



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

ILYA SHMULEVICH<sup>1</sup>, OLLI YLI-HARJA<sup>1</sup>, EDWARD COYLE<sup>2</sup>, DIRK-JAN  
POVEL<sup>3</sup> and KJELL LEMSTRÖM<sup>4</sup>

<sup>1</sup>*Signal Processing Laboratory, Tampere University of Technology, Finland*  
(E-mail: ilya@cs.tut.fi yliharja@cs.tut.fi);

<sup>2</sup>*School of Electrical and Computer Engineering, Purdue University, USA*  
(E-mail: coyle@purdue.edu);

<sup>3</sup>*Nijmegen Institute for Cognition and Information, University of Nijmegen, The Netherlands*  
(E-mail: povel@nici.kun.nl);

<sup>4</sup>*Department of Computer Science, University of Helsinki, Finland*  
(E-mail: klemstro@cs.helsinki.fi)

**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 paper, 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 (1998). 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 in general and so, the ideas contained herein should be applicable to any music recognition system that uses content (pitch and rhythm) information.

The ability to retrieve music by content (and not by secondary information such as title, composer, lyrics, etc.) will have a major impact on the music industry. By incorporating research in music perception and cognition, a 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 such a 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 a *scanned* pattern from a database is to be minimized. In other words, the result of the query is the pattern or patterns that have the smallest error or distance to the target pattern. 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  $[q_1, q_2, \dots, q_n]$  of  $n$  notes, we define a *pitch difference vector*

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

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  $p_i$  is 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 the onsets of the notes within the metrical structure, rather than the durations of the notes. Because of this fact, we will prefer to use inter-onset intervals (IOI), which are defined to be the times between consecutive note onsets. To this end, for a sequence  $\mathbf{d} = [d_1, d_2, \dots, d_n]$  of IOIs, we define a *rhythm difference vector*

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

as an encoding of the sequence of IOIs.

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

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

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). Briefly, the rhythm error is defined as

$$e_r = \left( \sum_{j=1}^{n-1} \frac{\max(s_j, t_j)}{\min(s_j, t_j)} \right) - (n - 1), \quad (2)$$

where  $\mathbf{s} = [s_1, s_2, \dots, s_{n-1}]$  represents the rhythm difference vector of the scanned rhythm pattern (of length  $n$ ) and  $\mathbf{t} = [t_1, t_2, \dots, t_{n-1}]$  represent the rhythm difference vector of the target pattern. The term  $(n - 1)$  in the above expression is subtracted so that a perfect match of the rhythm difference vectors would result in  $e_r = 0$ .

The relative weight  $\sigma$  of the two error components, namely  $e_q$  and  $e_r$ , is determined on the basis of the complexity of the rhythm patterns in question. The idea behind this is that target patterns with relatively simple rhythm complexity, which occur rather frequently in music, should contribute less to the rhythm error  $e_r$  than more complex rhythms. A rhythm's complexity reflects the amount of information embedded in it. Consequently, if a music pattern contains relatively little "rhythmic content", the overall error between it and another candidate rhythm should be largely based on its "pitch content." After all, a music pattern with a distinctively rich rhythm content can often be recognized and identified even without resorting to pitch information. The next section is devoted to rhythm complexity and its measures.

### 3. Rhythm Complexity

The representation of information, and of rhythms in particular, is achieved via coding. When a human being enters the equation, however, care must be taken in interpreting the notion of complexity, which necessarily becomes subjective. Moreover, depending on the context, only certain types of codes may be perceptually significant and hence coding efficiency or complexity must be considered within such constraints (Chater, 1996). This is well known, for example, in the field of visual perception (Leeuwenberg, 1971).

In Shmulevich and Povel (1998), three new measures of rhythm complexity are examined. We argue here that a 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. This measure presupposes the existence of a “temporal grid” or time scale consisting of isochronic intervals, which is selected among a set of possible grids according to the “economy principle” (Povel, 1984).

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, all rhythms are assumed 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-Reynaud 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 (1994) 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 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 true vectors (lower units) of a monotone Boolean function. It is well known (Gilbert, 1954) that the maximum number of minimal true vectors 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 (Shmulevich and Povel, to appear).

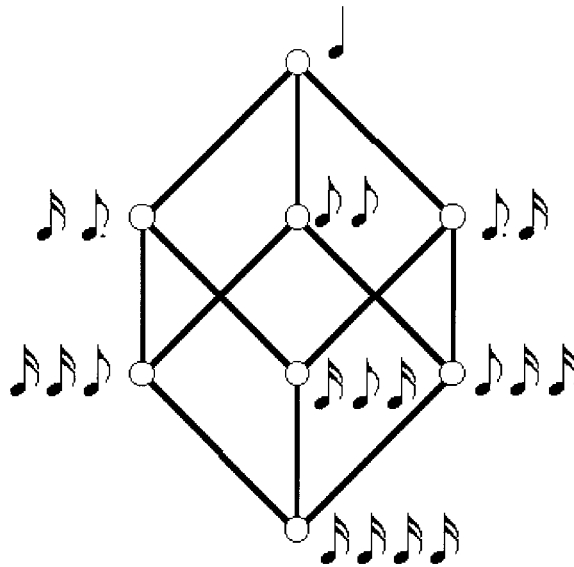


Figure 1. Elaborations of a quarter note.

### 3.2. LZ-MEASURE (LEMPERL-ZIV MEASURE)

Another approach for quantifying complexity of rhythms is to use the popular measure proposed by Lempel and 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 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 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. Presumably, this temporal segmentation serves to reduce the

coding complexity of the stimulus, which would be consistent with the Gestalt simplicity principle, implying that sensory input is coded in the simplest possible way (Chater, 1996). The induction of the clock is determined by the distribution of accents in the sequence (Povel and Essens, 1985). 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, Povel and Essens (1985) 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.” 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:

$$C = W \cdot s_e + u_e, \quad (3)$$

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  $d_1, \dots, d_4$  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:

$$D = \sum_{i=1}^n c_i + m \cdot d_5, \quad (4)$$

where  $c_i \in \{d_1, \dots, d_4\}$  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:

$$P = \lambda \cdot C + (1 - \lambda) \cdot D, \quad (5)$$

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  $\lambda$ , where  $W$  is the weight used in (3) to compute  $C$  while  $\lambda$  represents the relative importance of clock induction strength and coding efficiency.

All parameters were determined by utilizing the results of an experiment reported by 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 by 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 was  $r = 0.83$ . This measure was subsequently tested by applying it to a new set of data containing complexity judgments and was found to be reliable (Shmulevich and Povel, to appear).

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 (Shmulevich and Povel, to appear). Therefore, the PS-Measure is a good candidate for determining the relative weights of the pitch and rhythm errors.

#### 4. Key Finding

Let us again return to the music pattern recognition system and focus on the pitch error  $e_q$ . For the pitch error component of the overall error 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, respectively. 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 by invoking transpositional invariance. Consequently, a relatedness rating can be defined solely in terms of intervals.

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  $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_{n-1}]$  as well as  $\beta = [\beta_1, \beta_2, \dots, \beta_{n-1}]$ , the vector of relatedness ratings for the target pattern in the same tonal context. Each  $\alpha_i$  and  $\beta_i$  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 vectors of relatedness ratings for the scanned and target patterns, we can define the perceptual pitch error to be  $e_p = \|\alpha - \beta\|_1$ . We can combine the objective and perceptual errors into a pitch error

$$e_q = \lambda \cdot e_p + (1 - \lambda) \cdot e_o. \quad (6)$$

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. In fact, the problem of automated key finding in music is a prerequisite to successful automation of music analysis, since the determination of key is necessary for meaningful coding of melodic and harmonic events (Krumhansl, 1990, p. 77). Such an algorithm was developed by Krumhansl (1990) and is based on the fact that “most stable pitch classes should occur most often”

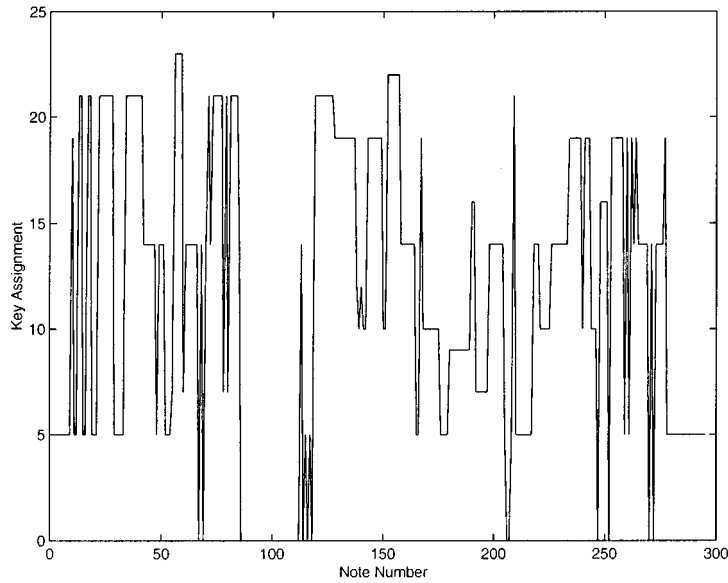


Figure 2. Typical sequence of key assignments.

(Takeuchi, 1994). We now make certain modifications to this algorithm and present a method for determining the parameter  $\lambda$  in equation (6).

The algorithm produces a 24-element vector of correlations,  $\mathbf{r} = [r_1, \dots, r_{24}]$ , 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 music 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  $\mathbf{t} = [t_1, t_2, \dots, t_{m-n+1}]$  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, 1997a, b). 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 the tonal context produced by the key-finding algorithm. Since the values of the assigned key sequence often appear arbitrary in the regions of oscillation, the perceptual pitch error is distorted in these regions. Therefore, the preferable action is to smoothen 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, 1997b).

One difficulty with using such filters is due to the class quality of the input data. In the field of psychology, this type of data is referred to as nominal scale data (Bartz, 1988, pp. 1–21). Sequences of key assignments are examples of class data, since there is no natural numerical ordering of the keys. Suppose that 24 possible classes (keys) are numbered or ordered arbitrarily, and filtered with the recursive median filter, an order statistic filter. One property common to all such filtering schemes is that they inherently depend on some ordering of the data to be filtered. If this worked satisfactorily, we would get the same result in all  $24!$  possible input class permutations. However, this is not the case. To address this problem, Shmulevich and Yli-Harja (to appear) propose a method of smoothing sequences of key assignments using graph theoretic norm estimates.

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 a 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. The operation of the graph-based norm estimate for localized key finding, applied in a sliding window fashion, is explained in the Appendix. 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 result of applying this method to the sequence of key-assignments shown in Figure 2.

Having obtained a reliable estimate of the local key or tonal context, we now briefly discuss how the parameter  $\lambda$  – the weight given to the perceptual pitch error – can be determined. It has been shown that the maximum correlation,  $r_{\max}$ , is strongly correlated with the degree of tonal structure (Takeuchi, 1994). Therefore, if  $r_{\max}$  is small, indicating a low degree of tonal structure, we should have less faith in the output of the localized key-finding algorithm and so, the perceptual pitch error,  $e_p$ , should have less significance. Thus, it is reasonable to relate  $\lambda$  directly to

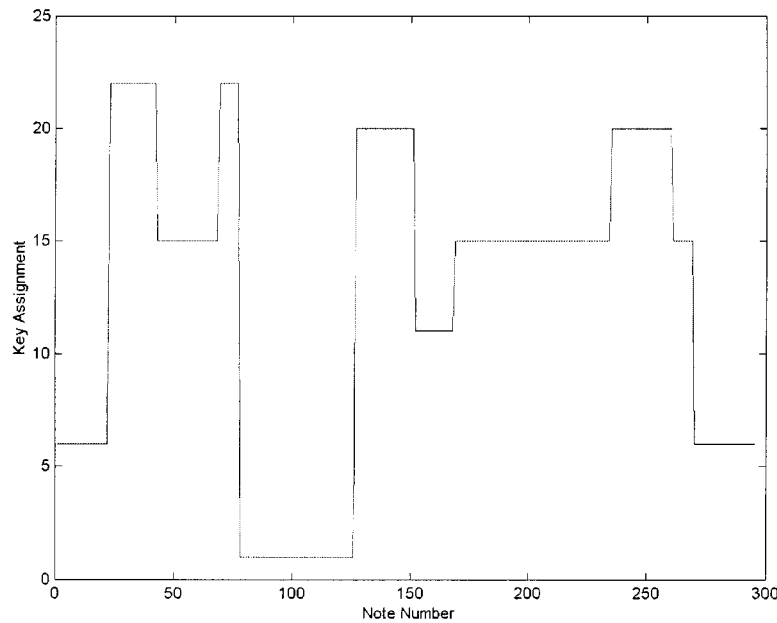


Figure 3. Graph-based  $L_1$ -norm estimates of key assignments.

an appropriately scaled  $r_{\max}$ . This is discussed by Shmulevich and Coyle (1997a), where a cubic smoothing spline is first applied to the sequence  $r_{\max}(i)$  of maximum correlations.

## 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 in order to be useful in practice. 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 localized key-finding algorithm. This algorithm, in turn, incorporates perceptually and musically meaningful information about key-distances derived from empirical studies.

## Appendix: Graph-based Smoothing of Class Data

The present method was introduced in Yli-Harja et al. (1999) and is motivated by an analogy with the median filter, while taking into account the class-quality of the data,

namely, the keys. In the case of real numbers, it is well known that the median of  $(X_1, X_2, \dots, X_n)$ ,  $X_i \in \mathbb{R}$ , is the value  $\beta$  minimizing

$$\sum_{i=1}^n |X_i - \beta|.$$

More formally,

$$\text{med}\{X_1, X_2, \dots, X_n\} = \arg \min_{\beta \in \{X_1, \dots, X_n\}} \sum_{i=1}^n |X_i - \beta|. \quad (7)$$

Similarly, the mean value of  $(X_1, X_2, \dots, X_n)$  is

$$\text{mean}\{X_1, X_2, \dots, X_n\} = \arg \min_{\beta \in \mathbb{R}} \sum_{i=1}^n (X_i - \beta)^2. \quad (8)$$

Equations (7) and (8) are both estimates of location, using the  $L_1$ - and  $L_2$ -norm, respectively. This, of course, presupposes the existence of a metric space and the standard properties of distances necessarily hold; i.e. distance from  $A$  to  $B$  is equal to the distance from  $B$  to  $A$  (symmetry), distance from  $A$  to itself is zero, and distance from  $A$  to  $B$  plus the distance from  $B$  to  $C$  is not less than the distance from  $A$  to  $C$  (triangle inequality). One of our goals, however, is to relax the requirement of a metric space while still being able to make estimates. Thus, even though our samples may possess no numerical properties in that they arise from class data, we can still allow arbitrary “distances” between them and the above metric rules need not apply. This idea is formalized below, where the classes from which the samples come are represented by vertices on a graph and the distances between the classes are weights given to edges joining two vertices. In the context of key finding, these weights or distances would represent interkey distances.

Consider a complete undirected weighted graph  $G(V, E)$  with vertex set  $V$ , edge set  $E$  and a weight function  $w : V \times V \rightarrow \mathbb{R}$ . Let us suppose that  $w(v, v) = 0$  for all  $v \in V$ . Suppose now that we have some set of samples  $A = \{V_1, V_2, \dots, V_n\}$ ,  $V_i \in V$  of graph  $G$ . In a similar manner to (7) and (8), we can define

$$\text{graph-}p(A) = \arg \min_{\beta \in A} \sum_{i=1}^n w(V_i, \beta)^p \quad (9)$$

to be the graph-based  $L_p$ -norm estimate. The values of  $p = 1$  and  $2$  correspond to graph-based median and mean, respectively. Note that the estimate is necessarily one of the vertices under consideration. Also, vertices may be repeated; that is, it is possible that  $V_i = V_j$  for  $1 \leq i < j \leq n$ .

Similarly to the median filter, we can define a sliding window filtering operation based on (9) as

$$Y_i = \text{graph-}p[X_{i-k}, \dots, X_i, \dots, X_{i+k}], \quad (10)$$

where  $\{X_i\}$  is the sequence of input class data and  $\{Y_i\}$  is the sequence of output class data, with  $n = 2k + 1$  being the filter window width. Analogously to the recursive median filter, the graph-based filter can also be applied recursively as

$$Y_i = \text{graph-}p[Y_{i-k}, \dots, Y_{i-1}, X_i, \dots, X_{i+k}].$$

## References

- Bartz, A.E. "Some Thoughts on Measurement". In *Basic Statistical Concepts*, 3rd edn. New York, NY: MacMillan, 1988.
- Chater, N. "Reconciling Simplicity and Likelihood Principles in Perceptual Organization". *Psychological Review*, 103 (1996), 566–581.
- Coyle, E.J. 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.
- Essens, P. "Structuring Temporal Sequences: Comparison of Models and Factors of Complexity". *Perception and Psychophysics*, 57(4) (1995), 519–532.
- Gilbert, E.N. "Lattice Theoretic Properties of Frontal Switching Functions". *Journal of Mathematical Physics*, 33(1) (1954), 57–67.
- Hulse, S.H., A.H. Takeuchi and R.F. Braaten. "Perceptual Invariances in the Comparative Psychology of Music". *Music Perception*, 10(2) (1992), 151–184.
- Krumhansl, C.L. and R.N. Shepard. "Quantification of the Hierarchy of Tonal Functions Within a Diatonic Context". *Journal of Experimental Psychology: Human Perception and Performance*, 5 (1979), 579–594.
- Krumhansl, C.L. *Cognitive Foundations of Musical Pitch*. New York: Oxford University Press, 1990.
- Leeuwenberg, E.L. "A Perceptual Coding Language for Visual and Auditory Patterns". *American Journal of Psychology*, 84(3) (1971), 307–349.
- Lempel, A. and J. Ziv. "On the Complexity of Finite Sequences". *IEEE Transactions on Information Theory*, IT-22(1) (1976), 75–81.
- Mont-Reynaud, B. and M. Goldstein. "On Finding Rhythmic Patterns in Musical Lines". *Proceedings of the International Computer Music Conference*. San Francisco, CA, 1985.
- Nodes, T.A. and N. Gallagher. "Median Filters: Some Modifications and Their Properties". *IEEE Trans. Acoust., Speech, Signal Process.*, 31 (1983), 739–746.
- Povel, D.J. "A Theoretical Framework for Rhythm Perception". *Psychological Research*, 45(4) (1984), 315–337.
- Povel, D.J. and P.J. Essens. "Perception of Temporal Patterns". *Music Perception*, 2 (1985), 411–441.
- Shmulevich, I. and E.J. Coyle. "Establishing the Tonal Context for Musical Pattern Recognition". *Proceedings of the 1997 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. New Paltz, N.Y., 1997a.
- Shmulevich, I. 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, 1997b.
- Shmulevich, I. and O. Yli-Harja. "Localized Key-Finding: Algorithms and Applications". (to appear in) *Music Perception*.
- Shmulevich, I. 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.
- Shmulevich, I. and D.J. Povel. "Measures of Temporal Pattern Complexity". (to appear in) *Journal of New Music Research*.
- Takeuchi, A.H. "Maximum Key-Profile Correlation (MKC) as a Measure of Tonal Structure in Music". *Perception and Psychophysics*, 56 (1994), 335–346.
- Tanguiane, A. "A Principle of Correlativity of Perception and Its Application to Music Recognition". *Music Perception*, 11(4) (1994), 465–502.
- Yli-Harja, O., I. Shmulevich and K. Lemström. "Graph-Based Smoothing of Class Data with Applications in Musical Key Finding". *Proceedings of IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing*. Antalya, Turkey, 1999.

