Chapter 17
Using Geometric Symbolic Fingerprinting to Discover Distinctive Patterns in Polyphonic Music Corpora

Tom Collins, Andreas Arzt, Harald Frostel, and Gerhard Widmer

Abstract Did Ludwig van Beethoven (1770–1827) re-use material when composing his piano sonatas? What repeated patterns are distinctive of Beethoven's piano sonatas compared, say, to those of Frédéric Chopin (1810–1849)? Traditionally, in preparation for essays on topics such as these, music analysts have undertaken inter-opus pattern discovery—informally or systematically—which is the task of identifying two or more related note collections (or phenomena derived from those collections, such as chord sequences) that occur in at least two different movements or pieces of music. More recently, computational methods have emerged for tackling the inter-opus pattern discovery task, but often they make simplifying and problematic assumptions about the nature of music. Thus a gulf exists between the flexibility music analysts employ when considering two note collections to be related, and what algorithmic methods can achieve. By unifying contributions from the two main approaches to computational pattern discovery—viewpoints and the geometric method—via the technique of symbolic fingerprinting, the current chapter seeks to reduce this gulf. Results from six experiments are summarized that investigate questions related to borrowing, resemblance, and distinctiveness across 21 Beethoven piano sonata movements. Among these results, we found 2–3 bars of material that occurred across two sonatas, an andante theme that appears varied in an imitative minuet, patterns with leaps that are distinctive of Beethoven compared to Chopin, and two potentially new examples of what Meyer and Gjerdingen call *schemata*. The chapter does not *solve* the problem of inter-opus pattern discovery, but it can act as a platform for research that will further reduce the gap between what music informaticians do, and what musicologists find interesting.

Tom Collins
Faculty of Technology, De Montfort University, Leicester, UK
e-mail: tom.collins@dmu.ac.uk

Andreas Arzt · Harald Frostel · Gerhard Widmer
Department of Computational Perception, Johannes Kepler University, Linz, Austria
e-mail: {andreas.arzt, harald.frostel, gerhard.widmer}@jku.at

17.1 Introduction

The topic of borrowing, between composers or within a single composer's oeuvre, has long been a concern for musicologists studying various periods and genres (Burkholder, 2001). George Frideric Handel's (1685–1759) music has received much attention in this regard, so it seems appropriate to begin this chapter with an example of Handel's borrowing from Reinhard Keiser (1674–1739), given in Fig. 17.1 (Roberts, 1986; Winemiller, 1997). In Fig. 17.1(a), Keiser's seven-note pattern occurs first in the oboe and is then sung by Clotilde. Shown in Fig. 17.1(b), Handel uses this same sequence of pitches, again in the oboe, but with a different rhythmic profile. Whereas Handel is often mentioned in connection with borrowing between composers, a composer well known for reworking of his own compositions is Beethoven:

> More than a third of Beethoven's compositions reworked his existing music in some way.
>
> (Burkholder, 2001)

Lutes (1974) identifies a pattern from the first movement of Beethoven's Piano Sonata no. 5 in C minor, op. 10, no. 1 (Fig. 17.2(a)) that recurs in the first movement of the Piano Sonata no. 6 in F major, op. 10, no. 2 (Fig. 17.2(b)). Beethoven, however, was also apt to borrow from other composers and Lutes (1974) credits Radcliffe (1968) with identifying the pattern in Fig. 17.2(a) as an instance of borrowing from Joseph Haydn's (1732–1809) Symphony no. 88 in G major, Hob.I:88 (Fig. 17.2(c)).

What do the pattern occurrences in Fig. 17.2 have in common, and how do we define the term *pattern*? Commonalities first (and see Sect. 17.3 for a definition of

Fig. 17.1 (a) Bars 1–5.1 of 'Mit einem schönen Ende' from *La forza della virtù* by Keiser. Two occurrences of a seven-note pattern are highlighted. (b) Bars 1–4.1 of 'Must I my Acis still bemoan' from *Acis and Galatea* by Handel. An occurrence of a seven-note pattern is highlighted. Throughout this chapter, 'bar *x.y*' means 'bar *x* beat *y*'

Fig. 17.2 (a) Bars 233–240 of the first movement from Beethoven's Piano Sonata no. 5 in C minor, op. 10, no. 1. One occurrence of a twelve-note pattern is highlighted in blue. A second occurrence in bars 237–240 is not highlighted. Instead, a different ten-note pattern is highlighted in red and discussed later on with reference to Fig. 17.11(c). (b) Bars 18–26 with upbeat of the first movement from Beethoven's Piano Sonata no. 6 in F major, op. 10, no. 2. One occurrence of a twelve-note pattern is highlighted in blue. A second occurrence in bars 25–26 with upbeat is not highlighted. (c) Piano reduction of bars 1-4.2 with upbeat of the second movement from Symphony no. 88 in G major by Haydn. An occurrence of a ten-note pattern is highlighted in bars 1–2, followed by a second occurrence in bars 3–4

pattern): the bass moves from scale degree $\hat{1}$ to $\hat{2}$. Simultaneously, the melody—doubled at the octave—outlines a rising arpeggio, beginning on $\hat{3}$ and ending on scale degree $\hat{1}$, followed by a fall to scale degree $\hat{7}$. This pattern provides the antecedent of an antecedent-consequent formula typical of the period. In each excerpt, the consequent consists of the bass moving from $\hat{7}$ to $\hat{1}$, while the melody—still doubled at the octave—outlines a rising arpeggio beginning on $\hat{5}$ and ending on $\hat{4}$, followed by a fall to $\hat{3}$. Putting the octave arpeggio to one side, the bass movement of $\hat{1}$–$\hat{2}$–$\hat{7}$–$\hat{1}$ and melodic movement of $\hat{1}$–$\hat{7}$–$\hat{4}$–$\hat{3}$ was referred to by Meyer (1980) as a *schema*—instances of simultaneous bass and melodic movement to be found across many pieces. Gjerdingen (1988) identified numerous instances of this particular schema across the period 1720–1900, and then other different categories of schemata (Gjerdingen, 2007). So the subdiscipline of music analysis known as *schema theory* was born, and remains popular to this day (Byros, 2012). So prevalent was the use of Meyer's schema in the Classical period, and so abstract the definition, that to mix it with remarks on borrowing is perhaps inappropriate. It seems neither Meyer (1980) nor Gjerdingen (1988, 2007) were aware of the examples identified in the earlier work of Radcliffe (1968) and Lutes (1974), which are perhaps better described as *borrowing* rather than *schemata* due to the specificity of octaves and rising arpeggios in each case. Still, Radcliffe (1968, p. 38) cautions,

> it is very dangerous to attach too much importance to thematic resemblances of this kind, especially in music written at a time when there were so many familiar turns of phrase used by all and sundry. Haydn's tune is slow and majestic, and Beethoven's recollections of it all move at a quicker pace, sometimes with the suggestion of a dance.

Whether referred to as pattern or schema, evidently the highlighted content of Figs. 17.1 and 17.2 and surrounding discussion are of interest to music analysts and musicologists more broadly. Processes of musical variation (more about which below) are in evidence in these figures, but instances of literal borrowing—e.g., where a number of bars are reused more or less verbatim—are of interest also (Winemiller, 1997). This chapter explores computational methods for identifying resemblances between pieces of music (involving both literal borrowing and more complex variation). The methods are described, applied to 21 movements from Beethoven's piano sonatas (which, given the above discussion, seem a sensible place to begin), and the results are presented and discussed.[1] It is remarkable how much existing literature on Beethoven's piano sonatas focuses on intra-movement as opposed to inter-movement or inter-piece analyses (Caplin, 2013)—an exception being Marston (1995). Even so, Marston's (1995) mix of sketchbook, biographical, and Schenkerian analysis is quite apart from what follows here. The question of whether Beethoven intended any discovered resemblance will not be considered. We focus instead on the likelihood of the pattern occurring in other Beethoven sonatas, or in the piano works of another composer such as Chopin. If the pattern occurs more often in the works of Beethoven than some other composer(s), then it can be

---

[1] The movements used are: op. 2, no. 1, mvts. 1–4; op. 10, no. 1, mvts. 1–3; op. 10, no. 2, mvts. 1–3; op. 10, no. 3, mvts. 1–4; op. 26, mvts. 1–4; and op. 109, mvts. 1–3. Movements were selected on the basis of frequent mentions in the analytic literature.

said to be distinctive of Beethoven's style. Books and articles abound on the topic of Beethoven's sonatas, but one motivation for this chapter is to see what light can be shed on the sonatas from the point of view of computational music analysis.

It is also remarkable how much existing (predominantly non-computational) work on borrowing involves musical excerpts that are either at the beginning of pieces or already known to be themes (Barlow and Morgenstern, 1948). One of the advantages of taking a computational approach is that it can be made more democratic in terms of detecting borrowing beyond incipits and themes. The apparent bias towards incipits and themes in existing work also raises the question: is thematic material inherently more *distinctive* than excerpts drawn from elsewhere in a movement? This is a question that we also seek to address in the current chapter.

17.2 Select Review of Computational Pattern Discovery

While the current chapter focuses on the music of Beethoven, it is part of a wider literature on *inter-opus pattern discovery*. That is, given a corpus of music, define an algorithm that returns musically interesting patterns occurring in two or more pieces. With regards to work on inter-opus pattern discovery, the major contribution of this chapter is a method capable of being applied to polyphonic music—polyphonic in the most complex sense of the term, where any number of voices may sound at a given point in time. After processing of the symbolic representations in Fig. 17.1 to extract individual melodic lines, there are computational methods capable of retrieving the type of patterns shown (Conklin, 2010; Knopke and Jürgensen, 2009). At present, however, no computational method exists capable of discovering the type of patterns shown in Fig. 17.2. The lack of such a method goes some way towards explaining why musicologists do not, in general, employ computational methods as part of their research into borrowing: practitioners of music computing have tended to import algorithms from other fields such as bioinformatics, which work well for melodic representations but do not apply to polyphonic music where voices can appear and disappear. Should practitioners of music computing be in any doubt about the need to look beyond melody-only representations, then let us consider Caplin (2013, p. 39):

> Although it is easy to focus attention on the melody, it is important to understand that the basic idea is the *complete* unit of music in all of its parts, including its harmonic, rhythmic, and textural components. *The basic idea is much more than just its "tune".*

Broadly, there are two approaches to the discovery of patterns in symbolic representations of music: (1) string-based or *viewpoint* methods (Cambouropoulos, 2006; Conklin, 2010; Conklin and Bergeron, 2008; Knopke and Jürgensen, 2009; Lartillot, 2005) (see also Chaps. 11, 12 and 15, this volume); (2) point-set or *geometric* methods, such as those described in the current chapter and Chap. 13 in this volume (see also Collins, 2011; Collins et al., 2013, 2011; Forth, 2012; Janssen et al., 2013; Meredith et al., 2002). As the name suggests, the viewpoints approach involves treating musical events as *sequences* considered from different perspectives

(e.g., sequences of MIDI note numbers, sequences of intervals, durations, etc.) and in different combinations. The geometric approach, on the other hand, involves representing numerical aspects of notes in a given piece as *multidimensional points*. The two approaches diverge when more than two notes sound at the same point in time, because in the viewpoints approach the sequential ordering of features of those notes becomes ambiguous. Viewpoints have been applied in both intra- and inter-opus pattern discovery scenarios, but up until this point, geometric methods have been applied in *intra*-opus discovery scenarios only.

In this chapter, we describe the first application of geometric pattern discovery algorithms in an inter-opus scenario. We discuss the challenges involved, present results from the piano works of Beethoven, and suggest possible directions for future work in this domain. The geometric method has some advantages over the viewpoint approach: first, the geometric method can be applied conveniently to both polyphonic and monophonic representations. Viewpoints have been applied to polyphonic representations before (Conklin and Bergeron, 2010), but rely on extracting a fixed number of voices from each piece in the chosen corpus; second, the geometric method is more robust to interpolated events in pattern occurrences. For instance, the six notes highlighted in red and labelled H_2 in Fig. 17.3(c) are a diatonic transposition of the six notes highlighted in red and labelled H_1 in Fig. 17.3(a). In between the C♯4 and A3 of H_2, however, there is an interpolated B3 (similarly, there is an interpolated G♯3 between the following A3 and F♯3). The sequential integrity of C♯4 → A3 is broken by the interpolated B3, compared with G♯4 → E4 of H_1, and so the viewpoint method will not recognize the evident similarity of the melodies in bar 1 of Fig. 17.3(a) and bar 21 of Fig. 17.3(c). We refer to this as the *interpolation problem* of the viewpoint method—a problem that also affects models of music cognition derived from the viewpoint method (e.g., Pearce et al., 2010).[2] The geometric approach is more robust to this type of variation (Collins et al., 2013, 2011). Therefore, the application of geometric pattern discovery algorithms in an inter-opus scenario described in this chapter constitutes an important advance for computational music-analytic methods.

17.3 Method

This section begins with a mathematical definition of the term *pattern*. As in Chap. 13 of this volume, in the current chapter we represent notes in a given piece of music as multidimensional points. For example, a note has a start time that might be assigned

---

[2] Advocates of the viewpoint method might say that defining a so-called threaded viewpoint to take pitch values only on quarter note beats would address this interpolation problem, but then offbeat notes that do belong to a pattern are overlooked also. Since we have mentioned an advantage of the geometric approach over the viewpoint approach, it is fair to state an advantage in the other direction also: if one is interested primarily in patterns that consist of substrings (as opposed to subsequences) in monophonic voices, then such patterns can be found more efficiently using string-based representations than they can be using point-based representations.

Fig. 17.3 (a) Bars 1–2 of the third movement from Piano Sonata no. 30 in E major, op. 109, by Beethoven. One occurrence of a six-note pattern is highlighted in red and labelled H_1. Taken together, the red and blue notes form a fifteen-note pattern that occurs inexactly in (b), which shows bars 31–37 of the same movement. In (b), one inexact occurrence can be seen in bars 33–34, and a second inexact occurrence in bars 35–36, with the same colour scheme as in (a) being maintained. (c) Bars 20–22 of the third movement from Piano Sonata no. 7 in D major, op. 10, no. 3 by Beethoven. An inexact occurrence of the six-note pattern from Fig. 17.3(a) is highlighted in red and labelled H_2

to the *x*-value of some point, and a numeric pitch value (e.g., MIDI note number) that might be assigned to the *y*-value of the same point, to give $\mathbf{d} = (x, y)$. (Using two dimensions is typical, but more are admissible, and later in the chapter we represent chord labels as points rather than notes.)

In a point-set representation of a given piece of music, there may be a collection of points P_1 that are perceived as similar to some other collection of points P_2, heard either elsewhere in the same piece or in another piece. In general, there could be *m* so-called *pattern occurrences* $P_1, P_2, \ldots, P_m$ across a corpus of pieces. Sometimes it is convenient to group these together into an *occurrence set*, denoted $\mathcal{P} = \{P_1, P_2, \ldots, P_m\}$. The term *pattern* is used rather loosely to refer to a member $P_i \in \mathcal{P}$, normally the member that is most typical of the occurrence set (often but not always the first occurrence, $P_i = P_1$).

17.3.1 Calculating the Distinctiveness of a Pattern

Rather than seeing viewpoint and geometric approaches to pattern discovery as two opposing camps, this chapter seeks to unify the methods to some extent, by developing geometric equivalents of the viewpoint technique for measuring pattern distinctiveness in inter-opus scenarios (Conklin, 2010). This technique is based on the concept of likelihood ratio. In statistics, the likelihood ratio test gives the best chance of occurrence of an observation under some null hypothesis, divided by its best chance of occurrence overall. Common uses include testing goodness of fit of observed data to some hypothesized underlying distribution (Pielou and Foster, 1962), and testing dependencies between variables such as crime and drinking (Pearson, 1909). In viewpoint pattern discovery, the likelihood ratio appears in various guises, e.g., for estimating the interest of an observed pattern in some corpus (Conklin and Bergeron, 2008). Conklin (2010) uses another likelihood ratio to measure the distinctiveness of an observed pattern P for one corpus of pieces Θ versus another *anticorpus* of pieces Θ', written

$$d(P, \Theta, \Theta') = p(P|\Theta)/p(P|\Theta') . \qquad (17.1)$$

In these settings, the statistics are based on either piece counts (the number of pieces in which the pattern occurs (Conklin, 2010)) or a zero-order model (Conklin and Bergeron, 2008). Piece counts can be problematic if a pattern occurs note for note (or feature for feature) in some pieces but only partially in others. Only counting exact occurrences leads to underestimation of the probability, whereas counting inexact occurrences on a par with exact occurrences leads to overestimation. In a zero-order model, a pattern is defined as a sequence of musical features, and its probability is proportional to the product of the relative frequencies of occurrence of the constituent features. Temporal order of features does not impact on the calculated probabilities in a zero-order model, which is a shortcoming (i.e., because B4, G4, C5 might be more probable in a certain style than the same pitches in different order, C5, B4, G4, say). We refer to this as the *zero-order problem* of the viewpoint method. An extension of these likelihood calculations to polyphonic textures has been proposed (Collins et al., 2011), but it too assumed a zero-order model.

To develop a geometric equivalent of the distinctiveness measure, it is necessary to calculate the empirical probability of a given pattern occurrence in a piece or across multiple pieces, $p(P|\Theta)$, preferably using a model that is: (1) less reliant on the sequential integrity of pattern occurrences and so addresses the interpolation problem, which is important since variation is such a central concept in music; (2) more realistic than one based on zero-order distributions, and so addresses the zero-order problem. Central to this development will be the technique of symbolic fingerprinting, which enables us to estimate the likelihood of occurrence of a pattern across one or more pieces of polyphonic music.

17.3.2 Symbolic Fingerprinting

Symbolic fingerprinting consists of calculating, storing, and retrieving differences between local pairs or triples from a point-set representation of a piece or pieces, denoted D (Arzt et al., 2012). It enables us to take some point-set query Q and find occurrences in D of Q that have been transposed, time-shifted and time-scaled. For readers familiar with music theory, the definition of a fingerprint will be reminiscent of Lewin's (1987, Chapter 4) generalized interval systems. Independently of Lewin's work, Wang and Smith (2012) developed an efficient fingerprinting storage and retrieval method that enabled automatic, fast recognition of music audio, known as Shazam.

   In the general case, we have a piece of music represented as an ordered point set $D = \langle \mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_n \rangle$.[3] To begin with in this chapter, each point $\mathbf{d}_i \in D$ represents a note from the piece, and is a pair $\mathbf{d}_i = (x_i, y_i)$ consisting of an ontime x_i and a morphetic pitch y_i (MPN, see Meredith, 2006). Ontime is the time in the piece in quarter note beats, counting from zero for bar 1 beat 1, and MPN is the height of the note on the staff, with C4 = 'middle C' = 60, C♯4 = 60, D♭4 = D4 = D♯4 = 61, etc.[4] Other choices about how to represent time and pitch, and how many dimensions to include in one point set, have been explored (Collins et al., 2010), but for the sake of simplicity we will use ontime and MPN at present. As an example, Beethoven's op. 109, mvt. 3 (see Fig. 17.3(a)) would be represented as

$$
\begin{aligned}
D = \big\langle &(0,48),(0,59),(0,64),(1,50),(1,59),(1,62),\ldots, \\
&(192,48),(192\tfrac{1}{4},59),\underline{(192\tfrac{1}{2},64)},(192\tfrac{3}{4},55), \\
&(193,57),\underline{(193\tfrac{1}{4},62)},(193\tfrac{1}{2},59),(193\tfrac{3}{4},50), \\
&(194,51),\underline{(194\tfrac{1}{4},60)},(194\tfrac{1}{2},63),(194\tfrac{3}{4},58), \\
&(195,59),(195\tfrac{1}{4},61),(195\tfrac{1}{2},56),(195\tfrac{3}{4},52), \\
&(196,53),(196\tfrac{1}{4},55),(196\tfrac{1}{2},59),(196\tfrac{3}{4},60), \\
&(197,61),(197\tfrac{1}{4},59),(197\tfrac{1}{2},56),(197\tfrac{3}{4},54),\ldots, \\
&(891,59),(891,62),(891,64) \big\rangle \, .
\end{aligned}
\tag{17.2}
$$

The first chord, consisting of pitches E2, B3, and G♯4, has ontime 0, and it can be verified that the MPNs of these pitches are 48, 59, and 64 respectively. The next excerpt of the piece given in (17.2) corresponds to bars 33–34 (the beginning of variation II, Fig. 17.3(b)). The beginning of bar 33 has ontime 192 (although calculating this is nontrivial, given some intervening repeat marks and first/second

---

[3] The order is called *lexicographic order*. It is most easily explained in relation to (17.2). For instance, $(0,64)$ is lexicographically less than $(1,50)$ because $0 < 1$. If there is a tie in the x-dimension, it is broken by the values in the y-dimension, which is why $(1,50)$ is lexicographically less than $(1,59)$. And so on.

[4] Note that Meredith (2006) defines the morphetic pitch of $\ldots$, A♭0, A0, A♯0, $\ldots$ to be 0, so that middle C has a morphetic pitch of 23.

time bars), and its first note is E2, or MPN 48. The movement ends on bar 293, beat 3, ontime 891, with a five-note chord, of which the top three notes are B3, E4, and G♯4, or MPNs 59, 62, and 64 respectively.

In what follows, we want to be able to perform matching that is invariant to time-shifting, time-scaling, transposition, or any combination of these operations, so we will use triples of points $(\mathbf{d}_i, \mathbf{d}_j, \mathbf{d}_k)$ (where i, j and k are the indices of the points in the lexicographically ordered dataset) such that:

1. the points are local, obeying $i < j < k$, with $j - i < 5$ and $k - j < 5$;
2. the ontimes are not simultaneous, i.e., $x_i \neq x_j$ and $x_j \neq x_k$;
3. the ontimes are proximal, with $x_j - x_i < 10$ and $x_k - x_j < 10$; and
4. the MPNs are proximal, with $y_j - y_i < 24$ and $y_k - y_j < 24$.

These criteria were selected based on previous work (Arzt et al., 2012). A *fingerprint*— that is, the information stored for each $(\mathbf{d}_i, \mathbf{d}_j, \mathbf{d}_k)$—is a quadruple, ⟨token, piece ID, ontime, ontime difference⟩, where each token is itself a triple:

$$\left\langle \underbrace{\left\langle y_j - y_i, \; y_k - y_j, \; \frac{x_k - x_j}{x_j - x_i} \right\rangle}_{\text{token}}, \; \underbrace{\text{``beethoven123''}}_{\text{piece ID}}, \; \underbrace{x_i}_{\text{ontime}}, \; \underbrace{x_j - x_i}_{\text{ontime difference}} \right\rangle. \quad (17.3)$$

For the three underlined points in (17.2), which form a legal triple according to criteria 1–4 above, the fingerprint is

$$\left\langle \left\langle 62 - 64, \; 60 - 62, \; \frac{194\frac{1}{4} - 193\frac{1}{4}}{193\frac{1}{4} - 192\frac{1}{2}} \right\rangle, \; \text{``beetOp109Mvt3''}, \; 192\frac{1}{2}, \; 193\frac{1}{4} - 192\frac{1}{2} \right\rangle$$

$$= \left\langle \left\langle -2, \; -2, \; 1\frac{1}{3} \right\rangle, \; \text{``beetOp109Mvt3''}, \; 192\frac{1}{2}, \; \frac{3}{4} \right\rangle. \quad (17.4)$$

For each legal triple $(\mathbf{d}_i, \mathbf{d}_j, \mathbf{d}_k)$ in the point set D, a fingerprint is calculated and stored in a so-called *fingerprint database*.

Given a query point set Q, which represents some known theme or otherwise-interesting excerpt (from the same piece or from another piece), the fingerprint database calculated over the point set D can be used to find ontimes $t_1, t_2, \dots, t_m$ in D at which events similar to the query Q occur. First, it is necessary to calculate the fingerprints of triples $(\mathbf{q}_i, \mathbf{q}_j, \mathbf{q}_k)$ from Q, in an analogous fashion to the calculations over D. Then tokens from the query are matched against tokens from the database. When there is a match, the ontime u_l of the matching fingerprint in the database and the ontime v_l of the matching fingerprint in the query are recorded as a pair (u_l, v_l). These ontimes are readily accessible, being stored as the third element in a fingerprint (see (17.3)).

Let the set of ontime pairs of matching tokens be denoted

$$U(Q, D, \alpha) = \{(u_1, v_1), (u_2, v_2), \dots, (u_L, v_L)\}, \quad (17.5)$$

where α is a parameter to be described in due course.

If the piece contains a transformation of the query Q, then an arbitrary point $\mathbf{q} \in Q$ will be expressible as $\mathbf{q} = (ax_i + b, y_i + c)$ for some $(x_i, y_i) \in D$, where a is the time scale, b is the time shift, and c is the transposition. Substituting this expression for some triple $(\mathbf{q}_i, \mathbf{q}_j, \mathbf{q}_k)$ from Q in (17.3), these operations will cancel, and so the query and database tokens will match:

$$
\left\langle (y_j + c) - (y_i + c),\ (y_k + c) - (y_j + c),\ \frac{(ax_k + b) - (ax_j + b)}{(ax_j + b) - (ax_i + b)} \right\rangle
$$
$$
= \left\langle y_j - y_i,\ y_k - y_j,\ \frac{x_k - x_j}{x_j - x_i} \right\rangle .
$$
(17.6)

While being able to match queries to instances that have undergone such transformations is useful, composers often write more complex variations of themes into their works than can be expressed in terms of the transformations considered above (i.e., time-shifting, time-scaling and transposition). As an example of more complex variation, let us take the opening two bars of Beethoven's op. 109, mvt. 3, as a query (Fig. 17.3(a)), the transition into the complex variation shown in Fig. 17.3(b) as database, and see what would be required to match the two via symbolic fingerprinting. The query is

$$
Q = \langle (0,48), (0,59), \underline{(0,64)}, (1,50), (1,59), \underline{(1,62)}, (2,51), \underline{(2,60)}, (2\tfrac{1}{2}, 63),
$$
$$
(3,52), (3,56), (3,61), (4,53), (4,56), (4,58), (4,59), (5,54) \rangle .
$$
(17.7)

Taking the underlined triple in (17.7), which corresponds to the underlined triple in (17.2), the fingerprint token would be $\langle -2, -2, 1 \rangle$. Comparing with $\langle -2, -2, 1\tfrac{1}{3} \rangle$ from (17.4), the disparity between the two is in the final element—the time difference ratio of 1 in the query token versus $1\tfrac{1}{3}$ in the database token. If however, we permit some percentage error, $\alpha = 40\%$ say, when matching tokens' time difference ratios, then the query token $\langle -2, -2, 1 \rangle$ would be considered a match to the database token $\langle -2, -2, 1\tfrac{1}{3} \rangle$, and the corresponding ontimes would be included in U from (17.5).

Plotted in Fig. 17.4(a) are the ontime pairs of matching tokens $U(Q, D, \alpha = 40)$ for the query Q from (17.7) and the point set D from (17.2). As there are coincident points, we use marker size to indicate the relative number of matches at a particular coordinate, with larger circles indicating more matches. The presence of approximately diagonal lines in this plot means that there are multiple subsequent matches between query and database (i.e., that there is a more or less exact occurrence of the query in the database). Two such occurrences are indicated by the two thick dashed transparent lines in Fig. 17.4(a). To summarize this plot properly, affine transformations are applied to the points (indicated by the arrows and straight vertical lines) and they are binned to give the histogram shown in Fig. 17.4(b). The histogram shows two peaks—one at ontime 192 (or bar 33) and another around ontime 198 (bar 35). There is an occurrence of the theme from Fig. 17.3(a) at each of these times,
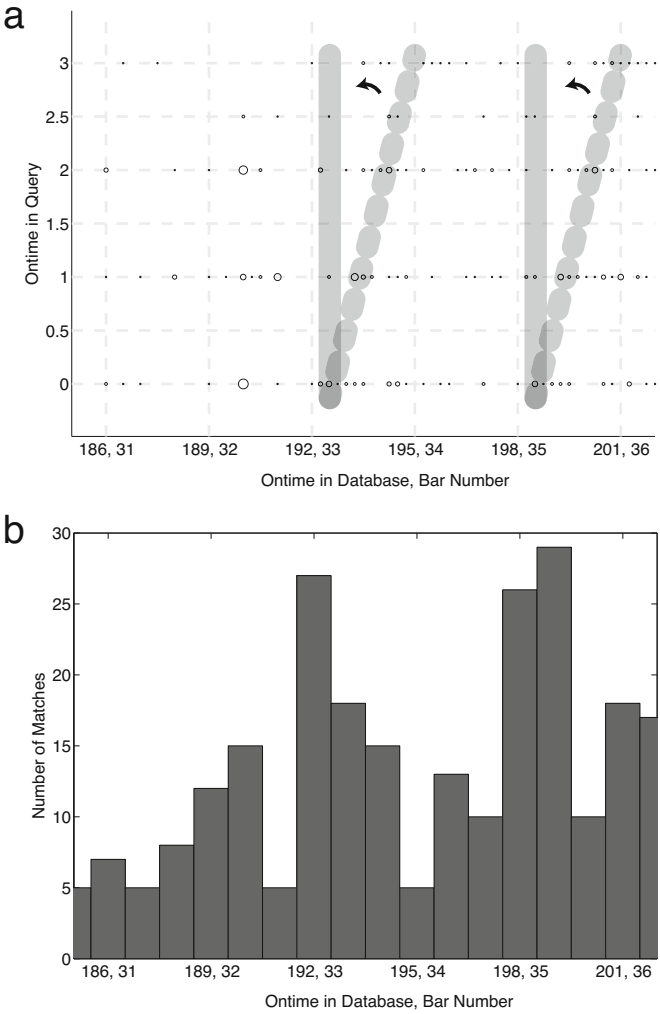
Fig. 17.4 (a) Plot of time stamps for matching query and database fingerprint tokens. Size of circular markers indicates the relative number of matches coincident at a particular point, with larger circles meaning more matches. Two occurrences of the query in the database are indicated by the two thick dashed transparent lines. The arrows and straight vertical lines allude to an affine transformation. (b) Fingerprint histogram indicating the similarity of the piece to the query as a function of time. This plot results from application of an affine transformation to the points in Fig. 17.4(a), followed by binning the transformed points to give the histogram. The rotation in the affine transformation is influenced by the value of the ontime difference, stored as the final element of a fingerprint (see (17.3), and, for more details, Arzt et al. (2012))

subject to quite complex variation as shown in Fig. 17.3(b) (where red and blue highlighting indicates likely contributors to the two peaks in the histogram).

In summary, symbolic fingerprinting can be used to identify occurrences of a given query in a point-set representation of a piece. The stronger the resemblance to the query at a particular time in the piece, the larger the number of matches in a fingerprint histogram such as Fig. 17.4(b). In what follows, we refer to the fingerprint histogram as $f(t)$, and use it as a measure of the similarity of the piece to the query as a function of time t. For the purposes of comparing different queries and different pieces, it is also convenient to normalize the fingerprint histogram so that the y-axis is in the range $[0, 1]$. For the purposes of analysing occurrences of a query across multiple pieces, we will also concatenate point sets $D_1, D_2, \ldots, D_N$ representing N pieces into one point set D. That is, we set $D = D_1$ and then for $i = 2, \ldots, N$, the set D_i is shifted to begin shortly after D_{i-1} ends, and then appended to D. To distinguish between the fingerprint histogram for a query calculated over some collection of pieces $\Theta = \{D_1, D_2, \ldots, D_N\}$ as opposed to some other collection $\Theta' = \{D'_1, D'_2, \ldots, D'_{N'}\}$, we will write $f_\Theta(t)$ and $f_{\Theta'}(t)$ respectively.

Among the advantages of symbolic fingerprinting are its speed and robustness (Arzt et al., 2012). As demonstrated, it is capable of identifying query occurrences that have had time shift, time scale, and transposition applied, as well as more complex transformations. Symbolic fingerprinting is not necessarily a definitive solution to the problem of modelling perceived music similarity, however. For example, based on the ontime-MPN representation, it is unlikely that the α-parameter could be increased sufficiently to identify Lutes' pattern occurrences without also returning many false positive matches. The concision of the fingerprint histogram can also be a double-edged sword. Especially with the α-parameter increased, sometimes there is a peak in the histogram (say, at ontime 190 in Fig. 17.4(b)), but, when referring back to the music, one is hard-pressed to justify the peak's existence.

17.3.3 Calculating the Probability of a Pattern Occurrence

To develop geometric equivalents of the distinctiveness measure, given in (17.1), we must be able to calculate the empirical probability of a given pattern occurrence in a piece or across multiple pieces. Symbolic fingerprinting, described in the previous section, will be central to this development. If we calculate fingerprints for a short pattern occurrence, as well as fingerprints for a whole piece or collection of pieces, it is possible to construct a histogram $f_\Theta(t)$ measuring the similarity of the pattern occurrence to the piece(s) as a function of time. Strong matches appear as large global peaks in the histogram, whereas partial/weaker matches appear as smaller local peaks or are indistinguishable from chance matches. Examples of such histograms are given in Fig. 17.5. The dashed curve shows $f_\Theta(t)$ for the query highlighted in red in Fig. 17.3(a). Intuitively, this is quite a specific query, with a relatively low likelihood of occurrence across Beethoven's piano sonatas (apart perhaps from its host piece, op. 109, mvt. 3). In accordance with this intuition, $f_\Theta(t)$ for this *specific* query is
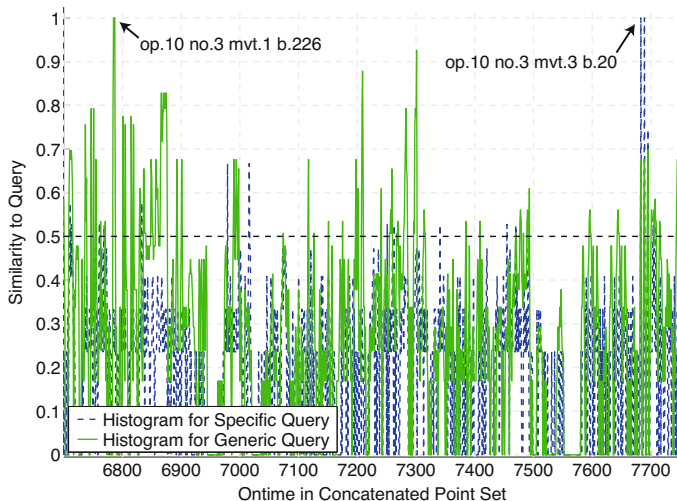
Fig. 17.5 Fingerprint histograms $f_\Theta(t)$ for a specific query (*dashed blue line*) and $g_\Theta(t)$ for a generic query (*solid green line*). The *x*-axis, time in the concatenated point set (database of Beethoven piano sonatas), extends from the end of op. 10, no. 3, mvt. 1, continuing through op. 10, no. 3, mvt. 2, and ending shortly after the beginning of op. 10, no. 3, mvt. 3

generally *below* the indicated similarity level of .5 in Fig. 17.5. For the sake of clarity, the time axis in Fig. 17.5 is restricted to a subset of our database, beginning towards the end of op. 10, no. 3, mvt. 1, continuing through op. 10, no. 3, mvt. 2, and ending shortly after the beginning of op. 10, no. 3, mvt. 3. Beyond op. 109, mvt. 3, two strong occurrences of the specific query stand out, in bars 18 and 20 of op. 10, no. 3, mvt. 3 (see arrow on the right of Fig. 17.5 and Fig. 17.3(c)).

The solid green curve, $g_\Theta(t)$, in Fig. 17.5 is the fingerprint histogram for a query consisting of a seven-note descending scale. Intuitively, this is quite a generic query, with a relatively high likelihood of occurrence across Beethoven's piano sonatas. That is, whilst listening to or studying Beethoven's piano sonatas, we would not be particularly surprised if a descending scale or scale fragment appeared. In accordance with this intuition, $g_\Theta(t)$ for the *generic* query in Fig. 17.5 is most often *above* $f_\Theta(t)$ for the specific query, and also quite often above the indicated similarity level of .5. A descending scale in the recapitulation of op. 10, no. 3, mvt. 1, is particularly noticeable (see arrow on the left of Fig. 17.5), but multiple other descending scale figures occur across these movements. Even though the generic query contains more notes than the specific query, the former appears to have a higher likelihood of occurrence than the latter.

In this chapter we use the superlevel set of a fingerprint histogram $f_\Theta(t)$ to formalize the intuitive sense of a query point set's likelihood. The superlevel set of $f_\Theta(t)$ is defined by

$$L_c^+(f_\Theta) = \{t \,|\, f_\Theta(t) \ge c\}\,, \tag{17.8}$$

which is the set of timepoints for which the histogram is equal to or exceeds some threshold similarity level c. The cardinality of the superlevel set (the number of timepoints it contains), divided by the total number of timepoint bins in the histogram, denoted $|L_0^+(f_\Theta)|$, can be used as a proxy for the empirical probability of observing the pattern occurrence across some piece or pieces.

Returning to the example queries and histogram excerpts shown in Fig. 17.5, the superlevel set (with parameter $c = .5$) for the specific query contains 638 timepoints, out of a total 13,003 timepoint bins. Thus the specific query has an empirical likelihood of $|L_{.5}^+(f_\Theta)|/|L_0^+(f_\Theta)| = 638/13,3003 = .049$. Meanwhile, the more generic query has an empirical likelihood of $|L_{.5}^+(g_\Theta)|/|L_0^+(g_\Theta)| = 1,345/13,003 = .103$. Thus the empirical likelihoods confirm our intuition: the highlighted note collection in Fig. 17.3(a) is less likely to occur in Beethoven's piano sonatas than a seven-note descending scale.

But what about *distinctiveness*? To bring this section to its natural conclusion, we turn back to Sect. 17.3.1 on distinctiveness, and substitute the above likelihoods into (17.1), writing

$$p(P|\Theta) = |L_c^+(f_\Theta)|/|L_0^+(f_\Theta)|\,, \tag{17.9}$$

where P is a point set representing some query, Θ is a collection of pieces in point-set representations, and $f_\Theta(t)$ is the fingerprint histogram of P across Θ. To measure how distinctive some pattern P is of some corpus Θ, relative to some anticorpus Θ', it follows that we can use

$$d(P,\Theta,\Theta') = \frac{|L_c^+(f_\Theta)| \times |L_0^+(f_{\Theta'})|}{|L_c^+(f_{\Theta'})| \times |L_0^+(f_\Theta)|}\,. \tag{17.10}$$

To avoid division by zero, we set $|L_c^+(f_{\Theta'})|$ equal to a minimum of 1.[5] Completing the worked example, we can take Chopin's piano sonatas as an anticorpus, and calculate the distinctiveness of the specific and generic patterns for Beethoven's piano sonatas, relative to Chopin's.[6] The specific query has empirical likelihood of .075 in Chopin's piano sonatas, and so the distinctiveness of the specific query for Beethoven's piano sonatas relative to Chopin's is $.049/.075 = .660$. The generic query has empirical likelihood of .117 in Chopin's piano sonatas, and so the distinctiveness of the generic query is $.103/.117 = .887$. Importantly, this example demonstrates that specificity and distinctiveness are not the same thing. According to intuition, the specific query has lower probability of occurrence in Beethoven's piano sonatas than the generic query. The specific query does not appear to be more distinctive of Beethoven's sonatas than the generic descending scale, however. This is because the specific query has a relatively high likelihood of occurrence in Chopin's compared to Beethoven's sonatas (cf. .075 and .049), and so its distinctiveness is low. Distinctiveness values

---

[5] Division by zero arises if $L_c^+(f_{\Theta'})$ is empty, either because c is too high and/or $f_{\Theta'}$ too low (Conklin, 2010).

[6] Chopin's sonatas are: op. 4, mvts. 1–4; op. 35, mvts. 1–4; and op. 58, mvts. 1–4. It is worth noting that even though there are allusions to Beethoven in these works (cf. Chopin, op. 35, mvt. 1, and Beethoven, op. 111, mvt. 1), the focus in this chapter is Beethoven–Beethoven resemblances, not Chopin–Beethoven resemblances.

greater than one indicate that a pattern is more probable in the corpus than the anticorpus, and so more distinctive of the corpus.

Choice of anticorpus will affect the results, so, for the first time here, we consider the impact of this choice by reporting results for a second anticorpus also: Chopin's mazurkas.[7] These corpora (21 Beethoven sonata movements, twelve Chopin sonata movements, 49 Chopin mazurkas) may appear to differ in size, but they are comparable (to within 1,000) in terms of number of notes.

17.4 Experimental Results

Although we were not necessarily expecting to detect instances of literal borrowing across Beethoven's piano sonatas (because they are not already known, suggesting perhaps there are none), it is prudent to at least check. Experiment 1 was designed primarily with this aim in mind. Experiment 2 increased the temporal inexactness parameter α to investigate less literal resemblances, and Experiments 3 and 4 were similar to 1 and 2 but for melodic rather than polyphonic queries. The last two experiments apply a pattern discovery algorithm SIARCT-CFP (Collins et al., 2013) to pairs of Beethoven sonata movements to find—in an unsupervised manner—inter-opus resemblances between note collections (Experiment 5) and chord sequences (Experiment 6). Apart from discovering resemblances, the experimental results also shed some light on: (1) patterns that are distinctive of Beethoven's piano sonatas compared to Chopin's; (2) whether thematic material is inherently more *distinctive* than excerpts drawn from elsewhere in a movement; and (3) the impact of anticorpus choice.

17.4.1 Experiment 1

In Experiment 1, we defined twelve-note queries using the full polyphonic representation of each movement, beginning at the start of each bar. This method of query definition is not exhaustive: if there are more than twelve notes in a given bar, then some content will be overlooked. Nor is the method always musically appropriate: if some phrase begins with an upbeat and/or contains fewer/more than twelve notes, then this will not be segmented appropriately. Taking a fixed number of notes is preferable, however, to taking a fixed time window and the (variable) number of notes appearing in this window, because the latter leads to variable-length queries, which could introduce biases in similarity calculations. In any case, our fixed-length queries are only intended to identify the *kernel* of some inter-piece resemblance, and then we can review the excerpts in question, to hear/see whether the resemblance extends

---

[7] Chopin's mazurkas include the following: op. 6, nos. 1–4; op. 6, nos. 1–4; op. 7, nos. 1–5; op. 17, nos. 1–4; op. 24, nos. 1–4; op. 30, nos. 1–4; op. 33, nos. 1–4; op. 41, nos. 1–4; op. 50 nos. 1–3; op. 56, nos. 1–3; op. 59, nos. 1–3; op. 63, nos. 1–3; op. 67, nos. 1–4; and op. 68 nos. 1–4.

Fig. 17.6 (a) Bars 64–70 of the third movement from Piano Sonata in F minor, op. 2, no. 1, by Beethoven. One occurrence of a twelve-note pattern is highlighted. (b) Bars 1–10.1 with upbeat of the first movement from Piano Sonata in D major, op. 10, no. 3, by Beethoven. One occurrence of a twelve-note pattern is highlighted, indicating a second occurrence of the pattern from Fig. 17.6(a)

over a longer time period, and to consider other contextual factors that need to be taken into account. In inter-opus pattern discovery in general, rarely will it suffice to present the algorithm output and say nothing more. Reviewing and interpreting the excerpts in question are vital steps towards producing a musical analysis.

Each twelve-note query was subject to fingerprinting analysis against the database of Beethoven piano sonata movements (as well as two anticorpora— Chopin's piano sonatas and mazurkas). Taking the fingerprint histogram $f(t)$ calculated for a query over the Beethoven piano sonata movements (see Sect. 17.3.2), we could determine the location of the strongest match to the query (other than in the piece where the query originated), as well as the strength of this strongest match. Queries that provide strong matches to segments from *other* movements may indicate instances of literal borrowing. The strongest-matching query to a segment from another movement is indicated in Fig. 17.6, and some summary statistics for the pattern, labelled A_1 and A_2, are given in Table 17.1.

As alluded to above, it so happens that A_1 and A_2 comprise the kernel of resemblance that extends over a longer time period: bars 65–66 of Fig. 17.6(a) appear in bars 5–7 of Fig. 17.6(b). The query alone is not particularly distinctive, consisting of a first-inversion triad played at successively lower scale steps twice and then one scale step higher. It is more interesting, however, that the resemblance between the two pieces extends over two bars in the first instance and three bars (because of

Table 17.1 Distinctiveness of Beethoven patterns relative to two anticorpora

Figure	Label	Anticorpus of Chopin Sonatas	Anticorpus of Chopin Mazurkas	Query Definition, Time Tolerance $\alpha =$
17.6(a), 17.6(b)	A_1, A_2	0.430	1.376	Polyphonic segment, 15%
17.7(a), 17.7(b)	B_1, B_2	0.476	1.188	Polyphonic segment, 15%
17.7(c), 17.7(d)	C_1, C_2	0.536	3.557	Polyphonic segment, 15%
17.8(a), (b)	E_1, E_2	0.799	1.870	Polyphonic segment, 40%
17.8(c), (d)	F_1, F_2	0.660	0.957	Polyphonic segment, 40%
17.9(a), 17.9(b)	G_1, G_2	12.880	2.084	Melodic segment, 15%
17.3(a), 17.3(c)	H_1, H_2	0.660	1.953	Melodic segment, 15%
17.10(a), 17.10(b)	I_1, I_2	2.036	4.305	Melodic segment, 40%
17.9(a), 17.9(c)	J_1, J_2	5.400	8.663	Note discovery, 15%
17.11(a), 17.11(b)	K_1, K_2	0.193	44.101	Note discovery, 15%
17.2(a), 17.11(c)	M_1, M_2	41.092	78.121	Note discovery, 15%
17.12(c), 17.12(d)	N_1, N_2	1.515	2.544	Chord discovery, 15%

the differing time signature) in the second instance. Of what does this extended resemblance consist? The extension consists of A_1 heard twice more at successively lower scale steps. In op. 2, mvt. 3, this causes a strong hemiola effect, with six beats of music being perceived as three groups of two (as opposed to the prevailing two groups of three). Beethoven arrives on chord V in bar 67, V^7 in bar 69, and then the theme from the trio returns with chord I in bar 70 (Barlow and Morgenstern, 1948). The context of the borrowing in op. 10, no. 3, mvt. 1, is different. According to Barlow and Morgenstern (1948), the theme of op. 10, no. 3, mvt. 1, covers bars 1–4.1 with upbeat. Bar 5 with upbeat is likely heard as a variation of the theme's opening, with the melody D–C♯–B–A in both cases. Therefore, the instance of borrowing in op. 10, no. 3, mvt. 1, has a different function from that in op. 2, no. 1, mvt. 3. In op. 10, no. 3, mvt. 1, it is a variation of the theme, followed by a perfect cadence in bars 9.2–10.1. Taken as a whole, bars 5–10.1 with upbeat act as a consequent to the antecedent of bars 1–4.1 with upbeat, with bars 1–4.1 concluding on scale degree $\hat{5}$.

As we did not expect to find instances of inter-piece resemblance stretching 2–3 bars, the result shown in Fig. 17.6 is surprising and, to our knowledge, novel. Other results from the first experiment did not extend to create a longer period of resemblance, but two more are included in Table 17.1 (labelled B and C) and given in Fig. 17.7. The third movement of op. 2, no. 1, features in all three patterns A, B, C, suggesting it contains a stock of patterns—albeit not particularly distinctive according to Table 17.1—that appear in later compositions.

Fig. 17.7 (a) Bars 1–5 with upbeat of the third movement from Piano Sonata no. 1, in F minor, op. 2, no. 1, by Beethoven. One occurrence of a twelve-note pattern is highlighted. (b) Bars 8–10 of the first movement from Piano Sonata no. 30 in E major, op. 109, by Beethoven. One occurrence of an eleven-note pattern is highlighted, indicating a partial second occurrence of the pattern from Fig. 17.7(a). (c) Bars 24–28.1 of the third movement from Piano Sonata no. 1 in F minor, op. 2, no. 1, by Beethoven. One occurrence of a twelve-note pattern is highlighted. (d) Bars 15–16 of the fourth movement from Piano Sonata no. 7 in D major, op. 10, no. 3, by Beethoven. One occurrence of a twelve-note pattern is highlighted, indicating a second occurrence of the pattern from Fig. 17.7(c)

17.4.2 Experiment 2

The previous experiment aimed towards identifying instances of literal repetition or borrowing between movements, so in that experiment it was sensible to keep the temporal inexactness parameter α quite low ($\alpha = 15\%$). As symbolic fingerprinting can be used to identify non-rigid variations such as between Fig. 17.3(a) and (b), however, in the second experiment, α is increased to 40% to enable such discoveries. Everything else from Experiment 1 is kept the same.

Many of the patterns discovered in Experiment 2 were similar to or the same as those discovered in Experiment 1. But beyond these, two examples of the inexact inter-opus patterns discovered in Experiment 2 are given in Fig. 17.8. Pattern occurrence E_1 in Fig. 17.8(a) consists of a rising major triad, F3, A3, C4, with each note played

Fig. 17.8 (a) Bars 1–5 with upbeat of the third movement from Piano Sonata no. 6 in F major, op. 10, no. 2, by Beethoven. One occurrence of a twelve-note pattern is highlighted. (b) Bars 172–173 of the third movement from Piano Sonata no. 30 in E major, op. 109, by Beethoven. One inexact occurrence of the pattern from Fig. 17.8(a) is highlighted. (c) Bars 47–50 of the second movement from Piano Sonata no. 6 in F major, op. 10, no. 2, by Beethoven. One occurrence of a twelve-note pattern is highlighted. (d) Bars 1–4 with upbeat of the third movement from Piano Sonata no. 12 in A♭ major, op. 26, by Beethoven. One inexact occurrence of the pattern from Fig. 17.8(c) is highlighted

three times following a lower member of the triad. In occurrence E_2 (Fig. 17.8(b)), the three notes are D♯5, F♯5, A5, with each note played twice, and on this occasion they form the upper three notes of a dominant seventh chord that has B1 in the bass. There are fewer notes in E_2 than in E_1, and the time difference ratios between triples of notes in each occurrence are not always the same, but with $\alpha = 40\%$ the two occurrences bear sufficient resemblance to cause a local maximum in the fingerprint histogram. The same observation applies to pattern occurrences F_1 (Fig. 17.8(c)) and F_2 (Fig. 17.8(d)). This pattern consists of a chordal progression, with each occurrence

having very similar voice-leading. The progression is I, I, Vb in Fig. 17.8(c), and i, i, V^7b in Fig. 17.8(d).

Fig. 17.9 (a) Bars 122–137.1 of the first movement from Piano Sonata no. 7 in D major, op. 10, no. 3, by Beethoven. One occurrence of a six-note pattern is highlighted in red. A different nineteen-note pattern is highlighted in blue. (b) Bars 43–44 of the second movement from Piano Sonata no. 7 in D major, op. 10, no. 3, by Beethoven. One inexact occurrence of the pattern from Fig. 17.9(a) is highlighted. (c) Bars 54–57.2 of the first movement from Piano Sonata no. 1 in F minor, op. 2, no. 1, by Beethoven. A second occurrence of the nineteen-note pattern from Fig. 17.9(a) is highlighted

17.4.3 Experiment 3

Since Handel sometimes borrowed melodies rather than full textures (see Sect. 17.1), it is sensible to make the same checks for Beethoven. Accordingly, Experiment 3 consisted of a melodic version of Experiment 1. We defined six-note queries using the highest-sounding notes in the right hand, beginning at the start of each bar. Everything else from Experiment 1 is kept the same.

The pattern H_1 from Fig. 17.3(a), recurring as H_2 in Fig. 17.3(c), is among the results of this experiment. In op. 109, mvt. 3, H_1 is the opening of a theme and variations. In op. 10, no. 3, mvt. 3, H_2 is imitative, passing between left and right hands at the beginning of the second section of the minuet. According to Table 17.1, however, pattern H_1 is not particularly distinctive of Beethoven's sonatas ($d = .660$) compared, say, to pattern occurrences G_1 and G_2 from Figs. 17.9(a) and (b), respectively ($d = 12.880$). The leap of six scale steps E3 to D4 in G_1 is the reason for this pattern's high distinctiveness. In op. 10, no. 3, mvt. 1, the first three notes of G_1 belong to a lower-octave repetition of the development's opening (see bar 125 with upbeat of Fig. 17.9(a)). This is the opening theme beginning on scale degree $\hat{5}$ instead of the original $\hat{1}$ (see Fig. 17.6(b)). The last three notes of H_1 belong to a minor statement of the opening theme, now beginning again on $\hat{1}$. Beethoven reuses these three-note scale fragments separated by a leap of six scale steps in the second movement of the same piece (Fig. 17.9(b)), just before the return of the main theme.

Fig. 17.10 (a) Bars 152–157 of the first movement from Piano Sonata no. 12 in A♭ major, op. 26, by Beethoven. One occurrence of a six-note pattern is highlighted. (b) Bars 126–128 of the third movement from Piano Sonata no. 30 in E major, op. 109, by Beethoven. One inexact occurrence of the pattern from Fig. 17.10(a) is highlighted

17.4.4 Experiment 4

Just as with Experiment 2, where we increased the temporal inexactness parameter of Experiment 1 from $\alpha = .15$ to $\alpha = .4$ to enable discovery of more complex variations on a pattern, so we did for melodic queries in Experiment 4. Many of the patterns discovered in Experiment 4 were similar to those discovered in Experiment 3, but an example of an additional pattern discovered in Experiment 4 is given as I_1 and I_2 in Fig. 17.10. Just as G_1 (Fig. 17.9(a)) was distinctive due to a large leap, so I_1 contains two compound sevenths. The inexactness of the relationships between notes belonging to I_1 and I_2, added to the large number of interpolated notes in I_2, make the resemblance between these passages difficult to discern. This was often the case for results from Experiment 4, inviting the observation that $\alpha = 40\%$ may be too high for melody-only queries.

Coming to the end of our experiments involving fixed-length queries, we address the issue of whether thematic material is inherently more *distinctive* than excerpts drawn from elsewhere in a movement. In each of the 21 Beethoven movements selected, it is the case that the first theme as annotated by Barlow and Morgenstern (1948) begins in bar 1. Therefore, to address the issue, we can calculate the distinctiveness of the opening six-note query from each movement, and compare this value to the distinctiveness of some other six-note query selected at random from the same movement. This procedure can be simulated many times for each movement to derive a proportion ρ of the times that the distinctiveness values of themes are significantly higher than those of randomly-selected queries. If a query Q belongs to piece D_i, then it makes sense to set the corpus Θ from (17.1) to this piece, $\Theta = \{D_i\}$, and the anticorpus Θ' to all other Beethoven movements apart from this piece, $\Theta' = \{D_1, D_2, \ldots, D_{i-1}, D_{i+1}, \ldots, D_{21}\}$. Then $d(Q, \Theta, \Theta')$ indicates how distinctive the query Q is for piece D_i, relative to its prevalence in other movements.

Following this procedure for $\alpha = 15\%$ and 1,000 simulations, we found that only on 18 occasions ($\rho = 1.8\%$) was distinctiveness of themes significantly higher according to the Wilcoxon signed-rank test than distinctiveness of queries selected at random from elsewhere in the same piece. The proportion ρ was still small (i.e., always below 5%), whether using $\alpha = 40\%$, twelve-note polyphonic queries, or Chopin as the anticorpus instead of other Beethoven movements. These results suggest that Beethoven's thematic material is not more distinctive of his style—at least quantitatively—compared to excerpts drawn from elsewhere in a movement.

17.4.5 Experiment 5

In the last two experiments reported in this chapter, we move away from predefined, fixed-length queries, and toward queries generated in an unsupervised manner by running the pattern discovery algorithm SIARCT-CFP (Collins et al., 2013) on concatenated pairs of movements $D = \text{conc}(D_i, D_j)$, where $1 \leq i, j, \leq 21$. That is, for instance, when running SIARCT-CFP on a point set consisting of the concatenation

Fig. 17.11 (a) Bars 103–107 with upbeat of the first movement from Piano Sonata no. 12 in A♭ major, op. 26, by Beethoven. One occurrence of a 28-note pattern is highlighted. (b) Bars 41–44 of the third movement from Piano Sonata no. 30 in E major, op. 109, by Beethoven. A second occurrence of the pattern from Fig. 17.11(a) is highlighted. (c) Bars 9–11 of the second movement from Piano Sonata no. 7 in D major, op. 10, no. 3, by Beethoven. A second occurrence of the pattern from Fig. 17.2(a) is highlighted (in red)

of op. 10, no. 3, mvt. 1, and op. 2, no. 1, mvt. 1, the pattern occurrences labelled J_1 (Fig. 17.9(a)) and J_2 (Fig. 17.9(c)) are discovered, and we can calculate the distinctiveness of all such output patterns. Until recently, it would not have been feasible to run geometric pattern discovery algorithms across all pairs of movements from a 21-piece corpus, but Collins (2011) introduced a parallel version of SIA (Meredith et al., 2002) called SIAR, whose runtime reduces with the number of available processors.

Pattern occurrences J_1 and J_2 provide a kernel of resemblance that could point to a more pervasive Beethovian stylistic trait: both bars 133–137.1 of Fig. 17.9(a) and bars 55-57.1 of Fig. 17.9(c) consist of alternating octave eighth notes in the left hand; both consist of a global descending melodic contour in the right hand with a concluding upturn; both involve dynamic emphases at beginning and end; both excerpts appear early on in the development sections of op. 10, no. 3, mvt. 1, and op. 2, no. 1, mvt. 1 (although the latter has been heard before as the second theme). So it would be interesting to know whether further instances appear in other pieces by Beethoven, Haydn, etc. Potentially, this could be an example of a new Beethovian developmental schema.

Similar remarks about schematic potential could be made regarding K_1 (Fig. 17.11(a)) and K_2 (Fig. 17.11(b)), since both constitute chordal textures in

theme and variations movements. An interesting aspect of this pattern is that while it does not appear to be distinctive of Beethoven's piano sonatas relative to Chopin's ($d = 0.193$ in Table 17.1), it is distinctive of Beethoven's sonatas relative to Chopin's mazurkas ($d = 44.092$). This trend in distinctiveness values was present across many discovered patterns (e.g., see M_1 and M_2 in Table 17.1 and Figs. 17.2(a) and 17.11(c) respectively), but it was most marked for K_1, K_2. Overall, the choice of anticorpus does not seem to create too much volatility in results: if pattern X has distinctiveness d_X for the Chopin sonata anticorpus and e_X for the Chopin mazurka anticorpus, pattern Y has distinctiveness d_Y for the Chopin sonata anticorpus and e_Y for the Chopin mazurka anticorpus, and there is some distinctiveness ordering for one of the anticorpora (e.g., $d_X < d_Y$), then apart from a couple of exceptions in Table 17.1, this ordering holds also for the other anticorpus (e.g., $e_X < e_Y$).

17.4.6 Experiment 6

The final experiment explored a new representation in the context of pattern discovery: geometric encoding of chord symbols. To create such a representation, the HarmAn algorithm (Pardo and Birmingham, 2002) was run on all pieces in our database. For given symbolic note input, HarmAn produces ontimes of chord labels, the root pitch classes, chord types, chord durations, and ratings of the confidence with which each label was assigned. It does not take into account modulations, inversions, or functional harmonic labels. Focusing on chord ontimes and root pitch classes, an example of HarmAn output is shown in Fig. 17.12(b) for the input of Fig. 17.12(a). SIARCT-CFP was used to identify collections of points that occur repeated and/or varied in the ontime-root space of Fig. 17.12(b), just as it had done in ontime-MPN space for Experiment 5.

An example of the type of harmonic pattern discovered is given in Figs. 17.12(c) and (d), where chord labels have been converted manually to Roman numerals for clarity. Both N_1 (Fig. 17.12(c)) and N_2 (Fig. 17.12(d)) begin by moving from i to V and back to i. Whereas N_2 then mimics this movement to reach the dominant minor (V, II, v), in N_1 the pattern occurrence finishes on V^7 and then cadences back on to i a bar later. Pattern discovery on a large scale in this harmonic space is still in its infancy, because robust algorithms capable of labelling chords to the standard of a human expert still do not exist (Collins, 2014).

17.5 Conclusions and Future Work

This chapter has described a computational method for discovering resemblances between excerpts from different pieces of polyphonic music, and summarized the results of this method applied to a collection of piano sonatas by Beethoven. The method combines formulae for calculating the distinctiveness of repeating patterns—
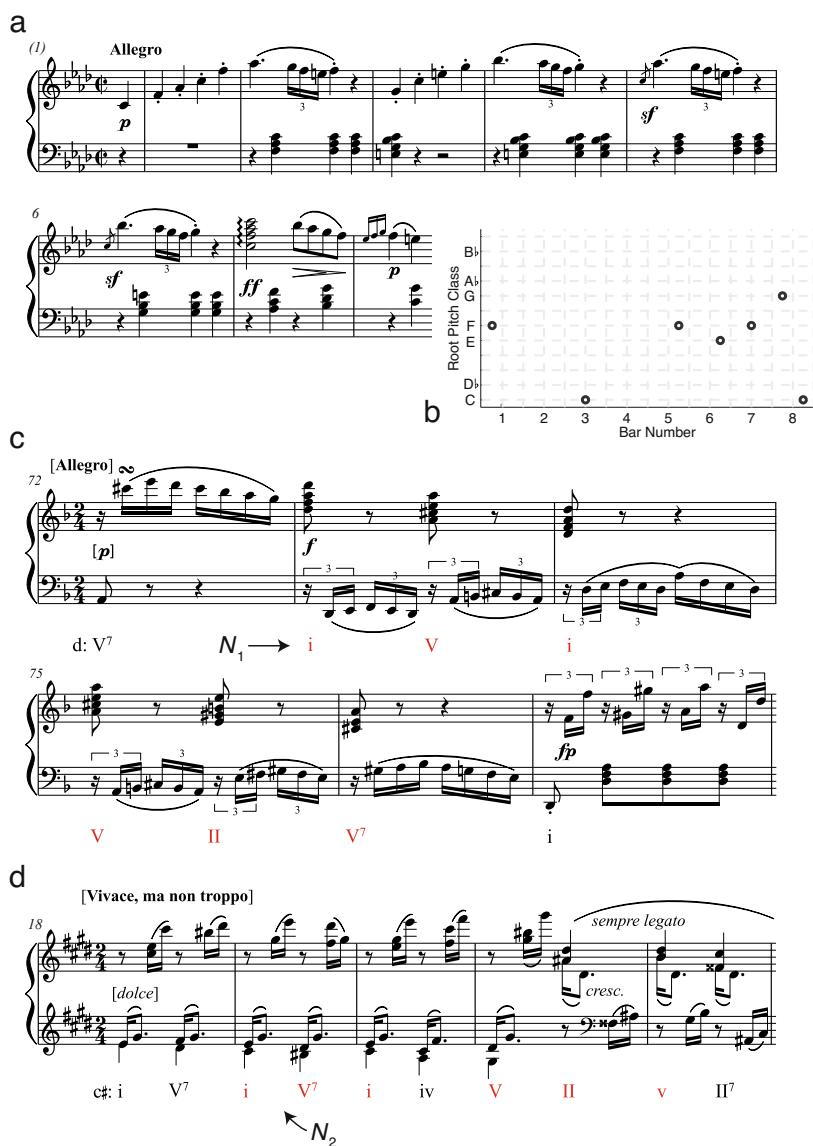
Fig. 17.12 (a) Bars 1–8.1 with upbeat of the first movement from Piano Sonata no. 1 in F♭ minor, op. 2, no. 1, by Beethoven. (b) Plot of root pitch class against ontime of chords, as estimated by the HarmAn algorithm. (c) Bars 72–77 of the first movement from Piano Sonata no. 6 in F major, op. 10, no. 2, by Beethoven. An occurrence of a six-chord pattern is highlighted below the staff. (d) Bars 18–22 of the first movement from Piano Sonata no. 30 in E major, op. 109, by Beethoven. A second occurrence of the six-chord pattern from Fig. 17.12(c) is highlighted. Chord labels were converted to Roman numerals in these plots to aid comparison

developed in the viewpoint approach to pattern discovery—with representations that apply conveniently to melodic or polyphonic (voiced or unvoiced) representations—developed in the geometric approach to pattern discovery. The combination or unification was made possible by symbolic fingerprinting, which enabled us to estimate the likelihood of occurrence of a music query across a piece or database of pieces.

Of particular interest among the results was an instance of literal resemblance between op. 2, no. 1, mvt. 3 and op. 10, no. 3, mvt. 1 (Fig. 17.6), a melodic resemblance between an andante theme in op. 109, mvt. 3 and an imitative passage in the minuet, op. 10, no. 3, mvt. 3 (Fig. 17.3), and two new candidates for schemata (Figs. 17.9 and 17.11). A quantitative analysis of the distinctiveness of Beethoven's thematic material versus material drawn from elsewhere in the same pieces found that the themes are not significantly more distinct. One should be wary of interpreting null effects, but this result is perhaps not so surprising when one considers the content of Beethoven's motivic designs. For instance, the themes in both Figs. 17.8(a) and 17.12(a) consist of rising triadic pitches followed by a descending scale fragment. Such musical building blocks have relatively high likelihood across many corpora, and so may not be distinctive of Beethoven. Based on this result, we might tentatively suggest that the apparent bias in existing analytic work toward discussing incipits and themes could be due to cognitive load (easier to remember and maintain incipits and themes than entire movements) rather than this material being particularly distinctive of a composer's style. Computational music analysis may serve to democratize intra- and inter-opus pattern discovery somewhat, giving equal consideration to less well-known passages from longer pieces, which could lead to interesting new findings.

At the beginning of the chapter, we stated that computational methods exist capable of discovering the type of melodic resemblance shown in Fig. 17.1, but that no computational method exists capable of identifying the more abstract resemblances in Fig. 17.2. As yet, therefore, music informaticians have not developed sufficiently flexible methods for identifying the types of patterns that musicologists find interesting (Burkholder, 2001; Byros, 2012; Gjerdingen, 1988, 2007; Lutes, 1974; Radcliffe, 1968; Winemiller, 1997). The method proposed and applied in the current chapter has made some progress in this regard, because it can handle melodies as well as any sort of polyphonic texture, and it addresses the interpolation and zero-order problems, which are simplifying and problematic assumptions about the nature of music made by previous approaches. If, however, the temporal inexactness parameter α of our fingerprinting method is increased to the point where occurrences of schema such as in Fig. 17.2 become detectable, then many false positive results are returned also. The method proposed in the current chapter is not a complete solution, therefore, and one idea for future work would be to query point-set representations of notes and chord labels simultaneously, to help filter out the number of false positives resulting from increases in the inexactness parameter.

Taking a step back, a more fundamental suggestion for future work is to create high-quality encodings in kern or MusicXML format, synchronized to audio recordings, of specific works that can support computational research into borrowing and distinctive pattern discovery. For instance, while there are a good number of Handel's works available in kern format, there are far fewer encodings of works by

contemporaries such as Keiser, let alone encodings of human-annotated instances of borrowing between these works. If we want to rigorously evaluate methods for identifying borrowing and resemblances—so that these methods can be applied more widely and with more confidence to other music—then testing against known instances of borrowing is a sensible first step. This testing requires that digital encodings of specific works and annotations of those works exist. Synchronizing audio recordings achieves two things: first, while it may sound trivial, it makes browsing of the discovered patterns a far more enjoyable process, because the researcher can listen to a human performance while doing so, rather than, say, inspecting piano-roll plots without sound or (possibly worse for some individuals) listening to mechanical MIDI files. This is important when trying to engage musicologists with the technology. Second, if audio and symbolic representations of a piece are synchronized, and there exist expert music-theoretical annotations of the symbolic data (e.g., functional harmonic analyses, cadence locations, textural categories, occurrences of schemata, occurrences of motifs, themes, and other repetitive elements), then via the synchronization this can act as a ground truth for evaluating audio-based algorithms (in addition to symbolic-based algorithms) that attempt to produce these annotations automatically.

We would like to see more joint work between technologists and musicologists on the development of interfaces that make it possible to run and browse the output of such algorithms across corpora and for representations of interest. The results presented in this chapter suggest that our knowledge about music *can* be advanced by computational music analysis—the extent to which it *will* be advanced in coming decades depends on researchers in music informatics getting to grips with some higher-level music-theoretic concepts, and musicologists being prepared to help guide and engage with this work.

Supplementary Material See http://www.tomcollinsresearch.net for code and data in support of the research reported in this chapter.

References

Arzt, A., Böck, S., and Widmer, G. (2012). Fast identification of piece and score position via symbolic fingerprinting. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR 2012)*, pages 433–438, Porto.

Barlow, H. and Morgenstern, S. (1948). *A Dictionary of Musical Themes*. Crown Publishers.

Burkholder, J. (2001). Borrowing. In Sadie, S. and Tyrrell, J., editors, *The New Grove Dictionary of Music and Musicians*. Macmillan, 2nd edition.

Byros, V. (2012). Meyer's anvil: Revisiting the schema concept. *Music Analysis*, 31(3):273–346.

Cambouropoulos, E. (2006). Musical parallelism and melodic segmentation. *Music Perception*, 23(3):249–267.

Caplin, W. (2013). *Analyzing Classical Form: An Approach for the Classroom*. Oxford University Press.

Collins, T. (2011). *Improved methods for pattern discovery in music, with applications in automated stylistic composition*. PhD thesis, Faculty of Mathematics, Computing and Technology, The Open University.

Collins, T. (2014). Stravinsqi/De Montfort University at the MediaEval 2014 C@merata task. In *Proceedings of the MediaEval 2014 Workshop*, Barcelona, Spain.

Collins, T., Arzt, A., Flossmann, S., and Widmer, G. (2013). SIARCT-CFP: Improving precision and the discovery of inexact musical patterns in point-set representations. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR 2014)*, pages 549–554, Curitiba, Brazil.

Collins, T., Laney, R., Willis, A., and Garthwaite, P. H. (2011). Modeling pattern importance in Chopin's mazurkas. *Music Perception*, 28(4):387–414.

Collins, T., Thurlow, J., Laney, R., Willis, A., and Garthwaite, P. (2010). A comparative evaluation of algorithms for discovering translational patterns in baroque keyboard works. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, pages 3–8, Utrecht, The Netherlands.

Conklin, D. (2010). Discovery of distinctive patterns in music. *Intelligent Data Analysis*, 14(5):547–554.

Conklin, D. and Bergeron, M. (2008). Feature set patterns in music. *Computer Music Journal*, 32(1):60–70.

Conklin, D. and Bergeron, M. (2010). Discovery of contrapuntal patterns. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, pages 201–206, Utrecht, The Netherlands.

Forth, J. C. (2012). *Cognitively-motivated geometric methods of pattern discovery and models of similarity in music*. PhD thesis, Department of Computing, Goldsmiths, University of London.

Gjerdingen, R. (1988). *A Classic Turn of Phrase: Music and the Psychology of Convention*. University of Pennsylvania Press.

Gjerdingen, R. (2007). *Music in the Galant Style*. Oxford University Press.

Janssen, B., de Haas, W. B., Volk, A., and van Kranenburg, P. (2013). Discovering repeated patterns in music: state of knowledge, challenges, perspectives. In *Proceedings of the International Symposium on Computer Music Multidisciplinary Research (CMMR 2013)*, pages 225–240, Marseille.

Knopke, I. and Jürgensen, F. (2009). A system for identifying common melodic phrases in the masses of Palestrina. *Journal of New Music Research*, 38(2):171–181.

Lartillot, O. (2005). Multi-dimensional motivic pattern extraction founded on adaptive redundancy filtering. *Journal of New Music Research*, 34(4):375–393.

Lewin, D. (1987). *Generalized Musical Intervals and Transformations*. Yale University Press.

Lutes, L. (1974). *Beethoven's re-uses of his own compositions*. PhD thesis, Department of Music, University of Southern California.

Marston, N. (1995). *Beethoven's Piano Sonata in E, op. 109*. Clarendon Press.

Meredith, D. (2006). The *ps13* pitch spelling algorithm. *Journal of New Music Research*, 35(2):121–159.

Meredith, D., Lemström, K., and Wiggins, G. A. (2002). Algorithms for discovering repeated patterns in multidimensional representations of polyphonic music. *Journal of New Music Research*, 31(4):321–345.

Meyer, L. (1980). Exploiting limits: creation, archetypes, and style change. *Daedalus*, 109(2):177–205.

Pardo, B. and Birmingham, W. P. (2002). Algorithms for chordal analysis. *Computer Music Journal*, 26(2):27–49.

Pearce, M., Ruiz, M., Kapasi, S., Wiggins, G., and Bhattacharya, J. (2010). Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation. *Neuroimage*, 50:302–313.

Pearson, K. (1909). On a new method of determining correlation between a measured character *a*, and a character *b*, of which only the percentage of cases wherein *b* exceeds (or falls short of) a given intensity is recorded for each grade of *a*. *Biometrika*, 7(1–2):96–105.

Pielou, E. and Foster, R. (1962). A test to compare the incidence of disease in isolated and crowded trees. *Canadian Journal of Botany*, 40:1176–1179.

Radcliffe, P. (1968). *Beethoven's String Quartets*. E.P. Dutton.

Roberts, J. (1986). Handel's borrowing from Keiser. *Göttinger Händel-Beiträge*, 2:51–76.

Wang, A.-C. and Smith, J. (2012). System and methods for recognizing sound and music signals in high noise and distortion. Patent US 8,190,435 B2. Continuation of provisional application from 2000.

Winemiller, J. (1997). Recontextualizing Handel's borrowing. *The Journal of Musicology*, 15(4):444–470.