A SYSTEM FOR MACHINE RECOGNITION OF MUSIC PATTERNS

Edward J. Coyle

Electrical and Computer Engineering
Purdue University
W. Lafayette, IN, USA

Ilya Shmulevich

NICI
University of Nijmegen
The Netherlands

ABSTRACT

We introduce a system for machine recognition of music patterns. The problem is put into a pattern recognition framework in the sense that an error between a target pattern and scanned pattern is minimized. The error takes into account pitch and rhythm information. The pitch error measure consists of an absolute (objective) error and a perceptual error. The latter depends on an algorithm for establishing the tonal context which is based on Krumhansl's key-finding algorithm. The sequence of maximum correlations that it outputs is smoothed with a cubic spline and is used to determine weights for perceptual and absolute pitch errors. Maximum correlations are used to create the assigned key sequence, which is then filtered by a recursive median filter to improve the structure of the output of the key finding algorithm. A procedure for choosing weights given to pitch and rhythm errors is discussed.

1. PITCH AND RHYTHM REPRESENTATION

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 [1]. 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 difference of pitch 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 [1]. 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 difference of rhythm 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.

2. MUSIC PATTERN RECOGNITION

Suppose that a human user has memorized a musical pattern (melody) and wishes to locate it in a large set of musical compositions. We will refer to the memorized pattern as the target pattern. The pattern to which the target pattern will be compared for purposes of classification will be referred to as the scanned pattern. Our goal is to minimize the error between the target pattern and the scanned pattern being considered. The overall error, under the chosen norm, is comprised of the pitch error and the rhythm error. Perceptual information will play a role in the computation of the pitch error. We will consider these two errors separately.

2.1. Pitch Error

As a first step toward computing the error between the target and scanned pattern, 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 $e_o = \|\mathbf{p} - \mathbf{p_0}\|_1$. 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 following fact. All intervals of equal size are not perceived as being equal when the tones are heard in tonal contexts [4]. For example, the notes B C played in succession heard in the context of C Major (for instance, after hearing a strong key-defining sequence of notes) would be perceived as being more natural and stable than the same two notes heard in the context of D Major. Such phenomena cannot be embodied by the objective pitch error alone. Perceptual information has been successfully incorporated into the design of error criteria for various applications. For example, visual error criteria based on the human visual system was used in [2].

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" [3, p. 283]. 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 [3, p. 125]. 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 $\beta = \left[\beta_1, \beta_2, \cdots, \beta_{n-1}\right]$, 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 vectors 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

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

2.2. Establishing the Tonal Context

The above discussion assumed that in the computation of the perceptual pitch error, we had knowledge of the tonal context of the scanned pattern. Before proceeding, we should pose the question: "What exactly is the meaning of the tonal context of a pattern?" Surely, if the pattern is of short length (1 note, for example), then speaking about its tonal context is meaningless. Similarly, if the pattern is very long, it may consist of several tonal contexts and the transitions between them are called modulations. Finally, quite often, a tonal context is a matter of degree in that for a given pattern, there are several possible candidates for tonal context. So, just because the key-signature of a given composition happens to be F major, for example, it does not imply that the relatedness rating vectors α and β must be chosen for that particular tonal context, since modulations and shifting tonal centers are likely to occur.

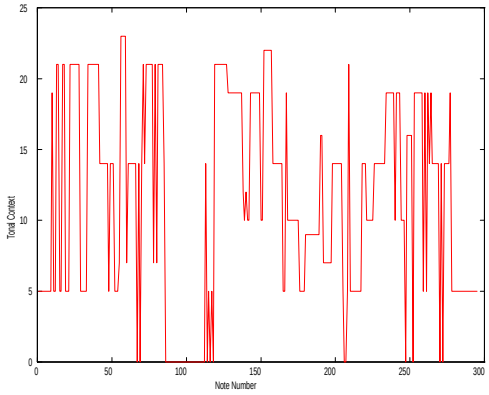

Figure 1: Assigned Key Sequence

We thus see the need for a key-finding algorithm which will present us with a most likely tonal context for a given musical pattern and this tonal context will be subsequently used for the relatedness rating vectors. Such an algorithm was developed by Krumhansl [3, p. 77] and is essentially based on the fact that "most stable pitch classes should occur most often" [7]. That is, tones that are sounded most frequently are the ones with high probe tone ratings, in a given tonal context (e.g., in a C major context, C, G, and E occur most often). We now make certain modifications to this algorithm and present a method for determining the parameter λ. We refer you to the algorithm described in [3, p. 77].

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 $t_i = \arg\max(\mathbf{r}_i)$. See Figure 1 for an assigned key sequence. The composition used in that example is Invention #8 by J.S. Bach with a target pattern $\mathbf{p} = [7, -8, 8, -5, 4, 1]$.

It turns out that there is quite a bit of variation in certain regions of the sequence of key assignments. Moreover, some impulses last only one note, which would seem to indicate that the tonal context changes for one note and then changes back - a very unlikely circumstance. This exposes a "weakness" of the key-finding algorithm in that it may be sensitive to window length as well as the distribution of pitches within the window. Nevertheless, whatever the tonal context is, it makes little sense to think of two modulations occurring one note apart. Besides this, there are small areas of oscillations, especially those close to edges between two flat regions. These edges signify modulations

and as the window slides across them, the key-finding algorithm is unable to determine a prevalent tonal context due to the presence of pitches that have high probe tone ratings in two different profiles. As a result, the assigned key values oscillate until a prevalent tonal context is established. Such small oscillations and impulses are undesirable, not only because they do not reflect our notions of modulations, 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 appears arbitrary in the regions of oscillation, the perceptual pitch error is distorted in these regions.

As a solution to the above problem, we employ the recursive median filter [5] with a large enough window to remove not only the impulses but also the small regions of oscillations. The output of the recursive median filter is defined as

$$y_i = med\left(y_{i-\nu}, \ldots, y_{i-1}, x_i, \ldots, x_{i+\nu}\right)$$

where the samples $y_{i-\nu}, \ldots, y_{i-1}$ have already been computed during previous positions of the sliding window. It has been shown that the recursive median filter has a higher immunity to impulsive noise than the standard median filter. This makes it a better choice for our purpose than the standard median filter. Moreover, the output of the recursive median is more correlated than the output of the standard median. This is due to the fact that is dependent on previous output values. This correlation in the output is advantageous since the tonal context at a particular position is more strongly dependent on previous values of the assigned key sequence than on future values. Finally, it is well known that the recursive median filter is idempotent. This property implies that any signal is reduced to a root signal after one pass; i.e., it is invariant to further passes of the same filter. This assures us that the assigned key sequence cannot be improved by more filter passes. The window width of the recursive median filter is a parameter that needs to be chosen. If we are to employ the recursive median filter in order to remove oscillations in the regions of modulation, we must establish a high measure of tonal structure prior to and after the region of modulation. The number of notes necessary to establish this, of course, depends on key membership of the notes as well as their relationship to the tonal center (i.e., stability). However, it has been shown that the maximum correlation, $r_{\max}$, is strongly correlated with the degree of tonal structure [7]. Therefore, if $r_{\max}$ is small, indicating a low degree of tonal structure, we should expect to use more notes to establish the latter. This implies that the window width of the recursive median filter should be inversely related to $r_{\max}$. Recall that for every window position of the key-finding algorithm, we have a maximum correlation, thus giving rise to the sequence $r_{\max}(i)$ of maximum correlations. We would like the window width, W, to be a function of the lowest of the maximum correlations. That is, $W = f\left(\min\left[r_{\max}(i)\right]\right)$ and one possible function is

$$f(r) = \left\lceil \frac{k}{r^{1/\nu}} \right\rceil$$

where $\lceil \cdot \rceil$ is the next odd integer. Experiments show that

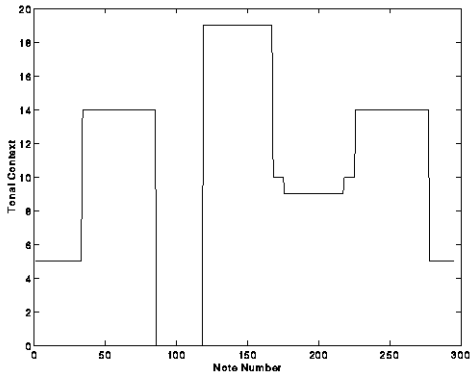

Figure 2: Recursive Median Filtered Assigned Key Sequence

values of $k = 17$ and $\nu = 8$ give good results. The parameter ν simply controls the rate of growth of the window width with respect to the lowest maximum correlation. Figure 2 shows a median filtered assigned key sequence using window width 19 for the recursive median filter, where $\min\left[r_{\max}(i)\right] = 0.46$. As can be seen, the impulses and oscillations are completely removed and yet the key assignments (and modulations) reflect what would be expected upon a visual inspection of the composition.

Now that we can successfully generate the assigned key sequence $\mathbf{t}$, all that remains is the determination of parameter λ. Essentially, λ is directly related to the maximum correlation $r_{\max}$. However, prior to that, it is necessary to smooth the sequence $r_{\max}(i)$ while maintaining peaks and removing small oscillations. This is accomplished with a cubic smoothing spline and is described in [6]. If $\hat{r}_{\max}(i)$ is the spline-smoothed maximum correlation sequence, the definition of λ becomes

$$\lambda(i) = m \cdot \left(\hat{r}_{\max}(i) - \max\left(\hat{r}_{\max}(i)\right)\right) + b$$

where

$$m = \frac{b-a}{\max\left(\hat{r}_{\max}(i)\right) - \min\left(\hat{r}_{\max}(i)\right)}$$

making $\lambda(i)$ just a scaled version of $\hat{r}_{\max}(i)$.

2.3. Rhythm Error

At this point, we are ready to compute the rhythm error between the target pattern and the scanned pattern. Recall that $\mathbf{r} = [r_1, r_2, \cdots, r_{n-1}]$ represents the difference of rhythm vector of the scanned rhythm pattern (of length n). Let $\mathbf{r_0} = [s_1, s_2, \cdots, s_{n-1}]$ represent the difference of rhythm vector of the target pattern. Since our difference of rhythm vectors are logarithmic (e.g., a quarter note followed by a half note produces 2 in the vector), it would not be appropriate to use the 1-norm as we have done for the pitch error. Consider the following example (S=sixteenth note, E=eighth note, Q=quarter note):

Scanned Pattern	Target Pattern
E (1) E (2) Q	S (2) E (2) Q
E (1) E (2) Q	Q (0.5) E (2) Q

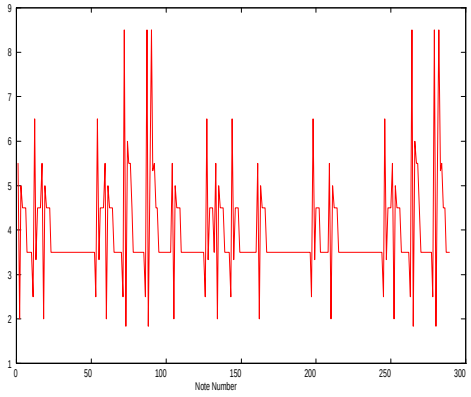

Figure 3: Rhythm Error Sequence

The numbers between the note values represent the components of the difference of rhythm vectors. If we were to use $\|\mathbf{r}-\mathbf{r_0}\|_1$ to compute the rhythm error, then in the first case, the error would be equal to 1 while in the second case, it would be equal to 0.5. However, there is no reason for penalizing S followed by E any more than Q followed by E, when the scanned pattern is E followed by E. The two errors should be equal. To accomplish this, we define the rhythm error to be

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

For the above example, our error (in both cases) is equal to $e_r = 1$. We subtract the term $(n-1)$ in the above expression so that a perfect match of the difference of rhythm vectors produces a zero rhythm error. Finally, the scanning window gives rise to the rhythm error sequence $e_r(i)$, an example of which is shown in Figure 3.

2.4. Overall Error

Having defined pitch and rhythm errors, we can combine them into one single error using a weighted combination of both. Let $e = \sigma \cdot e_q + (1-\sigma) \cdot e_r$ be the overall error and $e(i)$ be the overall error sequence. One way to set the parameter σ would be to ask the user to input his/her level of confidence in the pitch/rhythm information. For instance, if the user remembered a complex rhythm pattern, but did not quite remember the exact pitch pattern, then he/she might choose to give a low value to σ. Another way to set this parameter, and one which we employ here, is to make it dependent on the length of the target pattern. If the length, n, of the target pattern is short (several notes), then rhythm information contained in the difference of rhythm vectors is of little significance and should only be used to distinguish between identical difference of pitch vectors. For instance, the vector $\mathbf{r} = [1, 1, \cdots, 1]$ occurs very often in music. So, one possible way of setting σ is

$$\sigma = \begin{cases} \frac{100-n}{100}, & \text{if } n \leq T \\ \frac{T}{100}, & \text{if } n > T \end{cases}$$

For example, if we do not want the rhythm error to ever outweigh the pitch error, we could set $T = 50$. Then, if the target pattern happens to be longer than 50 notes long, σ would stay at 0.5.

2.5. Future Work

A rule for selecting the parameter σ, which is the weight given to the pitch error e_q, could be based on the complexity of the rhythm information contained in the target pattern. The complexity of the difference of rhythm vector could be expressed in one of many different ways and this needs to be investigated further. Finally, the recursive median filter used for the assigned key sequence could be replaced by a variable window length filter, the length of which would depend on the maximum correlation values provided by the key-finding algorithm; the effects of applying such an operation to the maximum correlation sequence need to be studied. Finally, models that incorporate human beat perception should be used to improve the robustness of the rhythm error, in an analogous way to the method used to compute the pitch error.

3. REFERENCES

[1] S. H. Hulse, A. H. Takeuchi, R. F. Braaten, "Perceptual Invariances in the comparative psychology of music," Music Perception, vol. 10, No. 2, pp. 151-184, 1992.

[2] J. J. Huang, E. J. Coyle, G. B. Adams, "The effect of changing the weights in the mean absolute error criterion upon the performance of stack filters," Proceedings of the 1995 Workshop on Nonlinear Signal Processing, Halkidiki, Greece, June 1995.

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

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

[5] I. Shmulevich, 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

[6] I. Shmulevich, 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., 1997

[7] A. H. Takeuchi, "Maximum key-profile correlation (MKC) as a measure of tonal structure in music," Perception & Psychophysics, vol. 56, pp. 335-346, 1994.