Perceptual Evaluation of Automatically Extracted Musical Motives

Oriol Nieto[1], Morwaread M. Farbood[2]

Dept. of Music and Performing Arts Professions, New York University, USA
[1]oriol@nyu.edu, [2]mfarbood@nyu.edu

ABSTRACT

Motives are the shortest melodic ideas or patterns that recur in a musical piece. This paper presents an algorithm that automatically extracts motives from score-based representations of music. The method combines perceptual grouping principles with data mining techniques, using score-based representations of music as input. The algorithm is evaluated by comparing its output to the results of an experiment where participants were asked to label representative motives in six musical excerpts. The perceptual judgments were found to align well with the motives automatically extracted by the algorithm and the experimental data was further used to tune the threshold values for similarity and strength of grouping boundaries.

I. INTRODUCTION

In order to understand how various structural features contribute to the listener's cognition of a piece, it is essential to be able to parse a musical surface into perceptually valid segments. One such method is described by Lerdahl and Jackendoff (1983), who present a series of grouping principles to formulate a formal analysis system that defines a musical grammar. They describe the process of grouping auditory stimuli as psychologically analogous to the one of visual perception, and many of their rules are based on Gestalt principles of grouping. This psychological approach to analysing musical structure and patterns provides a framework for understanding the perception of musical motives.

A motive can be defined as the shortest melodic idea or pattern that recurs in a musical piece, and it is often one of the most characteristic elements of the piece. Moreover, listeners can often identify a piece by just hearing its primary motives. Considerable prior work has been done on motive identification, much of which has its origins in the study of melodic similarity. Hewlett and Selfridge-Field (1998) provide a significant collection of articles that discuss various methods of assessing similarity between melodies given a symbolic representation of music. Other relevant prior work includes cognitive studies on the perception and modeling of melodic similarity (Ferrand, Nelson, & Wiggins, 2003; Martínez, 2001) and perspectives on rhythmic similarity (Aloupis et al., 2006; Martins, 2005; Toussaint, 2004).

Automatic methods of extracting motives given score-based and audio-based representations of music make use of both melodic and rhythmic similarity (Lartillot, 2005; Jiménez et al., 2010). Weiss & Bello (2011), on the other hand, use a probabilistic approach based on non-negative matrix factorization applied to audio signals.

What these approaches have in common is that they are based on repetition of material. Our approach to automatically extracting motives from score-based representations of music combines a similar data mining approach filtered by perceptual factors based on Gestalt grouping principles. These rules are mostly based on Chapter 3 of Lerdahl and Jackendoff's *Generative Theory of Tonal Music* (1983) and take into account the Gestalt principles of proximity, similarity and good continuation. Having a perceptual framework to extract possible motives should ideally lead to a better understanding of what makes a particular melodic segment more plausible as a coherent motive than others. Previous work has employed similar Gestalt grouping strategies for automatic segmentation of music and melodic similarity (Ferrand et al, 2003; Hamanaka, Hirata, & Tojo, 2004; Temperley, 2001), however, our goal is to combine both the data mining and perceptual approaches to the particular analytical task of *motive extraction*.

In order to evaluate the algorithm, an experiment was conducted in which musically trained participants were asked to determine the representative motives in six musical excerpts and rate each chosen motive based on its relative prominence. The empirical judgments were then compared with the output of the algorithm.

II. ALGORITHM DESCRIPTION

The input of the algorithm consists of monophonic, score-based representations of music (e.g., MIDI, Music XML). Two dimensions are considered when comparing melodic sequences: the diatonic interval between notes and the rhythmic values of the notes. The L1-norm is used to determine the "distance" in each dimension. Formally, this distance metric is defined as

$$d(x,y) = |x^r - y^r| + |x^{di} - y^{di}| \qquad ,$$

where x^r represents the rhythmic information of the symbol (i.e., note event) x, and x^{di} is the diatonic interval between the previous symbol and the current symbol x. The algorithm can be divided into two parts that make use of this metric and that are discussed in the following subsections.

A. Extraction of Potential Motives

The first stage of the algorithm is to try to identify all sub-sequences that are potential motives within a sequence of symbols that contains the pitch and rhythmic information of all the notes of the monophonic piece. To do so, we look for sub-sequences that meet the following criteria:

1. A potential motive must be at least three notes long.
2. A potential motive must repeat at least once; exact repetition is not necessary, but the distance of the repetitions must be less than a given threshold τ that can be adjusted.
3. A potential motive cannot have a rest that is longer than 25% of its length.
4. A potential motive must have a uniform contour shape.

These rules are mostly based on Gestalt principles of similarity, proximity and good continuation. The similarity rule is exemplified by the second rule listed above. The third and fourth rules mix the principles of proximity and good continuation. Finally, for each one of the potential motives, we store the number of times they appear within the piece (a minimum of two).

B. Clustering and Selection of the Motives

Once the set of potential motives that meet these criteria have been extracted and the frequency counts for each of them have been recorded, they are clustered into different groups based on how different they are from each other. To do so, a distance matrix is defined to find overlaps, where the size of the matrix is equal to the square of the number of potential motives previously found.

To compute the distance between each pair of potential motives, they are aligned based on their downbeats and the distance metric described above is computed for all possible shifts of alignment between the two motives. The distance value that is minimal across all possible shifts on the downbeats is the one stored in the matrix.

The matrix values are then used to group similar motives: if the distance between two potential motives is below a certain threshold θ, they are clustered in the same group. For each group, one final motive is selected: the one that has the median length across all the motives of that cluster. The output of the algorithm is a set of filtered motives, one for each one of the clusters.

C. Implementation

An implementation that extracts the set of potential motives given the rules defined in section II.A can be implemented using an algorithm that has a quadratic complexity in time, $O(n^2)$, where n is the number of notes contained in the melody. The space complexity will vary depending on the amount of potential motives found m, but it is low enough to be negligible. The clustering and filtering described in section II.B has a time complexity of $O(m^2k^2)$, where k is the average length of the potential motives (i.e., $k << m$).

The current implementation used to evaluate the algorithm was written in Python and makes use of the music21 framework (Cuthbert & Ariza, 2010) in order to easily read and parse music XML files.

III. METHOD

An experiment conducted to evaluate the quality of the algorithm. The goal of the study was to evaluate the results of the algorithm by comparing the automatically extracted motives with the perceptual judgments made by musicians. Furthermore, these findings would help tune the thresholds τ and θ of the algorithm as described in the previous section.

A. Participants and Task

Fourteen musical trained subjects were asked to identify motives for six different monophonic excerpts. Subjects were all graduate students at New York University and had an average of 10 years of formal musical training (*SD* = 2.3). They were asked identify all motives in an excerpt and to rate them on overall relevance. The rating choices were “Not so relevant”, “Relevant”, and “Highly relevant”. The rating values were important in determining the highest ranked motives for each excerpt, with different weights assigned to each choice (1 for “Not so relevant”, 2 for “Relevant”, and 3 for “Highly relevant”).

B. Stimuli

The excerpts used in the experiment were taken from the following pieces:

1. Bach – Cantata BWV 1, Movement 6, Horn
2. Bach – Cantata BWV 2, Movement 6, Soprano
3. Beethoven – String Quartet, Op. 18, No. 1, Violin I
4. Haydn – String Quartet, Op. 74, No. 1, Violin I
5. Mozart – String Quartet, K. 155, Violin I
6. Mozart – String Quartet, K. 458, Violin I

Some of these excerpts were intentionally chosen because they were particularly hard for humans to analyse given the structural ambiguity of some of the musical material. For example, the Bach chorale had very little rhythmic variation or clear grouping cues aside from phrase ending points. In general, Excerpts 1, 5 and 6 proved to be particularly difficult to parse for humans, and as we see in the Results section, there are some interesting discrepancies in the data. The data resulting from these “difficult” excerpts enable us to ascertain the amount and type of overlap that frequently occurs in motive perception. Excerpt 3 (shown in Figure 1), on the other hand, has more clearly defined motives that are readily apparent from a quick glance at the score. This excerpt will be used to discuss the results in detail.

IV. RESULTS

A. Quantifying the Experimental Results

The first step in evaluating the relative importance of the motives indicated by the subjects was to quantify each response/selection by the importance weighting described in the previous section. It was common to find a high degree of overlap between motives across subjects; however, there was often disagreement about the start and end points. The motives were thus manually clustered into groups based on overlap in a manner similar to the process described in Section II.B. Once grouped, all of the weighted responses for each motive were summed for each cluster, and the clusters were then sorted based on these values. Finally, a representative motive for each cluster was selected by choosing a version that had the median length with respect to the other motives in that cluster.

B. Evaluating the Experimental Results

For the purposes of this paper, Excerpt 3 of the experiment, taken from Beethoven’s Op. 18 No. 1 string quartet, will serve as the focus of the evaluation. The excerpt is shown in its entirety in Figure 1. Motives in this excerpt were relatively easy to discern and the empirical results indicate a significant degree of agreement. Figure 2 shows the most frequently chosen motives from Excerpt 3, ordered from the highest to lowest-rated in importance. It is interesting to observe the subtle differences in these selections. While some motives shown in Figure 2 are simply chromatic or diatonic transpositions of another (e.g., motives 2 and 3), there are other selections that differ with regard to start and end points (e.g., motives 7-8 and 10-11). These choices indicate that even in a piece with clearly defined motivic material, there is still disagreement concerning the most representative versions of each motive.

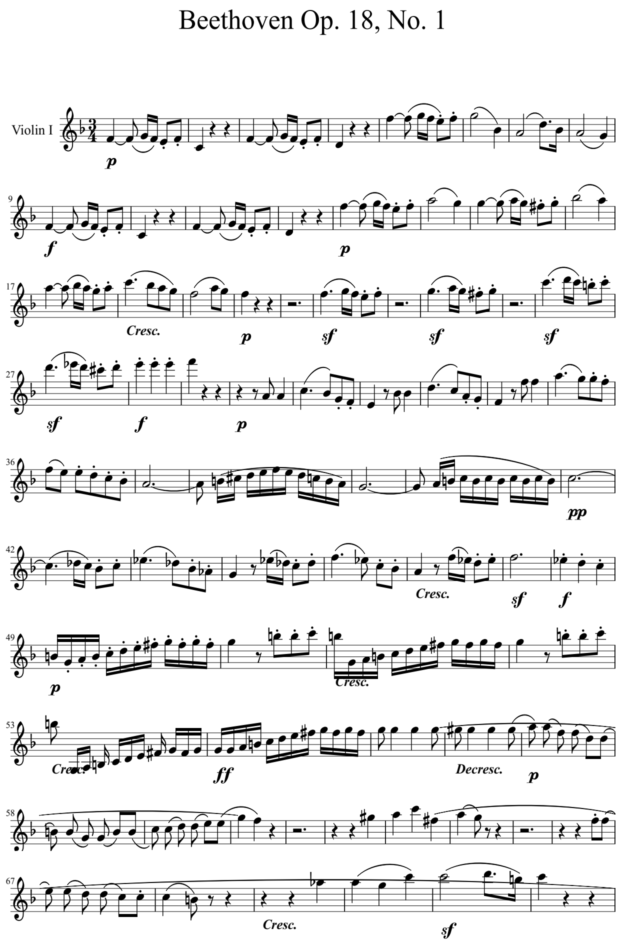

Figure 1. Excerpt 3, from Beethoven's string quartet Op. 18, No. 1.

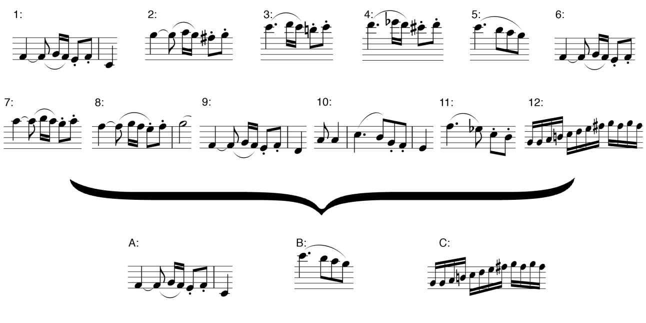

Figure 2. Motives most frequently chosen by human listeners for Excerpt 3. Top: All of the most relevant motives identified by subjects. Bottom: Motives representing the three major clusters of responses. Motives are ordered from highest to lowest importance.

Most of the motives in Figure 2 can be clustered into a single group encompassing motives 1-4 and 6-9. Motives 5, 10 and 11 form another group, leaving motive 12 in a group of its own. This results in three primary motives selected by the subjects who took the experiment. These primary motives, labelled A, B, and C, can be seen in the bottom section of Figure 2.

As noted, many of differences between similar motives selected by subjects concerned designated start and end points. However, additional variations included difference in contour as well as the duration of certain notes. The difficulty for both a human and a computer program lies in determining the threshold between simple variations of a primary motive versus a significant difference that results in a new perceptual motive category altogether.

C. Tuning the Parameters

Given the empirical data, the next step was to use these results to tune the two different thresholds (τ and θ) of the algorithm. This was accomplished by maximizing the amount of overlap between the output of the algorithm and the results of the experiment. Interestingly, tuning these thresholds given data from one of the excerpts not only improved the results for that particular excerpt (as expected), but also for other excerpts. This makes a good case for the perceptual validity of the method, given that it generalizes well. Excerpt 3 was chosen for the purposes of tuning the thresholds because it resulted in the highest degree of agreement across subjects. The values for the thresholds that maximized overlap with Excerpt 3 results were $\tau = 1$ and $\theta = 0.65$. $\tau = 1$ means that there can be a maximum of one differing diatonic interval or rhythmic value when determining whether repetition of that motive exists elsewhere (when finding similarity using the L1-norm). $\theta = 0.65$ means there must be a minimum of 65% overlap in order to cluster two motives into the same group.

It is important to realize that no specific set of thresholds will optimally work for all the pieces. However, setting these parameters by using the empirical data does tune the algorithm based on multiple human judgments rather than arbitrary values.

The representative motives extracted by the algorithm for Excerpt 3 (post-tuning) is shown in Figure 3.

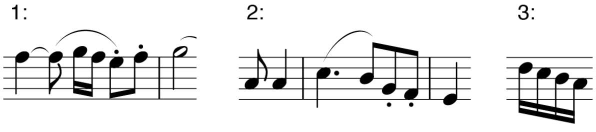

Figure 3. Motives extracted from Excerpt 3 by the automatic algorithm.

As can be seen in Figure 3, motive 1 is representative of the set of motives that previously categorized as cluster A in the experimental results. More specifically, this motive is identical to motive 8 from Figure 2; it was selected by the algorithm from a cluster containing 57 automatically extracted potential motives. This group of potential motives encompasses all of the cluster A motives shown in the top section of Figure 2. There is clearly a high degree of overlap between the first automatically extracted motive and the cluster A motives from the experimental results.

The second motive (Figure 3) is also selected from a cluster contained that highly similar to the motives from cluster B in the experimental results. The automatically extracted motive representing this group is in fact identical to motive 10 from the experimental results. If the motives contained within the cluster are examined (in this case, 29 potential motives), both motives 5 and 11 from the experimental results are found among them.

Finally, the third automatically extracted motive represents a cluster formed by the repetitions found in the sixteenth note passages represented in the empirical results by motive 12 in Figure 2. In this case, the algorithm considers the descending diatonic scale composed of four notes the best way to represent this cluster. Unfortunately, this scale only appears once in the piece, but the algorithm designates it as similar to the last four notes of motive 12 because there is only one directional difference in their contours. Despite this discrepancy, the algorithm is still able to capture 11 out of 12 motives selected by the subjects. This leads to an overall 91.6% overlap between the experimental results and automatic output for Excerpt 3.

C. Results for All Excerpts

Table 1 shows the overall comparison results between the empirical data and the algorithm output for all of the excerpts. The scores are computed following the same methodology as used for Excerpt 3: the most relevant motives of the empirical results are clustered and the amount of overlap they have with respect to the automatically extracted motives is computed. Each experimental cluster is weighted depending on the number of motives contained in the group (e.g. motive A from Figure 2 encompasses eight motives, thus has a higher weight than motive B); thus overlap with the automatic results are scaled by those weights.

Excerpt	**1**	**2**	**3**	**4**	**5**	**6**	**Average**
Score	80.0	100	91.6	81.1	75.0	50.0	79.6

Table 1. Results for each one of the excerpts from the empirical experiment

Excerpt 1 (Bach BWV 1) has a long melody that repeats twice in the beginning of the excerpt; some subjects chose the entire melody as a motive (over 40 notes long). Overall, the algorithm captured four out of the five clusters that resulted from the experimental data. The one motive it didn't find was missed due to reasons similar to case of motive 3 in Excerpt 3.

For Excerpt 2 (Bach BWV 2), there was a strong agreement with the empirical experiment results. Even though this piece does not have a clear motivic structure, its length is quite brief, providing few choices for human analysts. Subjects agreed on two primary motives, both so brief that the automatic algorithm selected both of them as one long motive (as some of the subject subjects did as well).

Excerpt 4 (Haydn Op. 74, No. 1) does not contain any representative motives that can be identified easily. This is reflected in the results, which show that there was considerable disagreement among subjects when selecting the motives. The output of the algorithm, however, corresponds with the results experimental results in general. There are four motives that are automatically extracted, and they contain the 11 primary motives that were selected by the subjects.

Excerpt 5 (Mozart K. 155) is also difficult to analyse. There are four primary motives, and two of them differ only in contour. The algorithm does not differentiate between these two due to the similarity thresholds employed. In all, the algorithm extracted four motives, three of them corresponding with the experimental results.

Excerpt 6 (Mozart K. 458) is another difficult-to-parse excerpt. Subjects agreed on four main motives. The algorithm also produced four motives, however three of them formed parts of the primary motive selected by the subjects. The two motives selected by the subjects that were not captured by the algorithm have different contours but similar diatonic intervals and rhythmic durations; the algorithm placed them into the same cluster as one of the other automatically selected motives. The thresholds in this case were too generous and ignored the smaller dissimilarities; given this problem, the algorithm only matched two out of four main motives in Excerpt 6.

Across all excerpts, there was a mean matching score of 79.6%, which was deemed successful given the difficulty and subjectivity of the task at hand.

V. CONCLUSIONS AND FUTURE WORK

This paper presents an algorithm that automatically extracts musical motives from symbolic representations of monophonic music by combining data-mining techniques with perceptual grouping rules. An experiment was described in which musically trained subjects were asked to label motives in six musical excerpts. These data were then used to evaluate and improve the algorithm. Using the results from one musical excerpt, thresholds in the algorithm were tuned to maximize agreement with human judgments. The algorithm was then evaluated for all excerpts by comparing its output to the empirical data. Results of this comparison indicate a high degree of agreement with human analysis.

One of the main issues explored was finding the right threshold values for the algorithm in order to successfully characterize perceptual similarity between melodic fragments. Future work can further improve understanding of this issue. A more exhaustive experiment can be conducted with a wider variety of musical excerpts. This might lead to a better understanding of what makes a particular motive distinctive in differing textural and stylistic contexts.

Another future step would be to run the algorithm on a large collection of scores. It would be interesting to see if it is possible to find motives that are not only repeated across a piece, but also across the entire oeuvre of a composer, or to compare motive variations between different composers.

Ultimately, this work could lead to the foundations of an algorithm that works on audio recordings as well. Whether the input to the algorithm is symbolic or signal based, higher-level hierarchical analysis of a piece can be aided by understanding the occurrence and recurrence of motivic material.

ACKNOWLEDGMENTS

Special thanks to Kwan Kim for helping compile the results of the experiment. This work is funded in part by the Caja Madrid Fellowship.

REFERENCES

Aloupis, G., Fevens, T., Langerman, S., Matsui, T., Mesa, A., Nuñez, Y., Rappaport, D., & Toussaint, G. (2006), Algorithms for Computing Geometric Measures of Melodic Similarity. *Computer Music Journal*, *30,* 67–76.

Cuthbert, M. S., & Ariza, C., (2010). Music21: A toolkit for computer-aided musicology and symbolic music data. In *Proceedings of the 2010 International Society for Music Information Retrieval Conference*, 637–642. Miami, FL.

Ferrand, M., Nelson, P. & Wiggins, G. (2003). Memory and Melodic Density: A Model for Melody Segmentation. In *Proceedings of XIV CIM 2003*, 95-98.

Hamanaka, M., Hirata, K., & Tojo, S. (2004). Automatic generation of grouping structure based on the GTTM. In *Proceedings of the 2004 International Computer Music Conference*.

Hewlett, W. B. & Selfridge-Field, E. (1998), *Melodic Similarity: Concepts, Procedures, and Applications*. Cambridge, MA: MIT Press.

Jiménez, A., Molina-Solana, M., Berzal, F., & Fajardo, W. (2010), Mining Transposed Motifs In Music. *Journal of Intelligent Information Systems, 36*, 99-115.

Lartillot, O. (2005), Multi-dimensional motivic pattern extraction founded on adaptive redundancy filtering. *Journal of New Music Research, 34,* 375–393.

Lerdahl, F., & Jackendoff, R. (1983). *A Generative Theory of Tonal Music*. Cambridge, MA: MIT Press.

Martínez, I. C. (2001), Contextual Factors In The Perceptual Similarity of Melodies. *The Online Contemporary Music Journal, 7*.

Martins, J., Gimenes, M., Manzolli, J., & Maia, A. Jr (2005), Similarity Measures For Rhythmic Sequences. In *Proceedings of the 10th Brazilian Symposium on Computer Music* (SBCM).

Temperley, D. (2001). *The Cognition of Basic Musical Structures*. Cambridge, Massachusetts: MIT Press.

Toussaint, G., (2004) A Comparison of Rhythmic Similarity Measures. In *Proceedings of the 5th International Conference on Music Information Retrieval,*. 242–245. Barcelona, Spain.

Weiss, R. J., & Bello, J. P. (2011), Unsupervised Discovery of Temporal Structure in Music, IEEE. *Journal of Selected Topics in Signal Processing*, *5*, 1240-1251.