An Approach Towards A Polyphonic Music Retrieval System

Shyamala Doraisamy
Dept. of Computing
Imperial College,
London SW7 2BZ
+44-(0)20-75948230

sd3@doc.ic.ac.uk

Stefan M Rüger
Dept. of Computing
Imperial College
London SW7 2BZ
+44-(0)20-75948355

srueger@doc.ic.ac.uk

ABSTRACT

Most research on music retrieval systems is based on monophonic musical sequences. In this paper, we investigate techniques for a full polyphonic music retrieval system. A method for indexing polyphonic music data files using the pitch and rhythm dimensions of music information is introduced. Our strategy is to use all combinations of monophonic musical sequences from polyphonic music data. 'Musical words' are then obtained using the n-gram approach enabling text retrieval methods to be used for polyphonic music retrieval. Here we extend the n-gram technique to encode rhythmic as well as interval information, using the ratios of onset time differences between two adjacent pairs of pitch events. In studying the precision in which intervals are to be represented, a mapping function is formulated in dividing intervals into smaller classes. To overcome the quantisation problems that arise with using rhythmic information from performance data, an encoding mechanism using ratio bins is also adopted. We present results from retrieval experiments with a database of 3096 polyphonic pieces.

1. INTRODUCTION

Music documents encoded in digital formats have rapidly been increasing in number with the advances in computer and network technologies. Managing large collections of these documents can be difficult and this has consequently motivated research towards computer-based music information retrieval (IR) systems. Music documents encompass documents that contain any music-related information such as music recordings, musical scores, manuscripts or sketches and so on [1]. Many studies have been carried out in using the music-related information contained in these documents for the development of content-based music IR systems. Such systems retrieve music documents based on information such as the incipits, themes and instrument families. However, most of these content-based IR systems are still research prototypes. Music IR systems that are currently in wide-spread use are systems that have been developed using meta-data such as file-names, titles and catalogue references.

One common approach in developing content-based music IR systems is with the use of pitch information. Examples of such systems are Themefinder [2] and Meldex [3]. However, these systems were developed using monophonic musical sequences, where a single musical note is sounded at one time, as opposed to polyphonic music where more than one note is sounded simultaneously at any one point in time. With vast collections of polyphonic music data available, research on polyphonic music IR is on the rise [4].

Our aim is the development of a polyphonic music IR system for retrieving a title and performance of a musical composition given an excerpt from a musical performance as a query. For content-based indexing, we use the pitch and rhythm dimensions of music information and propose an approach for indexing full polyphonic music data. In this paper we present our approach and evaluate it using a database of polyphonic pieces.

The paper is structured as follows: Section 2 highlights some of the issues and challenges in content-based indexing. Section 3 presents the approach taken in using the pitch and duration information for indexing. The steps in constructing n-grams from polyphonic music data and the mechanism of extending the representation to include rhythm information are outlined. The empirical analysis performed and approach for encoding patterns derived from n-gramming is presented. Section 4 reports the retrieval experiments using ranked retrieval and evaluation results using the mean reciprocal rank measure of our polyphonic music IR system.

2. ISSUES IN CONTENT-BASED INDEXING AND RETRIEVAL OF MUSICAL DATA

The problem of varying user requirements is common to most IR systems. Music IR systems are no exception. Music librarians, musicologists, audio engineers, choreographers and disc-jockeys are among the wide variety of music IR users with a wide range of requirements [1]. For example, with a musical query where the user plays a recording or hums a tune, one user could possibly require all musical documents with the same key to be retrieved while another user's requirement might be to obtain all documents

of the same tempo. Looking at another example where a musical composition's title is queried, one user could require the composer's full-name and another user might need to know the number of times the violin had a solo part in the composition. Knowledge of user requirements is an important aspect in developing useful indexes, and with music IR systems this challenge is compounded with others such as the multiple dimensions of music data and digital music data formats.

Music data are multi-dimensional; musical sounds are commonly described by their pitch, duration, dynamics and timbre. Most music IR systems use one or two dimensions and these vary based on types of users and queries. Selecting the appropriate dimension for indexing is an important aspect in developing a useful music IR system. Indexing a system based on its genre class would be useful for a system that retrieves music based on mood but not for a system where a user needs to identify the title of a music piece queried by its theme.

The multiple formats in which music data can be digitally encoded present a further challenge. These formats are generally categorised into a) highly structured formats such as Humdrum [5] where every piece of musical information on a piece of musical score is encoded, b) semi-structured formats such as MIDI in which sound event information is encoded and c) highly un-structured raw audio which encodes only the sound energy level over time. Most current music IR systems adopt a particular format and therefore queries and indexing techniques are based upon the dimensions of music information that can be extracted or inferred from that particular encoding method.

There are many approaches for the development of music IR systems. Some of these include the use of approximate matching techniques in dealing with challenges such as recognising melodic similarity [6], the use of standard principles of text information retrieval and exact matching techniques that demand less retrieval processing time [7,8].

3. A TECHNIQUE FOR INDEXING POLYPHONIC MUSICAL DATA
3.1 Pattern extraction

The approach we take for indexing is full-music indexing, similar to full-text indexing in text IR systems. This approach was studied by Downie [8], where a database of folksongs was converted to an interval-only representation of monophonic 'melodic strings'. Using a gliding window, these strings were fragmented into length-n subsections or windows called n-grams for music indexing.

With polyphonic music data, a different approach to obtaining n-grams would be required since more than one note can be sounded at one point in time (known as the onset time in this context). In sorting polyphonic music data with ascending onset times and dividing it into windows of n different adjacent onset times, one or more possible monophonic 'melodic string(s)' can be obtained within a window. The term melodic string used in this context may not be a melodic line in the musical sense. It is simply a monophonic sequence extracted from a sequence of polyphonic music data.

Various approaches in deriving patterns from unstructured polyphonic music for computer-based music analysis have been investigated in a study by Crawford et. al. [9]. The approach taken for our study would be a musically unstructured but an exhaustive mechanism in obtaining all possible combinations of monophonic sequences from a window for the n-gram construction. Each n-gram on its own is unlikely to be a musical pattern or motif but a pattern amenable for digital string-matching. The n-grams encoded as musical words using text representations would be used in indexing, searching and retrieving a set of sequences from a polyphonic music data collection.

The summary of steps taken in obtaining these monophonic musical sequences is as follows:

Given a polyphonic piece in terms of ordered pairs of onset time and pitch sorted by onset time,

1. Divide the piece using a gliding window approach into overlapping windows of n different adjacent onset times
2. Obtain all possible combinations of melodic strings from each window

N-grams are constructed from the interval sequence(s) of one or more monophonic sequence(s) within a window. Intervals (the distance and direction between adjacent pitch values) are a common mechanism for deriving patterns from melodic strings, being invariant to transpositions [10]. For a sequence of n pitches, an interval sequence is derived with $n-1$ intervals by Equation (1).

$$Interval_i = Pitch_{i+1} - Pitch_i \qquad (1)$$

To illustrate the pattern extraction mechanism for polyphonic music data, the first few bars of Mozart's Alla Turca, as shown in Figure 1, is used. The performance data of the first two bars of the piece was extracted from a MIDI file and converted into a text format, as shown in Figure 2(a). The left column contains the onset times sorted in ascending order, and the corresponding notes (MIDI semitone numbers) are on the right column. The performance visualised on a time-line is shown in Figure 2 (b).

Figure 1. Excerpt from Mozart's Alla Turca

0	71
150	69
300	68
450	69
600	57
600	72
900	64
900	60
1200	74
1200	64
1200	60
1350	72
1500	71
1500	64
1500	60
1650	72

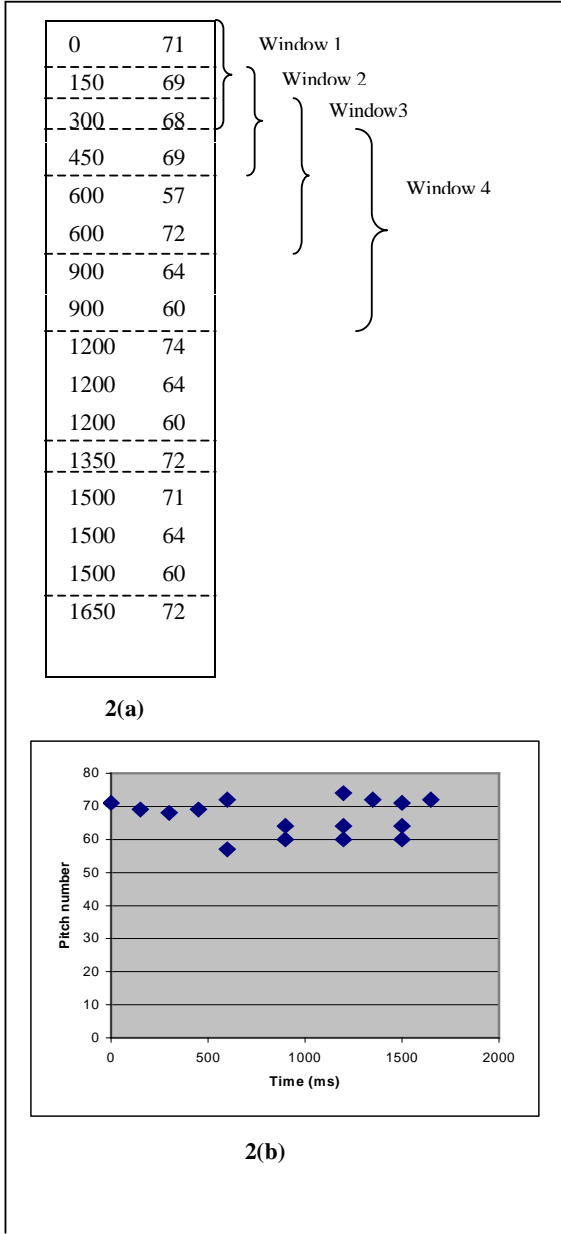Window 1
Window 2
Window3
Window 4

2(a)

2(b)

Figure 2. (a) Onset times and pitch events for Mozart's
All Turca (b) performance visualized on a time-line

Following the steps outlined in obtaining the n-grams and applying Equation (1) in pattern derivation, the interval sequences from the first 3 windows of length-3 onset times of the performance data in Figure 2 are:

Window 1: [-2 -1]

Window 2: [-1 1]

Window 3: [1 –12] and [1 3]

To add to the information content of the n-grams constructed using interval sequences, the duration dimension of music information is used. Numerous studies have been carried out with the use of patterns generated from various combinations of the pitch and duration dimensions. These studies either used pitch information [8, 11], rhythm information [12] or both pitch and rhythm information simultaneously [4, 13]. In using the duration dimension for pattern derivation, a common mechanism is to use the relative duration of a note to a designated base duration such as the quarter or the sixteenth note. Relative durations are widely used as they are invariant to changes of tempo [10]. However, the choice of base durations such as the quarter or the sixteenth note could pose quantisation problems with performance data compared to data obtained from score encodings. With performance data, one option for the selection of a base duration could be the time difference between the first two notes of a given performance. However, with errors such as timing deviations of these two notes or recordings being slightly trimmed off at the beginning, this error would be duplicated in obtaining rhythmic information of the whole performance data.

In our approach, we look at the onset times pattern based on the timeline - the times at which pitch events occur. The approach in using time between consecutive note onsets has been studied by I. Shmulevich et. al. [14]. For pattern derivation using rhythm information, the ratios of time difference between adjacent pairs of onset times form a rhythmic ratio sequence. With this approach, it is not necessary to quantise on a predetermined base duration, to use the duration length of a note (which can be difficult to determine from audio performances) and we do not assume any knowledge of beat and measure information. For a sequence of n onset times, a rhythmic ratio sequence is derived with $n-2$ ratios obtained by Equation (2).

$$Ratio_i = \left(\frac{Onset_{i+2} - Onset_{i+1}}{Onset_{i+1} - Onset_i} \right) \qquad (2)$$

In obtaining n-grams that incorporate interval and rhythmic ratio sequences using n onset times and pitches, the n-gram would be constructed in the pattern form of:

[Interval$_1$ Ratio$_1$ … Interval$_{n-2}$ Ratio$_{n-2}$ Interval$_{n-1}$]

Using the example of Figure 2, the combined interval and ratio sequences from the first 3 windows of length 3-onset are:

Window 1: [-2 1 -1]

Window 2: [-1 1 1]

Window 3: [1 1 -12] and [1 1 3]

Note that the first and last number of each tuple are intervals while the middle number is a ratio.

3.2 Pattern encoding

In order to be able to use text search engines we need to encode our n-gram patterns with text characters. One challenge that arises is to find an encoding mechanism that reflects the pattern we find in musical data. With large numbers of possible interval values and ratios to be encoded, and a limited number of possible text representations, classes of intervals and ratios that clearly represent a particular range of intervals and ratios without ambiguity had to be identified. For this, the frequency distribution for the directions and distances of pitch intervals and ratios of onset time differences that occur within the data set were obtained. A data collection of 3096 MIDI files of a classical music collection was used in obtaining these frequencies. These were mostly classical music performances obtained from the Internet. [http://www.classicalarchives.com]

For the pitch encoding, firstly the data set was analysed for the range and interval distances that occur within the data set and the frequency at which these occur. The frequency distribution versus interval (in units of semitones) graph obtained is shown in Figure 3.

According to Figure 3, the vast bulk of pitch changes occurs within one octave (i.e., -12 to +12 semitones). A good encoding should be more sensitive in this area than outside of it. We chose the code to be the integral part of a differentiable continuously changing mapping function (3), the derivative of which approximately matches the empirical distribution of intervals in Figure 3.

$$Code = \text{int}\left(X \tanh\left(\frac{Interval_{n-1}}{Y} \right) \right) \qquad (3)$$

In Equation (3), X is a constant set to 27 for our experiments as a mechanism to limit the codes range to the 26 text letters. Y is set to 24 to obtain a 1-1 mapping of semitone differences in the range [-13, 13]. In accordance with the empirical frequency distribution of Figure 3, less frequent semitone differences (which are bigger in size) are squashed and have to share codes. Based on the property of the tanh curve, Y determines the rate at which class sizes increase as interval sizes increase. This is a trade-off between classes of small (and frequent) versus large (and rare) intervals. The codes obtained are then mapped to the ASCII character values for letters. In encoding the interval direction, positive intervals are encoded as uppercase A-Z and negative differences are encoded with lower case a-z and in the centre code 0 being represented by the numeric character 0.

In using duration ratios, most studies have assumed *quantised rhythms, i.e., rhythm as notated in the score [14]* owing to simplicity and timing deviations that could occur with performance data. To deal with performance data, we adopt ratio bins for our study.
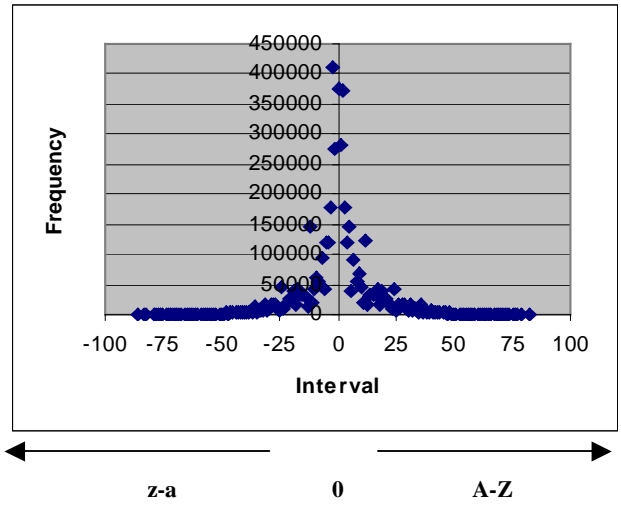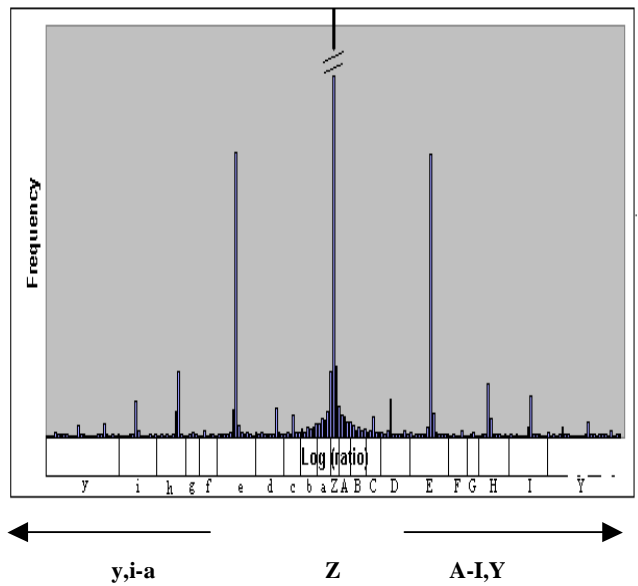

Figure 3. Interval Histogram

Figure 4. Ratio Histograms and Ratio Bins

Figure 4 shows the frequency versus the log of the ratios (onset times were obtained in units of milliseconds). We analysed the frequency distribution of ratio values of the data collection in order to provide quantisation ranges for the bins that reflect the data set. The peaks clearly discriminate ratios that are frequent and bins for ratio values for encoding can be established. Mid-points between these peak ratios were then used to construct bins which provided appropriate quantisation ranges in encoding the ratios. Ratio 1 has the highest peak as expected and other peaks occur in a symmetrical fashion where for every *peak ratio* identified, there is

a symmetrical peak value of 1/*peak ratio*. From our data analysis, the peaks identified as ratios greater than 1 are 6/5, 5/4, 4/3, 3/2, 5/3, 2, 5/2, 3, 4 and 5.

The ratio 1 is encoded as Z. The bins for ratios above 1 as listed above are encoded with uppercase alphabets A-I and any ratio above 4.5 is encoded as Y. The various bins for ratios smaller than 1 as listed above are encoded with lowercase alphabets a-i and y respectively. The ranges identified with this symmetry and corresponding codes assigned are visualised in Figure 4.

4   IMPLEMENTATION

4.1 Database development

One of the main aims of this study is to examine the retrieval effectiveness of the musical words obtained from n-grams based on the pitch and duration information. The experimental factors investigated for this initial study were a) the size of interval classes and bin ranges for ratios, b) the query length and c) the window size used for the n-gram construction. We use the same data collection of 3096 classical MIDI performances for the database development as in Section 3.

6 databases P4, R4, PR3, PR4, PR4CA and PR4CB were developed. The minimum window size is 3, as at least 3 unique onset times would be required in obtaining one onset time difference ratio. A description of each database and its experimental factors follows:

P4: Only the pitch dimension is used for the n-gram construction with the window size of 4 onset times. Each n-gram is encoded as a string of 3 characters corresponding to 3 intervals. Y is set to 24 to enable a 1-1 mapping of codes to most of the intervals within a distance of 20. The theoretical maximum of possible index terms is $148,877 = (26*2+1)^3$.

R4: Only the rhythm dimension is used for the n-gram construction with the window size of 4 onset times. All bin ranges identified as significant ratio ranges were used in encoding. The theoretical maximum of possible index terms is $441 = (10*2+1)^3$.

PR3: The pitch and rhythm dimensions are used for the n-gram construction in the combined pattern form stated in Section 3 with the window size of 3 onset times. Y is assigned 24 to enable similar interval class encoding as P4. All bin ranges identified as significant ratio ranges are used in encoding. The theoretical maximum of possible index terms is $58,989 = (53*21*53)$.

PR4: The pitch and rhythm dimensions are used for the n-gram construction as above but with the window size of 4 onset times. All bin ranges identified as significant ratio ranges are used in encoding. The theoretical maximum of possible index terms is $65,654,757 = (53^3*21^2)$.

PR4CA: The pitch and rhythm dimensions are used for the n-gram construction as above. To study the effects of the interval class sizes within the range of 2 octaves for a 2-1 mapping for most intervals for most intervals smaller than 20 semitones, Y is set to 48. Although one character now covers at least 2 semitones (as

opposed to 1 semitone above), still all alphabets are used with this encoding, i.e. 26 uppercase and 26 lowercase letters and 0 for no change. The encoding for the ratios was made coarser as well : where we previously used the codes A-I,Y and a-I,y we now use the codes A-D, Y and a-d,y respectively, now A covers what used to be represented by A and B, B covers what used to be C and D , C covers what used to be E and F etc. The theoretical maximum of possible index terms is $18,014,177 = (53^3*11^2)$.

PR4CB: The pitch and rhythm dimensions are used for the n-gram construction as above. To study the effects of the interval class sizes within the range of 2 octaves for a 3-1 mapping for most intervals up to around 20 semitones, Y is set to 72. Coarse ratio encoding with bins used as in PR4CA.

The summary of databases and experimental factors are shown in Table 1.

Table 1. Databases and experimental factors

Database	Pitch	Rhythm	n	Y	# R.Bins	#Terms
P4	Y		4	24		148,877
R4		Y	4		21	441
PR3	Y	Y	3	24	21	58,989
PR4	Y	Y	4	24	21	65,654,757
PR4CA	Y	Y	4	48	11	18,014,117
PR4CB	Y	Y	4	72	11	18,014,117

4.2 Retrieval Experiments

In examining the retrieval effectiveness of the various formats of musical words and to evaluate the various experimental factors, an initial run, R1, was performed on the 6 databases. For query simulation, polyphonic excerpts are extracted from randomly selected musical documents of the data collection. Query locations were set to be the beginning of the file. In simulating a variety of query lengths, lengths of the excerpts extracted from the randomly selected files were of 10, 30 and 50 onset times. These excerpts were then pre-processed and encoded to generate musical words with similar formats to the corresponding 6 databases: P4, R4, PR3, PR4, PR4CA and PR4CB. The ranked retrieval method was used for run R1 averaged over 30 queries. In ranking the documents retrieved, the cosine rule used by the MG system was adopted [15] and in evaluating our retrieval using the known item search of our query excerpt, the Mean Reciprocal Rank (MRR) measure was used. The reciprocal rank is equal to 1/r where r is the rank of the music piece the query was taken from. In using the known item search, the rank position of the document that the query was extracted from was used in obtaining the reciprocal rank measure. These were averaged over the 30 queries. This MRR measure is between 0 and 1 where 1 indicates perfect retrieval. The retrieval results are shown in Table 2.

Table 2. MRR measures for run R1

	10	30	50
P4	0.60	0.77	0.81
R4	0.03	0.11	0.15
PR3	0.46	0.74	0.81
PR4	0.74	0.90	0.95
PR4CA	0.71	0.83	0.71
PR4CB	0.47	0.68	0.73

The results clearly indicate that using n-grams with polyphonic music retrieval is a promising approach with the best retrieval measure 0.95 being obtained by musical words of the PR4 format and a query length of 50 onset times. Comparing the retrieval measures of P4 and PR4 for all 3 query lengths, it can be said that the addition of rhythm information to the n-gram is a definite improvement to widening the scope of n-gram usage in music information retrieval.

The length of a window for n-gram construction would require further study, as there are clear improvements of measures between PR3 and PR4 for all query lengths. Further experiments will be needed to obtain the optimal length. In looking at the class size of the intervals and bin range of ratios, measures clearly deteriorate from smaller class sizes of PR4 to larger sizes of PR4CA and PR4CB. The class sizes require further investigation to determine its usefulness in providing allowances for more fault-tolerant retrieval.

In general, and as expected, the measure improves with the length of the query for all databases although retrieval using only ratio information with R4 is almost insignificant. Clearly, the 441 possible different index terms are insufficient to discriminate music pieces.

4.3 Error Simulation

A second run, R2, was performed by simulating errors in the queries to study the retrieval behaviour under error conditions. Error models used in monophonic music described in [3, 8] were not adopted for this study as the range of intervals was significantly different. As there were no error models available with polyphonic music, we adopted the Gaussian error model for intervals as shown in Equation (4) and for ratios as shown in Equation (5). ε is the Gaussian standard random variable and D_i is the mean deviation for an interval error and D_r is the mean deviation for an error in the ratio.

$$NewInterval_k = Interval_k + (D_i * \varepsilon) \qquad (4)$$

$$NewRatio_k = Ratio_k * \exp.(D_r * \varepsilon) \qquad (5)$$

As an initial attempt to investigate retrieval with error conditions, we arbitrarily selected two sets of error deviation values D1 and D2. With D1, D_i was assigned 3 and D_r assigned as 0.3. For the second set of mean error deviation values, D_i was assigned 2 and D_r was retained as 0.3. D_r was left unchanged, as the ratio bin range was not varied between PR4CA and PR4CB. All musical words generated for the similar queries used in R1 and with length 30 were modified by incorporating the error deviation for the pitch and duration dimensions correspondingly for the 3 databases PR4, PR4CA and PR4CB. The MRR measures are shown in Table 3.

Table 3. MRR measures for run R2

	D1	D2
PR4	0.24	0.50
PR4CA	0.30	0.65
PR4CB	0.27	0.50

The results clearly indicate that musical words encoded with a wider interval class size perform better with error conditions. A compromise between musical words encoded using larger interval class sizes and wider ratio bin ranges and smaller ones is clearly required. This can be seen from the improvement in measures obtained with run R2 and deviation set D2 of Table 3 where the measure of PR4CA is 0.65 and PR4 only 0.50. For the counterpart run, R1, with no query errors, it indicates deterioration in measure with the wider encoding (where a measure of 0.90 was obtained with PR4 and only 0.83 for PR4CA with query length 30).

This initial experiment under error conditions clearly identifies the need for a detailed analysis in obtaining optimal values for interval class size and effective retrieval in using n-grams in polyphonic music retrieval.

5   FUTURE WORK

Based on the experimental results and initial experimental factors investigated, this study will be continued with an in-depth study of the following experimental factors: a) query length b) window length c) ration bin range d) Y value for interval classification e) error model. Further issues for investigation are a) the development of error models with polyphonic music, b) a relevance judgment investigation in assessing the documents and finer retrieval measures, c) suitability of the ranking mechanism for musical words, d) an analysis of the search complexity of the algorithm in extracting all possible patterns

6   CONCLUSIONS

This study has proven the usefulness of using n-grams in polyphonic music data retrieval. An interval mapping function was utilised and proved useful in mapping interval classes over the text alphabetical codes. Onset time ratios have proven useful for incorporating rhythm information. With the use of bins for ranges of significant ratios, the rhythm quantisation problem in music

performance data has been overcome. The results presented so far for polyphonic retrieval are qualitatively comparable to published successful monophonic retrieval experiments [8] and, hence, very promising.

7 ACKNOWLEDGEMENTS

8 REFERENCES

[1] David Huron, *Perceptual and Cognitive Applications in Music Information Retrieval*, International Symposium on Music Information Retrieval, Music IR 2000, Oct 23rd - 25th, 2000, Plymouth, Massachussetts.

[2] Andreas Kornstadt, *Themefinder: A Web-Based Melodic Search Too*l, Computing in Musicology 11, 1998, MIT Press

[3] Rodger J. MacNab, Lloyd A.Smith, David Bainbridge and Ian H. Witten, *The New Zealand Digital Library MELody inDEX*, D-Lib Magazine, May 1997

[4] M. Clausen, R. Engelbrecht, D. Meyer, J. Schmitz, *PROMS: A Web-based Tool for Searching Polyphonic Music*, International Symposium on Music Information Retrieval, Music IR 2000, Oct 23rd - 25th, 2000, Plymouth, Massachussetts.

[5] David Huron, *Humdrum and Kern: Selective Feature Encoding*, Beyond MIDI: The Handbook of Musical Codes, pp 375-40.

[6] Eleanor Selfridge-Field, *Conceptual and Representational Issues In Melodic Comparison*, Computing in Musicology 11, 1998, pp 1-64

[7] Massimo Melucci and Nicola Orio, *Music Information Retrieval using Melodic Surface*, The Fourth ACM Conference on Digital Libraries '99, Berkeley, USA,pp 152-160

[8] Stephen Downie and Michael Nelson, *Evaluation of A Simple and Effective Music Information Retrieval Method*, SIGIR 2000, Athens, Greece, pp 73-80

[9] Tim Crawford, Costas S. Iliopoulus and Rajeev Raman, *String-Matching Techniques for Musical Similarity and Melodic Recognition*, Computing in Musicology 11, 1998, MIT Press, pp 73-100

[10] Kjell Lemström, Atso Haapaniemi, Esko Ukkonen, *Retrieving Music - To Index or not to Index*, ACM Multimedia '98, -Art Demos, Technical Demos - Poster Papers, September 1998, Bristol, UK

[11] Steven Blackburn and David DeRoure, *A Tool for Content-Based Navigation of Music*, ACM Multimedia '98, Bristol, UK, pp 361 – 368

[12] Chen, J.C.C. and A.L.P. Chen, *Query by Rhythm: An Approach for Song Retrieval in Music Databases*, In proc. Of IEEE Intl. Workshop on Research issues in Data Engineering, pp 139-146, 1998

[13] Shyamala Doraisamy, *Locating Recurring Themes in Musical Sequences*, M. Info. Tech. Thesis, 1995, University Malaysia Sarawak.

[14] I. Shmulevich, O. Yli-Harja, E. Coyle, D.-J. Povel, and K. Lemström, *Perceptual Issues in Music Pattern Recognition – Complexity of Rhythm and Key Finding*, In Proceedings of the AISB '99 Symposium on Musical Creativity, pages 64-69, Edinburgh, 1999

[15] Ian H. Witten, Alistair Moffat and Timothy C. Bell, *Managing Gigabytes: Compressing and Indexing Documents and Images*, 2nd edition, 1999, Morgan Kaufmann Publishers