-based are more accurate than the note-based ones. Drawback of this system is that the distance between the silence of humming and the song is not calculated and in doing so the system can withstand some segmentation errors. The NLS and NRA are not sensitive to rhythm of the humming due to scaling, however sensitive to the pitch values. The absolute pitch value is used in the system because of which the result is not very perfect in the situation like difference of pitch between the humming and the song. For example, different people may use dissimilar start tones, as a result comparing the humming and the song directly will produce inaccurate result.

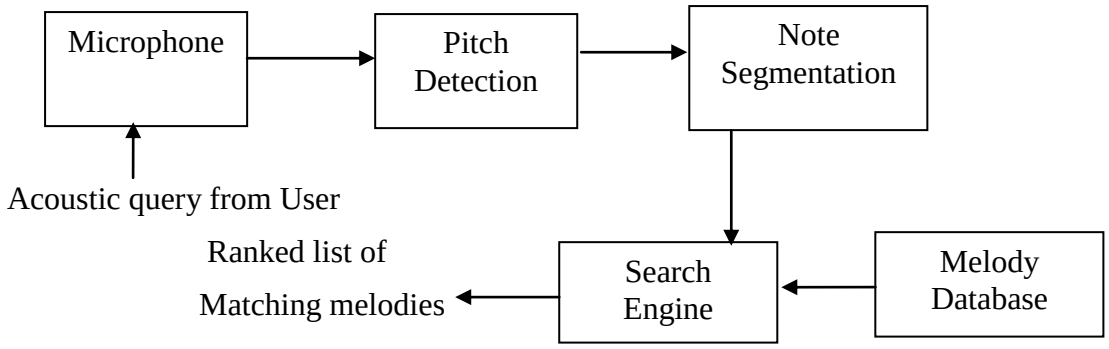

Figure 1: Basic blocks of query by humming/singing

Wennen Wang et al. [2] proposes an enhanced query by humming/ singing (QBH) system using melody and lyrics information together for achieving better performance. In this architecture, mobile users are able to retrieve intended songs by singing or humming a portion of the intended song in order to retrieve it. Most of the QBH researches so far utilize melody information as the only cue for retrieval. Lyrics are also an important part of a song which can be served as the cue for detecting the song's identity or its mood or genre. However, the use of lyrics for content-based music analysis appears much later. Singing/humming discrimination (SHD) is used in this system to differentiate singing and humming queries. If the query is classify as humming, the result is based on melody recognition only. On the other hand, if the input is classified as singing, lyrics recognition is performed to obtain a decoded sequence of lyrics. The output of the system then uses the combined scores of melody and lyrics. The main disadvantage of this system is SHD can misclassify humming clips as singing, which may generate erroneous output in lyrics recognition. Also initial error analysis indicates that several misclassified humming clips are caused by a variety of pronunciation during the humming. Also there are few famous songs with the same tune but they have different lyrics in different languages this can give erroneous output to the user.

The main challenges for QBH systems contain i) queries vary for different users, how to extract audio features which accurately represent music content, ii) how to illustrate the musical features, and iii) which technique to be implement for feature matching. In [3] challenges are being tried to be solved effectively. There are three levels of musical feature which are physical features, acoustic features and perceptual features. The physical features express audio content on the format of flow media. The acoustic feature mainly includes time and frequency domain features. They are the most expressive feature of audio and are usually used for different phases of speech recognition. Perceptual features reflect people's feelings such as pitch, rhythm, intensity, timbre, etc. A template with standard fundamental frequency range to estimated the input humming template is used and be matched with the standard template instead. A melody contour alignment algorithm based on GA (Genetic Algorithm) is suggested, which might be a linear shift to input templates, and seems to reserve the detail information rather than using normalization directly. LSH (locality sensitive hashing) NN search is employed for templates indexing and matching, the final ranked list seems to be improved. Euclidean distance is the most widely used similarity measure for time series similarity research. The benefit of LSH against Euclidean distance is that it can get a sublinear time complexity. The database used in this system is very small. Much work needs to be done to solve interference from high similarity of music fragments in large database.

M. Anand Raju et al. [4] proposed a TANSEN, a query-by-humming music indexing and retrieval system based on melody, or the "tune", of the music. The melody database used here is an ordered set of tracks. The audio query is generally a few notes whistled, hummed or sung by the user that are processed to identify its melody line. After processing this query the system returns a ranked set of matching melodies which can be used to retrieve the desired original soundtrack. The major modules in this system are extraction of a melody representation from the query and the melodic similarity distance computation. One of the main challenges arise in the implementation of this system is signal processing required to extract the melody from the stored music database and from the audio query, also the pattern matching algorithm to get proper ranked result. To recognize melodies relative frequency intervals between the notes are calculated. This relative variation of pitch in time is known as the "pitch contour".

For note segmentation it is required that the query to be sung using a syllable such as "ta" and the stop consonant "t" causes the local energy of the waveform. The instantaneous energy of query waveform is computed. This energy contour requires smoothing because of the energy spikes that are created due to improper recording, stray mic clicks. In this system time domain autocorrelation function is used for pitch extraction since it is computationally simple and fast. It is computed on non-overlapping frames of fixed duration. Typical inaccuracies in the query given by the users are (i) insertion of new notes (ii) replacement by different note (iii) deletion of notes. These inaccuracies can be taken care of by a dynamic programming (DP) based "edit distance" algorithm. For matching purpose, DP is used to obtain minimum edit distance between two sequences. If minimum edit distance between two sequences is 0, then it is an exact match. If the minimum distance is high, then the sequences are considered to be very dissimilar. The main disadvantage is to sing the query using a syllable such as "ta" and the stop consonant "t". The results of this system showed a 95% success rate.

Trisiladevi C. Nagavi et al. [5] proposed MIR system which uses content based search for music and require no musical acquaintance. The system provides an analysis on QBH through query excerpt. The design of QBH system consists of two distinguished phases. The first phase is training and the second is referred as operation or testing phase. Each of these phases performs different operations on the input signal such as Pre-processing, Vocal and Non-Vocal Separation, Feature Extraction and Query Matching. Pre-processing is useful to extract information required by the system. Songs are decoded into wave streams and converted to mono channel. In any music, human vocal component always acts as an important part in representing melody than its background music. Centre pan removal techniques used in many karaoke machines to

extract the lead voice from a song. Extracting important feature from an audio is a main task to generate a better retrieval presentation.

In this system Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Coefficients (LPC) and Linear Predictive Cepstral Coefficients (LPCC) features are favored as they show potential in terms of discrimination and robustness. MFCC is base on the information approved by low-frequency portions of the audio signal. The plan of MFCC is to generate finest approximation of the human auditory system's response. LPC examination is to characterize each section of the signal in the time domain by a linear combination of preceding values. LPC analysis uses the autocorrelation method. LPCCs are the coefficients of the Fourier transform demonstration of the logarithm magnitude spectrum. Euclidean Distance (ED), K-Nearest Neighbor (k-NN) and Dynamic Time Warping (DTW) are used for distance measures. The ED measure is the standard distance measure involving two vectors. k-NN is supervised learning algorithm for matching the query instance based on majority of k-nearest neighbor category. Minimum distance among query instance and each of the training set is designed to determine the k-NN category. The DTW is the third measure for finding likeness between two time series which may differ in time. DTW produced a little better retrieval presentation compared to other two distance measures since it minimizes the total distance between the respective points of the signal. Mean Reciprocal Rank (MRR) is a statistic for assessing any organism that produce a list of possible responses to a query, ordered by probability of correctness. The main advantage of this system is that song retrieval accuracy increases with increase in song/hum percentage. This system shows that the size of the databases and song/hum percentage are main factors that determine the success of the system.

Matti Ryynanen et al. [6] proposed a query by humming technique based on locality sensitive hashing (LSH). The QBH task can be broadly divided into two sub parts: i) converting a query into a format which enables robust searching and ii) matching the query with melodies in the database. The technique constructs a directory of melodic fragments (refers here to a melody pitch contour in a fixed-length time gap.) by extracting pitch vectors. A pitch vector stores an approximate representation of melody contour within a fixed-length time window. The similarity of melodic fragments is here measured using Euclidean distance between pitch vectors. Euclidean distance is not only simple but also appears to be very effective measure for similarity. For all query pitch vectors, the technique looks for nearest neighbors in Euclidean space from the directory of database melody fragments, which obtain melody candidates and their matching positions in time. This can be done by using locality sensitive hashing (LSH). LSH is a randomized algorithm for finding nearest neighbors approximately in high dimension spaces.

A sung query is first converted into a note sequence for this a melody transcription method is used. The method produces a sequence of notes in the specific format. The tuned query note sequence is then used to retrieve similar melodic fragments from the database by extracting pitch vectors from the query notes. For all query pitch vectors, the technique then search for identical melodic fragments in the database using LSH. The LSH returns the nearest neighbors and their distances to the query point as matches. The final list of retrieved melodies, the candidate melodies are ranked according to their distance to the entire query note sequence. The main drawback of this method is it only uses databases of MIDI melodies.

Most existing QBH systems have been developed on basis of MIDI files, instead of polyphonic music tracks such as MP3 files. Sungjoo Park et al [7] discusses various implementation issues in developing the QBH system for polyphonic music retrieval service. The pitch information extracted from the polyphonic music is not an accurate as one from MIDI files. To overcome this inaccuracy problem the matching engine in the QBH system should be design. The QBH system generally consists of feature extraction processes and matching engine. The design of the matching engine used here mainly focuses on reducing the inaccuracy problem of the feature extraction. The proposed QBH system uses the matching engine for the music retrieval based on pitch sequences of acoustic inputs: one is a polyphonic music input from DB and the other is a user's query. The proposed QBH system consists of three functional modules:

The first functional module is to extract features from users humming query and suppress noises in user singing/humming input. The second module is to extract features and build DB of the polyphonic music; this can be done by using the harmonic structure of vocal and musical instruments. The last module is the matching engine which is used to find the music which has the most similar feature with user singing/humming input. For any QBH System the matching engine that is used, it must be robust against the mismatch between the pitch sequences of user query and the pitch sequence stored in DB. Errors must be considered that are occurring in extracting features as well as reside in user query and music DB. To avoid these errors chroma-scale representation, compensation, and asymmetric DTW are adopted in the system. The matching engine based on the DTW (Dynamic Time Warping) algorithm is to find the shortest DTW path the distance metric.

Chai-Jong Song et al in [8] proposed a method for extracting the predominant melody of polyphonic music based on harmonic structure. Harmonic structure is an important feature of music signal that has spectral peaks at the integer multiples of its fundamental frequency. All fundamental frequency candidates contained in the polyphonic music signal are extracted by verifying the required condition of harmonic structure and also a rank is assigned to each by calculating its harmonic average energy. After this a pitch tracking based on the rank and the continuity of extracted fundamental frequency is run and finally the predominant melody is determined. The system proposes a matching engine based on the DTW (dynamic time warping) algorithm that provides better retrieval accuracy by adopting an asymmetric sense with weighting coefficient, chroma-scaling representation, compensation, and an optimal distance metric.

The first functional module suppresses the noise in user singing/humming signal and extracts pitch sequence as an input query. Spectro-temporal autocorrelation is used for this module. The second module extracts the main melody from the polyphonic music from DB. The proposed melody extraction method is based on harmonic structure of input signal and vocal signals that has spectral peaks. The melody extraction is done on a frame basis. Multi-pitch extraction module searches for all possible meaningful candidates contained in the signal. It first searches for all local peaks from the input spectrum, and those with magnitude larger than a given threshold are selected as the valid local peaks. Since music signal has different characteristics for each frame, the threshold for each band is decided adaptively according to the

spectral shape which is computed by the spectral skewness. There is a difference in the length of the query input and the original music track. Also there are some errors in the pitch sequences extracted from the query and the polyphonic music. The last module is the matching engine which finds music from the feature DB which is most similar to the input query. To reconcile these issues DTW matching algorithm is used. It is widely used as the matching engine in the QBH system, as it gives a robust matching result against a local timing variation and inaccurate tempo. The pitch information extracted from the polyphonic music and user's query are not perfectly accurate. This inaccuracy can cause a data distortion problem. To reduce this distortion, a distance metric that is not sensitive to the distortion should be used.

3. COMPARISON

Table 1. Comparison of different Query by humming/singing systems. (Courtesy [1, 2, 5, 8])

Sr. no.	Topic name/journal	Work description	Problems found	Publishing year
1	INTERSPEECH	Query by humming system based on notes	Absolute pitch value is used in the system because of which the result is not very perfect	2010
2	ISMIR	Enhanced query by singing/humming system using melody and lyrics information together	May misclassify humming clips as singing	2010
3	IJMA	Query by singing/humming through query excerpt	Small database used.	2012
4	IEEE	extracting the melody of polyphonic music based on harmonic structure	Inaccuracy in pitch extraction can cause a data distortion problem.	2013

4. CONCLUSION

Recently, a number of query by humming/singing (QBH) systems have evolved which can search for the song without manual input but each one of them has their own advantages and disadvantages. In this paper a brief overview of different techniques and methods used in the query by humming/singing (QBH) such as ways of segmentation and pitch tracking of music pieces from user as well as from the music database are presented. Also the different methods for matching the audio query from user and music from the database are presented that have been proposed in the literature for the QBH application.

5. REFERENCES

[1] Jia Liu, Yang Jingzhou and Weiqiang Zhang. "A fast query by humming system based on notes." In INTERSPEECH, pp. 2898-2901. 2010.

[2] Wang Chung-Che, Jyh-Shing Roger Jang, and Wennen Wang. "An Improved Query by Singing/Humming System Using Melody and Lyrics Information." In ISMIR, pp. 45-50. 2010.

[3] Qin Jing, Hongfei Lin, and Xinyue Liu. "Query by humming systems using melody matching model based on the genetic algorithm." Journal of Software 6, no. 12 (2011): 2416-2420.

[4] Raju M. Anand, Bharat Sundaram, and Preeti Rao. "TANSEN: A query-by-humming based music retrieval system." In Proc. National Conference on Communications (NCC). 2003.

[5] Nagavi Trisiladevi C. and Nagappa U. Bhajantri. "An Extensive Analysis of Query by Singing/Humming System through Query Proportion." arXiv preprint arXiv: 1301.1894 (2013).

[6] Ryynanen, Matti, and Anssi Klapuri. "Query by humming of midi and audio using locality sensitive hashing." In Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, pp. 2249-2252. IEEE, 2008.

[7] Park Sungjoo, and Kwangsue Chung. "Query by singing/hum ming (QbSH) system for polyphonic music retrieval." In Consumer Electronics (ICCE), 2012 IEEE International Conference on, pp. 245-246. IEEE, 2012.

[8] Song Chai-Jong, Hochong Park, Chang-Mo Yang, Sei-Jin Jang, and Seok-Pil Lee. "Implementation of a practical query-by-singing/humming (QbSH) system and its commercial applications." Consumer Electronics, IEEE Transactions on 59, no. 2 (2013).