Making Searchable Melodies: Human vs. Machine

Mark Cartwright, Zafar Rafii, Jinyu Han, Bryan Pardo

Electrical Engineering and Computer Science Department

Northwestern University

2133 Sheridan Rd, Evanston, IL 60208

mcartwright@u.northwestern.edu, zafarrafii@u.northwestern.edu, jinyuhan@u.northwestern.edu, pardo@cs.northwestern.edu

Abstract

Systems that find music recordings based on hummed or sung, melodic input are called Query-By-Humming (QBH) systems. Such systems employ search keys that are more similar to a cappella singing than the original recordings. Successful deployed systems use human computation to create these search keys: hand-entered MIDI melodies or recordings of a cappella singing. Tunebot is one such system. In this paper, we compare search results using keys built from two automated melody extraction system to those gathered using two populations of humans: local paid singers and Amazon Turk workers.

Introduction

Music audio is a popular category of multimedia content. Services like iTunes provide millions of songs to the public, but typically index their recordings with such metadata as title, composer, and performer. Finding the desired recording can be a problem for those who do not know the metadata for the desired piece. However, if the user can sing or hum some portion of the song, a query-by-humming (QBH) system (Huq, Cartwright, and Pardo 2010) can be used.

One key challenge for building an effective QBH system is the creation of a large database (perhaps millions) of relevant search keys that are effective for matching against sung or hummed queries. Creating searchable keys that can be queried by singing is non-trivial. Sung queries typically outline a melody drawn from the desired recordings. The vast majority of music recordings do not have machine-readable notated scores or MIDI versions available. Therefore, melodic keys must be created directly from the audio. Historically, automated approaches to extracting the main melody from a polyphonic recording have not been sufficiently robust to build melodic keys. Therefore, databases of searchable melodies have been created either by hand keyed-in melodies (e.g. Musipedia (Typke 2011)) or by persuading singers to sing solo melodies to the system (e.g. Tunebot (Huq, Cartwright, and Pardo 2010)).

In this paper we compare search keys built through human computation to those built using 2 promising new automated vocal melody extraction methods: probabilistic latent component analysis (PLCA) (Han and Chen 2011) and vocal melody isolation based on rhythmic repetition (REPET) (Rafii and Pardo 2011). To the authors' knowledge, no researchers have published such a comparison. We evaluate the difficulty of creating a searchable database using each method and the effectiveness of the resulting searchable database.

The approach we took to measuring the effectiveness of the searchable keys created using each of the four methods (local paid singers, Amazon Turk workers, the REPET vocal melody extractor, PLCA vocal melody extraction) was simple. We used them as search keys in the database of a currently deployed QBH search engine: Tunebot.

We selected a target set of 100 popular songs not currently in the Tunebot database of 13,271 melodies. For these 100 songs we inserted melodic search keys generated by one of the methods into the database. We then took 1200 sung queries (12 per target song) with known target songs from the list of 100 and searched for them using Tunebot. The search key generation method that generates better search rankings was deemed better.

Human Generated Search Keys

We tried two approaches to generating search keys using human input: Hiring local singers and using Amazon Mechanical Turk.

Local singers were interviewed and auditioned by a trained musician with a graduate degree in music, and they were hired based on their singing ability. Each singer worked between 5 and 10 hours a week at a rate of $9.00 per hour. It took each local singer an average of 12 minutes to contribute one song. We collect three singer's contributions (a total of six examples) per song. The average human computation time for each song was 36 minutes, and the average total cost of each song was $3.60. This translates into roughly 9.5 days to create keys for the 100 songs in our test set.

We also tried outsourcing the human computation of search keys using Amazon's Mechanical Turk service. We solicited singers by posting a "Human Intelligence Task" (HIT) with the title: "Sing Popular Songs". We set our price at $0.10 per successfully contributed singing example.

The HIT went as follows: a brief description of the task was outlined in the text of the HIT. Turkers that accept the HIT were redirected to our website site and presented with a list of popular songs to sing from and instructed to sing the "most memorable portion of the song (often the verse or melody)".

Because of the anonymity of Mechanical Turk, we know little about the workers except for their singing ability and possible gender. While we had 211 unique Turk workers begin our HIT, only 70 of them actually submitted a contribution. Of the 70 workers, the mean number of contributions per worker was 8.83 and the median was 5. Upon a listening inspection, we estimate that 25% are male, and 75% are female. It took 20 days to obtain the sufficient coverage for the 100 songs in our test.

Machine Generated Search Keys

The REpeating Pattern Extraction Technique (REPET) is a novel and simple approach for extracting the repeating musical background from the non-repeating musical foreground in an audio signal. REPET took 64 minutes to extract melodies from the 100 songs in the target set.

In Probabilistic Latent Component Analysis (PLCA), an audio signal is first divided into vocal and non-vocal segments using a trained Gaussian Mixture Model (GMM) classifier. A statistical model of the non-vocal segments of the signal is learned adaptively from this particular input music by PLCA. This model is then employed to extract the vocal components from the audio mixture.

Experimental Results and Conclusions

Figure 1 shows experimental results for 1200 queries on a database of 13271 melodic search keys. Lower values indicate better search rankings. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points the algorithm considers to be not outliers, and the outliers are plotted individually. Median values are shown in each box, just above the central mark.

It is clear that human computation still dominates machine computation in quality. Further, Mechanical Turk workers may be a promising, cheaper alternative to local singers.

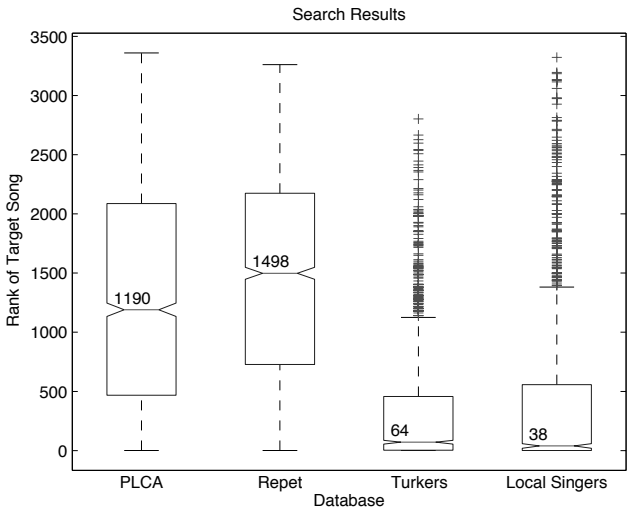

Figure 1. Search rank of the correct target song in a database of 13271 melodic search keys. N=1200 queries per boxplot. Lower numbers are better. Values in boxes are medians. Each boxplot shows search results using search keys generated with the specified method.

Acknowledgements

This work was funded by National Science Foundation Grant number IIS-0812314.

References

Han, J., and Chen, C.-W. 2011. Improving Melody Extraction Using Probabilistic Latent Component Analysis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011)*.

Huq, A.; Cartwright, M.; and Pardo, B. 2010. Crowdsourcing a Real-World On-Line Query by Humming System. In *Proceedings of the Sixth Sound and Music Computing Conference (SMC 2010)*.

Rafii, Z. and Pardo, B. 2011. A Simple Music/Voice Separation Method Based on the Extraction of the Repeating Musical Structure. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011).*

Typke, R. 2011. Musipedia Melody Search Engine. Retrieved April 28, 2011, from http://www.musipedia.org/.