Evaluating the Potential of Similarity Estimation Using Hyperloglog for Query by Humming Problem

Arthur Lubambo

A Thesis Submitted in Partial Fulfilment

of the requirements for the

Degree of

Master of Science in Data Analytics

Diagram

Description automatically generated with medium confidence

February 2023

Supervisor: Kislay Raj

# Abstract

To Do

# Introduction

## Query by Humming

The Query by Humming (QBH) problem is a significant challenge in the area of Music Information Retrieval (MIR), which consists of retrieving or identifying a piece of music based on a user's vocal singing or humming of a melody. This problem encapsulates the difficulty of transforming a non-verbal, subjective representation of a musical piece into a format suitable for comparison and retrieval within a computational framework.

QBH has an inherent complexity due to the variability of vocalizations reproduced by humans. Different from text-based queries, where keywords usually have good matching with the corpus of documents, the humming or vocals do not have by default a standardized notation or representation that is a good match with the songs in the database, to identify and retrieve the musical piece searched in the database. The natural variations of pitch, rhythm, tempo, and timbre reproduced by users make it particularly a hard problem.

This problem was originally studied by (Asif Ghias et al., 1995) who introduced a pioneering approach that utilised the UP, DOWN, and SAME notation  (for example, SUDUDDSUDUD) for representing the pitch movements of hummed melodies. This encoding was associated with an algorithm for string matching that tolerates mismatches to develop the core of an engine that retrieves a melody MIDI stored in the database that matches best with the humming.

The literature review will cover methods studied by other researchers in the area using a wide range of strategies such as dynamic time warping (DTW), Locality Sensitive Hashing (LSH), and others. There are different contexts and variants of this problem where each method might have a strength or weakness. A common bottleneck between the techniques is the computational resources involved in executing the information retrieval, in time and memory complexity.

With this context mentioned above this project aims to study the potential of Hyperlolog++ Data Structure usage in a method to calculate the similarity between query and songs for QBH problem.

## Cardinality Estimation

The cardinality is the measurement of the size of a set, in other words, it refers to the number of distinct elements in a set or a population. The computation of cardinality is relevant in multiple domains. Depending on the context it might be impractical to calculate the exact solution, due to the size of the dataset, or the nature of the distributed data across multiple nodes. Under this scenario, it was developed solutions that give up an exact result for controlled approximated estimation, using probabilistic data structures getting significant efficiency gains in memory and time complexity to execute the estimation. The literature review covers in more detail the different algorithms for cardinality estimation. It was chosen the Hyperloglog++ as an algorithm in this project due to its superior accuracy, efficiency, and versatility. Its advanced techniques and optimizations make it a preferred choice for many applications requiring fast and accurate estimation of set cardinalities, particularly in the context of large-scale datasets and memory-constrained environments.

## Research Objectives

This project assessed the potential of using Hyperloglog++ in the QBH problem. Associating each Query to a Hyperloglog instance (HLL) and each Song in the Database to another HLL, then computing the similarity by an overlap coefficient, using the estimation of the intersection size, using HLL operations.

### Develop a Hyperloglog-based Similarity Estimation Framework for Query by Humming

It consists of creating sets of encodings for queries and songs then using the jaccard overlapping index estimation based on Hyperloglog++ (HLL) data structure, developing a framework for HLL parameters tuning for query by humming problem.

### Investigate the relationship between memory usage vs performance in Hyperloglog++ based Similarity Estimation

it includes running experiments to discover the impact of memory reduction and performance, establishing a modelling for the relationship between the hyper-parameters and accuracy. Analysing the hyper-parameters impact on memory usage. Finetuning it will bring the key value for the proposed method and its benefits.

### Compare and contrast Hyperloglog++ based Similarity Estimation Framework with a baseline algorithms performance by metrics.

it includes experimentation and comparison about the weakness and strongness of each type of algorithm used in the experiments under multiple aspects beyond memory usage, and performance, but also analyses in what musical context they tend to work better or not.

## Hypotheses

**Hypothesis 1:** The memory usage of HLL-based algorithms is smaller than the exact set-based solution.

**Hypothesis 2:** The Mean Reciprocal Ranking (MRR) of HLL-based algorithms is equal to the MRR from the equivalent exact set-based solution.

## Scope and Limitations

The project focused on the QBH problem, so the algorithm developed is not appropriate for plagiarism detection. The dataset and techniques developed and studied are suitable for the Western music context (12 notes system), so it is not adequate to use it for microtonal music pitch less music, or other types of music. The proposed algorithm requires a discrete musical notation to fit in a set modelling. The project applied different variants of algorithms with these characteristics to analyse the HLL estimation version versus the exact Set version cardinality computation.

TO DO ADD VOCALS

## Report Overview

This project Chapter X….. TO DO

# Literature Review

## Query By Humming

Query by Humming (QBH) is an area of Music Information Retrieval that studies the content-based search in the song database. Humming serves as a natural and efficient method for searching a musical audio database by vocalizing the melody of a song. It was originally studied by (Asif Ghias et al., 1995), using the pitch movements “Up”,” Down”, and “Same”, to encode the query, and substring searching in a database of 183 songs was conducted using parameters to accept a controlled level of mismatch. The big challenge at that time was encoding the audio signal in a string (‘U’, ‘D’, ‘S’), due to computational power and algorithms available, taking up to several minutes to process 10 seconds of audio. This Research opens the area of QBH it was quite interesting how a simple method using a UDS encoding could produce good results.

Later, other approaches were introduced to this problem such as dynamic time warping (DTW) a technique used in Times Series, but it could be modelled specifically for the audio processing context. For example (Fu et al., 2007) used the pitch level over time to model it as a times series using the dynamic programming technique DTW to calculate the matching level. Even with the optimizations, it has a quadratic cost to calculate the match between two series where n is the length of the series.

(Tripathy et al., 2009) they followed a very similar approach to the original (Asif Ghias et al., 1995) it created a layer for Wav to MIDI from the query and then used an algorithm for string matching, but they used a different method using dynamic programming to calculate the edition distance between strings.

In general, all the methods try different approaches to deal with the mismatching between what the song is, with the humming sung by a human. It can be observed that it is not only a time alignment between the query and the song. It is also a pitch alignment problem as was highlighted by (Stasiak, 2012), who proposed a method inspired by auto-adaptive human behaviour for ignoring errors in sung melodies.

All the algorithms were focused on the accuracy of the method so far, but there is another fundamental aspect to enable the implementation of it in practical applications, scalability. It is a natural question, to understand how to apply it in a large database. (Guo et al., 2013) Introduced the application of Locality Sensitive Hashing (LSH), a technique that creates an index based on hashes operations considering the similarity between the content. They had applied this method developing other layers to tackle the problem of key transposing, it is when the music and the query are not in the same tone. So it was more scalable and improved the performance with better mean reciprocal rank (MRR).

(Alexios Kotsifakos et al., 2011) proposed a subsequence matching framework capable of dealing with gaps either in the query or the target song, performing better than the other Dynamic Programming methods, and maintaining the same time complexity.

Considering the diversity of queries, (Wang and Jyh-Shing Roger Jang, 2015) started to extract other type information from the queries that were not being used before, the lyrics. It is quite common besides humming the people to sing part of the song, or even whistle. So this study explored how the usage of lyrics in addition to the melody might impact the performance of QBH. They have combined speech recognition techniques to extract the lyrics of the humming combining it with the melody distance, reducing the error rates significantly. For obvious reasons, it would just impact music and queries with lyrics. The other challenge of this method is the language context, to make it generalizable it is necessary to have multiple languages trained, and lyrics structured for all the songs. So it implies extra effort in its implementation.

A common problem with the songs is the complex extraction of the melody in an automated way. The original studies were based on existing MIDI databases with the songs already in the symbolic format, annotated by humans. Producing it algorithmically through raw audio format is more complex, the song usually the song has multiple instruments, with harmony and a combination of more than one melody at the same time. With this challenge in mind, (Alfaro-Paredes, Alfaro-Carrasco and Ugarte, 2021) used a voice separation to improve the melody extraction from the songs, it demonstrated better results for the encoded melody for the song, consequently improving the matching algorithms.

Most of the studies were trying to use improvements in the string match algorithm to consequently have better results with the QBH, but (Velankar and Parag, 2018) changed the paradigm, introducing the matching using n-grams and inverted index.

They combined it with the Mean Normalized Frequency (MNF) Algorithm and developed a method called “Unified Algorithm for Melodic Music Similarity”. The big advantage of this method is that it tackles the information retrieval reducing the search space for the songs where the n-Grams matches. So, the pre-computed indexes help the reduction of the query time. The N-Gram contains each segment of intervals for the song, so even with imperfections on pieces of the query encoding is unlikely to compromise the whole result.

As a counterpoint to (Velankar and Parag, 2018) the work from (Ulfi and Mandala, 2022) highlighted that the “Unified Algorithm” has issues with performance and works slow for big datasets. They also implement the algorithm “Query by Humming System using Frequency-Temporal Attention Network” but apply enhancements to the Partial Matching of queries on it. In the end, it concludes the method has problems with scaling for big datasets. It might indicate a gap to be explored by applying adjustments in the algorithms for scalability efficiency.

In General It is possible to be seen this problem can be tackled in different layers: some researchers focused in the encoding and matching algorithm from the query and used symbolic MIDI database, while others managed to encapsulate as well the songs manipulation to extract the melodies from the raw signal format. It is a harder problem to cover both layers. It is possible to classify the types of musical encoding used to calculate the similarity in two classes ‘Discrete’ and ‘Continuous’. The Discrete encapsulate the original Up Down Same, or Rounded Pitch equivalents, where the signal is converted into sequence of characters. While the Continuous is the sequence of continuous numbers such as wave, frequencies, spectrogram, chromagram, raw pitch. There are also researchers that combined the traditional formats with textual data such as lyrics, and textual transcription of the queries to enrich the query song matching. In addition of these layers, an important aspect is the scalability of the methods with bigger databases, as it can be seen in this literature review it was covered by a few of the researchers but it still requires further investigation.

### Melody Extraction

The melody extraction or Transcription is the name of a task that transforms the musical audio signal in a temporal pitch notation with the sequence of notes. The Melody extraction on its own is already a sophisticated problem to be covered, when combined to query by humming it is an essential step for the QBH algorithms that deals with the raw format songs databases.

Traditionally, the methods used for these tasks are algorithmically based such as (de Cheveigné and Kawahara, 2002) used to fundamental frequency (F0) transcription. This approach might be good enough to get the transcription of a single instrument or singer recording, but when there are records with mixed sources making harmony, or multiple melodical lines, it gets more complicated to be covered with this approach.

The most recent application of Machine Learning techniques on it shifts the usage from fully algorithmic solutions to data-driven models. For instance, (Yu et al., 2021) used a method inspired by the human perception of frequency, time and intensity for audio, applying Convolutional Neural Networks to extract the melody. Or (Donahue, Thickstun and Liang, 2022) use generative models to create the synthetic dataset and combine it with annotated datasets to improve the melody transcription. Or (Bittner et al., 2022) who proposed a lightweight neural network for musical transcription, it is compatible with polyphonic outputs and could be applied to a wide range of instruments and vocals.  
  
We could see some of the methods, explicitly or implicitly used source separation for better results in the QBH, It is especially important as a pre-processing layer for the melody extraction phase. (Défossez, 2022) used a combination of Frequency Domain with Time Domain, introducing the method “Hybrid Spectrogram and Waveform Source Separation”, it is based on deep neural networks architecture.

The advantage under-explored from Source Separation for QBH is the ability to encode more than one simultaneous melody. It might improve the multi-vocal, counterpoint songs allowing the matching with the query to be on any of the melodic lines.

## Music Theory

Music theory is the academic discipline that explores the principles behind the composition, harmony, rhythm, and structure of music. It provides a framework for analysing and understanding the intricate relationships between musical elements, aiding both composers and performers in their artistic development. It is not a rigid set of rules to be followed instead, it is a dynamic framework that fosters creativity and facilitates deeper comprehension and communication of musical expression.

### Rhythm

According to (Schmidt-Jones, 2013) rhythm is a foundational element of music, emphasizing its role in temporal organization and providing structure to musical compositions. Alongside melody, harmony, timbre, and texture, rhythm forms the cornerstone of musical expression, with its placement of sounds over time being essential for music's unfolding and coherence. While melody and harmony often dominate discussions in music theory, (Schmidt-Jones, 2013) suggests that rhythm is equally indispensable, if not more so, given its intrinsic relationship with time. The hierarchical organization of rhythm, including concepts such as beat, meter, duration, and time signature, elucidates how rhythm is structured and perceived, offering frameworks for analysis and composition.

### Timbre

(Schmidt-Jones, 2013) elucidates the concept of timbre, also known as colour in music, as a fundamental element distinct from pitch, dynamics, and duration. Timbre encompasses the unique qualities of a musical sound that differentiate it from others, even when sharing identical pitch, duration, and volume. This distinction come from the complex waveforms produced by musical instruments, containing multiple frequencies that contribute to the perceived colour of the sound. The brain is capable to identify the pitch but it is also able to differentiate the mixture of other frequencies in the signal. The balance and interaction of these frequencies, particularly evident in the initial attack of a note, determine the distinctive timbre of each instrument or voice. Moreover, the text emphasizes the discerning ability of the human ear and brain to perceive subtle variations in timbre, enabling differentiation not only between instrument types but also between specific instruments or performers.

### Melody

(Schmidt-Jones, 2013) highlight that the melody consist of a sequence of notes with distinct pitch and duration that collectively form a cohesive musical line. Unlike mere successions of notes, a melody consists of those notes that prominently engage the listener's attention, delineating the primary musical theme. Additionally, the author introduces terminology pertinent to discussions of melody, such as the melodic line representing the core sequence of notes and ornaments or embellishments, which enrich the melodic texture without altering its essential structure. the book highlights the nuanced aspects of melody, and its role as a central and captivating element within musical compositions. In the context of QBH in this project it is possible to see challenges around different ways of singing with less or more ornaments. It will be covered in next chapters more details about it.

### Melodic Motion

The concept of melodic motion is explained in (Schmidt-Jones, 2013), delineating two primary forms: conjunct and disjunct. Conjunct motion characterizes a melody that ascends and descends gradually, with small pitch changes between successive notes, akin to step-wise or scalar motion. In contrast, disjunct motion describes a melody marked by rapid rises and falls, featuring large intervals between consecutive notes, often referred to as "leaps." The text also acknowledges that many melodies exhibit a blend of conjunct and disjunct motion, highlighting the dynamic interplay between these contrasting forms within musical compositions. The melody motion is the core concept used in the algorithms experimented in this project, due the fact that same melody in different pitch has the same relative motion. The humans can recognize as same melody two melodic lines with same motions but starting in different notes. It is an important characteristic in the melodical encoding explored in the next chapters.

### Harmony

In music harmony happens when multiple notes are executed simultaneously. It is important highlight that harmony is not necessarily "harmonious", depending on the intention of the composer it might be dissonant. The definition itself only refers to simultaneous notes. In this project it will not be explored to much once the focus is the melody and its movements.

### Counterpoint

The counterpoint is when there are more than one melody played at the same time, making a particular type of harmony. The counterpoint might be executed by same or different instruments or voices. It made a particular strong impact in this project, it will be discussed in future chapters the challenges around songs with counterpoint. How to distinguish and compute what are each independent melody that is played simultaneously.

### Texture

(Schmidt-Jones, 2013) explains musical texture referring to the density and complexity of elements within a piece of music at any given moment. Texture can be characterized as thick or thin, reflecting the presence of many or few layers of musical material. Various configurations contribute to texture, such as rhythm alone, a melody with chordal accompaniment, or multiple interlaced melodies. Understanding these terms can helps the musicians or audience appreciate deeper the musical structure. For this project this concept is useful to understand and give vocabulary to analyse the types of songs and how they perform with each type of algorithm and why.

### Types of Textures

**Monophonic:** is characterized by the presence of a single melodic line without of any accompanying harmony or counterpoint. While rhythmic accompaniment may exist, it does not detract from the singular focus of the melodic line, which consists of distinct pitches. This simplicity of texture allows for a clear and unobstructed presentation of the melodic material, making monophonic music a historically significant and aesthetically distinct form of musical expression.

**Homophonic:** is characterized by a single clearly melodic line that naturally captures the listener's attention, while other parts provide accompaniment or harmonies. Informally, references to chords, accompaniment, or harmony are often associated with homophonic compositions. These accompanying parts, while possessing their own melodic qualities, are distinguishable from the main melody by either sharing the same rhythm or serving primarily to fill in the chords or harmonies.

**Polyphonic:** is the type of texture characterized by multiple melodic lines at the same time. It is also called polyphony, counterpoint, or contrapuntal music. It can be executed by multiple instruments or vocals mixing independent melodies, that together creates a musical experience.

**Heterophonic: “**A heterophonic texture is rare in Western music. In heterophony, there is only one melody, but different variations of it are being sung or played at the same time.”, (Schmidt-Jones, 2013), It will not be covered in this project as the introduction already highlighted, the scope of this project is the Western Music.

In the book (Schmidt-Jones, 2013) there are references for examples of song from each texture type. Note that it is common the songs has multiple types of textures in the same song in different parts, so the examples are just guidance.

## Music Source Separation

TODO Demucs <https://github.com/adefossez/demucs?tab=readme-ov-file>

## MIDI

TO DO

### MIDI automated transcription

TO DO Basic

## Cardinality of Sets

TO DO

### Probabilistic Algorithms and Data Structures

In QBH a big challenge is to enable more accurate results, with algorithms that scale well for big datasets. For instance, Spotify Dataset has more than 100 Million Tracks (Spotify, 2023). It is necessary to have more efficient methods for QBH. This Project will explore the potential of Probabilistic Algorithms and Data Structures in the QBH Task.

One of the most impactful data structures created was the HyperLogLog. (Flajolet et al., 2007) did enhancements in the original loglog algorithm for the cardinality estimator, making it possible to estimate the size of sets beyond 10^9 with a standard error of 2% while using a memory of only 1.5 kilobytes. This data structure creates multiple registers of hashes and computes operations of add, count and merge (two HLLs). The extremely low memory has a cost with the accuracy, but it can be tunned by a parameter that increases the number of registers. The application of this Data Structure in Big Data was disruptive, making it possible the count distinct elements estimation in a distributed system extremely quickly. (Heule, Nunkesser and Hall, 2013) made improvements to the method increasing the accuracy and reducing even more memory usage.

As mentioned previously, there were some methods that used locality-sensitive hashing LSM (Guo et al., 2013) (Matti Ryynänen and Anssi Klapuri, 2008) and MinHash to index the songs for QBH. Although there are improvements, in the MinHash algorithm called HyperMinHash (Yun William Yu and Weber, 2020), there is a good opportunity to investigate if this new implementation would impact the performance of the QBH.

Also, one of the applications of HLL could be inclusion coefficient estimation as (Nazi et al., 2018) demonstrate in an efficient method applied in database columns schema analysis. This coefficient could be defined as the fraction of the intersection of two sets relative to the smaller set. This concept can be used as an inspiration for another context. For example in the QBH, the matching between the query and the song can be modelled as the inclusion coefficient index between the melodic encodings of the query and song, using it as a set of keys. So the challenge is to transform into how to encode it with relevant information for matching it. In addition to it, the HyperLogLog methods could be used to optimize the execution, creating a scalable implementation for it.

## Similarity Measures

TO DO

## Mean Reciprocal Ranking

Reciprocal ranking (RR) is a method used in information retrieval and machine learning to evaluate the effectiveness of ranking algorithms in presenting relevant results to users. It assesses the quality of rankings by considering the position of relevant items within the ranked list. In reciprocal ranking, the relevance of items is typically represented as binary values (relevant or non-relevant), and the reciprocal of the rank of the first relevant item is computed. This reciprocal value provides a measure of the effectiveness of the ranking algorithm, with higher reciprocal ranks indicating more relevant items appearing higher in the ranked list. Reciprocal ranking is commonly used in evaluating search engines, recommendation systems, and other applications where the goal is to present the most relevant content to users. The reciprocal Ranking is simply the reverse of the ranking index, considering the ranking starting in 1 instead of 0. For instance if the first relevant document for a query is in the ranking 1 it would be equivalent of 1 in RR. In Case it was at ranking 2 it would reflect into 0.5 (1/2) as RR.

Note on this project there is only 1 relevant document in the database per query, as there is no multiple versions of same song in the database.

When aggregated over multiple queries, this metric is referred to as the Mean Reciprocal Rank (MRR). [ADD equation for MRR]

[Add reference https://link.springer.com/referenceworkentry/10.1007/978-0-387-39940-9\_488]

# Methodology

## Proposed primary research methodology

The primary research methodology employed in this study centred around experimental investigations. The goal as mentioned previously is to investigate if the usage of Hyperloglog++ in the similarity estimation using the inclusion-exclusion principle will bring enhancements in the algorithm's execution and will not affect negatively the query-by-humming performance in a significant way.

In many other areas applications this type of data structure is demonstrating significant improvements in cardinality estimations, bring leaps in the algorithm efficiency. The Literature Review demonstrated a lack of studies with the HyperLogLog++ applied to Query By Humming, so this research aims to explore this approach and evaluate the results. The methodological framework was structured across 3 phases: data preparation, experimentation, and evaluation. In the [RMD] it is possible to see the breakdown of what is involved on each stage.

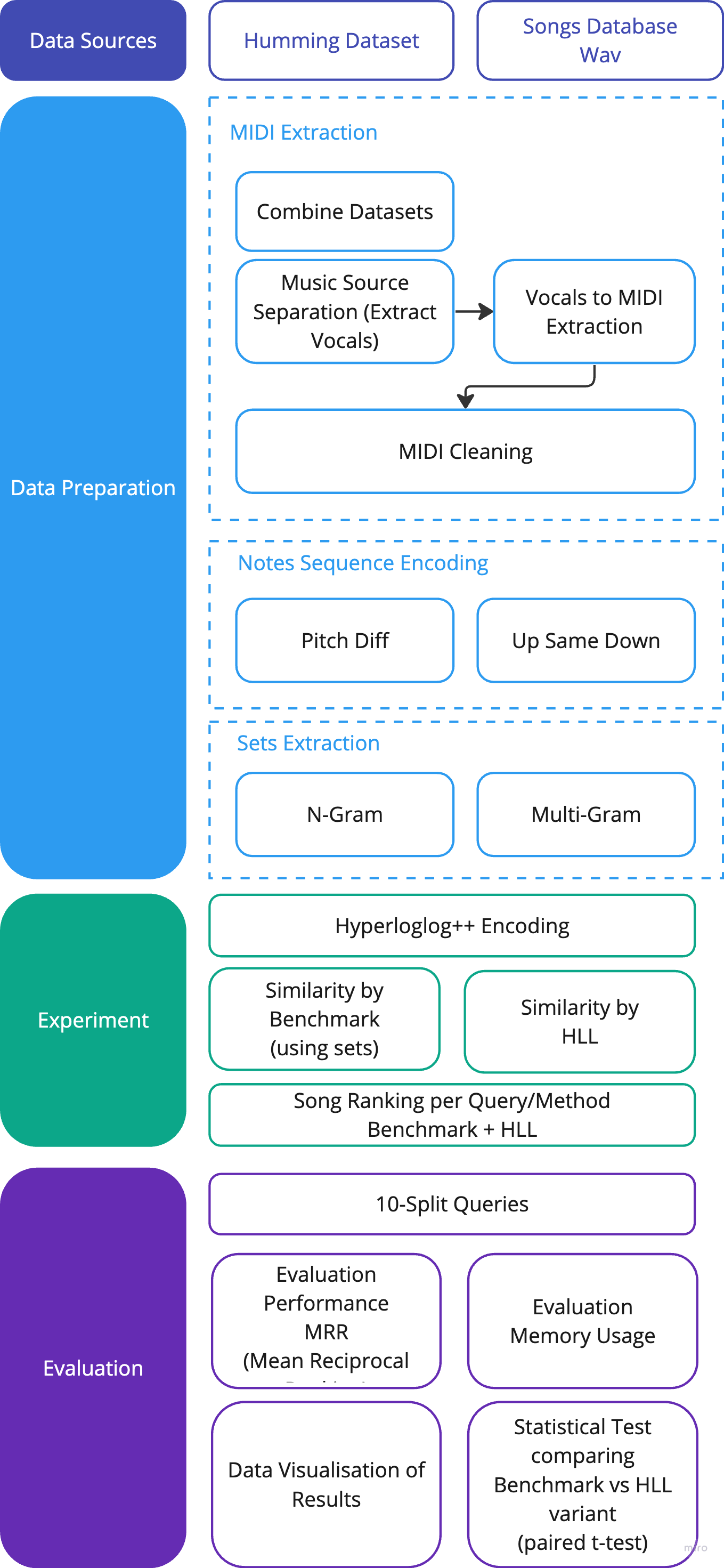


Fig .[RMD] Diagram with Research methodology

As Data Source It was used a suitable secondary datasets for the primary research of this capstone project. I was chosen reliable set data source [ref], with representativeness, and good references in other studies. It will be discussed in the next chapter more details about the data sources involved in this project.

The Data preparation phase, involved 3 main layers, MIDI extraction, Notes Sequence Encoding, then Sets Extraction. The data source was provided in WAV format, it is necessary process the audio signal to construct a symbolic representation of the melodies, for either queries and songs. In the next chapters covers details about each step on the MIDI extraction. The Notes Sequence Encoding is a Layer to convert the Notes into an format that is robust to variations in pitch, time, and duration between query and songs. Then the Sets Extraction transform the Sequences of Discrete Encoding into Sets where the similarity operations will be executed.

In the experimental phase Hyperloglog++ (HLL) data structures were created for each set variant query and song. Then then similarity is calculated for the baseline overlap coefficient using sets operations, followed by HLL based estimation using cardinality estimations and the inclusion exclusion principle, more details will be discussed about it in next sections. After calculating the similarity between each pair query-song, the ranking of the desired song is for each method and set combination.

In the evaluation phase, the queries were shuffled and split in partitions, with MRR (Mean Reciprocal Ranking) and memory calculated for each combination of method partition. Subsequently a comparison was made applying statistical test and data visualizations to assess the hypotheses mentioned previously in the introduction.

## Data Collection

The dataset for QBH in general might be a challenging to be produce. It is necessary to be done with a carefully designed procedure to avoid insertion of bias. It has been found (Salamon, Serrà and Gómez., 2012) a database with 118 recordings of sung melodies, used in multiples other studies. It was built from 17 subjects, keeping a good gender balance and a wide level of music knowledge level, from zero to amateur musicians. They were presented with a list of songs, and they were free to pick the ones they knew and were asked to sing/humming any part of the melody for recording. There was no restriction on time, or what part of the melody must be sung, they were free to sing with or without lyrics. They didn’t listen to the original song before recording it, what it quite important to avoid pitch bias, they could sing in any tone, so they just reproduce it by memory. All the records were made from a simple microphone from a laptop in order to simulate a realistic scenario for QBH. This dataset was chosen once it is robust and meets the good quality criteria and procedures on all the experimentation designs and it is widely used in other researches. So the sampling method and type to pick the dataset was respectively Non-probability sampling and judgment sampling.

It is important to highlight the QBH for copyrights reason, does not contains the songs database, however it is easy to build once it contains very popular songs. On the other hand it might make the results less directly comparable, because each version of the same song might have nuances that makes the whole QBH tasks easier or harder. For instance, if a researcher uses an annotated MIDI database from songs, with very clean melodies as a starting point, it is much simpler, and likely to have matches than a dataset the melody extractions made by machine. The QBH task in the second case accumulates the challenges and imprecisions of each layer of the data processing.

The duration queries in the dataset used has recordings with a duration of 26.8 seconds in average, varying from 11 to 98 seconds. Due cost efficiency for this experiment, it was used the canonical songs database 481 songs, instead of the full collection with 2125 songs that contains different versions of the same song. Note it might impact negatively the performance, once we are relying in a single version of the song in the database. For instance, the version might have particularities noise or extra challenges that makes harder to extract the high-level features for the matching mechanism.

## Data Preparation

This layer contains the steps necessary to transform raw signal data from sound WAV format to Sets.  
In this project, all audios from query and songs were provided in raw format.

### Midi Extraction

#### Music Source Separation

From the raw Wav songs and queries files it was made different processing flows. The songs contains in the audio file the whole track, with mixed signal of multiple instruments (including harmonical, melodical, and percussion), voice(s), crowd. Extract relevant high-level features from this signal is not an easy task. In western popular music, in general the main melody is produced by vocals. So It is natural try to separate the mixed track into channels with different sources.

For that it was used the current state-of-art model Demucs based in Hybrid Transformers for Music Source Separation [https://arxiv.org/abs/2211.08553] as mentioned in Literature Review, it can separate the sources between: ‘vocals’, ‘drums’, ‘bass’, ’other’. The vocals includes any voice, including multi-vocal songs, the drums, include any percussive sound, bass contains the low frequency bass instruments, and other has any other background accompaniment. As the queries has only vocals, it was necessary to execute this layer in the queries recordings, but it was quite useful to run it in the songs database. A the focus of this project was to analyse the Hyperloglog++ potential on the QBH, the scope of this project was delimited to vocal songs. For future works it would be interesting to extract in addition melodies from the non-vocal parts of the song.

#### Vocals to MIDI Transcription

The Vocals to MIDI operation was executed for both types of recordings the query and the vocals of songs extracted in previous by the Demucs model. Then it was applied the BASIC [https://arxiv.org/abs/2203.09893] mentioned in the literature review to extract the MIDI transcription automatically using a deep learning model. It was chosen due the good quality results, lightweight to run, open source. See in [ADD REFERENCE for MIDI sample TO DO] links for sample of MIDI created from BASIC. The resultant MIDI file, contains a list of instruments, where each instrument has a list of notes. Each note has a start time, end time, and pitch. The pitch follows a numeration form MIDI note number see table in the Appendices. Note the BASIC is capable to extract polyphonic music, in case there is multi-voice songs the midi generated has notes that is played simultaneously.

### Notes Sequences Encoding

From the low level signal with the two layers below it was extracted the MIDI that is a symbolic high-level representation of the song. Although it is not enough to match directly the part of the song with the query with both MIDIs, because the subjects might sing or humming the song from a different start, beat or even in a different. In addition to it subject and singers might different interpretations of the song adding extra notes, ornaments. Given all these circumstances, based on the concept explored in the literature review melodic motion is an important feature to be extracted, it makes the encoding invariant between recordings in different keys, also transform the notes in sequences helps the notion of time and duration being relativized. What matters in this approach is the order and the melodic movement, instead the precise time/duration of each note.

#### Pitch Diff

For capturing the melodic movements, the sequence of notes was processed to get the difference between the pitch of the current and the previous note.

TO DO: ADD FIGURE ILLUSTRATING IT

So a sequence of N notes will be converted in a sequence of N-1 pitch differences.

#### Up Same Down

The pitch difference the exact melodic motion in a scale of half-steps precision. It might bring problems if the subjects recording has inaccurate relative pitches in the humming. One approach the original research (Asif Ghias et al., 1995) used the Up, Same, Down representation of the sequence.  
It would be equivalent to transform the pitch diff sequence into a transformation: ‘Up’ if diff greater than zero, ‘Same’ if diff is equal to zero, and if diff less than zero ‘Down’. I can be observed that it is equivalent of reducing the resolution of the information in the pitch encoding, capturing the same melodic movement but transforming granular integer information into a just three classes encoding.

### Sets Extraction

After having all the sequences from songs and queries based in the two methods (UpSameDown and PitchDiff), it is necessary to match them. The sequence themselves are not directly comparable. There are traditional methods such as dynamic time warping, that could be used for this type of problem, but it usually requires quadratic complexity cost to compare two strings. To overcome it, this framework proposed a set extraction to compute parts of the melodic movements, it removes completely the time dependence of each part of the song. Each element of the set corresponds to an encoding of the melodic movement. Using this approach each query and song will be associated to a set of melodic movements.

TO DO Add Figure to Illustrate.

#### N-Gram

A natural way to extract the elements for the set creation is define a size for the melodic movements parts. In other words, a size of a substrings to be extracted from the original sentence. Inheriting the concepts from Natural Language Processing, it is equivalent to produce N-Grams from a sequence of tokens. Where each token is a PitchDiff or a UpSameDown from the Sequence Encoded in the previous step. Each N-gram Produced is a elements representing a part of the melody with its movements. In general it is similar to the Bag of words approach, were each element is equivalent to a term in the search mechanism.  
This method introduces a hyper parameter: N to be defined. How to fine tune it properly? If N is too small, the corpus of different terms would be very small, making it uncapable to retrieve the elements properly. For instance: N = 2 in the UpSameDown Encoding would create only 3^2 = 9 elements in the whole corpus ‘vocabulary’, making it harder to separate what right song in the database matching the query. On the other hand if it is too big, it makes less robust for subject mistakes, or ornaments, once it needed to get a perfect melodic encoding for longer, in addition the query set would have a limited set. For example: a query sequence with size k creates (k+1-N) elements based on N-Grams. There is a clear limit of N, it needs to be smaller or equal to k. In this extreme scenario, the query would have only 1 element in the set. This elements probably will not be enough to discriminate the songs that matches with it.

Analysing the query size, it was found that mean size of sequence is 33, minimum is 9, maximum 106. Considering this case it was tested N=5, 8, 10, 12, even with N=10;12 that it is above the minimum k, it was tested because there is a chance of having a positive impact in the performance of the other bigger queries compromising the few ones that are too small.

#### Multi-Gram

An interesting strategy to overcome the finetuning of the N hyperparameter is to create a set that is a union of multiple N-Grams with different N values. This approach was named as Multi-Gram. In this case the hyperparameter is a pair (k,j) where j>=k, representing the union the N-Gram from N=k until N=j.

To Do Add Equation for Union

The Multi-Gram proposal helps increasing the cardinality of query set, towards a better discrimination power with the songs in the database.

## Experiment

The experiment consists calculate the similarity between queries and songs using the baseline method and the alternatives methods using HLL encoding, then comparing the results of the ranking. Each section will cover more details of each layer.

### Similarity

In the previous stages it was created an equivalent set for each song and query with melodical movements information encoded. In the similarity part, it was created a data frame with all possible pairs between queries and songs, from that the similarity measurement is calculated based on the similarity of the respective queries and songs sets involved. The rationale behind it is if the melodic information between sets are similar the query and song has a good match, if they are not similar, probably it is a non-match.

There are several methods to calculate similarity between sets such as Sørensen–Dice coefficient; Jaccard index, overlapping coefficient and others. In the proposed framework it was chosen the overlapping coefficient, also called Inclusion coefficient (\*considering the query as a target set), because the other metrics has problems with the difference in size. For instance, the Jaccard index, because the queries sets are much smaller than the song set. [ADD AVERAGE SIZE COMPARISSON]

The Jaccard index would penalize bigger songs even if the whole query set is included in the song set. For this reason, it was chosen the overlapping coefficient, that would only consider the size of the query (\*the smaller set) in the denominator. Note the Sørensen–Dice coefficient has a similar problem.

### Similarity by overlapping coefficient

The overlapping coefficient, calculate the cardinality of the intersection between the query set and song set divided by the cardinality of the smaller set between query and song set (\*in practice the query is the smaller always). Analysing the Algorithmic complexity to make this calculation involves compute the intersection between sets and count the cardinality of the intersection, and the query set. For intersection calculation between A and B is the O(min(size(a),size(b))), while computing the size of a set is O(1). So the computation of overlapping coefficient between two given sets, requires O(size(query Set)) to be computed.

### Hyperloglog++ Encoding

As mentioned in the literature review the Hyperloglog++ is a data structure capable to estimate the count of distinct elements in list. It is equivalent to compute the estimation of a cardinality for the set. The idea behind this data structure is compromise the exact solution using a controlled error estimation but using less memory in the computation. As mentioned before, it has a hyper parameter: **p** it defines how registers will be used in the estimation (2^p). In this project it was tried multiple combinations of **p={4,8,12,16}** , as higher is p more accurate is the cardinality estimation, but it uses more memory.

### Similarity Estimation using Hyperloglog++

The HLL++ has the following functionalities: add an element to the sketch, merge two HLL, and estimate the cardinality. Note it is not possible to compute a HLL for the intersection given two HLL.

But in this project, it will be used principle of inclusion-exclusion equation to compute the intersection cardinality.[Eq Ref]

[EQ |A T B| = |A| + |B| − |A S B|.]

In this case the estimation of the cardinality of A intersection B the sum of the estimation of cardinality of A and B, subtracted by estimation of the union A,B (merged HLL).

[Add Eq for intersection HLL]

So the ratio of cardinality of intersection and min(card(A), card(B)) provides the estimation for overlapping coefficient.

[Add Eq for overlapping coeafficient HLL]

For estimating the cardinality of a given HLL is O(1), then to compute the merge between two HLL is O(R) where R is the number of registers. Consequently, the algorithm to compute the estimation of overlapping coefficient is O(R) in time.

In terms of memory, the HLL++ uses O(R) but there are some improvements for sparse sets. [Add reference]

### Song Ranking

After have the similarity calculated for all combination of pairs Query Songs for all methods including the combination between sequences types (PitchDiff, UpSameDown), the Sets Extraction Types (N-grams, Multigram), and similarity estimation method (Baseline, HLLs). For each query , method the respective ranking is computed, to evaluated.

## Evaluation Plan

The evaluation of the methods are based in the Mean Reciprocal Ranking (MRR), and Memory Usage. The queries were divided into shuffled 10 splits, so a statistical test were applied to compare the MRR between the samples. Additionally the memory size if the sets and HLL were measured and compared. All comparisons were paired between the baseline set based operation versus multiple variants of the HLL hyper parameters.

### Statistical Test

The statistical test applied was Paired Samples T-Test [Add Reference <https://link.springer.com/chapter/10.1007/978-94-6351-086-8_4>] , because for each query split we have equivalent pairs between baseline and a target hyperlolog++ variant. It can be assessed the performance of QBH of each split of query under the same circumstances using different methods.

## Tools and Technologies

It was used in the experiments a simple laptop MacBook Air, Apple M2, 8 GB of RAM. It was not required the usage of bigger cluster for heavier computational power to execute experiments. As it can be seen in the attachments in the git repository, it was used Jupyter Notebooks with the experiments scripts. In addition, There were open source libraries used in the experiments such as:

**Librosa**, for audio processing music feature extraction and manipulation; **Pandas**, for general data analytics tasks; **pretty midi** to manipulate midi data; **nltk** to compute NGrams; **seaborn** and **matplotlib** for data visualization; **Hyperloglog** it a python library with the Hyperloglog++ implementation; **HTDemucs** is a machine learning model for music source separation; **BASIC** a model for automated music transcription, it was used to create the MIDI from the wav.

For production usage of these algorithms, it would be necessary for bigger databases a distributed computing architecture such as Hadoop, Spark, or equivalent. The nature of the algorithms applied on the experiments are suitable for distributed computing, once it is implemented using the tools mentioned previously.

## Ethical and Risk Considerations

The ethical considerations for this project are based on Ethics and data protection (HAYES and KUYUMDZHIEVA, 2021). It is clear that the data involved in this research is NOT sensitive, as does NOT deal with data concerning children, or vulnerable people. The data used has a Creative Commons Attribution 4.0 International license. It is NOT invasive neither put at risk any kind of freedom of the participants. The source organization that collected the data, had the consent of the people to use the data for this type of research.

There is no sensitive Personal Identifiable Information (PII) such as Sex Orientation, Race, Religion, or similar. The data involved is already anonymised. Given the nature of the data involved, it is not necessary to implement of the DPIA process. The Data was collected in Europe and does not have any PII, so There is no Transfer of personal data to non-European countries at all.

It is important to highlight a possible risk of miss-usage of the algorithm developed in this research, such as plagiarism detection. It would be an unfair and unethical usage if it were implemented as an automated decision-making. Any decision made by this algorithm must be reviewed by a human. For example: In case someone uses the proposed similarity algorithm in an automated system for automated plagiarism detection, the responsible for operating it must ensure to have a human review for each positive detection, to consider this application ethical. The risk will mitigated by informing in the public repository of good practices of usage of this method.

Any software, or library used in this research will be following a proper license such as MIT, Apache License, BSD license or equivalent. In the case of proprietary software or cloud platform usage, the proper permission will be in place. Regarding the Sampling Method, The intrinsic bias involved will be managed by explicitly representativeness and relevance of data selection with clear criteria and justification. For each stage of the research will be reviewed the proper usage and reduction of bias using a proper technique from the data preparation, until data visualization and Results presentation.

After taking all these measures and in case any other ethical concern was not properly detected, any external person can report it through GitHub issue, and adequate measures will be put in place.

The context this project will be based on is Western music theory (12 pitches), so for example it would NOT be adequate to use it different types of music such ad microtonal music, pitch less or equivalent.

Applying the method developed on this project in real projects must be considered carefully once depending on the context it might exclude songs from other cultures to have accurate results. Due to the limitations of datasets and time constraints, it will be tackled with clear advice about the usage.

# Evaluation and Analysis

For the evaluation of the proposed Hyperloglog potential in the query by humming, the experiments were executed as mentioned in the Methodology section. It was used the MTG-QBH dataset and the songs database was gathered. Originally the database contained 481 songs and 118 queries. But after analyse the dataset, it was excluded from the collection the songs with non-vocal songs to respect the scope of the project. The resultant dataset contained 473 songs and 112 queries. After applying the split mentioned before, it was produced 10 splits with 2 of them with 12 queries, and 8 with 11 queries. Note that the song database was not split, so the whole 473 songs are searchable for all splits.

A diagram of a method

Description automatically generated

Fig. X Diagram Combination of Attributes for Methods

## Performance

Applying all the 50 methods to the 10 splits queries, result in 500 results. Analysing the MRR breakdown per Method and Similarity Type, applying the average, it is possible to the results in [TABLE RESULTS]. It highlights that even the baseline method just had acceptable results for the encoding for **pitchDiff-NGram-5** and **pitchDiff-MultiGram-5-10,** with respectively 0.142 and 0.136**.** For all the other types in baseline, the results got poor results (MRR < 0.04). Mainly the UpSameDown Sequence encoding, it is probably due to the excessive low granularity of information, for instance the UpSameDown-NGram-5 can allow in the maximum different ‘terms’ in the corpus, making it harder to discriminate the songs. It can be observed it has a slight improvement for higher number of N-Grams and Multi-Gram, but it still low. As the N number in N-Gram increase, the number of elements for each query is reduced as well. So in general the **UpSameDown** did not perform well.

A screenshot of a table

Description automatically generated

Table. [TABLE RESULTS] Results By SIMILARITY TYPE/SET TYPE [ADD PROPER TABLE]

A graph with different colored bars

Description automatically generated

Fig. X Mean of 10 Splits MRR per Method Type

For **PitchDiff** method the lower number of n-gram could make it work properly, because each item in the sequency has a wide range of possibilities as it can be seen in the histogram distribution [Hist PitchDiff Query & Song] these queries have the majority of the 94.5% of the pitches diff between -5 and 5, with -0.008 as average and 3.027 as standard deviation. So it provides bigger range of possibilities compared to the UpSameDown approach. But on the other hand, In case is has longer elements (bigger N parameter in N-Gram), it requires a good level of pitch precision for the subject’s humming or singing. Another interesting observation is the number of half-steps in the songs pitch diff, much wider, with a standard deviation of 7.649753 compared to the queries ones. Given the histogram shape, it probably indicates the presence of noise the melody transcription from the method. It indicates a point of investigation to be tackled for future works.

A graph of a column

Description automatically generated A graph of a line graph

Description automatically generated

Fig x [Hist PitchDiff Query & Song]

After applying the paired t-Test comparing each method using the Baseline Similarity versus HLL Similarity measurement.

## Memory

To DO

## Analysis

To do

# Conclusion and Discussion

To do

## Key Findings

To do

## Recommendations for future research

To do

# Appendix

## MIDI Pitch Table

TO DO

## MIDI sample from Basic Sample

To DO

# References

To do