University of Tampere

Department of Information Studies

Research Notes

RN • 1998 • 1

PEKKA SALOSAARI & KALERVO JÄRVELIN

MUSIR
A RETRIEVAL MODEL FOR MUSIC

Tampereen yliopisto • Informaatiotutkimuksen laitos • Tiedotteita
1998 • 1

MUSIR — A Retrieval Model for Music

Pekka Salosaari & Kalervo Järvelin
Department of Information Studies
University of Tampere, Finland

salosaar@hkkk.fi, likaja@uta.fi

Contents:

MUSIR — A Retrieval Model for Music

Pekka Salosaari & Kalervo Järvelin
Department of Information Studies
University of Tampere, Finland

salosaar@hkkk.fi, likaja@uta.fi

Abstract

Traditionally, information retrieval in music has been based on surrogates of music, i.e., bibliographic descriptions of music documents. This does not provide access to the essence of music, whether it is defined as the musical idea represented in the score, the gestures of the performer playing an instrument or the resulting auditive phenomenon - the sound. In this paper we develop a retrieval model for music content. We develop representations for music content and music queries, a matching method for the representations and show that the model has desirable properties for the retrieval of music content. Our model captures representative and memorable features of music in a simple representation, supports inexact retrieval, and ranks retrieved music documents. The MUSIR retrieval model is based on filtering the MIDI representation of music and n-gram matching.

1. Introduction

Music documents cover books about music, printed scores and recordings of music performances (CD discs) as well as electronic representations of music such as files created by composition software (e.g., MIDI representation; Loy, 1985). Traditionally, information retrieval in music has been based on surrogates of music, i.e., bibliographic descriptions of music documents. In such approaches, the music "content" is represented in terms of classes or keywords of specially designed documentation languages for music, e.g., the music classification within the UDC. Modern text retrieval methods can be used for the retrieval of textual music documents. However, neither approach provides access to the essence of music, whether it is defined as the musical idea of the composer represented in the score, the gestures of the performer playing an instrument or the resulting auditive phenomenon — the sound.

We believe that a retrieval mechanism for music content is needed for several reasons. The bibliographic content description is not always available nor sufficient for proper retrieval (McLane, 1996). Consumers or producers of music may have a tune in mind for which they

do not know the composer or other textual attributes. Moreover, MIDI and other digital representations of music have become a common means for storage and transfer of music in the Internet which nowadays provides many music archives in MIDI form. Modern music composition and production procedure often means using computers for collecting elements from several files into a collage, which may cause problems with data management if the amount of elements used increases radically. Finally, the use of musical ideas may be tracked by content retrieval methods. This may be valuable for the scholarly analysis of music and for supervising copyrights.

The audio content of music cannot be accessed in the sense texts can be accessed through the words they contain. This is because music does not refer directly to anything we can easily describe in natural language. Thus, music retrieval could be even more difficult than image retrieval — although pattern matching within images still is difficult, image content can often be verbalized fairly consistently for text-based retrieval. However, music has one advantage, the symbol system provided by common music notation (CMN). Music contains elements like melody, harmony and rhythm which may be formally represented and manipulated while some other features like tension, expectation and feeling are not formally representable (Dannenberg, 1993).

Although the problem of content-based retrieval of music has not been addressed a lot in IR research, various studies and projetcs in other fields of research exist, mainly in the areas of computing and musicology. Bakhmutova, Gusev & Titkova (1997) present string matching functions for melody retrieval. The MuseData-system, created for musicological analysis, supports searching of melodic patterns in a text-based environment (Selfridge-Field, 1994). There is also an operational system for melody retrieval, MELDEX, which is accessible via WWW (McNab *et. al* (1997). The system includes a database of 9,400 folk tunes and retrieval interface for acoustic input. Lemström, Haapaniemi & Ukkonen (1998) present a coding scheme of music which is invariant under different keys and tempos, and investigate the application of two approximate string matching algorithms to retrieve music.

In this paper we develop a retrieval model for music content. A retrieval model specifies the representations of documents and information needs, and how they are compared (Turtle & Croft, 1992). In the present case we develop (a) representations for music content and music queries, (b) a matching method for the representations and (c) show that the model has desirable properties for the retrieval of music content. Works of music may differ greatly and be very complex. Multiple representations like scores, audio presentations and spectral images of sound exist (Dannenberg, 1993, McLane, 1996). We therefore pose the following requirements on the retrieval model for music content:

- it must capture representative, usable and memorable features of music from the viewpoint of an inquirer — it must allow queries that are music, not just about music
- it must be a simple enough model for retrieval of rich music documents — it should hide the richness of scores and interpretation
- it must be based on an easily available digital representation of music
- it must support inexact retrieval — queries which are not "correct" with respect to the desired music document(s) — because inquirers cannot be expected to provide "correct" queries, e.g. as regards note pitch, note length, or their sequence in a melody
- it must rank the matched documents according to decreasing similarity.

We use the MIDI representation as a starting point for the retrieval model. MIDI is a widely applied standard for music representation and can easily be manipulated for the purposes of retrieval. However, the data structure of a MIDI-file is both too rich and compressed for retrieval purposes and does not support inexact retrieval and ranking. Therefore we shall present how a MIDI file may be filtered to a simpler form that supports matching. Our matching method is based on n-grams (Ashford & Willett, 1988). We will demonstrate how various n-gram representations can be filtered from the MIDI representation and what their retrieval effects are. We shall use J.S. Bach's Fugue VII as our sample of music and its parts as documents to be matched.

We shall focus on melody patterns as the basis for retrieval, because melodies are most easily recognized and remembered and often internally played in people's minds. A melody is a sequence of notes with varying pitch and duration (Kontunen, 1991a).

2. Music Representation and Matching

2.1. Music Representation Approaches

Computers have been used for many music-related purposes. Consequently, there are many music representation schemes which support computer processing but not necessarily music retrieval — for comprehensive reading, see Beyond MIDI: The Handbook of Musical Codes (1997). The representation schemes are used, roughly, for three purposes: recording, analysis and generation/composition. Wiggins *et al.,* (1993) used a two-dimensional matrix to analyze music representations: Structural generality refers to the means of representing and manipulating high level structures in music. Expressive completeness refers to the means of representing in detail and accurately the audio content of music. Honing (1992) considers temporal, structural, declarative and procedural representations of music.

The score (music notation) is a very well-known representation of music. It provides high structural generality. However, it is not an explicit representation and may be interpreted in various ways by an interpreter. The same music event may also be represented in varying ways through notes. For instance, the beams and stems of notes can be applied in various ways to represent similar musical events. Moreover, not all users of music are competent of applying the music notation.

The acoustic phenomenon of music can be represented through sound spectrograms which supply detailed data on note pitch, length, timbre and velocity. While the expressive completeness of spectrograms is rich, they are very weak in structural generality.

Several computer representations have been developed for the purposes of research projects in musicology. The DARMS representation (Hewlett & Selfridge-Field, 1991) was originally developed for producing music notation but has been used in many other projects in musicology. A query language for DARMS has been developed but to our knowledge there is no retrieval system. The MUSTRAN (McLane, 1996) representation was the first designed to facilitate the transcription of music performances. The Standard Music Description Language (SMDL; ISO, 1995) is an on-going effort by the American National Standard Institute for a generic and structural representation of music in various forms. It is based on the HyTime representation. At the moment there are no music databases nor systems based on the SMDL.

The MIDI representation (Musical Instrument Digital Interface; Loy, 1985) is a representation intended for use between music instruments. The MIDI specification defines both the physical connections between the devices of the system and software protocol for sending and receiving performance related messages. A MIDI-system constitutes of sound producing synthesizers or samplers, control devices like keyboards and MIDI-instruments, and software running on computers.

Basic elements in the representation are events like Note On and Note Off which provide data on note pitch, velocity and channel (each of 16 channels may correspond to a single instrument or a group of instruments). MIDI representations can be stored in standard MIDI file - form and manipulated by sequencer programs in computers. MIDI-data can be manipulated in various ways: for example, works can easily be transposed to another key or the event times and durations of events can be quantized. The latter means synchronization of events which may have been performed (temporally) inexactly. This feature is important for retrieval purposes since unquantized representations have shown to be difficult for searching (Selfridge-Field, 1994).

2.2. The MUSIR Retrieval Model

The MUSIR retrieval model is described in Figure 1. We assume that for works of music a MIDI representation can be created. This representation is then quantized to reduce the temporal deviation of event times. The quantization can be processed on arbitrary level of granularity, it only has to remain consistent throughout the collection. In the second step, the MIDI representation is filtered so that data on (a) the relative event pitch sequences and (b) relative event time interval sequences are provided. The former are computed from consecutive MIDI note numbers, e.g., <70, 67, 65, 67, 63, 68> representing the pitch of the notes 1 - 6. The relative pitch sequences for each notes 2 - 6 are derived by subtracting from each note number the note number of the preceding note. The latter are computed similarly from the event occurrence times in the MIDI file using the same subtracting procedure to represent the rhythmic pattern of music. These sequences are then transformed into an n-gram representation. We shall consider in this paper di-, tri- and tetra-grams as shown below. A music database is thus a database of records, which represent the pitch pattern and rhythmic pattern of events numerically through n-grams.

The process is the same on the retrieval side. We assume that the retrieval system user plays the request with a MIDI instrument (e.g., a keyboard or other kind of controller connected to a MIDI-system). Alternatively, the request may be derived from any MIDI file containing a monophonic sequence of note events. The MIDI representation of the request is thus created and captured and finally transformed into an n-gram representation. A representation based on relative pitch and/or time intervals is supported by musicology because intervals are meaningful syntactic unit in music which function on several abstraction levels of music (phonological, grammatical, lexical, discourse; Stefani, 1985).

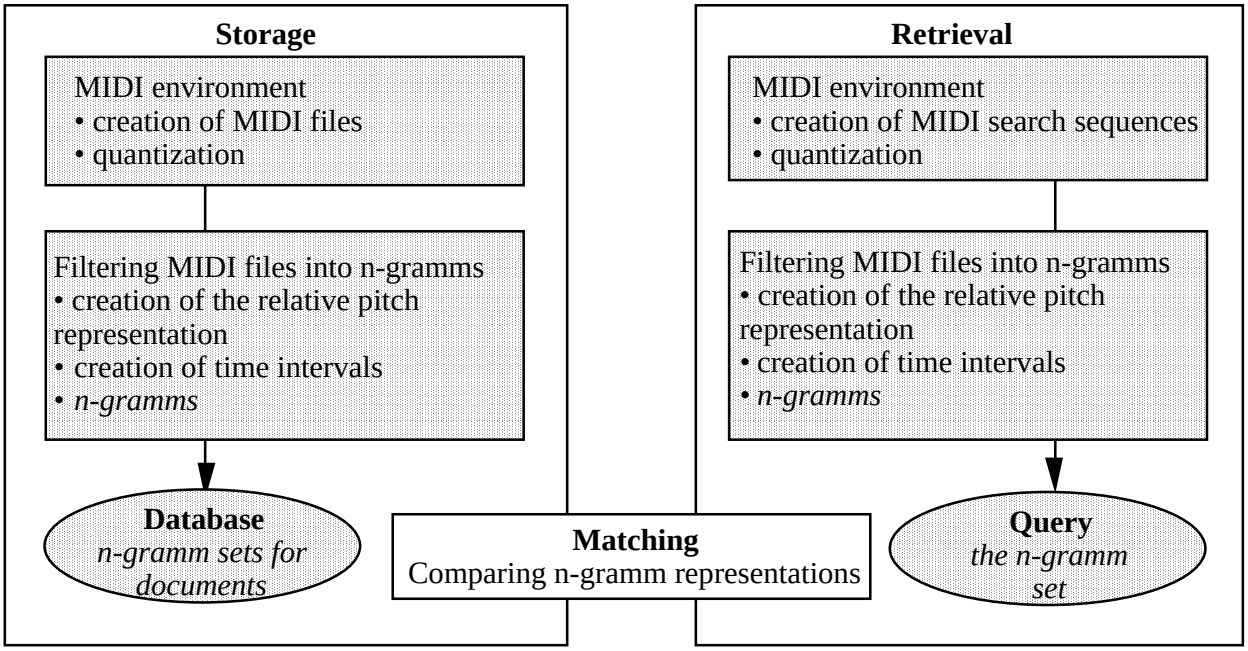

Figure 1. The MUSIR retrieval model

N-gram matching was done in the experiments reported later below by the simple formula:

$$match(G_Q, G_D) = | G_Q \cap G_D | / | G_Q |$$

where G_Q, G_D are the query and document n-gram sets, respectively.

In other words, the number of shared n-grams between the query and the document is compared to the number of n-grams in the query. The resulting score is a real number in the range [0, 1] which can be used for document ranking.

2.3. Construction of n-grams in MUSIR

Table 1 shows how n-grams are derived from a sequence of MIDI note numbers <70, 67, 65, 67, 63, 68>. Line two in the table gives the pitch interval between two consecutive note numbers and lines three and four the di-grams and tri-grams, respectively. The tetra-grams are derived in the same way. In the n-grams, all intervals are represented by two digits, possibly preceded by a minus sign and a vertical bar separating the components.

When the event time intervals were used with pitch di-grams, the time intervals were represented (by four digits) with the corresponding pitch intervals. Thus, a representation of pitch and time interval as a di-gram looks like -03 / 0060 | -02 / 0060. In this example the time base of the sequence is 240 ticks per quarter note and thus the time interval value of 0060 indicates the duration of a 1/16 beat.

MIDI note number	70 67 65 67 63 68
Interval	-3 -2 2 -4 5
di-gramms	-03 | -02 02 | -04 -02 | 02 -04 | 05
tri-gramms	-03 | -02 | 02 -02 | 02 | -04 02 | -04 | 05

Table 1. Computation of n-grams in MUSIR

Our n-gram representations are invariant with respect to key. One melody played in different keys has a single representation. By using n-grams, variations in the melody pattern (some "wrong" notes) do not prevent a document from being found. Both features facilitate fuzzy retrieval.

Table 2 gives the representation/matching methods tested for MUSIR development. In this paper we shall present test results for n-gram matching. Exact matching results are presented by Salosaari (1998).

	exact matching Methods		n-gram representation/matching methods			
			di-grams		Tri-grams	tetra-grams
Method features	pitch	time	Pitch + time	Pitch	Pitch	Pitch
Method symbol	*p*	*t*	*2pt*	*2p*	*3p*	*4p*

Table 2. The tested matching methods for MUSIR

3. Sample Case: Bach's Fugue VII

Fugue is a polyphonic musical form which was particularly popular in baroque. In a fugue, there are usually three to five voices or parts identified according to the human voices soprano, alto, tenor and bass. Fugues are based on a clear musical theme, called the subject, and its variations. The monophonic subject is introduced in the beginning and is then repeated and varied in different keys and patterns. Structurally, a fugue can be divided into sections, usually two to six. A section is a presenting musical coherence: key. In the first section, called the subject is introduced in all voices, in the middle parts it is modulated in different keys, and in

the end it is reiterated. By using segments of the subject as queries we can demonstrate how the variations of this theme in different keys and forms occurring later in the composition can be matched.

We shall use the Fugue VII by J.S. Bach (Bach, 1975; Fugue, 1995) as a retrieval example. Fugue VII has 37 bars and nine theme occurrences in three voices which we call soprano, alto and bass. Figure 2 shows the incipits of the themes in the fugue. For the sake of comparability they are notated in the same g-clave. Consider the relationship of theme 1 with the other themes:

- theme 3 is structurally identical but on a different pitch level
- themes 2, 4, 7, 8 and 9 are variations both with respect to individual pitches and/or rhythmic pattern
- the tonal structure of themes 5 and 6 differ from theme 1 in the sense that they are based on different scales; while the subject is introduced in major key, themes 5 and 6 are based on the minor scale.

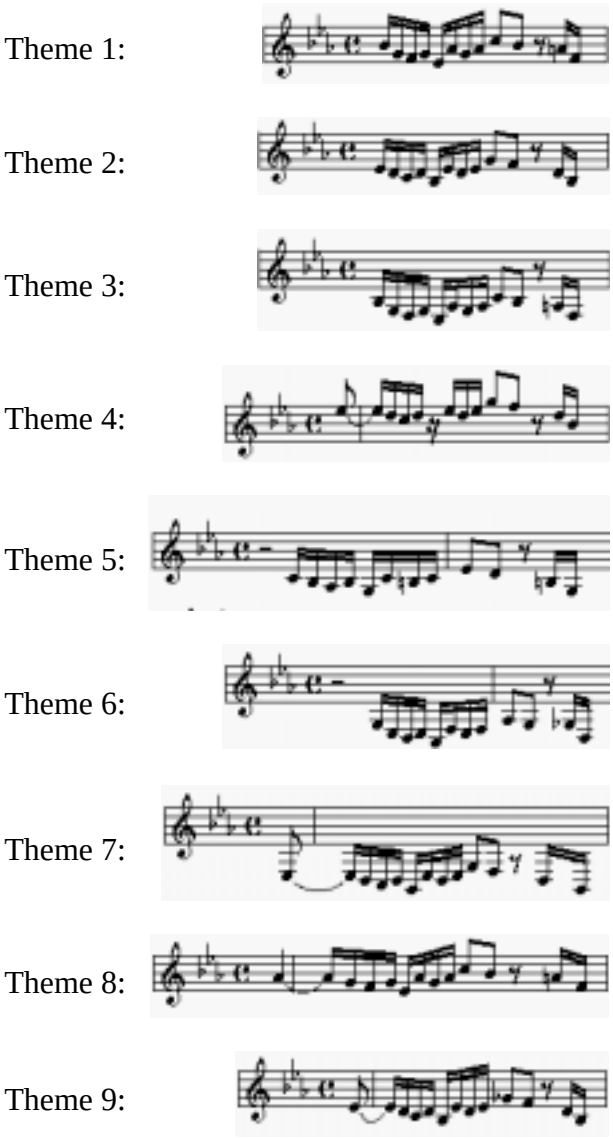

Figure 2. The incipits of the theme occurrences in Fugue VII

These variations of the theme can be considered as relevant parts of the fugue, or relevant documents, with respect to the theme 1. An extraction of the theme occurrences from the voices leaves eight melodic segments of the three voices, which can be considered as non-relevant parts of the fugue, or non-relevant documents. Thus we view the fugue as a database containing 17 documents, themes 1 - 9 as relevant documents and the remaining 8 parts as non-relevant documents.

Figure 3 depicts two queries based on the first theme, one shorter covering the first motif of the subject and the other the whole theme. Both queries and the fugue were represented by the methods *p, t, 2pt, 2p, 3p, 4p* for the retrieval tests.

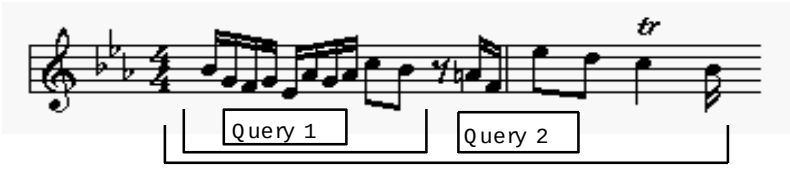

Figure 3. Query sequences formed from the first theme occurrence in Fugue VII

4. Test Results

We shall consider below the retrieval performance of the representation/matching methods *2p*, *2pt*, *3p* and *4p* in finding the nine relevant themes of Fugue VII. We shall also compare the performance of the two queries, the shorter and the longer query. In our small sample database we cannot say anything conclusive about the methods but we can, however, demonstrate the features of the MUSIR retrieval model and their effects as a rationale for further study.

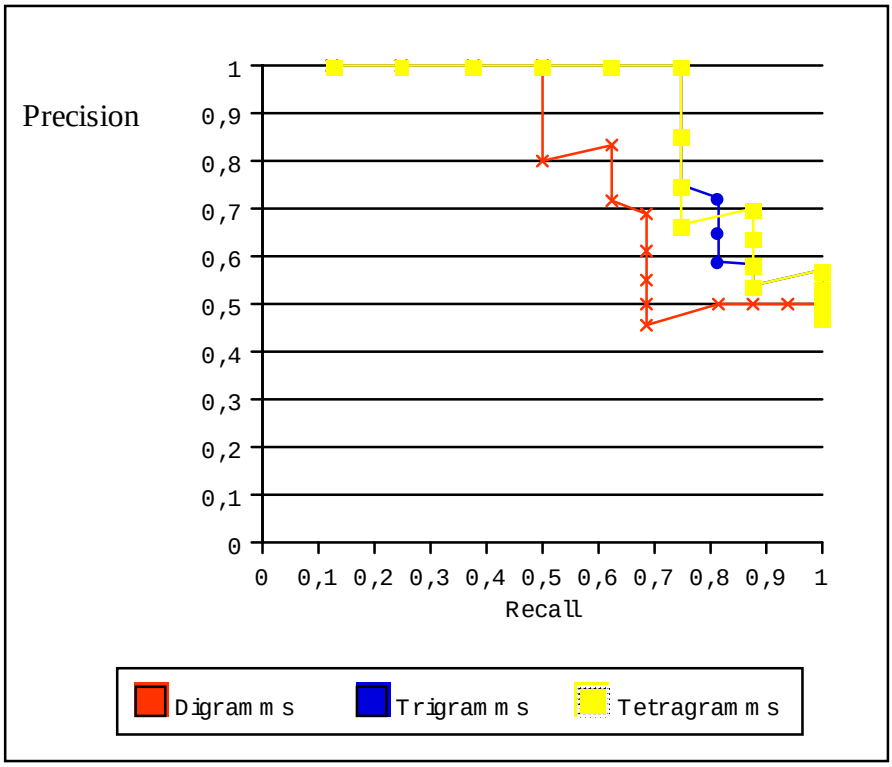

Figure 4. Recall and precision for n-grams *2p*, *3p* and *4p*

Figure 4 shows recall and precision for n-grams *2p*, *3p* and *4p*. The curves are averages for the two queries. It is obvious that tri- and tetra-grams perform better than di-grams although they failed to match all relevant documents of the database. This indicates that with the longer n-grams the risk of not matching some relevant documents at all increases. In this small data-

base there is no performance difference between tri- and tetra-grams. With large databases and long documents shorter n-grams would require longer queries.

Figure 5 shows recall and precision for di-grams *2p* and *2pt*. The curves are averages for the two queries. It is apparent that di-grams are enhanced by the time interval representation. Combining the two dimensions, pitch and time, for melody representation seems thus an interesting possibility. Whether a similar performance improvement would happen with tri- and tetra-grams remains to be seen in later studies. Again, adding time intervals to the already longer n-grams might lead to failing to retrieve relevant documents.

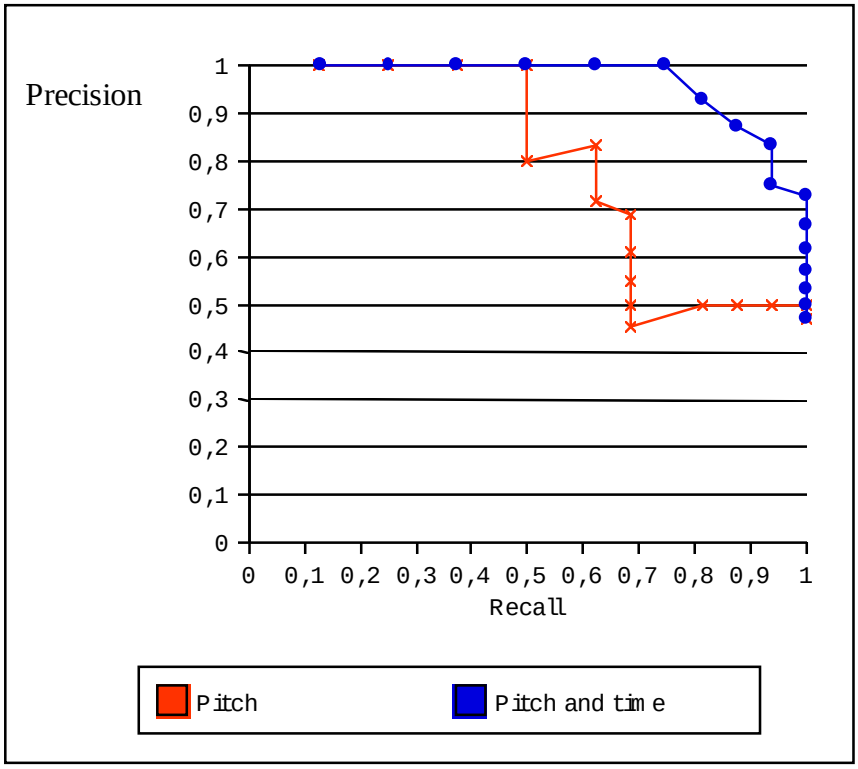

Figure 5. Recall and precision for di-grams *2p* and *2pt*

Figure 6 shows recall and precision by query length for the short Query 1 and the long Query 2. The curves are averages for all n-grams *2p*, *3p* and *4p* in each case. In this very small database containing variations of a theme and some non-relevant parts of one fugue, query length does not seem to affect retrieval performance. It is likely that query length plays an important role in a larger and more varied collection.

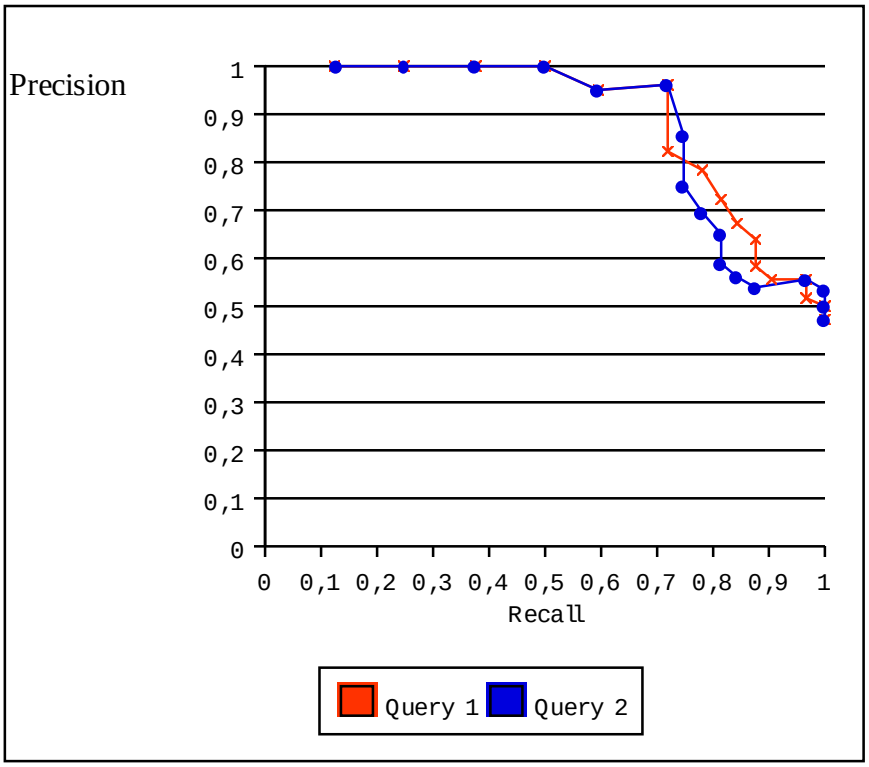

Figure 6. Recall and precision by query length: Queries 1 and 2

5. Discussion and Conclusions

We have presented the MUSIR retrieval model for music content. We used the MIDI representation for music documents and requests as our starting point and developed a filter mechanism for deriving automatically pitch and time interval sequences from the MIDI file. These were then converted into n-gram representations of various lengths for documents and queries. Matching the representations was simple n-gram matching. The MUSIR retrieval model has the following features:

- it supports retrieval of music by requests that are music (representations of musical performance)
- it represents melodic patterns of music which are representative and memorable features of music from the viewpoint of an inquirer
- it is a simple retrieval model employing only relative interval representations of pitch and time
- it is based on the widely available MIDI representation of music
- it supports inexact retrieval — queries which are not correct with respect to note pitch, length, or sequence in a melody
- it ranks the retrieved documents according to decreasing similarity using n-gram scoring.

There are several issues for further work with the MUSIR model.

Testing. The model has not been tested on a large collection. A test on a large and multifaceted collection containing all kinds of music is needed to learn about the relative strengths of various n-gram representations. This testing requires a fairly well-implemented prototype system.

Implementation. The test implementation of the model was based on using spreadsheets and word processing editors to filter the MIDI representation, construct the n-grams and search matching n-grams. This is not an operational system but demonstrates its feasibility. Signature files (Ashford & Willett, 1988) could be used for matching the n-grams efficiently.

Interfaces. Many composition programs which support the MIDI representation are widely available on PC platforms. These may be good environments for those who are familiar with the notation of music and/or have MIDI files available to supply with requests. A retrieval interface can be designed that allows copying a request sequence within composition program and pasting it into the interface's request window. Setting up a MIDI instrument as a request presentation and capture tool requires some engineering work. For people, like the second author, who cannot do better than hum or whistle melodies (incorrectly), a competent intermediary using the MIDI instrument might still be needed. Since contemporary MIDI-applications allow conversion from audio data to MIDI-form, it would also be possible to present queries by recording the user's singing. This is how the retrieval interface works in earlier mentioned MELDEX-system.

Musical limitations. Music that does not have (easily recognizable) melodies may prove difficult for the MUSIR model. In that case we can assume that the music content is embedded in other structures of musical work and requires different representational methods. Polyphonic music is another challenge. Although polyphonic music may be represented in MIDI as several synchronized monophonic tracks, sometimes they have shared events represented only in one track. Each track can be represented for MUSIR separately for matching by monophonic queries. Thus shared events may corrupt queries when parts of a melody are represented on another channel.

Representation. It is well-known that n-gram matching easily fails with long text documents. The same probably holds for music. We have not studied yet, how long music files should be split for retrieval. Neither do we know whether it would make sense to always represent all tracks (e.g., solo, accompaniment; exclude some instrument categories). This depends in part on the nature of the requests users would like to present.

Contemporary music retrieval systems provide access to textual attributes of music documents. We do not think that the MUSIR approach would replace such systems. Textual attributes (e.g., performer, composer, composition title, etc.) are effective for retrieval when they are known by the inquirer. The MUSIR approach helps when textual attributes are not available and complements contemporary systems when textual attributes are not precise enough.

Experience from Finnish public music libraries tells that the music librarians have developed indispensable expertise in memorizing and recognizing melodies often quite awkwardly presented by their clients. However, they still cannot serve all requests and, more importantly, are not digitally available in the web. Systems based on the MUSIR model may act as effective digital music intermediaries, especially in combination with possible textual attributes known by music consumers.

REFERENCES

Ashford, J. & Willett, P. (1988). Text Retrieval and Document Databases. Lund, Sweden: Chartwell-Bratt.

Bakhmutova, I., Gusev, V. & Titkova, T. (1997). The Search for Adaptations in Song Melodies. Computer Music Journal (21)1: 58-67.

Bach, J. S. (1975). Das Wohltemperierte Klavier 1 (BMV 846 - 869). London: Peters. [score]

Dannenberg, R.B. (1993). Music Representation Issues, Techniques, and Systems. Computer Music Journal, 17(3): 20-30.

Fugue (1995). Fugue n:o 7. In: Future Music CD (October 1995). Future Music, Future publishing. [a MIDI file]

Hewlett,W. & Selfridge-Field, E. (1991). Computing in musicology, 1966-91. Computers and the Humanities 25(6): 381-392.

Honing, H. (1992). Issues in the Representation of Time and Structure in Music. Desain, P. & Honing, H. Music, Mind and Machine. Studies in Computer Music, Music Cognition and Artificial Intelligence. Amsterdam: Thesis Publishers.

ISO (1995). International Organization for Standardization (ISO). Standard Music Description Language (SMDL)-standard. ISO/IEC DIS 10743. URL: ftp://ftp.techno.com/pub/SMDL/.

Kontunen, J. (1991a). The Language of Music 1. Basics. Helsinki, Finland: WSOY. [In Finnish]

Kontunen, J. (1991b). The Language of Music 2. Composition forms. Helsinki, Finland: WSOY. [In Finnish]

Lemström, K. & Haapaniemi, A. & Ukkonen, E. (1998). Retrieving music — to index or not to index. In Proc. ACM Multimedia '98 Conference — Art Demos — Technical Demos — Poster Papers, September 1998, Bristol, UK. Pp. 64-66.

Loy, G. (1985). Musicians Make a Standard: The MIDI Phenomenon. Computer Music Journal 9(4): 8-26.

McLane, A. (1996). Music as Information. Annual Review of Information Science and Technology (ARIST), vol. 31. Amsterdam: Elsevier Science Publishers, pp. 225-262.

McNab, R. J., Lloyd A., Smith, Bainbridge, D. & Witten, I. H. (1997). The New Zealand Digital Library MELody inDEX. D-Lib Magazine, May 1997. URL: http://www5.cnri.reston.va.us/dlib/may97/meldex/05witten.html

Salosaari, P. (1998). A music retrieval model based on signature files: n-gramms in the representation of melodies. Tampere: University of Tampere, Dept. of Information Studies, MSc. Thesis. [In Finnish].

Selfridge-Field, E. (1994). The MuseData Universe: A System of Musical Information. In: Computing in Musicology Vol 9. Menlo Park, CA: The Center for Computer Assisted Research in the Humanities.

Beyond MIDI: The Handbook of Musical Codes (1997). Selfridge-Field, E. (ed.). Cambridge (Mass.) : MIT Press.

Stefani, Gino (1985). Musical competence: How do we understand and produce music. Jyväskylä, Finland: University of Jyväskylä, Dept of Musicology, Report 3/1985. [In Finnish, Italian original: La Competenza Musicale, Cooperativa Libraria. Universitaria Editrice Bologna, 1982.]

Turtle, H. & Croft, W.B. (1992). A Comparison of Text Retrieval Models. The Computer Journal 35(3): 279–290.

Wiggins, G., Miranda, E., Smaill, A. & Harris, M. (1993). A Framework for the Evaluation of Music Representation Systems. Computer Music Journal 17(3): 31-42.