# Query by Humming: A Brief Overview

Information Retrieval System

Shivam Panchal
*Department of Computer Science*
*Nirma University*
*19bce150@nirmauni.ac.in*

Priyal Palkhiwala
*Department of Computer Science*
*Nirma University*
*19bce214@nirmauni.ac.in*

Aayush Shah
*Department of Computer Science*
*Nirma University*
*19bce245@nirmauni.ac.in*

***Abstract*—Query by humming (QBH) technique for music retrieval is a widely used and experimented method to achieve accurate matching of the queries. It is much different than the classification into title, artist, composer and genre. QBH is generally a song or a part of music with a single topic or tune. QBH system compares a user-hummed tune (input query) to an existing database. After that, the system delivers a ranked list of music that comes closest to the supplied query. Similarity search in large databases is a concern that has piqued the interest of many institutions, including music, databases, and data mining. There are many proposed solutions in the literature which perform well in several domains but there isn't one best method to solve the discussed QBH application. The goal of QBH is to find the songs that are most similar to a hummed query in a feasible manner. In this paper, we have given a brief overview of the matching methods and the previously proposed solutions for QBH system.**

***Index Terms*—Query-by-Humming, Similarity Matching, Music Information Retrieval, Dynamic Time Warping, Noise reduction, Machine Learning**

## I. Introduction

Many individuals listen to tunes on demand on their mobile devices. People utilise a variety of strategies to find their favourite songs, including search-by-text, which involves searching for a song based on a snippet of the lyrics, the artist's name, or other criteria. Some apps, such as 'Shazam,' allow users to record and search for a music playing in the background. However, there is a problem to this strategy, which happens when users forget the words to a new song or miss the song that is playing in the background. Humming can be used as a query to find songs as a solution to this problem. When humming is prolonged, it produces a low, monotonous sound similar to that of speaking. The proposed system would turn a hummed melody into music and then compare it to a music database to find the most comparable tune/song.

Because there are numerous ways for retrieving music, Music Information Retrieval (MIR) is a particularly interesting research subject. It's critical to select a match between the input query and the matching music while retrieving music. Query-by-Example (QBE) is one of the methods that has been presented and is presently being used in the various application domains. In QBE, a snippet of an audio recording playing in the background is fed as an input and the result is returned as the output. In a Query-by-Humming (QBH) application, however, there is no effective way to tackle this problem. The goal of a Query-by-Humming application is to quickly obtain music that is most comparable to the hummed query.

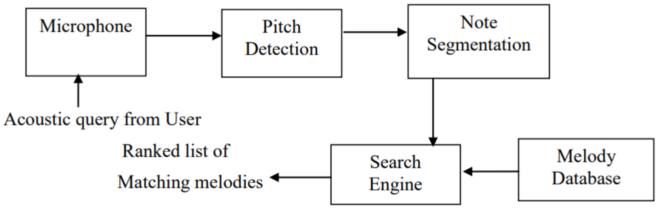

Figure 1. The primary stages of query by humming system

We'll examine the various music information retrieval strategies and associated system designs in this paper. In addition, We'll go through the Query-by-Humming method and its many approaches, which allows for a revolutionary manner of music retrieval [2].

## II. Techniques for Music Information Retrieval

MIR is a gradually advancing discipline with a possible future in quick information retrieval. This is due to the fact that it is quite similar to database retrieval; however, MIR uses a variety of approaches to retrieve music in a timely and effective manner. MIR encompasses multiple disciplines, including psychology, musicology, signal processing, optical music recognition, and machine learning (ML). Businesses and researchers employ MIR in applications like as automatic music transcription, recommender systems, automatic classification, and music production. The rest of this section goes into the approaches utilised by MIR systems.

### A. Query-by-Text (QBT)

A QBT approach searches for similarities between songs using conceptual metadata such as text queries. This functionality is used by applications such as 'Spotify', 'Apple Music' and 'Youtube Music' for their music retrieval mechanism. Because it relies on previously known text that can be searched through the database, this was the very first technique introduced in the field of retrieval.

Identify applicable funding agency here. If none, delete this.

### B. Query-by-Example (QBE)

The QBE technique, on the other hand, takes a segment of the original music recording and searches the database for the most similar song. 'Shazam' and 'SoundHound' are well-known examples of this type of technology in use in real-world applications. They employ a technique known as Audio Fingerprinting. This is the method of expressing an audio signal in a compact manner by extracting important features of the audio content.It is used to identify or search for an audio based on the fingerprint generated by the query sample. This is a well-known approach that is currently in use because it is quick and does not require the entire audio sample.

### C. Query-by-Humming (QBH)

To query the database, the QBH approach only employs the natural humming voice emitted by humans. Furthermore, this strategy is appropriate because humming occurs spontaneously and can be associated in the user's mind. The Google Search app can recognise a song based on your humming or whistling, making it easier to discover new music. This is something we can accomplish directly on our mobile devices. It might also provide a score to reflect the similarity in melody between the user's and the original artist's singing.

## III. Related Work

In this section, we have discussed about the system architectures proposed in the past, for query by humming system to improve the accuracy and the ease the process in general.

There are many works proposed to improve the QBH system and have proved to be much useful for the music/song retrieval. Ghias et al [7] developed a QBH system based on the contour approach in 1995. They adopted auto-correlation to track the sound's pitch and transform it to a string contour. They employed an alphabet of three potential pitch interactions ('U', 'S', and 'D'), each depicting a predicament in which a note is above, the same as the previous note, or below. In addition to it, they deployed a string-matching technique which helped in matching the query with the songs stored in the database. Nevertheless, their framework was not powerful enough for a large database, so the time it took to retrieve the song was prolonged. Another framework proposed by N. Kosugi [3] in 1999 devised a song retrieval system that segregated the original music data into sub-data and allowed users to sing/hum a certain piece of melody to retrieve it. To improve the accuracy of music retrieval, their system employs both tone distribution and tone transition. They had to deal with a number of issues, including reducing the partitioned sub-data and broadening the song dataset.

In 2003, Y. Zhu et al. [17] implemented DTW to examine the performance of the contour and time-series approaches. The analysis indicates that time-series information matching gave 80% retrieval accuracy. The previously discussed system designed by K. Adamska et al. [18] describes how the contour approach can be implemented to retrieve music in a QBH system. They asserted about converting the hummed file to MIDI format and further extracting melodic contour from it. The determination of elemental frequency for every time bit is known as a musical contour. A musical contour database can also be created by converting a MIDI file database. Finally, the song data matching and recognition algorithm will identify the song and obtain it from the database. The issue with this approach is that it is only feasible for a small database to produce accurate results. This, however, presents a challenge for a large music database, as is typically the case in real-world applications.

In 2005, the business 'SoundHound' developed a humming-based audio retrieval technology. They employed machine learning methods to train the neural network on the hummed music before extracting the song, which is a little different. They used a massive database of labelled audio samples to achieve this with high precision. They also have a database of hummed songs from a random sample of people, each of which is labelled with its original categorised tune. There was another database ready with them which included hummed songs obtained from random people which were already labelled with the original classified audio. The model they designed extracts the audio characteristics such as tone and rhythm. It determines the feature vectors by creating pairs of humming and the original songs. If the query and song were expected to be similar in terms of their respective features then their score increases. If they aren't similar to each other, the score drops. As a result, whenever a new hummed query appears, the trained model recognises what features it shares and will consequently retrieve the music.

However, after testing their framework, most attempts to retrieve the correct song fail. However, the framework majorly failed to retrieve the accurate song when the system was tested. For instance, it would often display the accurate results in their top five list and other times it would not display the result at all. This issue must have occurred if the model had over-fit on the dataset hence, it cannot accurately estimate and generalise the new data. Due to the computational load of the model, music retrieval using the ML or DL approach typically takes a long time.

## IV. Various Architectures for QBH system

In this section, we have discussed the various ways through which the outcome for the QBH system can be obtained with their applications performed before.

### A. Music Information Retrieval System Architectures

To accomplish the retrievement of music effectively through a system, one needs to follow the architectures structured of MIR system. Because of the various strategies employed throughout the system to strive better music retrieval than other systems, they have deployed diverse structures. A paperwork proposed by N. Kosugi et al. [3] concludes that we can use feature vectors [8] through a MIR architecture which are derived from the Musical Instrument Digital Interface (MIDI) file. In the same way, to transform the primary music sequences into feature vectors, the database needs to be modified. It is followed by a similarity search between the two components and the answer is retrieved. Another work proposed by R. Putri [9] presents that instead of using feature vectors, one can use a time-series data mapping algorithm named Dynamic Time Warping (DTW) approach. DTW is a time-series data similarity technique that proactively evaluates two different time-series data. Since individuals have various verbal skills, this is a prominent feature for domains like speech recognition or MIR (such as speed, tone, etc.). Furthermore, the amount of time would vary even if the same person spoke at different times during the day, as can be seen in Figure 2. This approach is built to be reliable and possess the potential to achieve greater retrieval accuracy. The DTW algorithm can compensate for out-of-sync tunes and sequences, as well as time warps. As a result, it is much more efficient than the feature vector method.

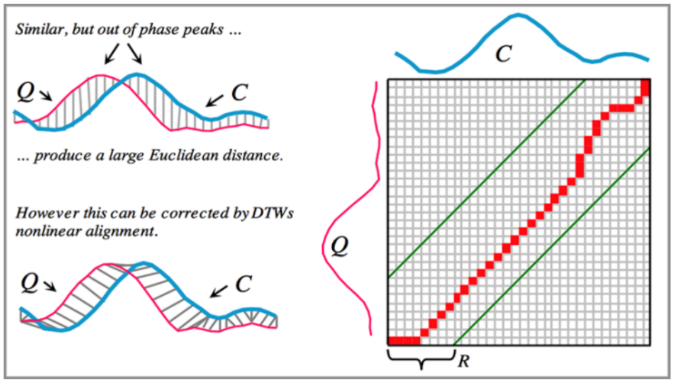
Similar, but out of phase peaks ...
Q
C
... produce a large Euclidean distance.
However this can be corrected by DTWs nonlinear alignment.
Q
C
C
Q
R

Figure 2. Dynamic Time Warping Theory [10]

### B. Speech Feature Extraction

Extraction of speech features is responsible for the conversion of speech signals into a sequence of feature vector coefficients that only entails information needed to identify a specific syllable. Because each speech has its own set of unique characteristics in pronounced terms, these characteristics can be recovered using a variety of feature extraction methodologies and used for speech recognition. However, when handling the speech signals, extracted features should meet specific criteria, such as: extracted speech characteristics should be easy to quantify, these features must also be stable over time, and they should be effective to noise and environment [12].

Spectral analysis techniques like Mel-frequency cepstral coefficients (MFCCs), linear predictive coding (LPC) wavelet transforms, Perceptual Linear Prediction (PLP) etc are commonly used to extract the feature vector of voice signal [11].

- Mel Frequency Cepstral Coefficients [11] (MFCCs): These are a small group of characteristics that characterise the overall shape of a signal's spectral contour. It's a popular method for extracting speech features, and it was first used to analyse seismic echoes caused by earthquakes. The Mel scale is used to analyse a note's reported frequency to its experimentally determined frequency. It adjusts the frequency to be as close to what the human ear can hear as possible.

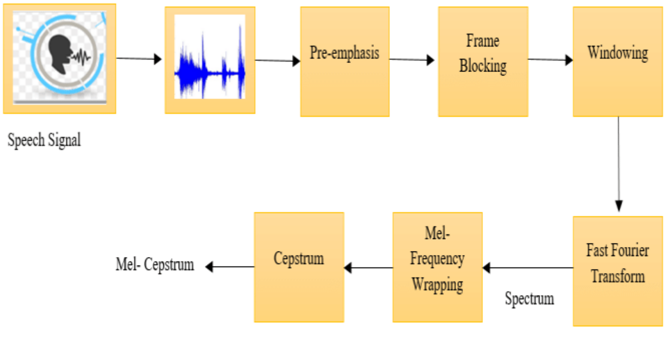

Figure 3. Process of MFCC [16]

- Linear Predictive Coding [11]: In this technique, the vocals of human vocal tract are imitated which produces an effective speech feature. It estimates the concentration and frequency of the left-over residue by estimating the formants [13], removing their effects from the speech signal, and evaluating the speech signal. Each sample of the signal is stated to be a direct assimilation of previous samples in the result. The formants explained here are originated from the coefficients of the differential equation. Therefore, we need to estimate the coefficients. LPC is a popular formant estimation method and a powerful speech processing tool [13].

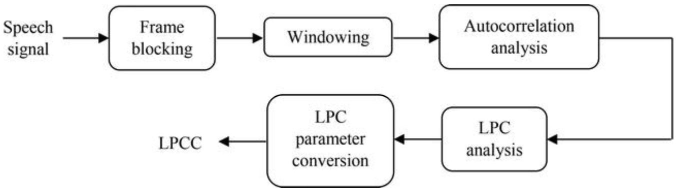

Figure 4. Process of Linear Predictive Coding

### C. Audio Fingerprinting

The technique of extracting essential information from an audio stream in order to express it in a compact manner is known as audio fingerprinting. It is based on the concept of human fingerprinting. It records the fingerprint of ingested audio content, which may later be compared to previously recorded audio or playlists on a phone, television, or other device. Audio fingerprinting allows audio files to be monitored regardless of their format or metadata. A robust acoustic fingerprinting identifies the audio file even after compression and sound quality loss. Audio retrieval based on content, broadcast monitoring, and other acoustic fingerprinting applications are only a few examples. On the market, Shazam is a well-known music-retrieval application.

The core feature behind all the techniques is the audio spectrum, where the frequency bins of each frame is plotted across time axis as it can be seen in the Figure 5.

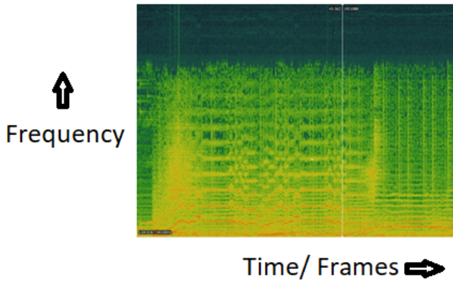

Figure 5. Spectogram generated by plotting frequency across time axis [1]

### D. Song Retrieval - Application of Machine Learning

ML is a constantly evolving field with active research and contributions. Because neural networks are replications of how the human brain processes information, ML can be used to train a neural network to perform tasks that would normally take a long time to complete. N. Mostafa et al. [15] envisaged a Deep Neural Network (DNN)-based note-transcription method that uses hummed notes as features to train the neural network [15]. The features extracted are made to pass on to the hidden layers present in DNN. Henceforth, these layers assist the neural network to train the input query 'deeply'. They integrated Hidden Markov Models (HMM) with Gaussian Mixture Models (GMM) also known as HMM-GMM, the accuracy of Mean Reciprocal Rank (MRR) reached to 0.7679 and when DNN-HMM-based acoustic model was used, the accuracy obtained was 0.8071. These readings were found on a small database containing 4431 songs with 116 Bollywood artists. The authors deduced that the accuracy score of transcription and retrieval system might increase when the DNN is trained on a comparatively larger dataset.

JQ Sun et al. [10] applied deep learning techniques on MIDI database to employ query by singing/humming (QBSH) system. Song retrieval is a broad field with numerous approaches to the specific problem. They implemented the aforementioned approach for the audio matching. They compared the before discussed DTW technique to the deep learning approach and obtained very little difference between the readings of MRR. They got a MRR of 0.79 under DTW approach and that of 0.82 on DL approach when implemented on a dataset having the size of 200. DL approach consisted of Deep Belief Networks [10] having multiple hidden layers where every layer is supposed to learn the hidden features of a higher order correlated process. Hence, they concluded that DL approach gave better accuracy between the humming query and the resulting song. But, here the challenge was that they implemented it on a very small dataset and it is proposed to be the future work on implementing DNN in a more generalised form on a larger dataset. As a result, the DTW approach is not limited to this problem and can be generalised to a larger dataset with minimal changes using the same system.

### E. Noise Reduction Techniques

To reduce noise from a signal, noise reduction techniques are used. This is constantly used in the field of Natural Language Processing (NLP) because it is the primary issue. Using well-known approaches, ambient noise can be reduced when audio input is coming from the microphone. For editing audio and removing noise, open-source software known as 'Audacity' can be utilised. However, because the noise in QBH systems must be decreased with live input, data cannot be post-processed using software. In QBH systems, on the other hand, this can be used to 'clean' the data before training the neural network. It's vital to know the frequency of the overall audio and the humming sound in order to figure out which noise to cancel.

Active Noise Reduction (ANR) is a way for controlling or reducing noise using live input. It is a method for reducing unwanted sound by adding a sound that is specifically designed to cancel out the noise. As a result, you'll have a clear audio input that can be processed further. Because the female humming frequency range is 350 Hz to 17 kHz, and the male frequency range is 100 Hz to 8 kHz, this feature is effective for QBH systems due to its resilience. As a result, this parameter in the QBH system can be tweaked to see whether there is any ambient noise that can be minimised.

## V. Conclusion

In this paper, we have provided a brief analysis of the similarity techniques based on different learning mechanisms. We have also given the overview of the past work which proposed solutions for QBH system and their contributions towards the better understanding and developing of the song matching with the hummed query. Hence, we conclude that QBH technique possess a potential to be mined deeper and improve the system as whole.

## VI. Future Work

We propose future work involving the use of this method in conjunction with an ML or DL model that may be trained on past humming data to recognise new/unknown hummed requests and the music associated with them. In order to enhance the accuracy, we can introduce hidden layers to the model and update the parameter values to match with the query of humming. Implementing the model on a larger dataset will also help for the same.

Furthermore, before the humming audio is parsed, it might pass through an ANR module to decrease noise in live input. The ANR module's output would be a crisp humming sound. Here, Active noise reduction (ANR) is a technique for minimising unwanted sound by adding a second sound that is especially designed to cancel out the first.

Before extracting the features from the MIDI file, data pre-processing could be employed to remove the undesirable parts of the original audio and the humming inquiry. As a result, the new system architecture would include an ANR and data pre-processing module to ensure that no extraneous sound is used to assess similarity, hence boosting the accuracy of the similarity search and retrieval.

Furthermore, the proposed system can be improved by devising an efficient technique, such as optimising database indexing. Indexes are a terrific way to easily organise and locate data. It improves the speed of data retrieval, making it more efficient. Indexes reduce the amount of data that must be scanned in order to discover the relevant song record.

As a result, indexing for the database can be used to construct a better QBH system in order to get the music more efficiently.

## VII. Acknowledgements

The authors would like to express their heartfelt appreciation to the course faculty Prof. Sapan Mankad for his continuous support and guidance. Authors also want to thank the computer science and engineering department for providing all the resources necessary to carry out the term paper and for providing this course on information retrieval system.

## References

[1] Adams, Norman H., et al. ”Time Series Alignment for Music Information Retrieval.” ISMIR. 2004.

[2] Patel, Parth, ”Music Retrieval System Using Query-by-Humming” (2019). Master’s Projects. 895. Research Paper

[3] N. Kosugi et al. “Music retrieval by humming-using similarity retrieval over high dimensional feature vector space,” 1999 IEEE PACRIM. Conf. Proc. (Cat. No.99CH36368), Victoria, BC, Canada, 1999, pp. 404-407. doi: 10.1109/PACRIM.1999.79956

[4] Pauws, Steffen. ”CubyHum: a fully operational” query by humming” system.” ISMIR. 2002.

[5] Dannenberg, Roger B., et al. ”A comparative evaluation of search techniques for query-by-humming using the MUSART testbed.” Journal of the American Society for Information Science and Technology 58.5 (2007): 687-701.

[6] M. Antonelli, A. Rizzi and G. del Vescovo, “A Query by Humming System for Music Information Retrieval,” 2010 10th ICISDA, Cairo, 2010, pp. 586-591. doi: 10.1109/ISDA.2010.5687200

[7] Ghias, Asif, et al. ”Query by humming: Musical information retrieval in an audio database.” Proceedings of the third ACM international conference on Multimedia. 1995.

[8] A. N. Silla Jr., A. L. Koerich and C. A. A. Kaestner, “A Machine learning approach to automatic music genre classification,” J. of the Brazilian Computer Society, vol.14, no.3, pp.7-18, 2008.

[9] R. A. Putri and D. P. Lestari, “Music information retrieval using Query-by-humming based on the dynamic time warping,” 2015 ICEEI, Denpasar, 2015, pp. 65-70. doi: 10.1109/ICEEI.2015.7352471

[10] Sun, J. Q., and Seok-Pil Lee. ”Query by singing/humming system based on deep learning.” Int. J. Appl. Eng. Res 12.13 (2017): 973-4562.

[11] Madan, Akansha, and Divya Gupta. ”Speech feature extraction and classification: A comparative review.” International Journal of computer applications 90.9 (2014).

[12] Radha, V., and C. Vimala. ”A review on speech recognition challenges and approaches.” doaj. org 2.1 (2012): 1-7.

[13] Sivaram, G.S.V.S., Hermansky, H.: ‘Multilayer perceptron with sparse hidden outputs for phoneme recognition’. 2011 IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP), Prague, 2011, pp. 5336–5339

[14] Cano, Pedro, et al. ”A review of algorithms for audio fingerprinting.” 2002 IEEE Workshop on Multimedia Signal Processing.. IEEE, 2002.

[15] N. Mostafa et al. “A machine learning based music retrieval and recommendation system,” Proc. 2016 Tenth Int. Conf. on Lang. Res. Eval. (LREC), pp. 1970-1977, May.2016.

[16] Zheng, Fang, Guoliang Zhang, and Zhanjiang Song. ”Comparison of different implementations of MFCC.” Journal of Computer science and Technology 16.6 (2001): 582-589.

[17] Zhu, Yunyue, and Dennis Shasha. ”Warping indexes with envelope transforms for query by humming.” Proceedings of the 2003 ACM SIGMOD international conference on Management of data. 2003.

[18] Adamska, Katarzyna, and Paweł Pełzyński. ”Melody recognition system.” 2012 Joint Conference New Trends In Audio Video And Signal Processing: Algorithms, Architectures, Arrangements And Applications (NTAV/SPA). IEEE, 2012.

[19] Shifrin, Jonah, et al. ”HMM-based musical query retrieval.” Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries. 2002.

[20] Baeza-Yates, Ricardo A., and Chris H. Perleberg. ”Fast and practical approximate string matching.” Annual Symposium on Combinatorial Pattern Matching. Springer, Berlin, Heidelberg, 1992.