Polyphonic Queries: A Review of Recent Research

Cory McKay
Faculty of Music, McGill University
555 Sherbrooke Street West
Montreal, Quebec, Canada H3A 1E3
cory.mckay@mail.mcgill.ca

The ever-increasing amount of digitally encoded music is making it more and more imperative to find efficient and effective ways of searching music databases. Queries composed of musical phrases or melodic fragments are particularly useful because of the ease with which humans remember such information. Melodic fragments also have the advantage of being relatively easy to enter into search systems, either through notation interfaces or query by humming. This last option can be particularly useful because it is fast and requires little musical knowledge. Of course, there is the disadvantage that only one note may be hummed at a time, but this problem can be solved by humming melodies that occur in different voices one after the other.

It is desirable for a system to return not only all occurrences of a given set of notes in a database, but also to be able to return records that contain similar sets of notes. This allows search systems to deal with incomplete or partially erroneous queries. Also, an ideal system should be able to search music stored in both symbolic and raw audio formats. Possible symbolic formats include MIDI recordings, KERN scores, manuscripts and sketches.

There has been some success in this field in the special case of monophonic music, such as in the cases of *Themefinder* and *Meledex*. Unfortunately, the task of searching polyphonic music is considerably more difficult.

One major difficulty with polyphonic music is that notes may begin simultaneously, making it impossible to outline an unambiguous sequence of note events. The presence of multiple voices, with their varying roles and relevance to particular queries, also increases the difficulty of the task.

Dealing with both symbolic and raw audio representations of music also becomes difficult with polyphonic music. Features that are easily derivable from the symbolic representations (e.g. pitches, note durations) do not correspond with the easily accessible features of music stored as raw audio (e.g. centroids, energy). Effective techniques for monophonic transcription make it possible to transform audio data directly into symbolic form in the case of monophonic music, thus circumventing this problem. Unfortunately, there are currently no reliable systems for transcribing polyphonic music.

Due to these problems, there are currently no content-based polyphonic search systems that are widely accepted as being sufficient for practical use. Although there are some polyphonic search systems in use that rely on meta-data, this reliance makes them inappropriate for content-based searches involving queries composed of melodic fragments. The need for a good content-based polyphonic search engine has led to the recent publication of a number of papers on the subject.

Wiggins, Lemström and Meredith (2002) have designed a new algorithm called SIA(M)ESE that can be used to make transposition-invariant queries on polyphonic records. This algorithm matches a query even if there are events in a score being searched that separate musical events in the query. This can be considered both an advantage and a disadvantage, depending on the needs of a particular search. Although the algorithm does appear to be powerful, it assumes that both queries and database files are in symbolic form and are accurate. This, unfortunately, severely limits the situations to which the algorithm can be applied. In addition, the authors fail to present an implementation of the algorithm.

Doraisamy and Rüger (2001) have implemented a system that uses the pitch and rhythmic dimensions of music to perform searches of polyphonic music using polyphonic queries. N-grams, which have proved useful for monophonic queries, are used to do this.

N-grams are produced by converting notes into an interval-based representation and grouping these intervals into subdivisions of length n using a gliding window. This approach leads to transposition-invariant data, which can be both an advantage and a disadvantage, depending on the requirements of a given search. In order to deal with the potential for simultaneous note onset times in polyphonic music, Doraisamy and Rüger constructed exhaustive "melodic strings" by first dividing pieces into overlapping windows of n adjacent onset times and then finding all possible combinations of onsets within each window.

Doraisamy and Rüger also incorporated rhythmic information into each n-gram window. This was done by calculating the ratio of the time differences between adjacent note onsets using the following formula:

$$Ratio_i = (Onset_{i+2} - Onset_{i+1}) / (Onset_{i+1} - Onset_i)$$

This approach avoids the need to quantize events based on a predetermined beat duration and, by using onsets only, avoids needing to determine the duration of notes, which can be a difficult task when analyzing raw audio recordings.

The n-grams were then converted into text-based representations so that text-based search engines could be used to perform searches. Interval and rhythmic ratio histograms were constructed in order to search for patterns in each piece.

The system was tested using a database composed of 3096 MIDI representations of classical music. The effects of varying window sizes, bin ranges and query lengths were all studied. A 95% success rate was achieved with window sizes of 4 onset times, variable bin ranges and queries involving 50 notes. Performance dropped to 74% when query lengths of ten notes were used. Tests were also done with queries containing errors, resulting in a performance drop to 65%.

Although this system does perform well under ideal conditions, it imposes certain undesirable limitations. Databases consisting of raw audio files were not considered, only transposition invariant searches were possible and undesirably long query lengths were necessary. In addition, the system was only tested with classical music records. This study was only intended as an initial investigation, however, and was successful in showing the potential utility of the modified n-gram approach, which has been used in an adapted form in most subsequent publications in this area.

Doraisamy and Rüger (2002) published further research a year later. Instead of discussing polyphonic queries, as was done in the 2001 paper, this paper concentrated on monophonic queries of polyphonic music. This was done with the aim of producing methods that could be used with query by humming.

Once again, n-grams were presented as an appropriate tool to use for this task. The authors argued that n-grams are particularly appropriate for error-prone queries, such as those resulting from query by humming, as one or two mistakes only lead to a few incorrect n-grams among a much larger number of correct ones.

The system discussed in the previous paper was updated to include more sophisticated error models to test the effects of query inaccuracies. The database was also expanded so that it included popular music as well as classical music. On average, 80% of the relevant compositions were returned in the first 15 hits.

This research seems to indicate that n-grams are an effective and error-tolerant tool for searching polyphonic music with monophonic queries, although improvements still need to be made. Unfortunately, the queries were still symbolic, so the applicability of the system to query by humming was not actually tested.

Pickens et al. (2002) implemented a system that takes in polyphonic queries in audio format, although the database was still derived from music represented symbolically. This system was designed to match pieces containing a given query as well as pieces containing variations on it.

This system dealt with queries in audio format by relying on transcription modules to transform them into symbolic form. In order to compensate for the error-prone nature of polyphonic transcription, it was hoped that the query system would show enough error tolerance to deal with transcription errors. Two types of polyphonic transcription systems were used, including a blackboard system.

The transcribed queries and the files in the database were analyzed and stored using a harmonic modelling module that characterized pieces by mapping chords to a probability distri-

bution. This module operated by first breaking the music into sequences of independent note sets. A smoothing procedure was then applied to these sets and Markov models were created from them.

In order to perform searches, a scoring function was used to compare query models with each of the models stored in the database. This resulted in dissimilarity scores that allowed search hits to be ranked.

The system was tested with a database containing 3150 pieces of classical piano music. Queries consisted of full-length audio recordings. On average, searches assigned a rank of between two and six to the correct database record. The system was also moderately successful at matching variations of a piece; on average, three out of the top five hits were relevant.

This system does provide the advantage of being able to deal directly with polyphonic audio queries. Its usefulness is, however, limited by the effectiveness of its transcription systems and the error tolerance of the query system. Considering that this system was only tested on piano music, results could seriously deteriorate if the system were exposed to polytimbral music. The use of only one feature to characterize pieces is also limiting, particularly with classes of pieces that are very similar harmonically but contain a great deal of rhythmic, timbral or melodic variation. The need to use full performances as queries to receive good results was also a serious limitation.

Song, Bae and Yoon (2002) have constructed a query system that can process both database queries and database records in raw audio form. Only monophonic queries were permitted, however, as this system was designed to be used with query by humming.

This system avoids the disadvantages of automated transcription by mapping audio data directly to a "mid-level" melody-based feature set description that can be used as is in searches. This is an interesting alternative to Pickens' approach of first transcribing audio data and then extracting features from the resulting high-level symbolic representation.

This "mid-level" representation was produced by processing audio frames using a five-step process: enhancement, harmonic sum, note strength calculation, note segmentation and note segment sequence construction. Instead of making definite decision on which notes were present, as a blackboard system would have done, vectors of all probable notes were kept for each audio segment. A DP-matching method were used during searches so that patterns of different lengths and with potential errors could be compared.

The system was tested by attempting to match 176 hummed melodies to 92 short extracts (15-20 seconds long) of Korean and Western popular songs. This resulted in exact matches roughly 43% of the time and matches in the top ten from 69% to 76% of the time (depending on window size). Search time varied from 3 to 14 seconds.

Although the performance of this system was somewhat poor relative to the other systems, and only monophonic queries were possible, it was the only one to use audio data for both queries and database records. Unfortunately, the system was only tested using a database of short recordings. Performance would likely deteriorate greatly if longer recordings were used.

Overall, it appears that no truly viable systems have been produced yet, although some potentially promising approaches have been proposed. There are a number of recurring problems that appear in the systems discussed here. None of these systems are able to deal with both symbolic and audio data, and none have been tested with both polyphonic and monophonic queries. These systems also tend to require fairly long queries to achieve good results, which may not be convenient in some cases. The systems also allow very little flexibility in terms of the kind of search that is needed. For example, none of these systems allow searches that are not transposition-invariant. It is also difficult to compare these systems because they each use different performance evaluation metrics and the testing that has been done has tended to use fairly limited data sets.

One possible area for improvement would be to use a greater number of features. Aside from the system of Doraisamy and Rüger, all of the systems discussed here limited themselves to one feature class. Harmonic, melodic, timbral and rhythm-based features could all prove useful in characterizing pieces, and a system which incorporates and compares all of these features would likely have improve success rates. This would also allow expanded types of searches. For example, searches based on rhythm only or on orchestration could potentially be performed. The combination of expanded feature classes and

the techniques experimented with in the papers discussed here could hopefully produce systems that are both more flexible and more accurate.

Bibliography

Doraisamy, S. and S. Rüger. 2001. An approach towards a polyphonic music retrieval system. *Proceedings of the International Symposium on Music Information Retrieval.* 187-193.

Doraisamy, S., and S. Rüger. 2002. A comparative and fault-tolerance study of the use of N-grams with polyphonic music. *Proceedings of the International Symposium on Music Information Retrieval.* Available on-line at http://ismir2002.ircam.fr/proceedings\02-FP04-1.pdf.

Pickens, J., et al. 2002. Polyphonic score retrieval using polyphonic audio queries: A harmonic modeling approach. *Proceedings of the International Symposium on Music Information Retrieval.* 140-9.

Song, J., S. Y. Bae, and K. Yoon. 2002. Mid-level music melody representation of polyphonic audio for query-by-humming system. *Proceedings of the International Symposium on Music Information Retrieval.* 133-9.

Wiggins, G., K. Lemström, and D. Meredith. 2002. SIA(M)ESE: An algorithm for transposition invariant, polyphonic content-based music retrieval. *Proceedings of the International Symposium on Music Information Retrieval.* 283-4.