EVALUATION OF VOCAL SOURCE SEPARATION MODELS FOR AUTOMATIC GENERATION OF QUERY BY HUMMING DATABASE

Vivek Vijayan, Domonkos Kostyál, Tanguy Lissenko
Music Technology Group
{vivek.vijayan,domonkos.kostyal,tanguy.lissenko}01@estudiant.upf.edu

ABSTRACT

In this study, we evaluate the effectiveness of vocal source separation models, *Spleeter* and *Open-Unmix*, in constructing a query-by-humming (QBH) database. We conducted an ablation study to assess their performance under varying conditions, comparing different input feature representations, specifically pitch and chroma features, to analyze their impact on the accuracy of similarity measures such as Dynamic Time Warping (DTW). Each combination of features and model is systematically evaluated using retrieval metrics, including mean reciprocal rank (MRR), and precision @ k ($p@k$). Our findings provide insight into the suitability of these separation models for Query by Humming (QBH) and highlight the influence of feature selection on retrieval performance.
URL to repository: https://github.com/enter-opy/qbh-catalog

1. INTRODUCTION

Query By Humming systems enable users to retrieve music by humming or singing a melody, offering an intuitive and accessible interface for music discovery. Despite advancements in music information retrieval, the accuracy of QBH systems heavily depends on the quality of the underlying database and the robustness of feature representations. A critical challenge in constructing such databases lies in isolating the vocal component from polyphonic audio recordings, as instrumental accompaniments can obscure melodic contours and degrade retrieval performance [1]. For this we need to assume the query is always going to be the voice melody and not the instrumental or rhythmical components of a song. Thus we only gether the data of songs which has singing voice and which is monophonic.

Our work makes two key contributions:

- We compare the performance of *Spleeter* and *Open-Unmix* against raw audio (no separation) to quantify the benefits of vocal separation for QBH.
- We assess the different feature representations (pitch and chroma) and separation methods, highlighting their combined effect on retrieval metrics.

By using standard retrieval metrics, including Mean Reciprocal Rank (MRR) and Precision@k ($p@k$), we offer a comprehensive evaluation framework for future research. Our findings aim to bridge the gap between source separation technologies and QBH applications, advancing the development of accurate and scalable humming-based retrieval systems.

2. RELATED WORKS

One of the foundational works in QBH is the study by Ghias [2], which proposed one of the earliest systems for querying an audio database by humming. Their system employed a contour-based approach, representing melodic sequences using three symbols: up, down, and same pitch. Using an approximate pattern matching algorithm, they were able to retrieve songs from a database of MIDI files based on relative pitch transitions. However, this early approach was limited by the need for a symbolic representation of music, making it less suitable for real-world audio recordings.

More recent research has sought to move beyond symbolic representations by extracting melodic information directly from audio recordings. Rocamora et al. [3] proposed an approach that automatically constructs a QBH database from music recordings rather than relying on manually transcribed symbolic data. Their method focused on extracting singing voice melodies from polyphonic music, utilizing a combination of note sequence matching and pitch time series alignment. This approach significantly improved scalability by eliminating the need for manually transcribed MIDI files, achieving an 85 % accuracy rate within the top-10 search results. Despite these improvements, challenges such as extracting accurate melodies from complex polyphonic textures remain.

Another significant contribution to the field comes from Salamon et al. [1], who explored the use of different tonal representations for music retrieval, including melody, bass line, and harmonic progression. Their study demonstrated that a melody-based approach could be successfully adapted for QBH by leveraging robust melody extraction algorithms. They compared different tonal descriptors and applied a dynamic programming algorithm for

Attribution: V. Vijayan, D. Kostyál, and T. Lissenko, "Evaluation of Vocal Source Separation Models for Automatic Generation of Query by Humming Database", submitted to *ISMIR*, 2025.

query matching, showing that combining different musical representations could enhance retrieval accuracy. However, their work also highlighted the limitations of current melody extraction algorithms, particularly in handling noisy or complex polyphonic backgrounds.

The evolution of QBH research demonstrates a clear trend toward fully automated audio-to-audio retrieval systems. While early systems relied on symbolic data, recent advances have shifted towards leveraging machine learning and signal processing techniques to extract melodies directly from recordings. The main challenges that persist include improving melody extraction accuracy, handling variations in user singing styles, and developing more robust matching algorithms to accommodate imperfect queries.

3. METHODOLOGY

3.1 Dataset

We used the MTG-QBH (Query-by-Humming) dataset [1], which consists of 118 sung melody recordings. These recordings were taken from 17 different subjects (9 women and 8 men) using a laptop microphone, without any additional processing applied.

In addition to the query recordings, the dataset includes three metadata files: one describing the queries and two detailing the music collections used for retrieval experiments in the original study. Although the query recordings are publicly available within the dataset, the audio files corresponding to the music collections are not included due to copyright restrictions. To obtain these reference music tracks, we used yt-dlp, a Python library that enables automatic search and downloading of YouTube recordings based on the song title and artist name.

3.2 Vocal Source Separation

We employed *Spleeter* [4] and *Open-Unmix* [5] to extract vocal stems from audio recordings. Only songs with a duration of less than 3 minutes and 30 seconds are included, resulting in a total of 210 recordings.

For *Spleeter*, we use the 2-stem model, which separates the audio into vocals and instrumental tracks. *Open-Unmix* separates the audio into vocals, drums, bass, and other instruments.

In addition to the processed vocal stems, we included the raw audio recordings as a baseline for comparison. The extracted vocal stems are used for feature extraction in the construction of the query-by-humming (QBH) database.

3.3 Feature extraction

We extracted chroma and predominant pitch tracks ($f0$) from both the vocal extractions and raw audio recordings for each song.

For chroma feature extraction, we used STFT-based chroma features [1] with a frame size of 2048 and a hop size of 512. These features are computed using the implementaion in librosa [7].

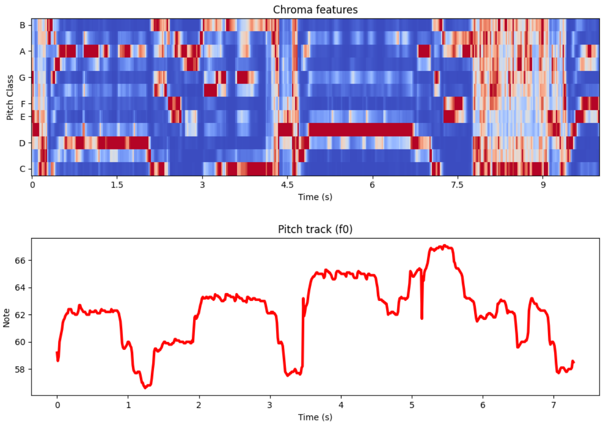

Figure 1. Chroma features (top) and Pitch track (bottom) of a query.

For $f0$ estimation, we use the pYIN algorithm [6] with the same frame size of 2048 and hop size of 512, as implemented in librosa.

Both chroma and pitch features were extracted for raw audio, and vocal extractions. The same feature extraction process is applied to the query recordings for evaluation.

3.4 Dynamic Time Warping

To address temporal mis-alignments in the query, which often arise due to inaccuracies in humming, we employ Dynamic Time Warping (DTW) [8] as the similarity measure. To constrain the warping path, we applied a Sakoe-Chiba band radius [9] of 0.1, which restricts the maximum allowable deviation from the diagonal. The accumulated cost is computed for each song. This process is performed for each source separation model and feature type.

Unlike traditional DTW implementations, we do not perform backtracking, as we do not require the alignment path. This significantly improves inference time while maintaining retrieval performance. Additionally, we apply a sliding window approach to identify the segment of the song that best aligns with the query.

For chroma features, we use cosine similarity as the distance metric. To mitigate transposition errors, we circularly shift the chroma features along the chroma axis.

For pitch tracks, we employed euclidean distance. Here, we center both the query and the window of the song by subtracting the median, ensuring that the median value is set to 0. This approach prevents transposition errors and avoids the need for repeated calculations of the cost for each transposition.

For each query, the costs associated with each song are computed.

4. EXPERIMENTAL SETUP

Once the similarity measures for each query have been computed, we evaluate the retrieval performance using standard ranking-based metrics: Mean Reciprocal Rank (MRR) and Precision@k ($p@k$). Additionally, we conduct an ablation study to assess the impact of different feature representations (chroma and pitch) and vocal source separation models (Spleeter, Open-Unmix, and raw audio). Results are reported in two separate tables, one for each feature type, with MRR, P@1, and P@10 scores.

4.1 Mean Reciprocal Rank

The Mean Reciprocal Rank (MRR) is used to evaluate the effectiveness of retrieval by measuring how early the correct song appears in the ranked results for each query. The metric is computed separately for chroma and pitch features across different vocal extraction methods. MRR is formally defined as:

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \tag{1}$$

where Q is the set of queries, and rank_i is the rank of the relevant song for query i.

4.2 Precision @ k

Precision@k ($p@k$) quantifies the proportion of queries for which the correct song appears within the top k ranked results. This metric is particularly useful for evaluating the effectiveness of ranking algorithms in scenarios where only the top-ranked candidates are relevant to the user. Precision@k is defined as:

$$P@k = \frac{1}{|Q|} \sum_{i=1}^{|Q|} 1(\text{rank}_i \leq k) \tag{2}$$

where $1(\text{rank}_i < k)$ is an indicator function.

For our evaluation, we report $p@1$ (i.e., whether the correct song appears as the top-ranked result), $P@3$, and $P@10$ (i.e., whether the correct song is retrieved within the top 10 results). P@1 is particularly stringent, measuring the exact retrieval accuracy, while P@10 provides a broader perspective on retrieval performance.

5. RESULTS

Model	MRR	P@1	P@3	P@10
Raw Audio	11.90	7.94	11.11	17.46
Spleeter	12.58	9.52	12.70	17.46
Openunmix	0.0	0.0	0.0	0.0

Table 1. Evaluation results for chroma features

Model	MRR	P@1	P@3	P@10
Raw Audio	1.16	0.00	0.0	0.0
Spleeter	26.41	20.63	26.98	34.92
Openunmix	0.0	0.0	0.0	0.0

Table 2. Evaluation Results for pitch track $f0$

63 queries were performed against 210 songs for each model-feature pair. From the results, it is evident that the best performance is achieved when using *Spleeter* with pitch tracks as the feature representation. In the case of chroma-based features, only a marginal improvement is observed. Notably, raw audio with chroma already achieves a mean reciprocal rank of approximately 11%, and the application of vocal separation models such as *Spleeter* and *OpenUnmix* leads to slight accuracy improvements.

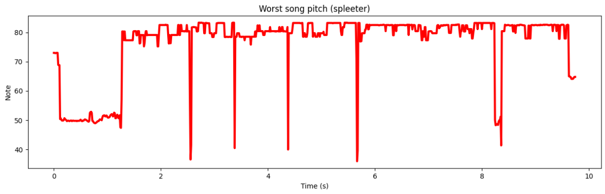

Figure 2. Pitch track ($f0$) of the worst performing song with *Spleeter*.

For pitch-based features, raw audio performs poorly, as expected, while vocal separation significantly enhances pitch accuracy. Further analysis of the worst performing songs reveals that inaccuracies in pitch estimation contribute to the reduced performance. These inaccuracies explain the observed inconsistencies in evaluation. Although the current results are not particularly strong, they are promising, suggesting that further research, particularly in improving pitch estimation and feature extraction could enhance the robustness of the proposed pipeline.

6. CONCLUSIONS & FUTURE WORKS

This study highlights the need for further research to improve the efficiency of automatic database generation for Query-by-Humming (QBH). While our proposed pipeline demonstrates the feasibility of leveraging vocal source separation and feature extraction for QBH tasks, several challenges remain.

Future work includes exploring more advanced vocal separation models and enhanced feature extraction techniques to improve the robustness and accuracy of the system. Vocal source separation and pitch extraction remain active areas of research in Music Information Retrieval (MIR), with continuous advancements that could significantly enhance the performance of QBH systems.

Additionally, developing more efficient similarity measures could drastically improve retrieval accuracy and computational efficiency. Optimizing these measures would contribute to the refinement of the proposed pipeline, making it more effective for large-scale databases.

7. REFERENCES

[1] J. Salamon, J. Serrà, and E. Gómez, “Tonal representations for music retrieval: from version identification to query-by-humming,” in *International Journal of Multimedia Information Retrieval*, vol. 2, pp. 45–58, 2013.

[2] A. Ghias, J. Logan, D. Chamberlin, and B. C. Smith, “Query By Humming – Musical Information Retrieval in an Audio Database,” in *ACM Multimedia 95 - Electronic Proceedings*, San Francisco, California, Nov. 1995.

[3] M. Rocamora, P. Cancela, and A. Pardo, “Query by humming: Automatically building the database from music recordings,” in *Pattern Recognition Letters*, vol. 36, pp. 272–280, 2014.

[4] R. Hennequin, A. Khlif, M. Voituret and F. Voituret, “Spleeter: a fast and efficient music source separation tool with pre-trained models,” in *Journal of Open Source Software* , 2019

[5] F. Stöter, S. Uhlich, A. Liutkus and Y. Mitsufuji, “Open-Unmix - A Reference Implementation for Music Source Separation,” in *Journal of Open Source Software* , 2019

[6] M. Mauch, and S. Dixon, “PYIN: A fundamental frequency estimator using probabilistic threshold distributions,” in *ICASSP 2014 - 2014 IEEE International Conference on Acoustics, Speech and Signal Processing* , 2014.

[7] B. McFee, C. Raffel, D. Raffel and D. Ellis, “librosa: Audio and Music Signal Analysis in Python,” in *Python in Science Conference* , 2015.

[8] M. Müller, “ Dynamic Time Warping. In: Information Retrieval for Music and Motion,” in *Springer*, Berlin, Heidelberg, 2007.

[9] T. Górecki, and M. Łuczak, “The influence of the Sakoe–Chiba band size on time series classification,” in *Journal of Intelligent & Fuzzy Systems*, vol. 36, pp.1-13, Berlin, Heidelberg, 2007.

[10] N. Craswell, “ Mean Reciprocal Rank,” in *Encyclopedia of Database Systems. Springer*, Boston, MA, 2009.

[11] S. Pothula, P. Dhavachelvan, “ Precision at K in Multilingual Information Retrieval,” in *International Journal of Computer Applications* , Vol. 24, 2011.