FORMALIZING INVARIANCES FOR CONTENT-BASED MUSIC RETRIEVAL

Kjell Lemström
Department of Computer Science
University of Helsinki

Geraint A. Wiggins
Department of Computing
Goldsmiths, University of London

ABSTRACT

Invariances are central concepts in content-based music retrieval. Musical representations and similarity measures are designed to capture musically relevant invariances, such as transposition invariance. Though regularly used, their explicit definition is usually omitted because of the heavy formalism required. The lack of explicit definition, however, can result in misuse or misunderstanding of the terms.

We discuss the musical relevance of various musical invariances and develop a set-theoretic formalism, for defining and classifying them. Using it, we define the most common invariances, and give a taxonomy which they inhabit. The taxonomy serves as a useful tool for idetinfying where work is needed to address real world problems in content-based music retrieval.

1. INTRODUCTION

To effectively perform content-based music retrieval (CBMR), the intrinsic features of music must be taken into account. Some of the most important features correspond directly with invariances. Invariances related to pitch, tempo and duration are widely used, but usually without proper definition or discussion of their inter-relationship. Indeed, a single term is sometimes used to name multiple phenomena, admitting confusion about its real meaning.

Western musical scales may be transformed, or *transposed*, to any other key so that the corresponding pitch intervals remain intact. Indeed, Western people tend to listen to music analytically, observing pitch intervals rather than absolute pitch values. Thus, musical works are identified regardless of the prevalent musical key. The same observation is valid for tempo: two pieces of music are considered the same if the other is just played slower than the other (i.e., a different time scale is used). So transposition and time-scale invariance are important in CBMR applications.

However, in some cases mere transposition and time-scale invariance are not enough. For example, in query by humming, untrained singers often cannot produce pitch intervals accurately enough to constitute a match. To address this, several pitch class generalizations have been suggested, such as pitch contour [9] and qpi classification [4]. Using these generalizations, only direction of interval (contour) or the order of magnitude of interval (small, medium or large) is observed, respectively.

In this paper, we will define what it means when a representation or a method (algorithm) is invariant under a given notion arising from a musical phenomenon. We will give definitions for widely used invariances related to three main dimensions of music: *pitch*, *onset time* and *duration*. The latter two are temporal features and, usually, the third is derivable from the second. However, it is sometimes useful to separate them since the invariances as applied, categorised by our taxonomy, may differ. We will also define a set of more abstract, structural invariances. All of these inhabit a taxonomy that shows the relationships between the invariances, and also serves as a tool for identifying areas where further work in CBMR is needed.

2. DEFINING THE INVARIANCES

2.1 The representation

Let us start by defining the notion of a *representation*. In this context, we are modelling an observed phenomenon (music perception), and it is important not to presuppose that the data *is* the phenomenon; therefore, making the representation explicit is important too.

Let the size of a set S be denoted by $|S|$. Let the set of ordered subsets of set S of size between n and m, inclusive, be denoted by $S^{n\dots m}$, and where $n = m$, S^n; S^* is the power set of S. Given a set of features, $f_i \in F$, each with a unique type, $\tau_i \in \tau$, identified by an injection $T : F \mapsto \tau$, an *abstract representation*, ρ, is a subset of F. The type of each feature should be a mathematical specification (e.g., linear Abelian group for pitch) which is chosen to model the corresponding reality appropriately [13]. Given an abstract representation, ρ, a *concrete representation*, r, is a set of tuples

$$\{\langle f, \Sigma_f, \succ_f, \Phi_f, \Pi_f \rangle \mid f \in \rho\}$$

where Σ_f is an alphabet adequate to express f, $\succ_f$ is a partial order on Σ_f, Φ_f and Π_f are sets of functions and predicates, respectively, which apply to members of Σ_f defining the operations and tests required for the algebra of $T(f)$. Wiggins et al. [13] give detailed examples of datatypes for

	Feature invariances			Structural invariances
	Pitch	Onset Time	Duration	
Weaker/more specific	transposition (2)			ω-permutation (11/1)
↓	pitch-transposition (3)	time-position (6)		strongly permutation (11/2)
↓	pitch-warp (4)	time-scale (7)	time-scale (7)	ω-concatenation (12/1)
Stronger/less specific	Parsons (5)	time-warp (9)	duration-warp (8)	strongly concatenation (12/2)

Table 1. A sparse taxonomy on considered invariances. An invariance in the table subsumes the invariances above it, if no horizontal line appears in between. The number in parenthesis is that of the associated definition in Sections 2.4 and 3.

pitch and time. In general, $\succ_f$ is needed for the working of our formalism, not for the representation itself (there would be a member of Π_f for this, where appropriate); it is kept separate so that it may be different from any orders that are internal to the feature implementation, if necessary.

Let $\hat{\ }$ be a function which maps a concrete representation to its corresponding abstract representation.

Given a concrete representation, r, let an element e, $e \in r$, be a set of values, e_i, with concrete datatypes corresponding with r.

Let a *dataset*, E, be a set of elements. E is in r iff each e_i in E is in r.

2.2 A concatenator

To define invariances, we use a *concatenator* constructor.

A *concatenator*, $C_\omega^{r'}(E)$, constructs a lexicographically ordered multiset [1] of elements from a dataset, E, represented in r. The lexicographical order is specified by the ordered set $\omega \in \hat{r}^{1 \dots |r|}$ and the $\succ_i$ of the members of r corresponding with the members of ω. The superscript of the concatenator $r' \subseteq r$, gives the dimensions to be displayed.

For example, given a dataset, E, in an (abstract) representation including $\{pitch, onset, duration\}$ features, the concatenator $C_{\{onset\}}^{\{pitch\}}(E)$ creates a set of pitches ordered by onset time; one might use it to extract the pitches in a monophonic melody. If we generalise this to arbitrary features and combinations thereof, and consider only sequences including the first note of a piece, we arrive at the *viewpoint* representation of Conklin and Witten [3].

Evidently, the projective properties of this operator account for representational invariances where the invariant feature is an explicit feature in the representation, or a combination thereof. We use the term *capture* to denote this capacity: so projection to subsets of the existing feature set *captures* this kind of invariance.

For notational convenience we write operations applied to each member of an ordered set in order as operations on the set itself, where this is unambiguous, so, where A is a set of values and e is a value, $A \cdot e = \{a \cdot e \mid a \in A\}$; similarly, the elements of two sets of the same size, A, B may be combined pairwise in order under $\cdot$: $A \cdot B = \{a \cdot b \mid a_i \in A, b_i \in B\}$. Finally, to combine a value, $v \in \Sigma_f$, under an operation, $\cdot$, with one feature, f, of an element e, leaving other features unchanged, we write $e \cdot_f v$, so $e +_{pitch} k$ adds k to the $pitch$ feature of e.

In order neatly to specify a particular kind of derived invariance, we use $\langle S \rangle_f^{\cdot}$, where S is an ordered set, to denote the ordered set produced by ordered, pairwise operation on the feature f of elements $s_i \in S$ under $\cdot$. So, $\langle S \rangle_f^{-} = \{s_{i+1} -_f s_i \mid 1 \leq i < |S|\}$. This operation has consequences for the representation of the result: each feature type must be replaced by a derived type (corresponding with predefined ones where appropriate). For our concerns here, $pitch$ is replaced by $interval$, and $onset$ is replaced by ioi (inter-onset-interval), in the obvious way. We will need also a second-order derived invariance to be used with onsets (arriving at ioi proportions), thus: $\langle \langle S \rangle_{onset}^{-} \rangle_{ioi}^{\div} = \{ \frac{\langle s_{i+i} \rangle_{onset}^{-}}{\langle s_i \rangle_{onset}^{-}} \mid 1 \leq i < |S| - 1\}$.

2.3 Representational invariance

Some invariances can be captured by a change of representation. Whether or not this is possible depends on the representation used and on the nature of the phenomenon modelled. In many cases, a change of representation like this can usefully be thought of as indexing, and so it is helpful to know what remains invariant.

For example, because $pitch$ can be modelled by an Abelian group, it follows that for any set of pitches, E, thus modelled, there is another set formed by combining a constant member of Σ_{pitch} under the *plus* function in Φ_{pitch} with each member of E (the members of Σ_{pitch} are by definition in one-to-one correspondence with a partition of $\Sigma_{interval}$). It is implicit in the specification of the abstract representation that this operation, which is mathematically *translation*, models *musical pitch transposition*. Reversing this argment, it follows that any sequence of $pitch$es can be expressed as a sequence of $pitch$ differences, or *intervals*. Now, again because of the mathematical properties of the representation, it happens that each such interval is represented in Σ_{pitch}, and the algebra defined by Φ_{pitch} models the additive behaviour of intervals too: they also form an Abelian group. Thus, it is possible to produce a transposition-invariant version of any dataset, E, in any representation which contains $pitch$ and $onset$ information, by computing the ordered set whose members are computed by calculating $\langle C_{\{onset,pitch\}}^{\{pitch\}}(E) \rangle_p^{-}$. If the music modelled by E is monophonic, then this is the familiar interval sequence representation; however, if the music is not monophonic, care must be taken, because the *relative*

[1] This is a multiset because it is possible for the concatenator to map more than one element of E to any given element in the resulting representation; it may be necessary to know that this has happened.

nature of this representation makes its values dependent on their position in the sequence generated by the concatenator; therefore, one cannot apply many of the operations one would like. This fact is well-known to representers of music: an interval-based representation is not readily amenable to the representation of non-monophonic music. However, in this change of representation, relatively little information is lost: just one constant value, which tells us on what pitch the original dataset started; given that information, the entire original E may be reconstructed. In this sense, we say that the change is *structurally conservative*. However, though useful in itself, this property is neither necessary nor sufficient for a transformation to be useful.

For example, a familiar invariance transformation is that based on perceptual octave-equivalence, used in computing a chromagram. Here, perception maps exactly on to the mathematics, and so perceptual octave equivalence can be modelled by a chromatic equality function, defined as equality modulo n, where n is the number of divisions of the octave being used in the underlying scale of the pitch system. Here, Σ_{chroma} can very usefully be a contiguous subset of Σ_{pitch}, so $\mathbb{Z}_{12}$ does very nicely, and Φ_{chroma} and Π_{chroma} are equally easily defined. However, this representation change is also not, in general, structurally conservative, and it is mathematically evident why: the mapping from $\mathbb{Z}$ to $\mathbb{Z}_{12}$ is many-to-one, and so information is lost. The same principle, with a mapping to $\mathbb{Z}_8$, gives scale-degree representation, which is also octave-invariant.

A more interesting example is contour, an important aspect of melodic memory [8, §2.3]; Parsons coding [9] is a common way to represent the contour of music. However, $\Sigma_{Parsons} = \{-, 0, +\}$; it is not possible to give a fully defined $plus$ function over this set, while maintaining it as a model of musical contour, for obvious reasons. Therefore, we confirm that information is lost in changing to a representation whose pitch is based on Parsons coding, and one can argue this in advance because the abstract type of the Parsons code is not as expressive as a linear Abelian group. Thus, change of representation to Parsons code from, say, MIDI, is not structurally conservative. The same applies to comparable but more detailed interval-based representations such as the *qpi alphabet* [4].

Parsons coding captures an invariance which is *stronger* than transposition invariance in the sense that the equivalence classes it creates are fewer and larger. We will define two such invariances. In these, contour is preserved, but interval size is not—formal specifications are given in Definitions 4 and 5. Transposition from major to parallel minor is a (rather cautious) example of pitch warping; so, more generally, are interval augmentation and diminution in contrapuntal theory, or expansion and contraction in the music of Bartók. We note that among the passages captured by pitch warping lie also the equivalent transpositions, and this confirms that the stronger pitch-warp invariance is a indeed generalisation of transposition invariance. Therefore, a content-based music retrieval technique using Parsons coding can be seen as a filtering technique for finding transposed occurrences of a query (only filtering and not identifying, because false positives will be generated).

Our remaining common musical features, onset time and note duration, and the corresponding invariances (see Table 1) can be dealt with in the obvious way using the concatenator. For instance, given two datasets, $\mathcal{B}$ and $\mathcal{B}'$ in the same representation[2], two ioi sequences produced by the appropriate concatenator are time-scaled versions of each other if there is a number[3] d such that

$$C^{\{ioi\}}_{\{onset\}}(\mathcal{B}) = C^{\{ioi\}}_{\{onset\}}(\mathcal{B}') \times_{ioi} d.$$

A similar observation to that above, that time-warp invariance is stronger than time-scale invariance, applies here.

2.4 Algorithmic invariance

Music comparison is usually carried out in practice by an algorithm using a distance measure. Like representations, measures can be invariant under some property. At this level, we speak about *algorithmic invariances*. The following partial definition is a necessary but not sufficient condition to that end; it will be completed below.

Definition 1 *Let $\mathcal{M}$ be a CBMR method and P a property of a finite space*[4]*, where $|P|$ is the size of the space under consideration. $\mathcal{M}$ is* algorithmically P-invariant*, if working on datasets in representations in which the underlying datatype(s), explicit or implicit, of P does not introduce a factor into the computational complexity of $\mathcal{M}$ that is dependent from $|P|$.*

This definition rules out invariances achieved by discretizing a search space, enumerating it, and then searching exhaustively. Although such methods are sometimes called P-invariant in the MIR literature, this is really not the case; they are methods that merely appear to take advantage of invariance via brute-force calculation.

2.4.1 Pitch invariances

We now define the invariances in our taxonomy (Table 1), starting with pitch. Recall that our sets are by default *ordered* multisets. We omit duration, which is derivable from ioi, and abbreviate $\{pitch, interval, onset, ioi\}$ to $\{p, i, o, ioi\}$, respectively.

Definition 2 *Let r be a representation including $pitch$ and $onset$. A distance function $\mathcal{D}$ is* transposition-invariant *iff*

$$\forall a, b \in \Sigma_p.\forall \mathcal{A}, \mathcal{B}\ in\ r.\mathcal{D}(C^{\hat{r}}_{\{o\}}(\mathcal{A}), C^{\hat{r}}_{\{o\}}(\mathcal{B})) = \mathcal{D}(C^{\hat{r}}_{\{o\}}(\mathcal{A}) +_p a, C^{\hat{r}}_{\{o\}}(\mathcal{B}) +_p b).$$

It may be helpful to visualise Definition 2, as in Fig. 1. In this example, $\hat{r} = \{p, o\}$.

Note that Definition 2 captures the exact transposition invariance that a music theorist would expect of that property. At times, however, it is useful to have a more relaxed

[2] This restriction is not mathematically necessary, but to admit comparison between representations here would over-complicate the example.

[3] What kind of number depends on the kind of time representation: a metrical one would use $\mathbb{Z}$ or $\mathbb{Q}$; a real-time one might use $\mathbb{R}$.

[4] It may have been derived by quantizing a continuous space P'.

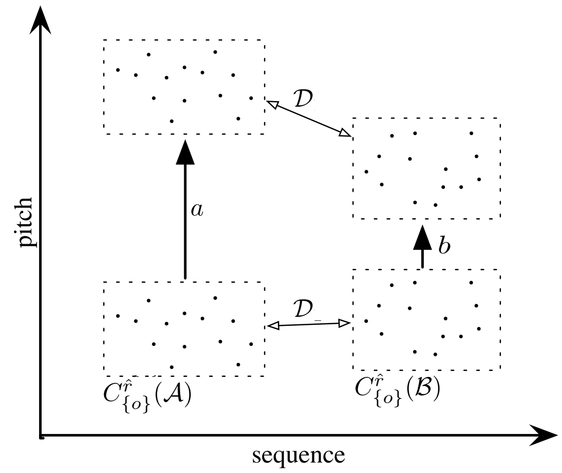
pitch
sequence
a
b
$\mathcal{D}$
$\mathcal{D}$
$C^{\hat r}_{\{o\}}(\mathcal{A})$
$C^{\hat r}_{\{o\}}(\mathcal{B})$

Figure 1. Visualisation of Definition 2. $\hat r = \{p, o\}$.

version of transposition invariance. Indeed, the following *pitch-transposition invariance*, which omits the exact onset times, is often used in music retrieval applications.

Definition 3 *Let r be a representation including pitch and onset. A distance function $\mathcal{D}$ is* pitch-transposition-invariant *iff*

$$\forall a, b \in \Sigma_p.\forall \mathcal{A}, \mathcal{B}\ in\ r.\mathcal{D}(C^{\hat r\setminus\{o\}}_{\{o\}}(\mathcal{A}), C^{\hat r\setminus\{o\}}_{\{o\}}(\mathcal{B})) = \mathcal{D}(C^{\hat r\setminus\{o\}}_{\{o\}}(\mathcal{A}) +_p a, C^{\hat r\setminus\{o\}}_{\{o\}}(\mathcal{B}) +_p b).$$

Stronger kinds of pitch invariance than the above (as defined in Section 2.3) are defined as follows.

Definition 4 *Let r be a representation including pitch and onset. A distance function $\mathcal{D}$ is* pitch-warp-invariant *iff*

$$\forall K_{\mathcal{A}} \in \mathcal{N}^{|\mathcal{A}|-1}.\forall K_{\mathcal{B}} \in \mathcal{N}^{|\mathcal{B}|-1}.\forall \mathcal{A}, \mathcal{B}\ in\ r.$$
$$\mathcal{D}\left(\langle C^{\hat r}_{\{o\}}(\mathcal{A})\rangle^-_p, \langle C^{\hat r}_{\{o\}}(\mathcal{B})\rangle^-_p\right) = \mathcal{D}\left(\langle C^{\hat r}_{\{o\}}(\mathcal{A})\rangle^-_p \times_i K_{\mathcal{A}}, \langle C^{\hat r}_{\{o\}}(\mathcal{B})\rangle^-_p \times_i K_{\mathcal{B}}\right)$$

where $\mathcal{N}$ is one of $\mathbb{Z}^+, \mathbb{Q}^+, \mathbb{R}^+$.

Note that the multiplication operation here needs to be duly definable in terms of functions in Φ_i. If we omit the onset information of that above, we get Parsons invariance:

Definition 5 *Let r be a representation including pitch and onset. A distance function $\mathcal{D}$ is* Parsons-invariant *iff*

$$\forall K_{\mathcal{A}} \in \mathcal{N}^{|\mathcal{A}|-1}.\forall K_{\mathcal{B}} \in \mathcal{N}^{|\mathcal{B}|-1}.\forall \mathcal{A}, \mathcal{B}\ in\ r.$$
$$\mathcal{D}\left(\langle C^{\hat r\setminus\{o\}}_{\{o\}}(\mathcal{A})\rangle^-_p, \langle C^{\hat r\setminus\{o\}}_{\{o\}}(\mathcal{B})\rangle^-_p\right) = \mathcal{D}\left(\langle C^{\hat r\setminus\{o\}}_{\{o\}}(\mathcal{A})\rangle^-_p \times_i K_{\mathcal{A}}, \langle C^{\hat r\setminus\{o\}}_{\{o\}}(\mathcal{B})\rangle^-_p \times_i K_{\mathcal{B}}\right)$$

where $\mathcal{N}$ is one of $\mathbb{Z}^+, \mathbb{Q}^+, \mathbb{R}^+$.

2.4.2 Temporal invariances.

We now move to temporal invariances. The first allows for linear time shifts. So, for instance, in musical pattern matching, the pattern may occur anywhere in the database, not just as an incipit. Being additive, it is usually easily combined with the first pitch invariances, above.

Definition 6 *Let r be a representation including pitch and onset. A distance function $\mathcal{D}$ is* time-position-invariant *iff*

$$\forall a, b \in \Sigma_o.\forall \mathcal{A}, \mathcal{B}\ in\ r.\mathcal{D}(C^{\hat r}_{\{o\}}(\mathcal{A}), C^{\hat r}_{\{o\}}(\mathcal{B})) = \mathcal{D}(C^{\hat r}_{\{o\}}(\mathcal{A}) +_o a, C^{\hat r}_{\{o\}}(\mathcal{B}) +_o b).$$

Note that the above invariance is not meaningful with durations. The next two temporal invariances are of multiplicative nature, the first of which, the time-scale-invariance, is applicable both with onsets and durations.

Definition 7 *Let r be a representation including pitch and onset. A distance function $\mathcal{D}$ is* time-scale-invariant *iff*

$$\forall F_{\mathcal{A}} \in \mathcal{N}^{|\mathcal{A}|}, F_{\mathcal{B}} \in \mathcal{N}^{|\mathcal{B}|}, K_{\mathcal{A}} \in \Sigma_o^{|\mathcal{A}|}, K_{\mathcal{B}} \in \Sigma_o^{|\mathcal{B}|}.$$
$$\forall \mathcal{A}, \mathcal{B}\ in\ r.\mathcal{D}(C^{\hat r}_{\{o\}}(\mathcal{A}), C^{\hat r}_{\{o\}}(\mathcal{B})) = \mathcal{D}(C^{\hat r}_{\{o\}}(\mathcal{A}) \times_o F_{\mathcal{A}} +_o K_{\mathcal{A}}, C^{\hat r}_{\{o\}}(\mathcal{B}) \times_o F_{\mathcal{B}} +_o K_{\mathcal{B}}).$$

where $\mathcal{N}$ is one of $\mathbb{Z}^+, \mathbb{Q}^+, \mathbb{R}^+$.

The next *duration-warp* invariance is most useful with duration sequences; it is "durational Parsons invariance", i.e., the one for which "shorter, longer, same" encoding is often used. To this end we use the second order derivation of sets $\mathcal{A}$ and $\mathcal{B}$ with ioi proportions, abbreviated ip below.

Definition 8 *Let r be a representation including pitch and onset. A distance function $\mathcal{D}$ is* duration-warp-invariant *iff*

$$\forall K_{\mathcal{A}} \in \mathcal{N}^{|\mathcal{A}|-2}, K_{\mathcal{B}} \in \mathcal{N}^{|\mathcal{B}|-2}.\forall \mathcal{A}, \mathcal{B}\ in\ r.$$
$$\mathcal{D}\left(\langle\langle C^{\hat r}_{\{o\}}(\mathcal{A})\rangle^-_o\rangle^{\div}_{ioi}, \langle\langle C^{\hat r}_{\{o\}}(\mathcal{B})\rangle^-_o\rangle^{\div}_{ioi}\right) = \mathcal{D}\left(\langle\langle C^{\hat r}_{\{o\}}(\mathcal{A})\rangle^-_o\rangle^{\div}_{ioi} \wedge_{ip} K_{\mathcal{A}}, \langle\langle C^{\hat r}_{\{o\}}(\mathcal{B})\rangle^-_o\rangle^{\div}_{ioi} \wedge_{ip} K_{\mathcal{B}}\right)$$

where $\wedge$ is the power operator and $\mathcal{N}$ is one of $\mathbb{Z}^+, \mathbb{Q}^+, \mathbb{R}^+$.

The last temporal invariance does not bother with the onset information, except in as far as order is preserved. This is the case, for instance, with CBMR methods based on string representations that omit explicit onset times. Note that, although it is temporal, there is no intuitive interpretation of this invariance to duration information.

Definition 9 *Let r be a representation including pitch and onset, and let $K_{\mathcal{A}} \in \mathcal{N}^{|\mathcal{A}|}, K_{\mathcal{B}} \in \mathcal{N}^{|\mathcal{B}|}$ be such that*

$$\begin{aligned} a_{i-1} +_o K_{\mathcal{A}}(i-1) &< a_i +_o K_{\mathcal{A}}(i) \ and \\ b_{i-1} +_o K_{\mathcal{B}}(i-1) &< b_i +_o K_{\mathcal{B}}(i) \end{aligned}$$

for $2 \leq i \leq |K_{\mathcal{A}}|, |K_{\mathcal{B}}|$. A distance function $\mathcal{D}$ is time-warp-invariant *iff*

$$\forall \mathcal{A}, \mathcal{B}\ in\ r.\ \mathcal{D}\left(C^{\hat r\setminus\{o\}}_{\{o\}}(\mathcal{A}), C^{\hat r\setminus\{o\}}_{\{o\}}(\mathcal{B})\right) = \mathcal{D}\left(C^{\hat r\setminus\{o\}}_{\{o\}}(\mathcal{A}) +_o K_{\mathcal{A}}, C^{\hat r\setminus\{o\}}_{\{o\}}(\mathcal{B}) +_o K_{\mathcal{B}}\right)$$

where $\mathcal{N}$ is one of $\mathbb{Z}, \mathbb{Q}, \mathbb{R}$.

Now, we can fully define algorithmic invariance.

Definition 10 *A method $\mathcal{M}$ is* algorithmically P-invariant *iff $\mathcal{M}$ satisfies Definition 1 and its similarity measure satisfies the definitions above corresponding with property P.*

3. STRUCTURAL INVARIANCES

Let us now consider a set of stronger invariances that relate primarily not to the music represented, but to the results proven using our order-based formalism. To be maximally useful, it is helpful to know how strongly the results apply: in particular, does the order imposed by our concatenator make a difference to the outcome? For example, in the following *permutation invariances*, when applied to contour-based melody comparison, onset-order matters, but in a pitch-class-distribution comparison, it probably does not.

Definition 11 *Let r be a representation and $\omega \subseteq \hat{r}$. A distance function $\mathcal{D}$ is ω-*permutation-invariant *iff*

$$\forall \mathcal{A}, \mathcal{B}\ in\ r.\mathcal{D}(C^{\hat{r}}_{\omega}(\mathcal{A}), C^{\hat{r}}_{\omega}(\mathcal{B})) = \mathcal{D}(\mathcal{P}(C^{\hat{r}}_{\omega}(\mathcal{A})), \mathcal{P}(C^{\hat{r}}_{\omega}(\mathcal{B})))$$

where $\mathcal{P}$ is any size-preserving permutation operator on ω. If $\omega = \hat{r}$, the distance function is strongly permutation-invariant.

Further, it may be useful to know that a distance is preserved no matter which dimension is used for ordering.

Definition 12 *Let r be a representation. A distance function $\mathcal{D}$ is ω-* concatenation-invariant *iff*

$$\forall \omega_1, \omega_2 \subset \hat{r}.\forall \mathcal{A}, \mathcal{B}\ in\ r.\mathcal{D}(C^{\hat{r}}_{\omega_1}(\mathcal{A}), C^{\hat{r}}_{\omega_1}(\mathcal{B})) = \mathcal{D}(C^{\hat{r}}_{\omega_2}(\mathcal{A}), C^{\hat{r}}_{\omega_2}(\mathcal{B})).$$

If $\omega_1, \omega_2 = \hat{r}$, the distance function is strongly concatenation-invariant.

For a strongly concatenation-invariant distance function the ordering does not make any difference at all. Note that a strongly permutation invariant distance function is also a strongly concatenation invariant, and vice versa.

4. INVARIANCES IN POLYPHONIC CONTENT-BASED MUSIC RETRIEVAL

4.1 Representations of non-monophonic music

The concatenated representations used here are evidently directly applicable when dealing with monophonic music. In the case of (discretely represented) polyphonic music, a geometrical representation [1, 11, 12, 14] is a more effective and natural choice [5]. An example of geometrical music matching (under transpositional equivalence, in this example) is given in Figure 2, where the common pitch-against time-representation, giving the onset times but not durations, is used. Several possible ways to represent durations have been suggested [10, 11, 12].

As Figure 2 suggests, the maximal subset match of the given query pattern of length m within the database of length n can be found by observing the translation vectors. Note that a translation corresponds to two musically distinct phenomena: a vertical move corresponds to pitch-shift while a horizontal move corresponds to aligning the pattern time-wise; the combination of these is what a musician calls "a transposition" (to be distinguished from the *process* of transposition, performed during performance). Thus, working on the translation vectors captures transposition and position invariances, in the terms defined here.

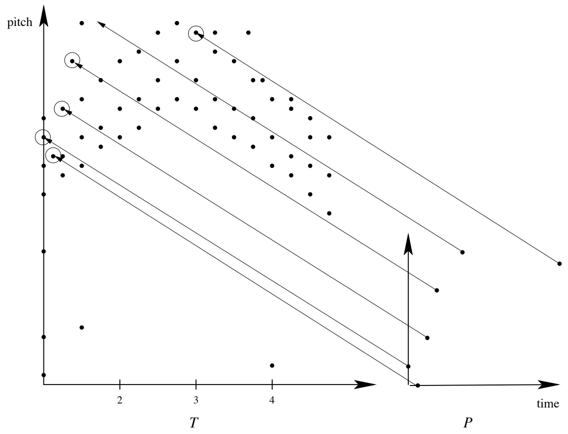

Figure 2. Pointset P, to the right, represents a pointset (musical) pattern to be matched against a pointset database to the left. The arrows represent translation vectors, from pattern to database, that give maximal occurrence.

Ukkonen et al. [12] gave an algorithm to solve the maximal subset matching problem in $O(mn \log m)$ time. It is still the fastest known deterministic algorithm for the problem. Clifford et al. [2] showed that quadratic running times are probably the best one can achieve for this problem by proving that the maximal subset matching problem is 3SUM-hard. They also gave a randomized algorithm for the problem that works in time $O(n \log n)$.

4.2 Combining invariances

When using the sequence (string) representation, pitch-transposition invariance is easily combined with time-warp invariance (and the latter serves as a filtering method for time-scale invariance). However, the explicit encoding of the onset times in the geometrical representation makes it difficult to combine transposition invariance with most of the temporal invariances, such as time-scale invariance. The difficulty of combining transposition invariance and time-scale invariance is due to the fact that the former is an additive property, while the latter is multiplicative.

Romming and Selfridge-Field [10] gave the only known non-brute-force algorithm capable of dealing with polyphonic music, transposition invariance and time-scale invariance. Their algorithm is based on geometrical hashing and works in $O(n^3)$ space and $O(n^2m^3)$ time. By applying a window on the database such that w is the maximum number of events that occur in any window, the above complexities can be restated as $O(w^2n)$ and $O(wnm^3)$, respectively. The algorithm works on all three of the musical features discussed here (pitch, onset time and duration), finding a maximal subset match in such a scenario. However, as with the SIA algorithm family [7], its applicability to real world problems is reduced due to the fact that matches are mathematically exact, and so performance expression and error is difficult to account for.

5. CONCLUSIONS

In this paper we have discussed invariances related to content-based music retrieval; they are central concepts in defining and developing effective representations, similarity measures and algorithms to that end. Because of their centrality to the matter, invariances are widely used in the literature—but very seldom are they properly defined or their relationship discussed which has occasionally resulted in misuse of the term and confusion.

We have given a sparse taxonomy of the invariances along three featural dimensions of music—pitch, onset time and duration. We also defined stronger invariances, intrinsic to our formalism. The taxonomy shows explicitly the relationships of these invariances to each other. Moreover, we have precisely defined them, minimizing confusion in future discussion. The taxonomy works also as a useful tool in discussing what has been done, and in identifying where there is still much space for future developments towards efficient and effective CBMR tools.

It seems that the geometrical framework provides the best (and most natural) representation when dealing with polyphonic music. Using this framework, however, it is not easy to combine translation and time-scale invariances in a computationally efficient way; there is still a huge gap to be bridged in this respect to be able to meet the real world requirements for responsive and error-tolerant database queries. One way to improve error-tolerance—as is evident in our taxonomy—would be to adapt the geometrical frameworks to work also on the level of the more general invariances. To date, there is next to no work in this direction, though Lubiw and Tanur [6] presented an algorithm that measures the distance between the desired pitches and observed pitches that are combined in a final similarity value. So, with respect to our taxonomy, their work resides somewhere in between the two ends. Their method, although built on discrete space, does not straightforwardly lend itself to a non-strict time-scale invariance.

We are currently studying how to adapt the geometrical approach to the more general classes of our taxonomy thus achieving more error-tolerant geometrical methods for content-based music retrieval. Another direction is to refine the definitions in order to be able to discriminate methods that allow“gaps” (as the geometrical methods usually do) from those that do not (for instance, methods based on exact string matching).

6. ACKNLOWEDGEMENTS

We gratefully acknowledge funding for Kjell Lemström from the Academy of Finland (Grants #108547 and #129909). The work was carried out during his visit to the ISMS group at Goldsmiths, University of London. We also acknowledge the intellectual contribution of the ISMS group, in particular Richard Lewis.

7. REFERENCES

[1] M. Clausen, R. Engelbrecht, D. Meyer, and J. Schmitz. Proms: A web-based tool for searching in polyphonic music. In *Proc. ISMIR'00*, Plymouth, MA, October 2000.

[2] R. Clifford, M. Christodoulakis, T. Crawford, D. Meredith, and G. Wiggins. A fast, randomised, maximal subset matching algorithm for document-level music retrieval. In *Proc. ISMIR'06*, pp. 150–155, Victoria, BC, Canada, October 2006.

[3] D. Conklin and I. H. Witten. Multiple viewpoint systems for music prediction. *J. New Music Research*, 24:51–73, 1995.

[4] K. Lemström and P. Laine. Musical information retrieval using musical parameters. In *Proc. ICMC'98*, pages 341–348, Ann Arbor, MI, 1998.

[5] K. Lemström and A. Pienimäki. On comparing edit distance and geometric frameworks in content-based retrieval of symbolically encoded polyphonic music. *Musicae Scientiae*, 4a:135–152, 2007.

[6] A. Lubiw and L. Tanur. Pattern matching in polyphonic music as a weighted geometric translation problem. In *Proc. ISMIR'04*, pages 289–296, Barcelona, October 2004.

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

[8] D. Müllensiefen, G. A. Wiggins, and D. Lewis. High-level feature descriptors and corpus-based musicology: Techniques for modelling music cognition. In A. Schneider, editor, *Systematic and Comparative Musicology: Concepts, Methods, Findings*, number 24 in Hamburger Jahrbuch für Musikwissenschaft. Peter Lang, Frankfurt am Main, 2008.

[9] D. Parsons. *The Directory of Tunes and Musical Themes*. S. Brown (Cambridge, Eng.), 1975.

[10] C. A. Romming and E. Selfridge-Field. Algorithms for polyphonic music retrieval: The hausdorff metric and geometric hashing. In *Proc. ISMIR'07*, Vienna, Austria, October 2007.

[11] R. Typke, P. Giannopoulos, R. C. Veltkamp, F. Wiering, and R.v. Oostrum. Using transportation distances for measuring melodic similarity. In *Proc. ISMIR'03*, pp. 107–114, Baltimore, MA, October 2003.

[12] E. Ukkonen, K. Lemström, and V. Mäkinen. Geometric algorithms for transposition invariant content-based music retrieval. In *Proc. ISMIR'03*, pages 193–199, Baltimore, MA, October 2003.

[13] G. A. Wiggins, M. Harris, and A. Smaill. Representing music for analysis and composition. In M. Balaban, K. Ebcioglu, O. Laske, C. Lischka, and L. Sorisio, editors, *Proc. 2nd IJCAI AI/Music Workshop*, pages 63–71, Detroit, Michigan, 1989.

[14] G. A. Wiggins, K. Lemström, and D. Meredith. SIA(M)ESE: An algorithm for transposition invariant, polyphonic content-based music retrieval. In *Proc. ISMIR'02*, pages 283–284, Paris, France, October 2002.