Perceptual Issues in Music Pattern Recognition - Complexity of Rhythm and Key Finding

Ilya Shmulevich, Olli Yli-Harja
Signal Processing Laboratory
Tampere University of Technology, Finland
E-mail: *ilya@cs.tut.fi*

Edward Coyle
School of Electrical and Computer Engineering
Purdue University, USA

Dirk-Jan Povel
Nijmegen Institute for Cognition and Information
University of Nijmegen, The Netherlands

Kjell Lemström
Department of Computer Science
University of Helsinki, Finland

February 23, 1999

Abstract

We consider several perceptual issues in the context of machine recognition of music patterns. It is argued that a successful implementation of a music recognition system must incorporate perceptual information and error criteria. We discuss several measures of rhythm complexity which are used for determining relative weights of pitch and rhythm errors. Then, a new method for determining a localized tonal context is proposed. This method is based on empirically derived key distances. The generated key assignments are then used to construct the perceptual pitch error criterion which is based on note relatedness ratings obtained from experiments with human listeners.

1 Introduction

In this work, we explore the ability of a computer to recognize patterns in music in a perceptually and musically meaningful manner. Our discussion here will be centered around a system for machine recognition of music patterns introduced by Coyle and Shmulevich (1997). Such a system is intended to be used for retrieval of music information from large music databases. However, as the title suggests, we will concentrate mostly on perceptual matters related to music pattern recognition.

The ability to retrieve music by content (and not by secondary information such as title, composer, lyrics, etc.) may someday have a major impact on the music industry. By incorporating research in music perception and cognition, the music recognition system becomes a bit more like a human being, using what is known about how humans perceive, memorize, and reproduce music patterns. When a human being attempts to reproduce a (possibly incorrectly) memorized piece of music, say by singing or humming it, he or she is likely to introduce errors. However, these errors are, more often than not, musically meaningful ones. It is precisely this type of knowledge that this system exploits.

We frame the problem of recognition of musical patterns as a classical pattern recognition problem in the sense that an error between a *target* (query) pattern and *scanned* pattern from a database is to be minimized. The main difference, however, between our approach and traditional approaches is that the error criterion used to judge the goodness of a match between the two patterns is derived from perceptual studies. This error takes into account pitch and rhythm information. Section 2 briefly reviews the components of the music pattern recognition system considered here. In Section 3, we discuss the role of rhythm complexity in determining relative weights of pitch and rhythm errors. Then, in Section 4, we focus on key-finding algorithms which are used in the pattern recognition system for the formation of the perceptual pitch error criterion.

2 The Music Pattern Recognition System

Melodies are perceptually invariant under a multiplicative transformation of frequencies; hence, pitch relations rather than absolute pitch features underlie the perceptual identity of a melody (Hulse et al., 1992). Since it is this relative information that is encoded, it is precisely that same information that needs to be represented on a computer. Taking this into account, we only need to represent the differences of notes, rather than the notes themselves. So, for a sequence $[q_1, q_2, \cdots, q_n]$ of n notes, we define a *pitch difference vector*

$$\mathbf{p} = [p_1, p_2, \cdots, p_{n-1}], \text{ where } p_i = q_{i+1} - q_i$$

as an encoding of the sequence of notes. Note that the q_i are absolute pitch values, defined according to, say, the MIDI standard and thus p_i are the number of semitones (positive or negative) from q_i to q_{i+1}.

Representation of rhythm information also relies on a perceptual invariance under a change of tempo. This type of invariance is linked to the fact that changes in tempo maintain constant durational ratios among structural elements (Hulse et al., 1992). Similar to pitch representation, we represent ratios of durations rather than the durations themselves. When encoding or memorizing rhythmic patterns, we register times of occurrence of the notes within the metrical structure, rather than the durations of the notes. Because of this fact, we will prefer to use a new notion referred to as the *term* of a note, which we will define to be the time between consecutive note onsets. To this end, for a sequence $\mathbf{d} = [d_1, d_2, \cdots, d_n]$ of terms, we define a *rhythm difference vector*

$$\mathbf{r} = [r_1, r_2, \cdots, r_{n-1}], \text{ where } r_i = \frac{d_{i+1}}{d_i}$$

as an encoding of the sequence of terms.

The overall error (distance) between a target pattern and a scanned pattern is a combination of both pitch and rhythm errors. Let us express this relationship as

$$e = \sigma \cdot e_q + (1 - \sigma) \cdot e_r$$

where e_q represents the pitch error, itself a combination of objective and perceptual pitch errors discussed in Section 4, while e_r represents the rhythm

error (see (Coyle and Shmulevich, 1997) for details). Consequently, the weights of the two error components need to be determined. The idea is that target patterns with relatively simple rhythm complexity, which occur frequently in music, should contribute less to the rhythm error e_r, than more complex rhythms. We now discuss this further.

3 Rhythm Complexity

In (Shmulevich and Povel, 1998), three new measures of rhythm complexity are examined. We argue that a successful and perceptually salient measure of rhythm complexity can be used in the music pattern recognition system described above by allowing it to determine relative weights of pitch and rhythm errors. The first measure is based on the work of Tanguiane (1994) and uses the idea that a rhythmic pattern can be described in terms of (elaborations of) more simple patterns, simultaneously at different levels. The second measure is based on the complexity measure for finite sequences proposed by Lempel and Ziv (1976), which is related to the number of steps in a self-delimiting production process by which such a sequence is presumed to be generated. Finally, the third measure proposed is rooted in the theoretical framework of rhythm perception discussed in (Povel and Essens, 1985). This measure takes into account the ease of coding a temporal pattern and the (combined) complexity of the segments resulting from this coding.

First, we must define the domain of rhythms studied. We restrict ourselves to quantized rhythms, i.e. rhythms as notated in a score, without timing deviations due to performance. Furthermore, the rhythms studied are supposed to repeat or loop infinitely and thus form an infinite sequence of events. We notate a rhythmic pattern as a string of ones and zeros, in which the symbol '1' represents a note onset and '0' represents no note onset. We now proceed to discuss the proposed measures.

3.1 T-Measure (Tanguiane measure)

Consider dividing the quarter note into elaborations (Mont-Reynayd and Goldstein, 1985) or rhythmic patterns of equal total duration. Such a subdivision forms a partially ordered set. In the case of dividing a quarter note into patterns containing notes with durations no smaller than a sixteenth, we form the Boolean lattice on 3 generators, E^3, shown in Figure 1. In this example, the elements of this lattice can be coded as binary strings of length 3. Tanguiane shows how a rhythmic pattern can be described by rhythmic configurations at several levels simultaneously, e.g., at the eighth note level, at the quarter note level, and so on. Of course, for each such level, we use an appropriate partially ordered set similar to the one above. At each such level, some patterns are elaborations of other patterns. The patterns which are not elaborations of any other pattern are called *root patterns*. The complexity of the rhythmic pattern is defined by taking the maximum number of root patterns, over all possible structural levels, required to generate the rhythmic pattern in question. A bijective mapping can be established between the set of root patterns and the set of minimal primes of a monotone Boolean function. It is well known that the maximum number of minimal primes of a monotone Boolean function of n variables is equal to $\binom{n}{\lfloor n/2 \rfloor}$ and hence determines the maximum possible number of root patterns and consequently the maximum complexity under the T-measure.

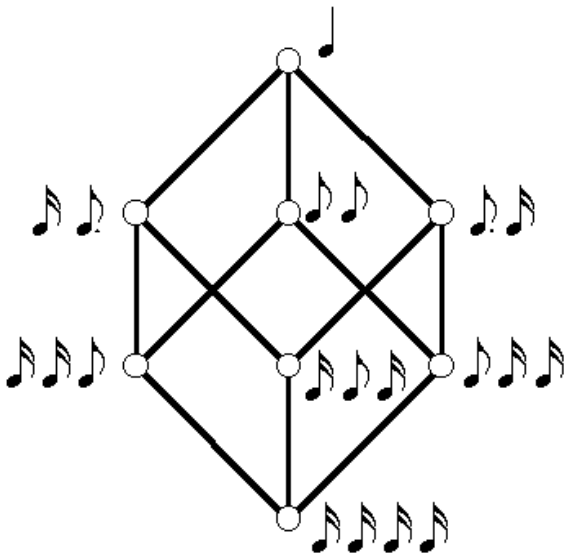

Figure 1: Elaborations of a quarter note

3.2 LZ-Measure (Lempel-Ziv measure)

Another approach for quantifying complexity of rhythms is to use the popular measure proposed by Lempel & Ziv (1976). Essentially, this complexity measure captures the number of "new" substrings found as the sequence evolves from left to right (as is the case in music). As soon as a new substring is found, the complexity increases by 1. The measure essentially takes into account repetitions of patterns on all structural levels. It should be pointed out, however, that the LZ complexity in general is not well suited for very short sequences and thus the assumption of cyclical rhythms is very useful. The measure is intended to capture the multi-level redundancy embedded in the rhythmic pattern without regard to any perceptual mechanisms involved in coding it. Thus, the measure does not take into account the possibility that some of the information embedded in the sequence may not be perceptually relevant to a human listener. Therefore, it can be used as a reference point for other measures that do incorporate perceptual constraints in that they should exhibit greater correspondence to subjective judgements of complexity than the LZ-Measure.

3.3 PS-Measure (Povel-Shmulevich measure)

The PS-Measure is rooted in the theoretical framework of rhythm perception discussed in (Povel and Essens, 1985). A basic notion of the model is that a listener attempts to establish an internal clock (beat) that segments the rhythm into equal intervals. The induction of the clock is determined by the distribution of accents in the sequence. For any given rhythm, a number of possible clocks can be induced. However, it is assumed that the clock which best fits the distribution of accents in the rhythm is the one actually induced. This clock is referred to as the best clock. Furthermore, the ease with which the best clock is induced depends on how well it fits the distribution of accents. After having chosen the best clock, the listener codes the segments produced by this clock.

Discussing the complexity of rhythms, the authors state that a "given temporal pattern will be judged complex when either no internal clock is induced or, where it is induced, when the coding of the pattern is relatively complex" (Povel and Essens, 1985). In light of that, the proposed measure of complexity should be a combination of the induction strength of the best clock on the one hand and the efficiency of coding of the rhythm on the other. The first part of the PS-Measure thus pertains to the induction strength of the best clock, which is captured by the C-score (Povel and Essens, 1985). The C-score is computed by taking into account a weighted combination of the number of clock ticks that coincide with unaccented events and with silence:

$$C = W \cdot s_e + u_e \tag{1}$$

where s_e stands for the number of clock ticks coinciding with silence and u_e with the number of unaccented events. The lower the score, the higher the induction strength of the clock; hence higher scores correspond to higher complexity.

The second part of the PS-Measure pertains to the efficiency of the code. In determining coding complexity, we distinguish between four types of possible segments: an empty segment (E), an equally subdivided segment (S_k, where k indicates the number of equal subdivisions), an unequally subdivided segment (U), and finally a segment which begins with silence (N). To compute the coding complexity, a different weight is associated with each type of segment. Weights $d_1, \cdots, d_4$ correspond respectively to the four types of segments distinguished above. Finally, a weight d_5 is used in order to account for repetitions of segments. Specifically, if a segment is different from the segment following it, a value of d_5 is added to the sum of all weights accumulated so far. The rationale behind this is that two different consecutive segments are likely to increase complexity. Now, the total coding complexity can be expressed as:

$$D = \sum_{i=1}^{n} c_i + m \cdot d_5 \tag{2}$$

where $c_i \in \{d_1, \cdots, d_4\}$ is the weight of the i^{th} segment, n is the number of segments, and m is the number of consecutive segment pairs containing different segments.

Finally, the PS-Measure is defined as the weighted combination of the induction strength of

the clock and the total coding complexity:

$$P = \lambda \cdot C + (1 - \lambda) \cdot D \tag{3}$$

where C is the induction strength of the best clock and D is the total coding complexity obtained by segmenting the rhythm with that clock. Two parameters which must be determined are W and λ, where W is the weight used in (1) to compute C while λ represents the relative importance of clock induction strength and coding efficiency.

All parameters were determined by utilizing the results of an experiment reported in (Essens, 1995). Experiment 3 of that work consisted of asking twenty human listeners to make complexity judgements on 24 rhythmic patterns, on a scale of 1 to 5. All parameters were optimized so as to increase the correlation between the average judged complexity reported in (Essens, 1995) and the PS-Measure. To achieve this, simplex search as well as quasi-Newton search methods were used. The resulting correlation between the average judged complexities and the PS-Measure complexities computed with these parameters is $r = 0.83$. This measure was subsequently tested by applying it to a new set of data containing complexity judgments and found to be reliable (Shmulevich and Povel, submitted).

The T-Measure, based on the work of Tanguiane, was the poorest performer. The LZ-Measure also performed poorly, but this was most likely due to the very short lengths of rhythms on which it was used. The PS-Measure is the most promising in that it incorporates perceptual information and is based on an empirically tested model of rhythm perception.

4 Key Finding

Let us again return to the music pattern recognition system. For the pitch error component, we wish to be able to reflect differences of contour - the direction of pitch change from one note to the next - in our error. Our objective pitch error is defined as $e_o = \|\mathbf{p} - \mathbf{p_0}\|_1$, where $\mathbf{p}$ and $\mathbf{p_0}$ are the pitch difference vectors of the scanned and target patterns. The L_1-norm is chosen (as opposed to L_p, $p \neq 1$) for lack of any apparent reason to bias the error in favor or against small or large increments in pitch. This norm, at this stage of the pitch error, reflects the differences of contour between the target and scanned patterns without bias. The bias will come into play when we incorporate quantified perceptual information.

Performing classification based solely on the objective pitch error would not take into account the fact that intervals of equal size are not perceived as being equal when the tones are heard in tonal contexts (Krumhansl and Shepard, 1979). Such phenomena cannot be embodied by the objective pitch error alone.

Since the ultimate goal is to recognize a target pattern memorized (possibly incorrectly) by a human being, it is important to consider certain principles of melody memorization and recall. For example, findings showed that "less stable elements tended to be poorly remembered and frequently confused with more stable elements." Also, when an unstable element was introduced into a tonal sequence, "... the unstable element was itself poorly remembered"(Krumhansl, 1990). So, the occurrence of an unstable interval within a given tonal context (e.g., a melody ending in the tones C C♯ in the C major context) should be penalized more than a stable interval (e.g., B C in the C major context) since the unstable interval is less likely to have been memorized by the human user. These perceptual phenomena must be quantified for them to be useful in the classification of musical patterns. Such a quantification is provided by the *relatedness ratings* found by Krumhansl (1990). Essentially, a relatedness rating between tone q_1 and tone q_2 ($q_1 \neq q_2$) is a measure of how well q_2 follows q_1 in a given tonal context. The relatedness rating is a real number between 1 and 7 and is determined by experiments with human listeners. Results are provided for both major and minor contexts. So, a relatedness rating between two different tones in any of 24 possible tonal contexts can be found due to invariance under transposition.

To this end, suppose we are scanning a sequence of n notes to which we compare a target pattern consisting of n notes. For the moment, assuming knowledge of the tonal context of the scanned pattern, we define its vector of relatedness ratings $\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_{n-1}]$ as well as $\beta = [\beta_1, \beta_2, \cdots, \beta_{n-1}]$, the vector of relatedness ratings for the target pattern in the same tonal context. Each α_i and β_i is the relatedness rating between pitches q_i and q_{i+1} in the given tonal context for the scanned and target patterns respectively. Having defined the vec-

tors of relatedness ratings for the scanned and target patterns, we can define the perceptual pitch error to be $e_p = \|\alpha - \beta\|_1$. It is worth noting that if $e_o = 0$, then $e_p = 0$, while the converse is not true. We can combine the objective and perceptual errors into a pitch error

$$e_q = \lambda \cdot e_p + (1 - \lambda) \cdot e_o \tag{4}$$

We have assumed that in the computation of the perceptual pitch error, we had knowledge of the tonal context of the scanned pattern. Thus, the need arises for a localized key finding algorithm which will present us with a most likely tonal context for a given musical pattern which will be subsequently used for the relatedness rating vectors. Such an algorithm was developed by Krumhansl (1990) and is essentially based on the fact that "most stable pitch classes should occur most often" (Takeuchi, 1994). We now make certain modifications to this algorithm and present a method for determining the parameter λ in equation (4).

The algorithm produces a 24-element vector of correlations, $\mathbf{r} = [r_1, \cdots, r_{24}]$, the first twelve for major contexts and the others for minor contexts. The highest correlation, $r_{\max}$, is the one that corresponds to the most likely tonal context of the musical pattern being scanned. Suppose a musical composition (or set of compositions) that we wish to scan for the purpose of recognizing the target pattern consists of m notes and the target pattern itself consists of n notes (typically, $m \gg n$). In our algorithm, we slide a window of length n across the sequence of m notes and for each window position, the key-finding algorithm outputs a key assignment. Thus, we have a sequence $\mathbf{t} = [t_1, t_2, \cdots, t_{m-n+1}]$ of key assignments such that $t_i = \arg\max(\mathbf{r}_i)$. Figure 2 shows a typical sequence of key assignments.

Unfortunately, in practice, there is quite a bit of variation in certain regions of the sequence of key assignments. Common artifacts are impulses and oscillations between modulations (edges). The reasons for this are described in detail in (Shmulevich and Coyle, 1997). This is due to the algorithm's sensitivity to the distribution of pitches within the window. These small oscillations and impulses are undesirable, not only because they do not reflect our notions of modulations and localized tonal context, but primarily because they affect the relatedness rating vectors, which inherently depend on

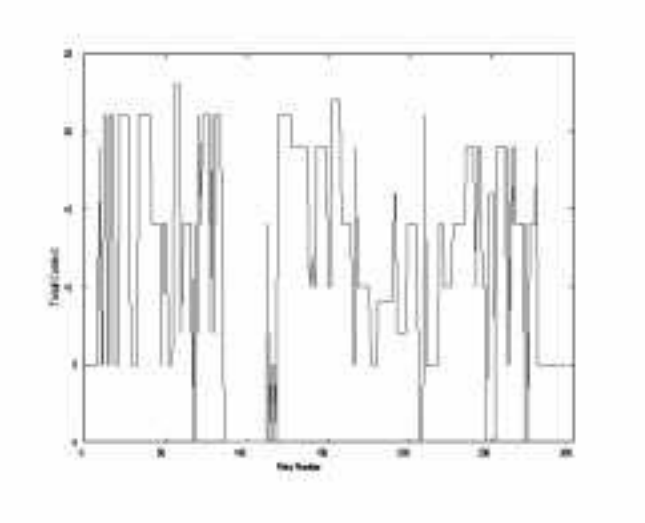

Figure 2: Typical sequence of key assignments

the tonal context produced by the key-finding algorithm. Since the values of the assigned key sequence often appears arbitrary in the regions of oscillation, the perceptual pitch error is distorted in these regions. Therefore, the preferable action is to smooth out those local oscillations. As a solution to this problem, various nonlinear filters, such as the recursive median filter (Nodes and Gallagher, 1983), have been employed (Shmulevich and Coyle, 1997). One difficulty with using such filters is due to the class quality of the input data. That is, there is no natural ordering of the tonal contexts. Suppose that 24 possible classes are numbered (ordered) arbitrarily, and filtered with the recursive median filter, an order statistic filter. If this worked satisfactorily, we would get the same result in all 24! possible input class permutations. Clearly this cannot be the case.

Addressing this problem, we start with key distances derived from experimentally determined tonal hierarchies (Krumhansl, 1990). These key distances provide a quantitative measure of similarity between all 24 tonal contexts. A multidimensional scaling solution then places the key distances into four dimensional Euclidean space. Two dimensions account for the circle of fifths while the other two account for parallel and relative major-minor relationships (Krumhansl, 1990). We would

like to emphasize that these derived key distances possess perceptual validity and musical meaning.

As a next step, we define a graph with 12 major and 12 minor tonal contexts as vertices, and set the edge values to distances from the multidimensional scaling solution. For example, the coordinate of C major is $[0.567, -0.633, -0.208, 0.480]$ and the coordinate of A minor is $[0.206, -0.781, -0.580, 0.119]$. Then, the Euclidean distance between these two keys is 0.6488, which is equal to the weight of the edge between those two vertices. We can then define L_1 (or L_2) norm estimates for the current tonal context as follows. It is well known that the median of n points is the point whose sum of distances to the n points is minimal. In an analogous manner, the "median" of some subset of vertices is the vertex whose sum of distances to other vertices in that set is minimum. Similarly, we can define the "mean" by using squared distances instead. Clearly, such an approach avoids the permutation problem, since no ordering is imposed on the graph, and is thus well suited for estimating the localized tonal context. Moreover, this idea can easily be extended to define other operations on the graph (e.g. weighted median). Let us consider an example.

Suppose that our window contains the following five key assignments: [C major; C major; C# major; C major, A minor]. We estimate the key assignment using the graph-based L_1-norm estimate. For each of the five keys, we compute and sum the distances to the other four keys. Then, we pick the vertex which had the minimum total distance to the rest of the vertices, which in this case is C major. Figure 3 shows the application of this method to the sequence of key-assignments shown in Figure 2.

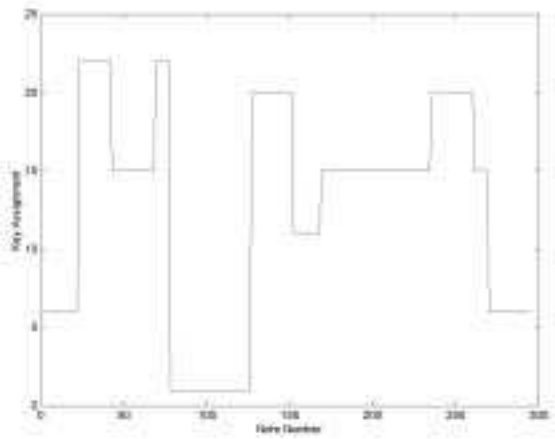

Figure 3: Graph-based L_1-norm estimates of key assignments

5 Conclusion

In this paper, we have considered several perceptual issues in the context of music pattern recognition. We argue that a successful implementation of a music recognition system must incorporate perceptual information and error criteria. For example, a measure of rhythm complexity based on an empirically tested model of rhythm perception and supported by experiments with human listeners is used for determining relative weights of pitch and rhythm errors. The pitch error, which also contains a perceptual error component, relies in part on a key-finding algorithm. This algorithm, in turn, incorporates perceptually and musically meaningful information about key-distances derived from empirical studies.

References

E.J. Coyle and I. Shmulevich. A System for Machine Recognition of Music Patterns. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Seattle, WA, 1998.

P. Essens. Structuring temporal sequences: Comparison of models and factors of complexity. *Perception & Psychophysics*, 57(4):519-532, 1995.

S. H. Hulse, A. H. Takeuchi, and R. F. Braaten. Perceptual Invariances in the comparative psychology of music. *Music Perception*, 10(2):151-184, 1992.

C. L. Krumhansl and R. N. Shepard. Quantification of the hierarchy of tonal functions within a diatonic context. *Journal of Experimental Psychology: Human Perception and Performance*, 5:579-594, 1979.

C. L. Krumhansl, *Cognitive Foundations of Musical Pitch*, New York: Oxford University Press, 1990.

A. Lempel and J. Ziv. On the complexity of finite sequences. *IEEE Transactions on Information Theory*, IT-22(1):75-81, 1976.

B. Mont-Reynaud and M. Goldstein. On Finding Rhythmic Patterns in Musical Lines. *Proceedings of the International Computer Music Conference*, San Francisco, CA, 1985.

T. A. Nodes and N. Gallagher. Median filters: some modifications and their properties. *IEEE Trans. Acoust., Speech, Signal Process.*, 31:739-746, 1983.

D. J. Povel and P. J. Essens. Perception of temporal patterns. *Music Perception*, 2:411-441, 1985.

I. Shmulevich and E.J. Coyle. The Use of Recursive Median Filters for Establishing the Tonal Context in Music. *Proceedings of the 1997 IEEE Workshop on Nonlinear Signal and Image Processing*, Mackinac Island, MI, 1997.

I. Shmulevich and D. Povel. Rhythm Complexity Measures for Music Pattern Recognition. *Proceedings of IEEE Workshop on Multimedia Signal Processing*, Redondo Beach, California, December 7-9, 1998.

I. Shmulevich and D.J. Povel. Measures of Rhythm Complexity. (submitted for publication)

A. H. Takeuchi. Maximum key-profile correlation (MKC) as a measure of tonal structure in music. *Perception & Psychophysics*, 56:335-346, 1994.

A. Tanguiane. A Principle of correlativity of perception and its application to music recognition. *Music Perception*, 11(4):465-502, 1994.