QBT-EXTENDED: AN ANNOTATED DATASET OF MELODICALLY CONTOURED TAPPED QUERIES

Blair Kaneshiro **Hyung-Suk Kim** **Jorge Herrera**
Jieun Oh **Jonathan Berger**
CCRMA, Stanford University
{blairbo,hskim08,jorgeh,jieun5,brg}@ccrma.stanford.edu

Malcolm Slaney
Microsoft Research
CCRMA
malcolm@ieee.org

ABSTRACT

Query by tapping remains an intuitive yet underdeveloped form of content-based querying. Tapping databases suffer from small size and often lack useful annotations about users and query cues. More broadly, tapped representations of music are inherently lossy, as they lack pitch information. To address these issues, we publish QBT-Extended—an annotated dataset of over 3,300 tapped queries of pop song excerpts, along with a system for collecting them. The queries, collected from 60 users for 51 songs, contain both time stamps and pitch positions of tap events and are annotated with information about the user, such as musical training and familiarity with each excerpt. Queries were performed from both short-term and long-term memory, cued by lyrics alone or lyrics and audio. In the present paper, we characterize and evaluate the dataset and perform initial analyses, providing early insights into the added value of the novel information. While the current data were collected under controlled experimental conditions, the system is designed for large-scale, crowdsourced data collection, presenting an opportunity to expand upon this richer form of tapping data.

1. INTRODUCTION

Query by tapping (QBT) is the process of identifying a musical excerpt based upon a tapped representation. QBT is a canonical Music Information Retrieval (MIR) task and an intuitive query to perform [16], yet the literature on this topic remains small relative to other query forms such as singing and humming [9]. This task is also among the least attempted in recent years of MIREX [3]. A number of retrieval systems and databases using rhythm or tapping have been published to date (Table 1). Some cite the need for larger datasets for testing and validation [4–6]. Some lack annotations, such as musical ability of the performer, or the performer's familiarity with the excerpt being queried; such annotations could prove useful in developing improved systems [11, 12, 16]. It is also not always clear how the performer was cued to perform the query (from short-term or long-term memory; or from score, lyrics, or audio), nor how the type of cue affected performance.

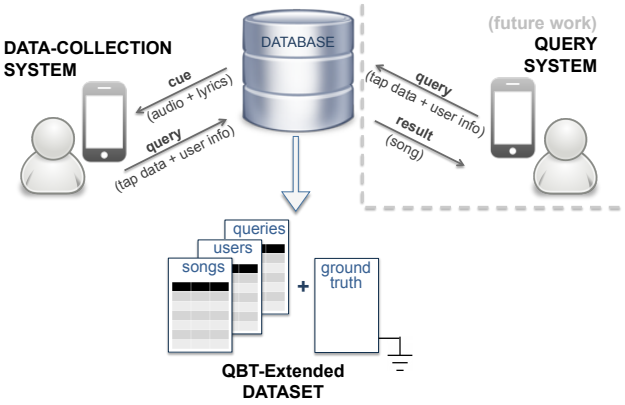


Figure 1: QBT-Extended system and dataset overview.

Study	Songs	Performers	Queries
Chen & Chen, 1998 [1]	102	NA	NA
Jang et al., 2001 [8]	NA	9	269
Eisenberg et al., 2004a [4]	9	3	27
Eisenberg et al., 2004b [5]	9	4	144/288
Peters et al., 2005 [15]	30	NA	NA
Peters et al., 2006 [16]	30	NA	518
Hanna & Robine, 2009 [6]	103	6	533
MIREX: MIR-QBT [1]	136	NA	890
MIREX: QBT_symbolic [1]	**143**	NA	410
Current study	51	**60**	**3,365** [2]

Table 1: Size of published QBT datasets to date. NA indicates that information was not available.

Another possible hurdle in QBT research is that tapping data are inherently lossy. A performer is likely replaying the pitches of a melody in his head as he taps its rhythm, but this experienced information is not captured in the output [14]. Confounding the issue further is the fact that musical excerpts can share similar or identical rhythms while being vastly different melodically (consider "Happy Birthday" and "The Star-Spangled Banner" as examples). On

[1] http://www.music-ir.org/mirex/wiki/Query_by_Tapping
[2] Numbers reported for the current dataset reflect those queries for which at least one tap event was registered. With zero-tap queries (for skipped tasks) included, the dataset comprises 3,943 queries.

a perceptual level, too, human recognition of musical excerpts is generally less successful using rhythm alone than melody alone, regardless of a listener's level of musical training [7]. Therefore, melodic information could be a useful addition to the QBT signal.

With the goal of facilitating future QBT research, we publish a dataset of annotated queries, and a system for collecting them. The queries were collected from 60 unique participants performing excerpts from a set of 51 songs. As the queries were performed on the 2D touchscreen of a mobile device, we were able to collect not only the timestamps of tap onset (touch on) and release (touch off) events, but also a rough melodic contour based upon the position of each tap on the screen. Participants were cued from either long-term memory (lyrics only) or short-term memory (lyrics and audio) for a given query task. Finally, the tapped queries are annotated with information about the performer, including musical training and the level of familiarity with each excerpt.

The current dataset was collected under controlled experimental conditions. However, the system, being mobile and open source, is easily extendable to crowdsourced data collection.

The remainder of this paper is structured as follows. We first explain how we devised the system and collected data (§2). We then describe the data (§3) and perform illustrative analyses for QBT application (§4). We conclude with a discussion of implications and next steps (§5).

2. DATA COLLECTION

2.1 Stimulus Set

We wished to maximize the number of queries that could be performed from long-term memory (cued by lyrics only). To assemble a set of songs that would be maximally familiar based upon lyrics alone, we conducted a survey to assess familiarity of lyrics excerpts from 120 top British and American pop songs from 1950–2010. Songs were chosen based upon their presence in a variety of Top 10 lists from each year, and lyrics were drawn from songs' main themes and choruses. Cued by lyrics, participants rated on a 3-point scale whether they knew the accompanying melody (Yes, Maybe/parts of it, No). We targeted the same demographic for the survey as we would for subsequent QBT data collection.

Fifty-five participants (born between 1929–1994; mean birth year 1985; 27 female) completed the survey. Each song was scored (# Yes responses - # No responses), and the 49 highest-scoring songs were retained for the QBT-Extended stimulus set. We additionally included "Happy Birthday" and "The Star-Spangled Banner" to illustrate the same-rhythm/different-melody phenomenon. The resulting 51 audio excerpts ranged in length from 9 to 28 sec (mean length 16.96 sec).

2.2 Participants

Tapping data were collected from 60 participants; participant information is summarized in Table 2. All participants were fluent in English, at least 18 years old, and reported normal hearing and no cognitive or decisional impairments. Informed consent was obtained from each participant at the start of his session. Both the survey and tapping study were approved by the local Institutional Review Board (IRB).

Data	Value Stored	Range Collected
age	integer	18-64 (mean = 30.8)
gender	male/female	33/27
native language	text	English = 53
music listening	0 (never) to 5 (all the time)	mean = 3.91
instrument training	0 (none) to 5 (professional)	mean = 2.53
theory training	0 (none) to 5 (professional)	mean = 2.08
handedness	left/right/both	3/55/2
tone-deafness	yes/no/don't know	1/58/1
arrhythmic	yes/no/don't know	0/59/1
specific training	time and instrument	varies

Table 2: Participant information collected from in-app questionnaire. Values for *music listening*, *instrument training*, and *theory training* are continuous in the given range. Multiple answers (instruments) per user were allowed for *specific training*.

2.3 System

The system comprises a front end for data collection and a back end for data storage and processing/analysis. An iOS application was developed for the front end so that participants could leverage the 2D touchscreen for queries, tapping higher on the screen for higher pitches, and lower for lower pitches. A screenshot of the tap screen is shown in Figure 2. Our internal tests show a tap-to-timestamp latency of 40 ms on average, with standard deviation of 10 ms. Because the start time of each recorded query is set to the onset of the first tap event, mean latency is not a factor. The front-end application was also used to obtain informed consent, and for the participant questionnaire (§2.4). The mobile implementation facilitated data collection and provided an ecologically valid apparatus that would extend easily to real-life use of a QBT system.

Figure 2: The data-collection application's tapping interface. Users can tap anywhere in the shaded gray area of the screen, using the vertical position of the tap to denote pitch height.

The back-end application was written in Ruby-on-Rails to receive and store the collected data. The data are stored in an SQLite3 database. The back end also stores the audio and lyrics files that are fetched by the front end when a new experiment is instantiated.

2.4 Data Collection Procedure

All data were collected using 4th Generation iPod Touch devices and Sony MDR-V6 headphones. Participants started the session by giving informed consent and filling out the questionnaire (Table 2). Following this, the participant completed 3 practice trials in order to learn how to use the application and perform queries, with the experimenter on hand to provide instruction and clarification. Once the participant was comfortable with the interface, he performed up to 51 trials in random order for the remainder of the 45-minute session. [3] Participants were given a $10 gift card at the end of the session. No authors contributed to the dataset.

A single trial is described as follows:

1. A lyrics excerpt, along with the song title, performer, and year, is presented on screen.
2. Long-term memory task: The participant is asked to tap the melody accompanying the lyrics if it is familiar, using the vertical axis to denote approximate pitch positions. If the user cannot recall the melody, he skips this step.
3. The participant is asked "How familiar was the song presented?" The answer is encoded as a continuous value from 0 to 5.
4. The participant listens to the audio accompanying the excerpt. The audio plays only once, and must be heard in its entirety. The lyrics and metadata are shown on screen while the audio plays.
5. Short-term memory task: The participant is taken back to the lyrics/metadata screen (described in Step 1) and taps the melody (regardless of whether he was able to do so from long-term memory).
6. The participant is asked "Did hearing the music help? Tap on the answer that fits best." The answer options, and distribution of responses, can be found in Table 3.

3. DATASET

3.1 Ground Truth

A ground truth was created for each excerpt by a single performer with 14 years of piano training. MIDI keyboard renditions of the sung melodies in the excerpts were converted to comma-separated value (CSV) files with MIDI note number, note-on, and note-off times. Both MIDI and CSV formats are included in the dataset.

[3] Some participants requested to perform queries for all of the songs, taking more than 45 minutes to complete the set.

3.2 Statistics of the Collected Data

A total of 3,365 queries were collected—1,412 tapped from long-term memory (cued by lyrics only) and 1,953 from short-term memory (cued by lyrics and audio). For each song, an average of 27.69 long-term memory queries (min = 16, max = 37) and 38.29 short-term memory queries (min = 31, max = 47) were collected. Each participant performed on average 23.53 queries from long-term memory (min = 0, max = 51) and 32.55 queries from short-term memory (min = 20, max = 51).

3.3 Structure

The database contains 3 tables: *songs*, *users* (participants), and *tasks*. The fields of each table are summarized in Table 4. The *songs* table contains information about the songs in the stimulus set. Due to possible copyright issues, the dataset does not include lyrics or audio, and only specifies the start and end times within the original song, as well as the song part from which the lyrics are derived (main theme, chorus, or other). The *users* table contains the participant information summarized in Table 2. The *tasks* table contains the information for each query including `user_id` and `song_title`, which can be used to identify the participant and song of a given query.

songs (7 fields)		
filename	song_title	artist
year	start_time	end_time
song_section		
users (11 fields)		
age	gender	listening_habits
instrument_training	theory_training	handedness
tone_deaf	arrhythmic	user_id
native_language	specific_training	
tasks (16 fields)		
version_number	song_title	user_id
session_id	experimenter_id	task_order
device_type	song_familiarity	with_music
audio_helpful	tap_data	tap_off_data
tap_x_data	tap_y_data	tap_off_x_data
tap_off_y_data		

Table 4: Database table fields. The `song_title` field connects the *songs* and *tasks* tables, while the `user_id` field connects the *users* and *tasks* tables.

The dataset is available as 1) an SQLite3 db-file; 2) comma-separated value (CSV) files; and 3) space-separated .onset files compatible with previous QBT datasets. In addition to .onset files for a given task, we provide in separate files the *x* and *y* screen coordinates and release times corresponding to each onset.

The dataset is distributed using the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 license, and is available for download at the following URL:

`https://ccrma.stanford.edu/groups/qbtextended`

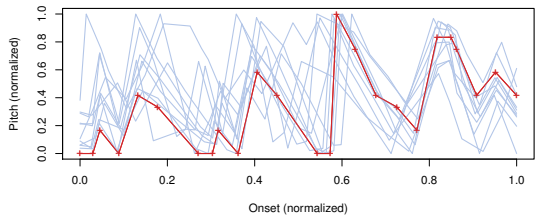
3.4 Observations of the Data

We present two example visualizations of the tapped queries. First, a rough melodic contour can be recon-

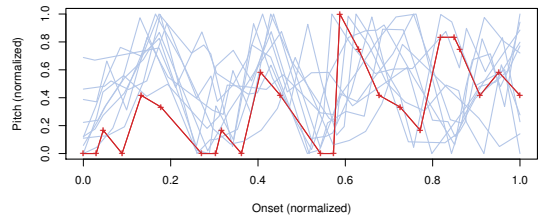
Did hearing the music help?	With long-term	Without long-term
Yes—it helped me remember more details of the song	70.0 %	30.6 %
Yes—I thought the lyrics were from a different song, but now I know which song it is	0.9 %	4.0 %
Yes—I had no idea of the song from just the lyrics, but listening made me recognize it	2.3 %	26.1 %
No—I already knew the song really well	24.7 %	1.5 %
No—this song is totally unfamiliar, so hearing it once didn't help	0.4 %	10.1 %
Yes—I didn't know the song at all, but I could tap it out after hearing it	1.6 %	27.7 %

Table 3: Distribution of answers to the question asked at the end of the short-term memory task (§2.4). The second column shows the distribution for short-term memory queries where the participant also did the long-term memory task; the third column presents the distribution for cases where the participant performed the query from short-term, but not long-term, memory.

structed by plotting each tap position as a function of its onset time. Figure 3 shows the short-term memory queries for "Happy Birthday" from participants in the top and bottom quartiles of musical instrument training for the song. The ground truth is overlaid in red. Both the query lengths (x-axis) and tap positions (y-axis) have been normalized to the length of each query and total vertical range of the screen used, respectively.



(a) Highest quartile of instrument training [3.5–5.0]



(b) Lowest quartile of instrument training [0.0–2.0)

Figure 3: Tapped pitch contours of "Happy Birthday" queries from short-term memory (blue) with ground truth (red). Pitch contours and timestamps have been normalized to the vertical range of the screen used and the total length of the query, respectively. Variance among queries appears to be lower for the highly trained participants, especially in the second half of the query.

In contrast, Figure 4 focuses solely on the temporal dimension of the queries. More variability in the temporal patterns is evident when absolute tap times are presented (Figure 4a), but the phrase structure becomes easier to discern across the set when each query is normalized by its length (Figure 4b).

We include some anecdotal observations from the experiment sessions:

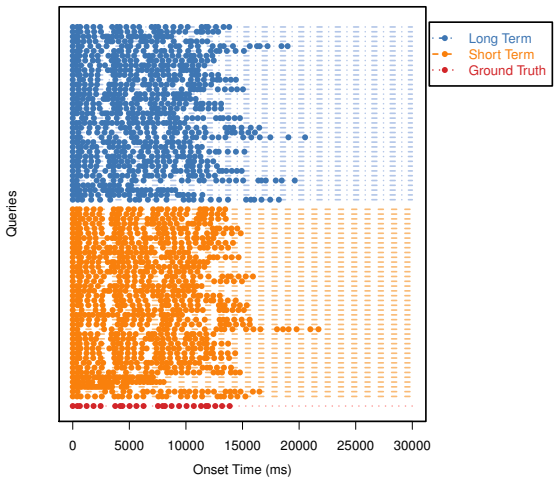
1. *The tapping queries were fun to do.* This observation is supported by some participants requesting to finish the full set of songs; but even participants who did not finish the full set reported that they had fun.
2. *The pitch positions are not exact representations of the melody being performed.* For example, repeated pitches in a melody were likely not tapped at the exact same position on the screen; similarly, recurring taps in a given position do not necessarily reflect the same pitch. In addition, the range of the melody relative to the range of the screen may have changed over the course of a query, as participants could encounter notes outside of the range they had accounted for initially.
3. *Participants with less musical training reported difficulty with the pitch dimension.* Some participants reporting this problem preferred to tap in the same general area of the screen, while others tapped monotonically up the screen for each line of lyrics. As evidenced in Figure 3, degradation of the pitch contour is observable among participants with lower reported levels of instrument training.
4. *No participants reported trouble with the rhythm dimension.* More analysis is needed to confirm this observation.
5. *Many participants reported often not knowing a melody from the lyrics alone, but recognizing it once the audio started playing.* Evidence of this can be seen in Table 3, as 26.1% of participants who could not do the long-term memory task actually did know the song once they heard the audio.
6. *Some participants reported that their instrument experience (e.g., guitar, drums) distracted them from the vocal line in the audio, and that they wanted to tap their instrument's part instead.* More analysis is needed to confirm this observation.

4. PRELIMINARY ANALYSIS

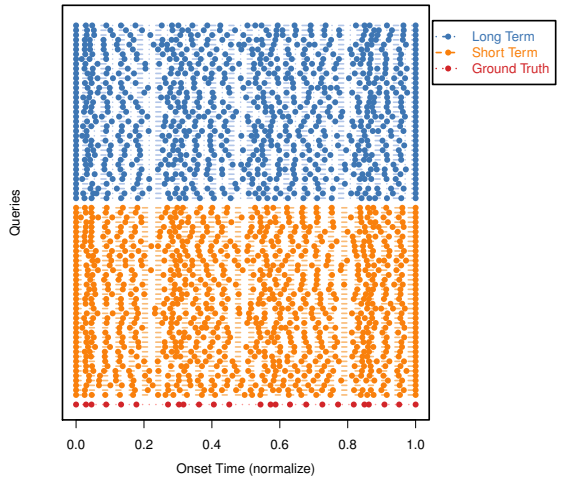
Although the main focus of this paper is to introduce QBT-Extended, in this section we perform some preliminary analyses of the dataset to build a rudimentary QBT system. We acknowledge that these analyses are very basic, and they are intended primarily for illustrative purposes and to validate the data.

We validated the temporal dimension of the dataset using Rhythmic Contour String encoding [16] for both the ground truth and the queries.[4] All queries, as well as

[4] This encoding normalizes onset-to-onset durations relative to mean inter-onset duration, then converts each inter-onset duration to S ("same"), U ("up"), or D ("down") using a threshold of "sameness" to define S. We used a threshold of 0.2.



(a) Not normalized



(b) Normalized

Figure 4: Temporal dimension of all queries collected for "Happy Birthday".

the ground truth, are represented as separate strings. We then used edit distance (or Levenshtein distance) between strings, computed by the approximate matching procedure [16], to rank the distance between a query and each of the ground-truth strings. The distances were computed over the full set of 51 candidate excerpts, and results were ranked by increasing order of distance. Table 5 contains the classification accuracy for the entire dataset against the ground-truth data. A query result is considered correct if the actual song was among the top N sorted matches.

We adapted the Rhythmic Contour String method to analyze the novel pitch dimension. Because the pitch information expressed in the queries is not exact (§3.4), we encoded pitch contours as separate strings based upon note-to-note variance relative to the overall position range of the query. Differences in position were computed between successive tap events, and the set of differences for a given query were normalized between 1 and -1. Following that, the same 0.2 thresholding was used. We then applied the approximate matching procedure to compare distances between a query and each ground truth, and ranked the results.

Each query therefore comprises two strings; one for rhythm and one for melody. As a preliminary attempt to make use of both representations, we computed the rhythmic and melodic distances separately and averaged them, giving equal weighting to each.

4.1 Assessing Performance with Added Pitch Dimension

The accuracy of our simple system using each dimension alone, and rhythm and melody combined, is shown in Table 5. These results were computed using the entire set of queries, and no filtering was applied based on user ability or familiarity with each song excerpt. Rhythm alone outperforms melody alone for all three ranking ranges.

	Accuracy (%)		
Condition	Top 1	Top 5	Top 10
Rhythmic contour	51.92	70.82	78.46
Melodic contour	37.68	54.27	64.67
Both contours	53.67	70.70	78.28
	Significance analysis		
$\chi^2(1, N = 3,365)$	4.69	0.017	0.060
p-value	0.030	0.90	0.81

Table 5: Accuracy of the simple QBT classification system using all 3,365 tapped queries (51-class problem). χ^2 and p-values, comparing classification using rhythmic contour alone versus rhythmic and melodic contours combined, were computed using McNemar's test.

To quantitatively assess the effect of adding the pitch dimension, we used McNemar's test [13], an established method of comparing two classifications of a single dataset [2]. We compared performance when the classifier used only rhythmic contour, versus rhythmic and melodic contours together. The tests show that adding melodic contour significantly improved accuracy for the Top 1 case, but did not significantly affect classifier performance in the Top 5 and Top 10 cases.

A specific case in which melodic information should boost accuracy is the 2-class problem of "Happy Birthday" versus "The Star-Spangled Banner", which share similar rhythms. As shown in Table 6, using rhythmic and melodic contour together significantly improved classifier accuracy over using rhythmic contour alone.

Condition	Accuracy (%)
Rhythmic contour	75.54
Melodic contour	72.66
Both contours	89.93
	Significance analysis
$\chi^2(1, N = 139)$	10.03
p-value	0.0015

Table 6: Comparison of accuracies in the 2-class problem classifying "Happy Birthday" and "The Star-Spangled Banner" queries. McNemar's test measures the significance of the change in classifier performance when both rhythmic and melodic contour are used, versus rhythmic contour alone.

5. DISCUSSION

The QBT-Extended dataset and system presents new possibilities for QBT research. By appropriately applying the new pitch position information, and by understanding how user background and memory cue affect performance, more effective QBT systems may be implementable. In addition, the touchscreen-based interface for data collection may prove useful for users who are not comfortable singing or humming, or who wish to query in situations where use of the microphone for acoustic input is not ideal (for instance, in a library or a noisy bar).

We acknowledge limitations of the current dataset. First, our choice of device for data collection imposed constraints upon the range of vertical space available, and users may have run out of room or needed to tap over the lyrics for some queries. In addition, the act of expressing the pitch dimension was confusing for some users, especially those who did not have musical training or could not read music. Therefore, it may be the case that having to focus on both timing and pitch degraded the quality of output along both dimensions. As we noticed that some sliders were not moved, nor instrument training fields filled out in the questionnaire, some users may not have entered their information completely. Finally, we acknowledge the demographic skew of the current participant population for this dataset, given the community that we targeted for the study [10].

5.1 Future Work

Many opportunities for future work are present. First, more analysis can be done to evaluate the usefulness of both the added pitch dimension and the annotations accompanying each query. For example, it may be useful to weight the temporal versus pitch dimensions of a query based upon users' musical expertise, experience, and familiarity with the specific excerpt. Alternate representations of queries, such as the normalized signals shown in Figure 3, could also provide feasible feature vectors for classification. Beyond the domain of query, the dataset is potentially useful for research on musical memory, expertise, and other aspects of music cognition.

Because the data-collection system is open source, a natural extension of the current implementation would be to port it to other platforms (e.g., Android, web [16]) and devices (e.g., tablets). The system is also well positioned for crowdsourced data collection, and the current dataset could then serve as a control to validate the quality of data collected via crowdsourcing.

6. REFERENCES

[1] James CC Chen and Arbee LP Chen. Query by rhythm: An approach for song retrieval in music databases. In *Research Issues In Data Engineering, 1998. Proceedings of Eighth International Workshop on Continuous-Media Databases and Applications*, pages 139–146. IEEE, 1998.

[2] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30, December 2006.

[3] J Stephen Downie, Andreas F Ehmann, Mert Bay, and M Cameron Jones. The music information retrieval evaluation exchange: Some observations and insights. In *Advances in music information retrieval*, pages 93–115. Springer, 2010.

[4] Gunnar Eisenberg, Jan-Mark Batke, and Thomas Sikora. BeatBank – an MPEG-7 compliant query by tapping system. In *Audio Engineering Society Convention 116*, 2004.

[5] Gunnar Eisenberg, Jan-Mark Batke, and Thomas Sikora. Efficiently computable similarity measures for query by tapping systems. In *Proceedings of the Seventh International Conference on Digital Audio Effects (DAFx'04), Naples, Italy, October*, pages 189–192, 2004.

[6] Pierre Hanna and Matthias Robine. Query by tapping system based on alignment algorithm. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 1881–1884. IEEE, 2009.

[7] Sylvie Hébert and Isabelle Peretz. Recognition of music in long-term memory: Are melodic and temporal patterns equal partners? *Memory & cognition*, 25(4):518–533, 1997.

[8] Jyh-Shing Roger Jang, Hong-Ru Lee, and Chia-Hui Yeh. Query by tapping: A new paradigm for content-based music retrieval from acoustic input. In *Advances in Multimedia Information ProcessingPCM 2001*, pages 590–597. Springer, 2001.

[9] Alexios Kotsifakos, Panagiotis Papapetrou, Jaakko Hollmén, Dimitrios Gunopulos, and Vassilis Athitsos. A survey of query-by-humming similarity methods. In *Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments*, page 5. ACM, 2012.

[10] Jin Ha Lee and Sally Jo Cunningham. The impact (or non-impact) of user studies in music information retrieval. In *Proceedings of the 13th International Society for Music Information Retrieval Conference*, pages 391–396, 2012.

[11] Micheline Lesaffre, Koen Tanghe, Gaëtan Martens, Dirk Moelants, Marc Leman, Bernard De Baets, Hans De Meyer, and Jean-Pierre Martens. The MAMI query-by-voice experiment: Collecting and annotating vocal queries for music information retrieval. In *Proceedings of the International Society for Music Information Retrieval Conference*, 2003.

[12] Mark Levy. Improving perceptual tempo estimation with crowd-sourced annotations. *Proceedings of the International Society for Music Information Retrieval Conference*, pages 317–322, 2011.

[13] Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, June 1947.

[14] Elizabeth Louise Newton. *The rocky road from actions to intentions*. PhD thesis, Stanford University, 1990.

[15] Geoffrey Peters, Caroline Anthony, and Michael Schwartz. Song search and retrieval by tapping. In *Proceedings of the National Conference on Artificial Intelligence*, volume 20, page 1696. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

[16] Geoffrey Peters, Diana Cukierman, Caroline Anthony, and Michael Schwartz. Online music search by tapping. In *Ambient Intelligence in Everyday Life*, pages 178–197. Springer, 2006.