Analysis of a Contour-based Representation for Melody

Abstract

Identifying a musical work from a melodic fragment is a task that most people are able to accomplish with relative ease. For some time now researchers have worked to give computers this ability as well, as it would be the cornerstone of any *query-by-humming* system. To accomplish this, it is reasonable to study how humans are able to perform this task, and to assess what features we use to determine melodic similarity. Research has shown that melodic contour is an important feature in determining melodic similarity, but it is also clear that rhythmic information is important as well. The goal of this research is to explore what variation of contour and rhythmic information can result in the most efficient, robust, and scalable representation for melody. We intend for this to be the basis of a query-by-humming system that will be used to test the validity of our proposed representation.

The importance of melodic contour

The literature suggests that a coarse melodic contour description is more important to listeners than strict intervals in determining melodic similarity. Experiments have shown that interval direction alone (i.e. the 3-level +/-/0 contour representation) is an important element of melody recognition. There is, of course, anecdotal and experimental evidence that humans use more than just interval direction (a 3-level contour) in assessing melodic similarity. In an experiment by Lindsay (1996), subjects were asked to repeat (sing) a melody that was played for them. He found that while there was some correlation between sung interval accuracy and musical experience, even musically inexperienced subjects were able to negotiate different interval sizes fairly successfully. From a practical standpoint, a 3-level representation will generally require longer queries to arrive at a unique match. Given the perceptual and practical considerations, we chose to explore finer (5- and 7-level) contour divisions for our representation.

Proposed melody representation

We used a triple <*T P B*> to represent each melody, where *T* is the time signature of the song, *P* is the pitch contour vector, and *B* is the beat number vector. The range of values of *P* vary depending on the number of levels of contour used, but follow the pattern of 0, +, -, ++, --, +++, etc. The first value of *B* is the location of the first note within its measure in beats (according to the time signature). Successive values of *B* are incremented according to the number of beats between successive notes. Values of *B* are quantized to the nearest whole beat. Additionally, we used a vector *Q* to represent different contour resolutions and quantization boundaries. The length of *Q* indirectly reveals the number of levels of contour being used, and the individual values of *Q* indicate the absolute value of the quantization boundaries (in number of half-steps). For example, *Q* = [0 1] represents that we quantize interval changes into three levels, 0 for no change, + for an ascending interval (a boundary at one half-step or more), and - for a descending interval. This representation is equivalent to the popular +/-/0 or U/D/R (up/down/repeat) representation. *Q* = [0 1 3] represents a quantization of intervals into five levels, 0 for no change, + for an ascending half-step or whole-step (1 or 2 half-steps), ++ for ascending at least a minor third (3 or more half-steps), - for a descending half-step or whole-step, and -- for a descent of at least a minor third. Thus far, we have assembled a data set of 50 multi-track MIDI files, containing a mixture of popular and classical music. The popular music selections span a variety of different countries. All selected songs had a separate monophonic melody sound track.

Results

In spite of anecdotal evidence, we wanted to explicitly verify the usefulness of rhythmic information in comparing melodic similarity. To test this, we used the simplest contour (3-levels, Q=[0 1]) for queries with and without the rhythmic information vector, B. Our results clearly indicate that rhythmic information allows for much shorter (and thus more efficient) queries. For the 5- and 7-level contours, we also examined a variety of quantization boundaries (different vectors Q_k). Our results showed that the performance of 5-level contours are generally better than the 3-level contour, and 7-levels is better than that. For quantization vectors, we limited our search to Q_k = [0 1 x …] cases only. Other values would have caused repeated notes (no interval change) to be grouped in the same quantization level as some amount of interval change, which does not make sense perceptually.

What is illuminating is that the best 5-level contour was able to equal the performance of the 7-level contour. This suggests that a 5-level contour may be an optimal tradeoff between efficiency and robustness to query variation (more levels will cause more variations in queries). Given this result, it is revealing to examine the histogram of interval occurrences in our data set. An optimal quantizer would divide the histogram into sections of equal area. This was approximately true for the Q = [0 1 3] case, which has the best performance. No interval change (0) occurs about 23% of the time. Ascending half-steps and whole-steps (+1 and +2) are about 21% of the intervals, whereas descending half- and whole-steps (-1 and -2) represent approximately 23%. Other choices for quantization boundaries clearly have less-optimal probability distributions, which is why they do not perform as well.

While this result is dependant on the statistics of the data set, it is worth noting that it also correlates well with our knowledge of melody perception. Others have noted the apparent correlation of statistical independence and perceptual importance in acoustic features, which supports a theory of perception evolving from statistical efficiency. Perhaps it is not surprising that these relationships may exist in higher-level features, such as melody, as well. Some surely will argue the reverse causality: that human perception has driven the statistics of melody, resulting in a distribution of intervals that is pleasing to human perception. Either way, it is a useful relationship that perhaps has not yet been fully exploited. The statistical features of this description for melody result in an efficient representation. And since the representation correlates well with our perception of melody, the representation becomes more robust since our queries are likely to be more accurate.

Author Information

Youngmoo E. Kim, Wei Chai, Ricardo Garcia, Barry Vercoe
Machine Listening Group
MIT Media Lab
{moo,chaiwei,rago,bv}@media.mit.edu

Suggested Readings

Lindsay, Adam T. 1996. *Using contour as a mid-level representation of melody*. Unpub. MS thesis. MIT Media Lab.

McNab, R. J. *et al*. 1996. "Toward the digital music library: tune retrieval from acoustic input." *Proc. ACM Digital Libraries,* Bethesda. http://www.nzdl.org.

MiDiLiB, University of Bonn, http://leon.cs.uni-bonn.de/forschungprojekte/midilib/english.

Themefinder™, Stanford University, http://www.ccarh. org/themefinder.

TuneServer, University of Karlsruhe, http://wwwipd.ira. uka.de/tuneserver.