7 Spontaneous User Behavior in "Vocal" Queries for Music-Information Retrieval

Micheline Lesaffre, Dirk Moelants, and Marc Leman

Institute for Psychoacoustics and Electronic Music
Department of Musicology
Ghent University
Blindijnberg 2, B-9000 Belgium
Micheline.LeSaffre, Dirk.Moelants, Marc.Leman@UGhent.be
Project website—*http://www.ipem.ugent.be/MAMI*

Abstract

We report the findings of a study of spontaneous behavior involved in vocal queries for music-information retrieval (MIR) systems. Two conditions were investigated: reproducing music purely from long-term memory, and imitation after listening to the piece. When participants are allowed maximal freedom in the formation of queries, a range of behaviors is found. We first give a global view on the characteristics of music queries, then consider the effects of long- and short-term memory. Our aim is to show that an informed knowledge of human approaches to music-information retrieval can provide valuable insights into appropriate development of user-friendly, content-based access to digital music collections.

Music Query: Methods, Models, and User Studies (Computing in Musicology 13). Published by CCARH and The MIT Press, 2004.

7.1 Introduction

In a vocal-query system, the imperfections of human memory and inaccuracies of amateur performance both require accommodation. The user may not correctly recall a theme, or may not have the vocal skills to produce a good imitation of it. Most MIR research (see e.g. ISMIR conference proceedings) is focused on systems that can automatically extract information from an audio signal. User performance of vocal queries has been little studied although it is necessary to determine the most comfortable way for a user to reproduce a tune in a MIR system.

7.2 Query-by-Humming vs. Query-by-Voice

In the ideal vocal-querying environment, users can freely use different vocal methods for expressing the target content. Possible methods include singing lyrics, singing syllables, humming, whistling, or even speaking. Ghias et al. (1995) introduced the concept of "query-by-humming" (QBH) and considered it a natural approach to content-based music retrieval. Yet humming is in fact a highly problematic query method. Singing without opening the lips is somewhat artificial for the user, and produces legato and low-volume performance, which makes automatic recognition relatively difficult. Instead, it seems more natural to request that users sing well-defined syllables (e.g. McNab 1997).

In practice, QBH became a generic term. Researchers talk about 'query-by-humming applications' and 'hummed queries' without distinguishing between the methods used as input. Given the very limited meaning of the word humming, we prefer to speak of vocal queries and thus of "query-by-voice" (QBV) systems. By vocal query we simply mean a query produced by the voice and the vocal organs (i.e. vocal cords, tongue, lips, teeth).

We are not aware of experiments described in the literature that directly investigate the spontaneous user behavior in vocal queries. Many vocal-query experiments are carried out in view of processing queries to symbolic transcription without any concern for the user's preferences. Although the output should be representative of what can be expected as input to a QBV-system, clear constraints such as singing a specific syllable are imposed and queries are made by a small number of subjects. Experimenters assume that users are willing to accept certain constraints. It is argued that these constraints are compensated by a higher reliability of the system. However, these assumptions have not been justified. We simply don't know which vocal-querying method users prefer, and how frustrated users may become by imposing certain constraints.

Most systems rely on queries that are made by imitating music, thus using direct memory responses. However, one may assume that a great percentage of users of a QBV system might rely on long-term rather than on short-term memory for melodies. Within the MIR community very few studies investigate the impact of long-term memory. In studying long-term memory of absolute pitch, Levitin (1994) asked subjects to reproduce the pitch of the first few notes of their favorite rock 'n' roll song prompted by visual (and tactile) memory. His findings showed that subjects' long-term memory preserved a stable and accurate representation of musical pitch of highly familiar music, but also showed that subjects who heard the correct pitches internally could not always reproduce them with the same accuracy.

The question to be asked is why any constraints should be imposed. And if so, what is the most natural vocal method for users? Would they prefer to hum, sing in syllables, use lyrics, or whistle instead? So, instead of imposing constraints based on the limitations of a recognition system, our approach starts from the side of the user. We present the results of an experiment in which 72 subjects were asked to perform queries with as few constraints as possible. This allows us to describe some important features of spontaneous vocal querying behavior.

7.3 Description of the Query-by-Voice Experiment

7.3.1 Setup

The QBV experiment was designed to give subjects a maximal freedom of expression. After checking participants' familiarity with the music, they were instructed to vocally imitate a part of a piece of music using any vocal-query method and any combination of methods. We maintained maximal freedom of spontaneous behavior by allowing queries to refer to any fragment of the original music and to any voice (i.e. any melody, bass line or any particular instrument).

7.3.2 Subject Profile

The seventy-two subjects were recruited from students and staff members at Ghent University. The subject pool can be described as follows:

- They ranged in age from 19 to 56, with an average of 27.7 years.
- Twenty-eight subjects (39%) were female and 44 (61%) male.
- Thirty-seven (51%) played a musical instrument, practicing between 0 and 30 hours a week with an average of 6.1 hours; 38 (53%) had no musical education, 30 (42%) basic musical education, and 4 (6%) higher musical education (conservatory).
- On the combination of the results for the playing of an instrument and musical education, the subjects were divided into a group of 30 (42%) musicians and 42 (58%) non-musicians.

The native language of all subjects was Dutch. The experiment was anonymous and the participants were rewarded with a cinema ticket.

7.3.3 Stimuli

The experiment was based on a selection of thirty pieces of music from the MAMI target test collection.[1] The selection of pieces has been based on the criteria that the pieces selected should be:

- maximally familiar to the subjects
- easy to remember
- easy to imitate (reflecting the heterogeneous musical landscape)

Thus it contained different genres: besides popular music songs, ranging from chanson to heavy metal, it includes well-known Flemish children's songs and classical music.[2]

7.3.4 Procedure

Four vocal query methods were suggested to (but not imposed on) these subjects:

- humming,
- singing the lyrics,
- singing nonsense syllables, or
- whistling.

Methods could also be combined. The query length was limited to 30 seconds.

In a preparatory stage, titles and performers (or composers) of the thirty pieces were presented textually. Subjects were asked to indicate whether they knew the piece, and if so, whether they could remember it and would be able to imitate it.

In the first part of the actual experiment, subjects were asked to perform a vocal query for ten titles previously indicated as known and imitable. After each performance, the subjects were offered a second chance to produce another query for the same piece. Before moving to the next title, they were offered the possibility to make an additional query using another method (by sound recording or by typing text), or to propose alternative query methods.

In the second part of the experiment, four entire pieces of music, previously indicated as unknown or not possible to recall, were aurally presented. If there were less than four songs in this list, pieces indicated as imitable in the preparatory stage but not included in the first part of the experiment were used. The subjects were first requested to listen to

[1] An audio database of 160 pieces collected for the MAMI (Musical Audio Mining) project at Ghent University (see: http://www.ipem.ugent.be/MAMI).

[2] Permission to use this collection for the purpose of the project has been given by SABAM, the Belgian author rights association. For a complete list of the musical stimuli, see the Appendix.

the entire piece. They were asked if it was familiar and then to perform a vocal query, still following the instructions of the first part. Also here, the possibility to perform a second query was given. Figure 7.1 gives a schematic overview of the procedure.

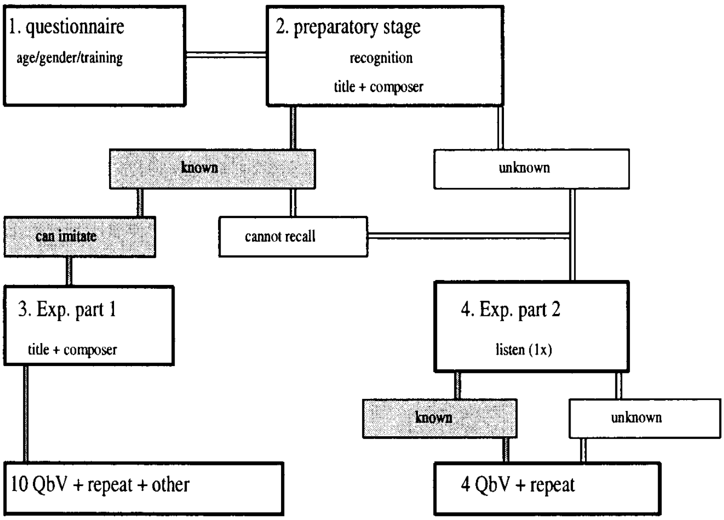

Figure 7.1. Schematic overview of the four steps in the procedure: (1) answering the questionnaire, (2) preparatory stage where the recognition of titles is checked, (3) first part of the experiment: queries relying on long-term memory, and (4) second part of the experiment: queries made immediately after listening to the piece.

7.4 Analysis Strategy

A team of musicologists manually annotated the 1,148 vocal queries resulting from the experiment. The annotation strategy was oriented towards general characteristics rather than a detailed analysis of, for example, issues of pitch and rhythm. The features investigated are: the length, beginning and end of the query, the vocal query method (singing lyrics, humming, whistling, etc.), performance style (melodic versus rhythmic), target similarity, and syllabic structure (what syllables are used). Some of these aspects are rather subjective, such as the assessment of the degree of similarity between query and original, and results are therefore rather tentative. Other aspects, such as temporal properties and method, allow a rather precise annotation.

Vocal queries could contain a mix of different query methods. In order to study the properties of the different methods, the queries were segmented according to the methods used. This resulted in a set of 2,114 segments, characterized by homogeneity of query method. In addition to the four suggested methods we found percussion (such as tapping along with the drum) and spoken comments.

7.5 Statistical Results

7.51 General aspects of the queries

General features are addressed using the original set of 1148 non-segmented queries.

7.5.1.1 BEGINNING AND LENGTH OF QUERIES

Seventy-eight queries (6.79% of the total) exceeded the 30-second limit of the recording time. In these cases the recording was stopped and the computer warned the performer. Twenty-six recorded fragments (2.26%) contained no query at all, as the subjects accidentally or deliberately stopped the recording before producing any kind of query. These files were not taken into account in our statistical analysis.

The average starting time of the query occurred 634 ms after the subjects started the recording, with 99.3% starting within 2 seconds. The mean length is 14.04 seconds, but the distribution is asymmetric, peaking towards 6 seconds and then slowly diminishing towards the maximum-allowed length of 30 seconds, with a peak between 5 and 15 seconds. However, large differences occur between subjects, with personal averages ranging from less than 5 seconds to close to the maximum-allowed recording time of 30 seconds. Also the average starting time shows huge differences between subjects, with personal averages varying roughly between 150 and 1500 ms.

7.5.1.2 NUMBER OF SEGMENTS, BASED ON QUERY METHOD

About 60% of the queries consisted of one homogenous query method and were treated as one single segment. Other queries contained different methods, as well as changes from one method to another and back (for example, lyrics, whistling, and back to lyrics, etc). The maximum number of segments observed in one single query was 12, but 97.8% of the queries contained a maximum of six segments. The distribution of the queries according to the number of segments is shown in Figure 7.2.

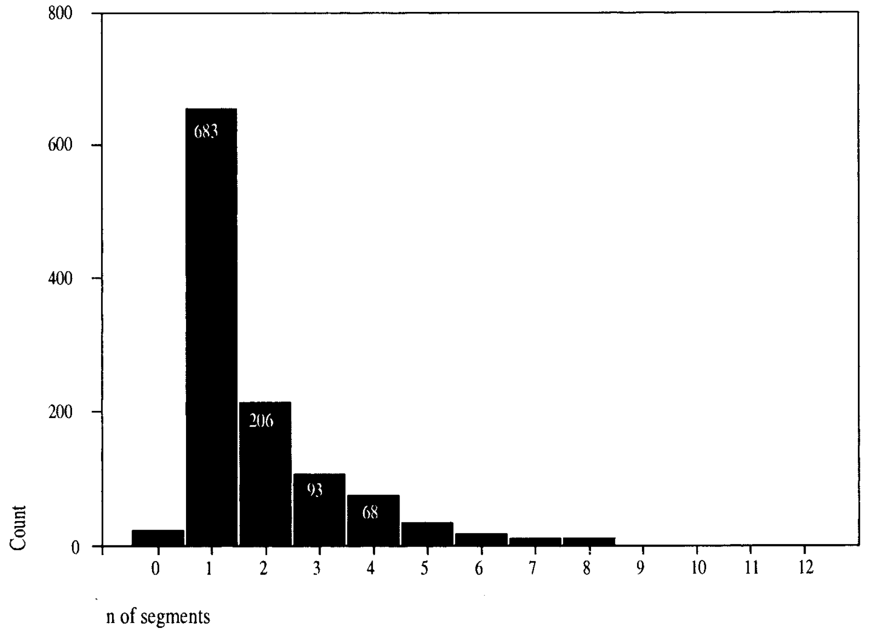

Figure 7.2. Distribution of the queries according to the number of segments. About 60% of the queries produced consist of only one query method.

7.5.2 Segment-Specific Aspects

For segment-specific aspects the set of 2114 segments was used.

7.5.2.1 LENGTH OF THE SEGMENTS

Analogous to the analysis of the queries as a whole, the timing of the segments was investigated. As for the duration of the full queries, a similar asymmetric distribution is shown, but the average length is shorter, with a mean of 7.42 seconds and a sharp peak around two seconds.

7.5.2.2 QUERY METHOD

Figure 7.3 and Table 7.1 give an overview of the different query methods in terms of number of segments and total time within the output. The two most prominent methods are singing on text and singing on syllables. Together they account for 83.4% of the queries, and 78.2% of the total time. The frequency of text segments is higher, but the syllabic method is more prominent within the total time, which shows that syllabic segments are in general longer. Among the other methods, whistling is prominent, particularly because the average duration of whistled segments is rather long (with a mean length of 14.63 seconds, almost double the average). Thus, although only 8.6% of the segments are whistled, they take 17.1% of the total time. By contrast, 8.1% of the segments contain humming, percussion, or comments, but these methods combined take only 4.5% of the total time.

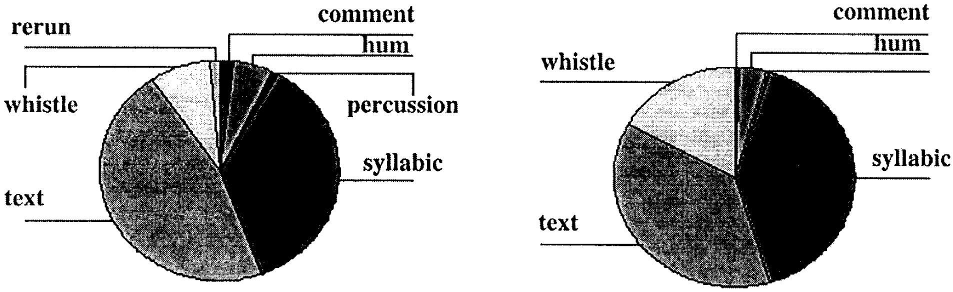

Figure 7.3. Pie charts showing the share of the different query types: textual, syllabic, whistled, hummed, percussive and comments. The left figure shows the types by count, the right figure by duration.

query method	n of segments	% of segments	total time	% of total time
text	926	45.60%	5558959	37.40%
syllabic	766	37.80%	6056644	40.80%
whistle	174	8.60%	2544864	17.10%
hum	101	5.00%	541815	3.60%
comment	42	2.10%	65108	0.40%
percussion	20	1.00%	77394	0.50%

Table 7.1. Occurrence of different query methods hierarchically ordered following the number of segments.

Within the syllabic segments, the syllables were analyzed according to their structural components: the onset (initial consonant or complex of consonants), the nucleus (vowel) and the coda (final consonant). Nucleus and coda together form the rhyme. In total 23 different onsets and 37 rhymes were found, but 99.3% of the total is organized in 11 onsets and 19 rhymes. The ten most common syllables are (in order of decreasing importance) [na], [n@], [la], [t@], [da], [di], [d@], [ta], [tu] and [ti], of which [t@], [na] and [d@] belong to the syllable repertoire of the largest number of subjects.[3]

[3] Syllables are transcribed using SAMPA (http://www.phon.ucl.ac.uk/home/sampa/home.htm), a machine-readable variant of the International Phonetic Alphabet.

7.5.2.3 PERFORMANCE STYLE

The annotation of performance style was aimed at distinguishing between melodic, rhythmic, and intermediate performances. A performance was considered to be melodic when a clear succession of different pitches, or melodic intervals, could be observed. A segment was labeled as rhythmic when no clear pitch intervals were noticed (such as in a spoken text or a percussive sequence). An intermediate category was necessary in order to classify segments where a sense of pitch was present, but without a clear melody (e.g. using a reciting tone). The distribution of these performance styles is given in Figure 7.4 and Table 7.2. Clearly, segments with melodic content are dominant, with 72.6% of the total number, and 76.7% of the total time. Intermediate and particularly rhythmic performance styles are much scarcer and in general also shorter than melodic segments.

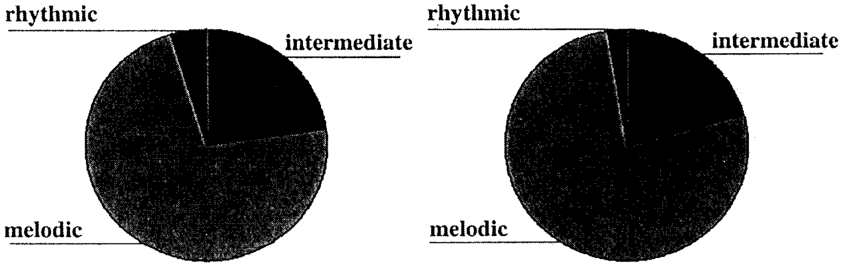

Figure 7.4. Pie charts showing the share of the performance styles: melodic, rhythmic and intermediate. The left figure shows the occurrence by counts and the right figure by length.

performance	n of segments	% of segments	total time	% of total time
melodic	1502	72.60%	11969478	76.70%
intermediate	469	22.70%	3264400	20.90%
rhythmic	98	4.70%	373249	2.40%

Table 7.2. Occurrence of performance types hierarchically ordered following the number of segments.

7.5.2.4 SIMILARITY BETWEEN SEGMENTS AND TARGETS

To each of the segments, a similarity rating was given on a six-point scale, from 0 (not recognizable) to 5 (sounds similar), focusing on melodic and rhythmic properties while, e.g., the timbre or the lyrics were neglected. The estimation of the degree of similarity between query segment and target, obviously, is subjective. Due to the subjectivity of the numbers, similarity measures are used only to compare large sets of data—to compare the efficiency of the different methods, the performance of different groups of users, and the effects of differences in memory recall.

7.5.3 Analysis of Query Methods by Subjects

The syllabic and textual query methods were the most widely spread among the subjects. Of the 71 subjects producing valid queries, 68 produced syllabic and 67 textual segments. Humming was used at least once by 39 of the subjects, whistling by 31 subjects. Twenty subjects gave comments, and 11 subjects produced percussive queries.

Looking at the distributions of the different methods for all subjects, different strategies can be distinguished. A small majority of the subjects (N = 38 or 54%) concentrated on one method (at least 60% of the query time). Eighteen subjects concentrated on text, 16 on syllabic queries, while a small group of 4 concentrated on whistling (2 produced only whistled queries). A quarter of the subjects (17), divided its query time between two methods (each taking between 30 and 60%, and together at least 80% of the time). In 15 of these 17 cases textual and syllabic queries are combined, while the combinations text/whistling and syllables/whistling each occur once. The remaining 16 participants used several methods (with three or four methods covering at least one eighth of the total query time); most common (N = 10) is a combination of textual, syllabic and whistled queries. In the other cases humming is added or replaces one of the three other methods.

Five different user profiles can thus be distinguished, four of which apply to about a quarter of the population each (text preference, syllable preference, two methods, mixed) and the fifth a small but rather distinctive group of 'whistlers.' The typical whistler is a young (mean = 24) male musician. This group produces the longest queries (mean = 19.3 s) with the highest overall similarity (mean = 4.12). They rarely switch methods within queries and if they use another method, they see the syllabic method as the only alternative. An ANOVA on the timing and segmentation aspects of the queries of the four remaining categories of subjects reveals significant effects on the mean number of segments [$F(3.63) = 1.418$, $p < 0.01$] and the mean segment length [$F(3,63) = 5.320$, $p < 0.01$]. As shown in Figure 7.5, subjects who divide their attention over two methods (average segment length = 6.60 s, average number of segments = 2.14) and subjects who use text as the dominant method (length = 5.92 s, number = 2.19) stand against subjects mixing more than two methods (length = 7.55 s, number = 1.59) and those who use syllabic queries as the dominant method (length = 9.29 s, number = 1.74).

7.5.4 Effects of Memory on the Timing and Accuracy of the Queries

A distinction was made between the queries from the first part of the experiment, based on songs indicated as "known, remembered and possible to imitate," entirely relying on long-term memory, and the queries from the second part of the experiment, made after hearing the complete song. Within the last category a distinction was made between songs indicated as "known" after listening and those indicated as "unknown."

For unknown songs, the subjects relied entirely on their short-term memory. For "known" songs they could combine with information recalled from long-term memory. Thus the experiment provided three classes of queries, each distinguished by a different use of memory:

- long-term memory (LTM) (833 queries from part 1),
- short-term memory (STM) (81 queries from part 2), and
- mixed (LTM+STM) (208 queries from part 2).

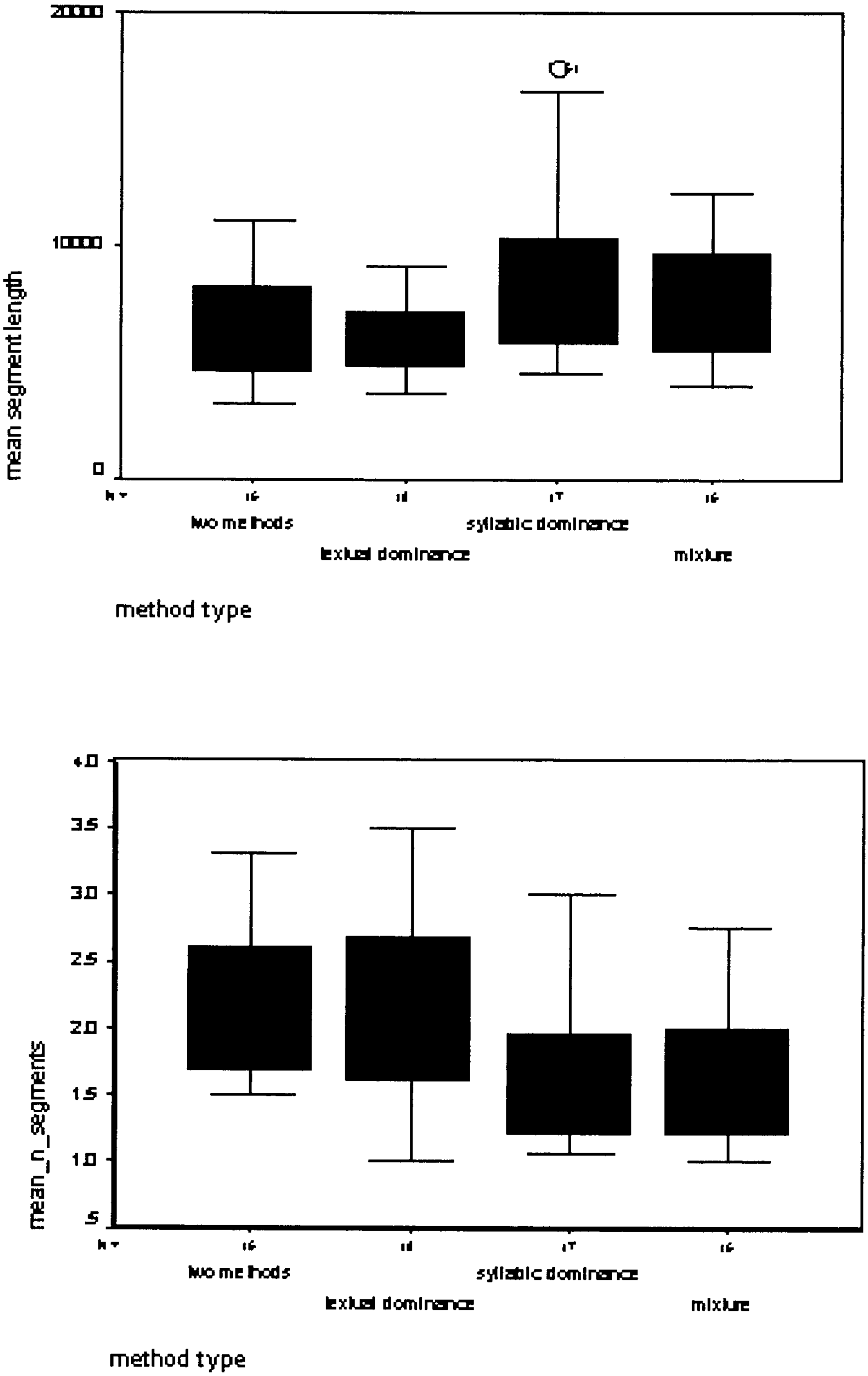

Figure 7.5. User categories according to the use of vocal query methods. Top: following the mean segment length. Bottom: following the mean number of segments.

The memory type had a significant effect on the start time and the length of queries, as well as on the similarity of the queries with the targets. There is a very clear effect on the query length [$F(2,1119) = 30.343$, $p < 0.001$], with these results:

- a mean of 13.21 s when using LTM,
- a mean of 12.78 s when using STM, but
- a mean of 17.87 s for LTM+STM (Figure 7.6).

This finding is reflected in a highly significant effect of memory category on the segment length [$F(2,2111) = 46.500$, $p < 0.001$]. The segments within the queries using LTM+STM are significantly longer (mean: 10.6 s) than those using only one type of memory (means: 6.8 s for LTM and 6.3 s for STM). Additionally, there is also a significant effect on the starting time of the queries [$F(2,1119) = 3.136$, $p < 0.05$]. When only relying on LTM (mean: 643 ms), subjects tend to start earlier than when using STM (mean: 685 ms), but when using LTM+STM they start still sooner (mean: 579 ms).

The type of memory used also has a highly significant effect on the similarity of the segments [$F(2,2111) = 10.571$, $p < 0.001$]. The distribution of the different similarity levels within the three categories (Figure 7.7) indicates only a small difference between LTM and LTM+STM but a drastic fall in similarity for the queries based on STM only.

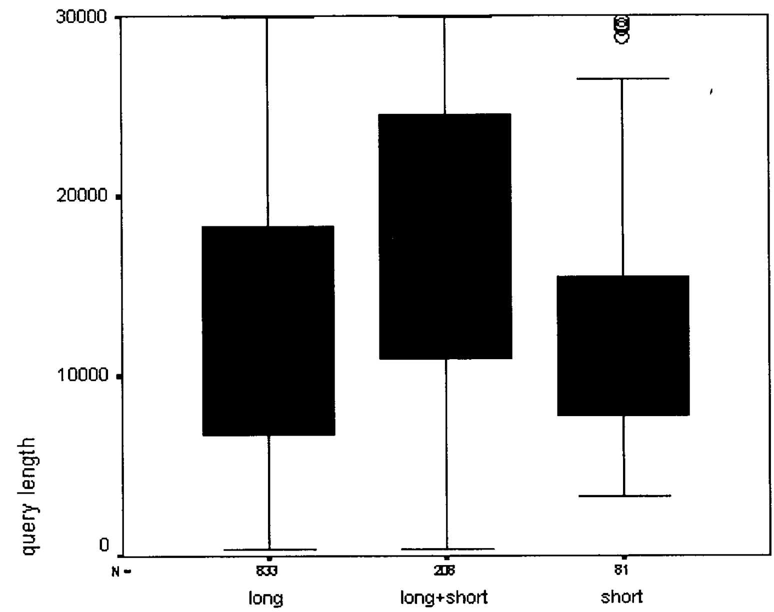

Figure 7.6. Effect of the three memory types on the query length.

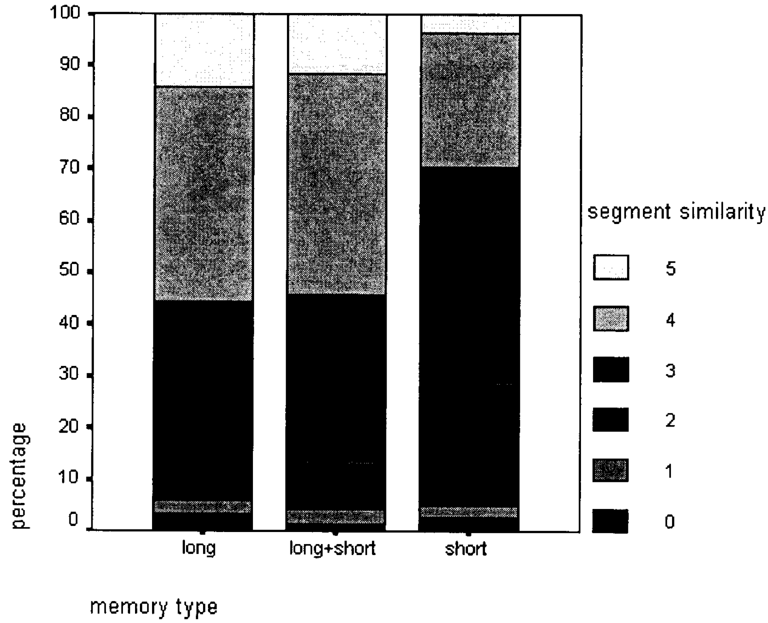

Figure 7.7. Distribution of the different similarity levels at the segment level within the three categories of memory.

Besides effects on timing and similarity, there is a clear influence of the memory type on query method and performance style. Figure 7.8 shows a change from a textual dominance to a syllabic dominance with a growing importance of short-term memory: for LTM, 48.7% of the queries are textual, taking 41.7% of the total time; for LTM+STM this becomes 39.7/33.3% and for STM 34.4/26.6%.

The importance of syllabic queries moves in the other direction: LTM 34.9/36.0%, LTM+STM 43.1/47.2%, and STM 49.1/58.3%. Parallel to this the importance of whistling decreases: LTM 8.6/18.0%, LTM+STM 9.5/15.3%, STM 4.3/8.0%. Also remarkable, finally, is the sudden increase in percussive queries when long-term memory is no longer present.

The influence of memory type on the performance style (melodic, rhythmic or intermediate) is shown in Figure 7.9. As with textual queries, the number of queries characterized by a melodic performance style decreases with a growing importance of short-term memory (LTM: 73.9/79.6%, LTM+STM: 69.0/73.7%, STM: 47.2/51.7%) in favor of intermediate (LTM: 19.1/18.2%, LTM+STM: 25.6/ 22.8%, STM: 45.5/41.9%) and, to a lesser extent, rhythmic style (LTM: 4.7/1.8%, LTM+STM: 3.7/3.2%, STM: 5.5/5.8%). The latter is only visible in the share of the total query time, which indicates that queries in a rhythmic style get relatively longer.

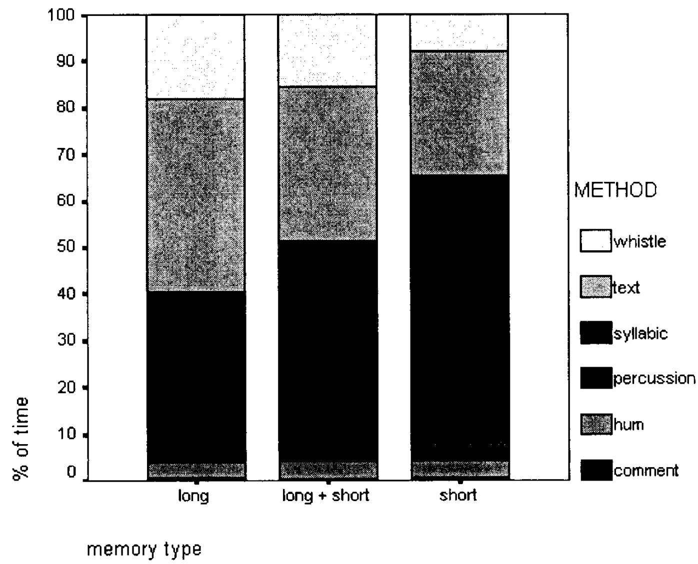

Figure 7.8. Change from textual dominance to syllabic dominance with a growing importance of short-term memory.

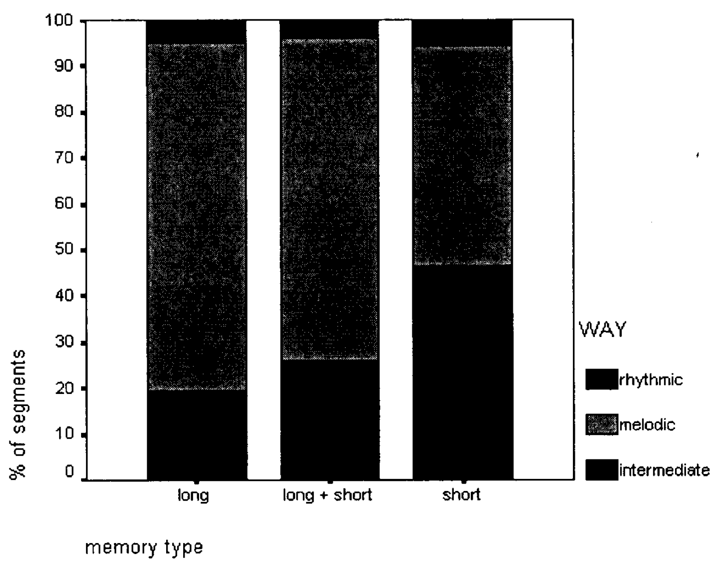

Figure 7.9. Influence of memory-type on performance style.

7.6 Discussion

Analysis of the timing characteristics of the queries shows a mean query length around 14 seconds; the start of the actual query occurs 634 ms average after the start of the recording. In both cases big inter-individual differences are found. Thus a user-friendly MIR system should be flexible in timing, expecting some people to start up to 2 seconds after the start of the recording, and expecting from a few seconds to over 30 seconds of information.

In classifying and segmenting the queries we distinguished between six methods: singing lyrics, singing syllables, whistling, humming, percussion and comments. In about 60% of the queries, subjects used only one method. In the other cases subjects used at least two methods, sometimes alternating two or more methods, but queries containing more than six segments are scarce.

The most common query methods are singing lyrics and singing syllables. Syllabic segments are in general longer than textual segments. Whistling is the third most popular method, while humming, percussion and comments occupy only a small share.

The use of certain methods is user-dependent. We could distinguish different types of users: (1) subjects that concentrate on one method (lyrics, syllables, whistling), (2) subjects that divide their query time between two methods (mostly a combination of textual and syllabic), and (3) subjects that mix several methods. These findings indicate that the ideal MIR system should be able to cope with changes in method. On the other hand, one could give users the choice of one specific method: text, syllables, or whistling.

Concentrating on one single method does not seem so attractive since some classes of users largely rely on one single method. This is most clearly illustrated by whistling. Less than half of the people used whistling as a query method, but a small group concentrated on whistling and moreover provided long, high-quality queries. Thus allowing only whistling (e.g. Prechelt and Typke 2001) does not seem to be a good choice, but excluding whistling would eliminate an interesting body of queries.

Around 75% of the query collection is performed in a melodic way, which supports the idea that melodic content is to be regarded as a salient feature of vocal queries (e.g. Chai 2001).

Since the use of syllables is one of the most important query methods, and most existing MIR systems require syllabic input, we further investigated the nature and structure of the syllables occurring in spontaneous queries. If a system requires specific syllables for ease of processing, it is desirable to choose some of the most common types. Some existing systems, however, ask the users to use syllables that hardly ever occur in spontaneous queries, e.g. [fa] (Pauws 2002). These syllables might be less comfortable for the users and thus might yield a larger number of errors, or diminish the attractiveness of the system for the public. Our finding that the most popular syllables end on a vowel and have no coda is in agreement with the syllabic constraints found in the experimental literature.

7.7 Conclusion

Although a large variety of vocal querying strategies was observed, we could establish some basic characteristics and distinguish some categories of users. These findings already generate some guidelines for the development of user-friendly MIR systems. The application of these guidelines would enhance the interactive process and open the way to make MIR systems more efficient and more attractive.

Acknowledgements

The research presented here has been carried out in the context of the MAMI (Musical Audio Mining) project (http://www.ipem.ugent.be/MAMI) at the Musicology Department at Ghent University. The aim of the project is to build a music-retrieval system using audio-mining techniques (Leman 2002, Leman et al 2002). The project is funded by the Flemish Institute for the Promotion of Scientific and Technical Research in Industry. It addresses research performed by an interdisciplinary consortium that is grounded in diverse fields such as signal processing, musicology and applied mathematics.

The authors gratefully acknowledge Koen Tanghe and Gaëtan Martens, members of the MAMI team, for the implementation of the application used for the experiment. We also wish to thank Jelle Dierickx and Liesbeth De Voogdt for their assistance in the annotation of the queries.

Appendix 1. Songs used in the experiments.

SONG TITLE	PART I			PART II		TOTAL
	recording			recording		
	1	2	3	1	2	
Blowin' in the wind	47	2	4	7	2	62
All I really want				23	1	24
Mooi, 't leven is mooi	68	12	7			87
De Marie-Louise	61	9	6	3		79
How do you do				19	1	20
Paloma Blanca				11	3	14
YMCA	67	10	4	2		83
When a man loves a woman	57	7	3	3		70
Don't worry, be happy	65	9	6	1		81
Walk on the wild side	44	7	1	6	1	59
Waterloo	60	3	3	5		71
Sunday, bloody Sunday	48	8	5	7	1	69
It's raining men	60	5	5	1		71
Without you				17	3	20
Ad mortem festinamus				22	2	24
My way	40	7	1	12	3	63
Rosa	17	2	2	16	4	41
Smells like teen spirit	10	1	1	20	3	35
Only happy when it rains	6	1		19	2	28
Temple of love	6		1	16	3	26
Highway to hell				13		13
Don't cry for me, Argentina	27	2	1	4	3	37
Klein klein kleuterke	17	1	2	6	2	28
'k zag twee beren	10	1		2		13

SONG TITLE	PART I			Part II		TOTAL
	recording			recording		
	1	2	3	1	2	
O Fortuna	4			11	1	16
Für Elise	2			12		14
TOTAL	716	87	52	258	35	1148

References

Chai, Wei (2001). "Melody Retrieval on the Web." M. S. Thesis, MIT Media Lab.

Ghias, Asif, Jonathan Logan, David Chamberlin, and Brian Smith (1995). "Query by Humming: Musical Information Retrieval in an Audio Database," *Proceedings of ACM Multimedia 1995*, pp. 213–236.

Leman, Marc (2002). "Musical Audio Mining," in *Dealing with the Data Flood: Mining Data, Text and Multimedia*, ed. J. Meij (Rotterdam: STT Netherlands Study Center for Technology Trends), chapter 5.5.2.

Leman Marc, Lieven P. Clarisse, Bernard De Baets, Hans De Meyer, Micheline Lesaffre, Gaëtan Martens, Jean-Pierre Martens, and Dirk Van Steelant (2002). "Tendencies, Perspectives, and Opportunities of Musical Audio-Mining," *Proceedings of Forum Acusticum Sevilla 2002*, special issue of Journal Revista de Acústica Vol. XXXIII, no. 3-4, abstract p. 79, full text: CD-ROM.

Levitin, Daniel J. (1994). "Absolute Memory for Musical Pitch: Evidence from the Production of Learned Melodies" in Perception & Psychophysics, 56, pp. 414–423.

McNab, Rodger J., Lloyd A. Smith, David Bainbridge and Ian H. Witten (1997). "The New Zealand Digital Library MELody inDEX," D-Lib Magazine, pp. 11–18.

Pauws, Steffen (2002). "CubyHum: A Fully Operational Query by Humming System" in *Proceedings of the Third International Conference on Music Information Retrieval*, ed. Michael Fingerhut (Paris: IRCAM–Centre Pompidou), pp. 187–196.

Prechelt, Lutz, and Rainer Typke (2001). "An Interface for Melody Input" in ACM Transactions on Computer-Human Interaction, pp. 133–149.