USER BEHAVIOR IN THE SPONTANEOUS REPRODUCTION OF MUSICAL PIECES BY VOCAL QUERY

Micheline Lesaffre[1], Dirk Moelants[1], Marc Leman[1], Bernard De Baets[3], Hans De Meyer[4], Gaëtan Martens[4] and Jean- Pierre Martens[2]

[1] Department of Musicology (IPEM), Ghent University, Belgium
[2] Department of Electronics and Information Systems (ELIS), Ghent University, Belgium
[3] Department of Applied mathematics, Biometrics and Process Control, Ghent University, Belgium
[4] Department of Applied Mathematics and Computer Science, Ghent University, Belgium

ABSTRACT

Background: This experiment is part of a broader research project in the field of Musical Information Retrieval. In order to realize a user-friendly system for searching musical pieces by vocal query, the behavior of subjects asked to imitate well-known songs from long-term memory and unfamiliar songs after a single hearing was investigated.

Aims: Our aim is to analyze the characteristics of the behavior of people who reproduce a piece of music from memory in an intuitive way. This should lead to a view of preferences for certain methods of vocal query.

Method: 72 subjects participated in an experiment in which they were asked to reproduce pieces of music in front of a microphone. No further restrictions were given. In the first part of the experiment subjects responded to titles of pieces they previously indicated as familiar. In the second part entire pieces of music, indicated as unfamiliar, were aurally presented before asking reproduction.

Results: In general, participants asked to reproduce music prefer a melodic use of the text or of specific syllables. Significant effects of gender and musical background were found as well as differences between the reproduction of unfamiliar melodies and the recall of known melodies. Clear relations between user behavior and musical content were found.

Conclusions: User preferences and general characteristics of vocal queries aimed at searching specific pieces in a music database are established. The findings generate some guidelines for the development of user-friendly systems for musical information retrieval based on vocal queries.

1. INTRODUCTION

Performing a vocal query from memory is an experience that is a direct result of memorization. The user of a Music Information Retrieval system may not correctly recall a theme, or may not have the vocal skills to produce a good imitation of a tune. In many experiments people's recognition of well-known melodies has been investigated. However, we must not only know how people recognize melodies. More investigation is needed that concentrates on the characteristics of naturally expressed vocal queries. By vocal query we simply mean a query produced by the voice and the vocal organs (i.e. vocal chords, tongue, lips, teeth).

2. EXPERIMENT DESCRIPTION

Our query-by-voice (QBV) experiment is designed such as to give subjects a maximal freedom of expression. After going through a preparatory stage in which their familiarity with music was checked, subjects were instructed to vocally imitate a part of a piece of music using any vocal query method, e.g. humming, singing lyrics, singing syllables or whistling and any combinations of these. Maximal freedom of spontaneous behavior was maintained by allowing queries to refer to any fragment part of the original music. The participants were also allowed to express any voice, i.e. any melody, bass line or any particular instrument.

72 subjects, students and staff members from Ghent University, participated in the experiment on a voluntary basis. The age ranged from 19 to 56, with an average of 27.7 years. 28 (39%) subjects were female and 44 (61%) male; 37 (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).

The experiment is based on a selection of thirty pieces [Musical_Stimuli.png] of music from the MAMI project target test collection (http://www.ipem.rug.ac.be/MAMI/). Permission to use this collection for the purpose of the research has been given by SABAM, the Belgian author rights association. The list contains different genres: besides popular music songs, ranging from chanson to heavy metal, well-known Flemish children songs and classical music.

The experimenter gave the subjects an outline of the general procedure and handed them a text with a detailed description. Before starting the actual experiment, subjects were asked to fill out a questionnaire, gathering information about age, gender and musical training. 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 would be able to imitate it, or would not be able to recall the musical content. 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 (again by sound recording or by typing text), or to propose alternative query methods. In

ISBN 3-931852-67-9
ISSN 1617-6847

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 are used. The subjects were first requested to listen to the entire piece. They were then asked whether they knew the piece 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.

3. ANALYSIS STRATEGY

A team of musicologists annotated the 1148 vocal queries resulting from the experiment. The features investigated are related to the beginning and ending time of the query, the vocal query method, performance, target similarity and syllabic structure.

At a first level of analysis, the queries where segmented according to the methods used. Vocal queries may contain a mix of different query methods such as humming, singing syllables, singing lyrics and whistling. In addition to those methods we find percussion (such as tapping along with the drum) and comments (spoken comments made by the subjects while performing a query). Studying these methods in more detail requires segmentation into homogeneous parts, and this resulted in a set of 2114 segments.

4. RESULTS

4.1. General aspects of the queries

The average starting time of a query occurred 634 ms after the subjects started the recording, with 99.3% starting within 2 seconds. The mean query 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 roughly between 5 and 15 seconds.

About 60% (683) 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 the other and back (for example, lyrics, whistling, and back to lyrics…). The maximum number of segments observed in one single query is 12, but 97.8% of the queries contains a maximum of six segments.

4.2. Segment specific aspects

Analogous to the analysis of the queries as a whole, the timing of the segments was investigated. A similar asymmetric distribution as for the duration of the full queries is shown but the average length is of course shorter, with a mean of 7.42 seconds. There is a sharp peak around two seconds, 50% of the segments are shorter than 8 seconds and 75% is shorter than 15 seconds.

The query method was used as a criterion for segmentation. The two methods standing out 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 in the total time, the syllabic method is more prominent, which indicates 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 the double of the average). Thus, although only 8.6% of the segments are whistled, they take 17.1% of the total time. Opposite, 8.1% of the segments contain humming, percussion, or comments, but these methods together take only 4.5% of the total time.

The annotation of performance style aims at distinguishing between melodic, rhythmic and intermediate performances. A performance style was considered to be melodic when a clear succession of different pitches, or melodic intervals, could be observed. A segment was annotated as a rhythmic segment when no clear pitch intervals were noticed (such as in a spoken text, a percussive sequence). An intermediate category was necessary in order to classify segments where a sense of pitch is present, but without a clear melody. Most often this annotation was applied to queries that use a kind of a reciting tone. Segments with a 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 segments characterized by one of these styles are in general also shorter than melodic segments.

To each of the segments, a relative similarity rating was given on a six-point scale, ranging from 0 (not recognizable) to 5 (sounds similar). Similarity annotation was focused on melodic and rhythmic properties and aspects of timbre or use of lyrics was neglected. Due to the fact that queries may contain more than one segment, an overall similarity rating for the queries, had to be defined. Due to the subjectivity of the numbers similarity measures are only used 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.

Syllabic queries form an important part: 766 segments out of 2114 had a syllabic content. These contain a total number of 14748 syllabic units, and additionally some 104 events such as tr, pf, and tongue clicks that were not considered as syllables. Analysis yields 179 different syllables, of which 45 occurred only once. Syllables were analyzed according to their structural components: the onset (initial consonant or complex of consonants), the nucleus (vowel) and the coda (final consonant). In total 23 different onsets and 37 rhymes were found. The most commonly used syllables are [na], [n@], [la], [t@] and [da], together they make up 52% of the total number of syllables within the results. Four onsets [n, d, t & l] initiate 86.9% of the syllables, as for the rhymes only two elements [a] and [@] stand out, together ending 67.7% of the syllables.

4.3. User behavior

Subjective aspects related to age, musicianship and gender are addressed. Analysis shows a large variance in the timing and segmentation behavior of the different subjects. The subject's average query length is between 4.49 and 26.84 seconds, with a mean of 13.94 seconds. The average start of the queries varies between 155 and 1510 ms, with a mean of 642.6 ms. The average number of segments within the queries lies between 1 and 3.5

ISBN 3-931852-67-9
ISSN 1617-6847

(mean =1.87 segments), which gives an average segment length varying between 3.33 and 23.54 seconds, with a mean of 7.98 seconds.

The syllabic and textual query methods are the most widely spread methods among subjects. Of the 71 subjects producing valid queries, 68 produce syllabic and 67 textual segments. Humming is used at least once by 39 of the subjects, whistling by 31 subjects. 20 subjects give comments, and 11 subjects produce percussive queries.

Looking at the distributions of the different methods for all subjects, different strategies are distinguished. A small majority of the subjects (54%) concentrates on one method (at least 60% of the query time). 18 subjects concentrate on text, 16 on syllabic queries, while a small group of 4 concentrates on whistling (2 of them only produce whistled queries). A quarter of the subjects (17), divides its query time between two methods (each taking between 30 and 60 %, and together at least 80% of the total query time of the subjects). In 15 of these 17 cases textual and syllabic queries are combined, the combinations of text with whistling and syllables with whistling each occur once. The remaining 16 participants use several methods (with three or four methods covering at least one eighth of the total query time), most common (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. From these results five user profiles are distinguished [Vocal_Query_Methods.png], four of which apply to about a quarter of the population and a small but rather distinctive group of 'whistlers'. The typical whistler is a young, male musician. This group produces the longest queries with the highest overall similarity. They hardly switch method within queries and if they use another method, they see the syllabic method as the only alternative.

Analyses showed a significant negative correlation between age and similarity and a significant positive correlation between age and the average starting time of the query. Correlation between age and syllable parts showed relations with the relative use of nuclei [a] and [@]and with onset [l]. As for the methods used, an increase of comment with age can be observed.

Some significant effects of gender on the choice of the syllables are revealed. For the onset, men tend to use [t] significantly more often than women do, with a 24.9% average against a 11.9% average. For the nuclei, a significant effect of gender on the use of the [a] is found, being a very dominant vowel in women's syllabic queries (56.1% average, against 37.5% for men). Men also tend to use a larger variety of syllables, an average of 20.19 different syllables, against 14.15 for women. Finally, women, tend to start their query later than men do (after 716 ms against 595 ms.

Also the distinction between musicians and non-musicians yields some significant effects, mainly on the methods used. Within the query method of musicians, text occupies a significantly smaller position, with an average of 35.3% of the time spent using text against 46.2% for non-musicians. This is compensated by a larger share of syllabic and whistled time in the musicians' queries (though the effects of each of these separately do not reach significance). Finally there is also an effect of musicianship on the average length of the segments: musicians make longer segments than non-musicians (means 9.18 and 7.12 seconds, this is a result of longer queries and less segmentation by musicians.

4.4. Effects of memory

For this analysis a distinction is made between the queries from the first part of the experiment, based on songs indicated as 'known, remembered and possible to imitate', but 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 is made between songs indicated as 'known' after listening and those indicated as 'unknown'. For unknown songs, the subjects rely entirely on their short-term memory. For 'known' they may combine queries with recalled information from long-term memory. Thus the experiment provides three classes of queries, distinguished by a different use of memory. There is the use of long-term memory (LTM) (833 queries from part 1), short-term memory (STM) (81 queries from part 2 based on songs indicated as unknown after listening), and mixed (LTM+STM) (208 queries from part 2 based on songs indicated as known).

The memory type has 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 with a mean of 13.21 s when using LTM, of 12.78 s when using STM, but of 17.87 s for LTM+STM. This finding is reflected in a highly significant effect of memory category on the segment length. 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. 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 still start sooner (mean: 579 ms). The type of memory used also has a highly significant effect on the similarity, both at the level of the query and at the level of the segment. The distribution of the different similarity levels at the segment level within the three categories indicates only a very small difference between LTM and LTM+STM but a drastic fall in similarity for the queries based on STM only.

Besides effects on timing and similarity, there is a clear influence of the memory type on query method and performance style. There is 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%, STM 49.1/58.3%. Parallel also the importance of whistling decreases: LTM 8.6/18.0 %, LTM+STM 9.5/15.3%, STM 4.3/8.0%. Remarkable is also the sudden increase in percussive queries when long-term memory is no longer present.

The amount of queries characterized by a melodic performance style diminishes with an increasing importance of short-term memory (LTM: 73.9/79.6%, LTM+STM: 69.0/73.7%, STM: 47.2/51.7%), this 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 in the total query time, which indicates that queries in a rhythmic style get relatively longer.

ISBN 3-931852-67-9
ISSN 1617-6847

5. DISCUSSION

Analysis of the timing characteristics such as query length and starting time of the queries shows that large differences occur between subjects. 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 only a few seconds of information in some cases and 30 seconds (or probably even more) in other.

In classifying and segmenting the queries, distinction is made 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. In general, the most common query methods are singing lyrics and singing syllables. Syllabic segments are in general longer and therefore more prominent than textual segments. Whistling is the third most popular method, while actual humming, percussion and comments occupy only a small share of the whole. Just like the timing characteristics, the use of certain methods is user dependent. We could distinguish different types of users: subjects that concentrate on one method (lyrics, syllables, whistling), subjects that divide their query time between two methods (mostly a combination of textual and syllabic) and subjects that mix several methods. Probably the choice for certain methods also seems to depend on the familiarity of the subjects with the language of the target song: the less familiar people are with the language, the more they tend to use syllables and whistle instead of using text. 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 for 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 concentrates on whistling and moreover provides 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 exclude 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 major 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, the nature and structure of the syllables occurring in spontaneous queries is investigated in further research.

Significant effects of age, gender and musical background were found as well as differences between the reproduction of unfamiliar melodies (STM) and the recall of known melodies (LTM). Younger people tend to start their queries sooner and have a better similarity score. The use comment as a query method increases with age. Men start their queries later than women and use a larger variety of syllables. Musicians make longer queries than non-musicians and use less text in favor of syllables and whistling. The timing of the queries as well as their similarity with the target is dependent on the type of memory used by the subjects. When known songs are 'refreshed' by presenting them aurally, the queries start sooner and last longer. The reproduction of unfamiliar melodies from STM has less quality than recall of known melodies even if the query only relies on LTM. The lesser degree of acquaintance is also reflected in the larger share of syllabic segments and the increased importance of rhythmic and intermediate performances. The present findings support the ability that people are quite good in reproducing tones of familiar songs from long-term memory (Levitin, 1994). Thus far, however, very few researchers have investigated the effect of long-term memory. Most experiments within the field of MIR research were set up in such a way that the participants practiced singing (e.g. McNab, 1997) or were requested to repeat a note sequence that was aurally presented (e.g. Lindsay, 1996).

6. ACKNOWLEDGMENTS

The authors gratefully acknowledge Koen Tanghe, member of the MAMI-team, for the implementation of the application used for conducting the experiment. We also wish to thank musicologists Jelle Dierickx and Liesbeth De Voogdt for their assistance in the annotation of the query corpus. This research and the experiment were conducted in the framework of the MAMI project, which is funded by the Flemish Institute for the Promotion of Scientific and Technical Research in Industry.

7. REFERENCES

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

2. Levitin, Daniel J. "Absolute memory for musical pitch: Evidence from the production of learned melodies" in *Perception & Psychophysics, 56*, (1994), pp. 414-423.

3. Lindsay, Adam. "Using contour as a mid-level representation of melody". M.S. Thesis, MIT Media Lab, 1996.

4. McNab, Rodger J., Lloyd A. Smith, Ian H. Witten, Claire Henderson, and Sally Jo Cunningham. "Towards the digital music libraries: tune retrieval from acoustic input", *Proceedings of Digital Libraries ACM* (1996), pp. 11-18.

5. Prechelt, Lutz, and Rainer Typke. "An interface for melody input" in *ACM Transactions on Computer-Human Interaction* (2001), pp. 133-149

ISBN 3-931852-67-9
ISSN 1617-6847