Chapter 6

Music information processing

Giovanni De Poli and Nicola Orio

6.1 Elements of music theory and notation

Music as well as language was long cultivated by aural transmission before any kind of systematic method of writing it down was invented. But the desire to record laws, poetry and other permanent statements gave rise the problem of how to write down music. In western tradition the focus is on a symbolic system which can represent both the pitch and the rhythm of a melody. In the following section the general principles of western notation will be presented.

In music the word *note* can mean three things: (1) a single sound of fixed pitch; (2) the written symbol of a musical sound; (3) a key on the piano or other instrument. A note is often considered as the atomic element in the analysis and perception of the musical structure. The two main attributes of a note are pitch and duration. These are the two most important parameters in music notation and, probably not coincidentally, the first ones to evolve. A functional piece of music can be notated using just these two parameters. Most of the other ones, such as loudness, instrumentation, or tempo, are usually written in English or Italian somewhere outside of the main musical framework.

6.1.1 Pitch

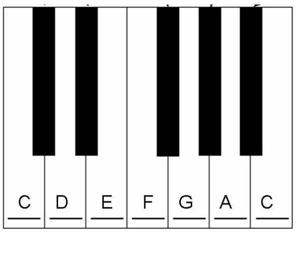


Figure 6.1: One octave in a piano keyboard.

In music, a scale is a set of musical notes that provides material for part or all of a musical work. Scales are typically ordered in pitch, with their ordering providing a measure of musical distance. Human pitch-perception is periodic: a note with a doubled frequency as another sounds very similar and is commonly given the same name, called *pitch class*. The interval (i.e. the span of notes) between these two notes is called *octave*. Thus the complete definition of a note consists of its pitch class and the octave it lies in. Scales in traditional Western music generally consist of seven notes (pitch classes) and repeat at the octave. The name of the notes of a scale is indicated by the first seven letters of the alphabet. For historical reasons the musical alphabet starts from C and not from A, and it is arranged thus: C D E F G A B, closing again with C, so producing an interval from C to C of eight notes. These eight notes are represented by white keys on the piano keyboard (Figure 6.1). In Italian the pitch classes are called, respectively, do, re, mi, fa, sol, la, si. The octaves are indicated by numbers. In general the reference is the fourth octave containing the C4 (the middle C) and A4 (the diapason reference) with frequency $f = 440$ Hz. The lowest note on most pianos is A0, the highest C8.

6.1.1.1 Pitch classes, octaves and frequency

In most western music the frequencies of the notes are tuned according the twelve-tone equal temperament. In this system the octave is divided into a series of 12 equal steps (equal frequency ratio). On a piano keyboard the steps are represented by the 12 white and black keys forming an octave. The interval between two adjacent keys (white or black) is called *semitone* or half tone. The ratio s corresponding to a semitone can be determined considering that the octave ratio is composed by 12 semitones, i.e. $s^{12} = 2$, and thus the semitone frequency ratio is given by

$$s = \sqrt[12]{2} \approx 1.05946309 \tag{6.1}$$

i.e. about a six percent increase in frequency. The semitone is further divided in 100 (equal ratio) steps, called cents. I.e.

$$1\text{cent} = \sqrt[100]{s} \approx 1.000577$$

The distance between two notes whose frequency are f_1 and f_2 is $12 \log_2(f_1/f_2)$ semitones $=$ $1200 \log_2(f_1/f_2)$ cents. The just noticeable difference in pitch is about five cents.

In the equal temperament system a note which is n steps or semitones apart the central A (A4) has frequency

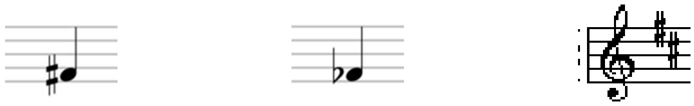
$$f = 440 \times 2^{n/12} \text{ Hz} = 440 \times s^n \text{ Hz} \tag{6.2}$$

For example middle C (C4) is $n = -9$ semitones apart from A4 and has frequency $f = 440 \times 2^{-9/12} = 261.63$ Hz. A convenient logarithmic scale for pitch is simply to count the number of semitones from a reference pitch, allowing fractions to permit us to specify pitches which don't fall on a note of the Western scale. This creates a linear pitch space in which octaves have size 12 and semitones have size 1. Distance in this space corresponds to physical distance on keyboard instruments, orthographical distance in Western musical notation, and psychological distance as measured in psychological experiments and conceived by musicians. The most commonly used logarithmic pitch scale is *MIDI pitch*, in which the pitch 69 is assigned to a frequency of 440 Hz, i.e. the A above middle C. A note with MIDI pitch p has frequency

$$f = 440 \times 2^{(p-69)/12} \text{ Hz} = 440 \times s^{p-69} \text{ Hz} \tag{6.3}$$

and a note with frequency f Hz has MIDI pitch

$$p = 69 + 12 \log_2(f/440) \tag{6.4}$$



(a) (b) (c)

Figure 6.2: Example of a sharp (a) and a flat (b) note. Example of a key signature (c): D major.

Because there are actually 12 notes on the keyboard, the 7 note names can also be given a modifier, called accidental. The two main modifiers are sharps (Fig. 6.2(a)) and flats 6.2(b)) which respectively raise or lower the pitch of a note by a semitone, where a semitone is the interval between two adjacent keys (white or black).

If we ignore the difference between octave-related pitches, we obtain the pitch class space, which is a circular representation. Since pitch class space is a circle, we return to our starting point by taking a series of steps in the same direction: beginning with C, we can move ”upward” in pitch class space, through the pitch classes C♯, D, D♯, E, F, F♯, G, G♯, A, A♯, and B, returning finally to C. We can assign numbers to pitch classes. These numbers provide numerical alternatives to the letter names of elementary music theory: 0 = C, 1 = C♯=D♭, 2 = D, and so on. Thus given a Midi pitch p, its pitch class pc and octave number oct are given by

$$\begin{aligned} pc &= p \mod 12 && (6.5) \\ oct &= \lfloor p/12 \rfloor - 1 && (6.6) \end{aligned}$$

and viceversa

$$p = pc + 12(oct + 1)$$

For example middle C (C4) has $p = 60$, and $pc = 0$, $oct = 4$. Notice that some pitch classes, corresponding to black keys in the piano, can be spelled differently: e.g. $pc = 1$ can be spelled as C♯ or as D♭.

6.1.1.2 Musical scale.

All humans perceive a large continuum of pitch. However, the pitch systems of all cultures consist of a limited set of pitch categories that are collected into ordered subsets called scales. In music, a scale is a set of musical notes that provides material for part or all of a musical work. Scales in traditional Western music generally consist of seven notes (diatonic scale) derived from an alphabet of the 12 chromatic notes within an octave, and repeat at the octave. Notes are separated by whole and half step intervals of tones and semitones. In many musical circumstances, a specific note is chosen as the ”tonic”–the central and most stable note of the scale. Relative to a choice of tonic, the notes of a scale are often labeled with roman numbers recording how many scale steps above the tonic they are. For example, the notes of the C diatonic scale (C, D, E, F, G, A, B) can be labeled I, II, III, IV, V, VI, VII, reflecting the choice of C as tonic. The term ”scale degree” refers to these numerical labels: in the previous case, C is called the first degree of the scale, D is the second degree of the scale, and so on. In the C diatonic scale, with C chosen as tonic, C is the first scale degree, D is the second scale degree, and so on. In the major scale the pattern of intervals in semitones between subsequent notes is 2-2-1-2-2-2-1; these numbers stand for whole tones (2 semitones) and half tones (1 semitone). The interval pattern of minor scale is 2-1-2-2-1-2-2. The scale defines interval relations relative to the pitch of the first note, which can be any one of the keyboard.

In the western music, the scale define also a relative importance of the different degree. The first (I) degree (called tonic or keynote) is the most important. The degree next in importance is the fifth (V), called dominant because of its central position and dominating role in both melody and harmony. The fourth (IV) degree (subdominant) has a slightly dominating role that the dominant. The other degree are supertonic (II), mediant (III), submediant (VI), leading note (VII). The numerical classification depends also on the scale: for example in the major scale the (major) third has $2 + 2 = 4$ semitones interval, while in the minor scale the (minor) third has $2 + 1 = 3$ semitones interval. There are five adjectives to qualify the intervals: perfect intervals are the I, IV, V, and VIII. The remaining intervals (e.g. II, III, VI, VII) in the major scale are called major intervals. If a major interval is reduced by a semitone, we get a minor interval. If a major or perfect interval is increased by a semitone, we get a corresponding augmented interval. Any minor or perfect interval reduced by a semitone is called diminished interval.

The scale made by 12 tones per octave is called chromatic scale.

6.1.1.3 Musical staff

Notation of pitch is done by using a framework (or grid) of five lines called a staff. Both the lines and spaces are used for note placement. How high or low a pitch is played is determined by how high or low the note head is placed on the staff.

Notes outside the range covered by the lines and spaces of the staff are placed on, above or below shorter lines, called leger (or ledger) lines, which can be placed above or below the staff. Music is read from 'left' to 'right', thus it is a sort of two dimensional representation in a time-frequency plane.

A piano uses two staves, each one covering a different range of notes (commonly known as register). They are read simultaneously–two notes that are in vertical alignment are played together. An orchestral score will often have more than ten staves. To establish the pitch of any note on the staff we place a graphical symbol called a clef at the far left-hand side of the staff. The clef establishes the pitch of the note on one particular line of the staff and thereby fixes the pitch of all the other notes lying on, or related to, the same staff (see Fig. 6.3 and 6.4).

Sometimes (but not always) accidentals are placed to the immediate right of the clef sign and before the Time Signature. This indicates the tonality (or key) the song should be played in. The Key Signature consists of a small group of sharps or flats and tells you if any note (more precisely, pitch class) should be consistently sharped or flatted (Fig. 6.2(c)). For example, if there is a sharp on the F and on the C in a key signature (as in Fig. 6.2(c)), it tells a musician to play all notes "F" as "F$\sharp$" instead and all C notes as as "C$\sharp$", regardless of whether or not they fall on that line. A flat on the B line tells a musician to play all notes "B" as Bb, and so on. The natural sign ($\natural$) in front of a note will signal that the musician should play the white key version of the note. The absence of any sharp or flats at the beginning tells you the song is played in the key of C, i.e. without any pitch modification (as Fig. 6.3).

6.1.2 Note duration

Music takes place in time, and so musiccians have to organize it in terms not only of pitch but also of duration. They must chose whether the sounds they use shall be shorter or longer, according to the artistic purpose they whish to serve.

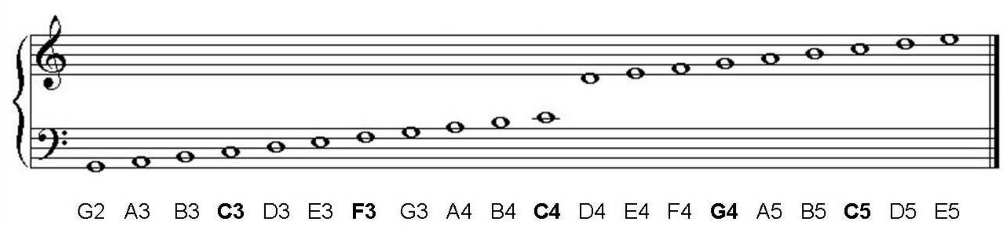


Figure 6.3: Staff and note names.

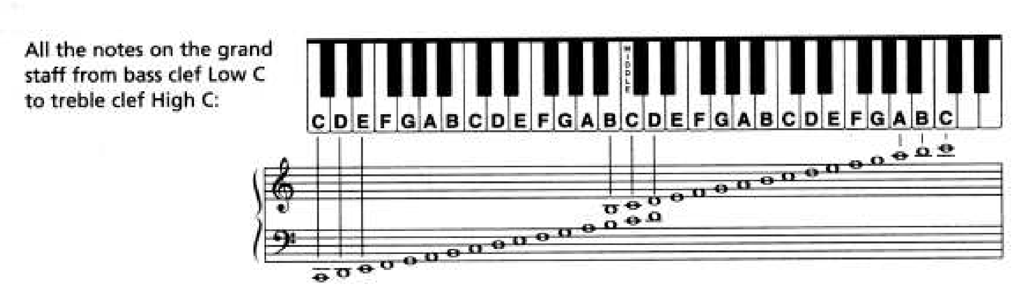

All the notes on the grand staff from bass clef Low C to treble clef High C:
MIDDLE
C D E F G A B C D E F G A B C D E F G A B C D E F G A B C

Figure 6.4: Correspondence of keys and notes on the staff.

When we deal with symbolic representation, the symbolic duration (or note length) refers to the perceptual and cognitive organization of sounds, which prefer simple relations. Thus the symbolic duration is the time interval between the beginning of the event and the beginning of the next event, which can also be a rest. Notice that the actual sound duration (physical duration) can be quite different and normally is longer, due to the decay time of the instrument. In this chapter when not explicitely stated, we will deal with symbolic duartion.

6.1.2.1 Duration symbols

In order to represent a sound, apart for naming it alphabetically, a symbol is used. Where the vertical position of a note on a staff or stave determines its pitch, its relative time value or length is denoted by the particular sign chosen to represent it. The symbols for note lengths are indicated in Table 6.1 and how sound lengths are divided is shown in Fig. 6.5. This is the essence of proportional time notation. The signs indicate only the proportions of time-lengths and do not give duration in units of time, minutes or seconds.

At present the longest note in general use is the whole note or semibreve, which serves as the basic unit of length: i.e. the whole note has conventional length equal 1. This is divided (Fig. 6.5) into two half notes or minims (*minime*), 4 quarters or crotchets (*semiminime*), 8 eighths or quarvers(*crome*), 16 sixteenths or semiquarvers (*semicrome*), 32 thirty-seconds or demisemiquarvers (*biscrome*) . The corresponding symbols for rests (period of silence) are shown in Figure 6.6.

Notice that when we refer to symbolic music representation, as in scores, the note length is also called duration. However symbolic duration does not represent the actual duration of a sound; instead it refers to the difference from beginning of the next event to the beginning of the actual event. The real sound duration depends on the instrument type, how it is played, etc., and normally is not equivalent.

A dot, placed to the immediate right of the note-head, increases its time-value by half. A second dot, placed to the immediate right of the first dot, increases the original undotted time-value by a

Note name:						
American	whole	half	quarter	heigth	sixteen	thirty-second
Italian	semibreve	minima	semiminima	croma	semicroma	biscroma
English	semibreve	minim	crotchet	quaver	semiquarver	demisemiquarver
Length	1	1/2	1/4	1/6	1/16	1/32
Note symbol	𝅝	𝅗𝅥	♩	♪	𝅘𝅥𝅯	𝅘𝅥𝅰
Rest symbol	𝄻	𝄼	𝄽	𝄾	𝄿	𝅀

Table 6.1: Duration symbols for notes and rests.

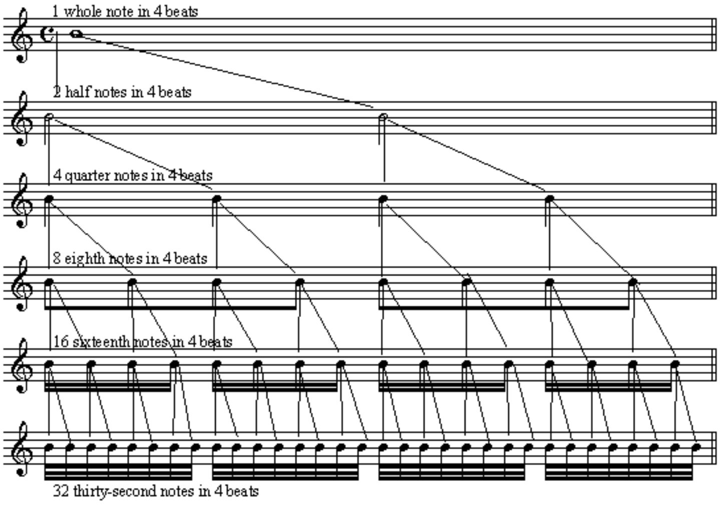

1 whole note in 4 beats
2 half notes in 4 beats
4 quarter notes in 4 beats
8 eighth notes in 4 beats
16 sixteenth notes in 4 beats
32 thirty-second notes in 4 beats

Figure 6.5: Symbols for note length.

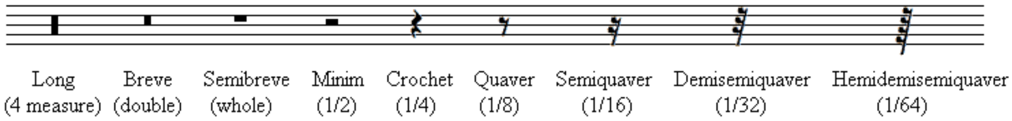


Figure 6.6: Symbols used to indicate rests of different length.

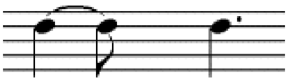


Figure 6.7: Tie example: crotchet (quarter note) tied to a quaver (eighth note) is equivalent to the dotted crotchet (dotted quarter note).

further quarter. Dots after rests increase their time-value in the same way as dots after notes. A tie (a curved line connecting the heads of two notes) serves to attach two notes of the same pitch. Thus the

(a) (b)

Figure 6.8: (a) Example of a time signature: 3/4 indicates three quarter note beats per measure. (b) Example of a metronome marking: 120 quarters to the minute.

sound of the first note will be elongated according the value of the attached note. This is illustrated in the example given in Fig. 6.7 where a crotchet (quarter note) tied to a quaver (eighth note) is equivalent to the dotted crotchet (dotted quarter note) that follows. To divide a note value into three equal parts, or some other value than two, tuplets may be used. The most common tuplet is the triplet: in this case the note length is reduced to 2/3 the original duration.

6.1.3 Tempo

The signs of Table 6.1 do not give duration in units of time, minutes or seconds. The relationship between notes and rests is formalized but the duration or time value in seconds of any particular note is unquantified. It depends on the speed the musical piece is played. *Tempo* is the word used to cover all the variation of speed, from very slow to very fast.

Until the invention of a mechanical device called the metronome, the performance speed of a piece of music was indicated in three possible ways: through the use of tempo marks, most commonly in Italian; by reference to particular dance forms whose general tempi would have been part of the common experience of musicians of the time; by the way the music was written down, in particular, the choice of note for the beat and/or the time signature employed. Many composers give metronome marks to indicate exact tempo. The metronome measures the number of beats per minute (BPM) at any given speed. The allegro tempo may correspond to 120 BPM, i.e. beats per minute. This value corresponds to a frequency of $120/60 = 2$ beats per second. The beat duration is the inverse of the frequency, i.e. $d = 1/2 = 0.5$ sec. However most musicians would agree that it is not possible to give beats per minute (BPM) equivalents for these terms; the actual number of beats per minute in a piece marked allegro, for example, will depend on the music itself. A piece consisting mainly of minims (half notes) can be played very much quicker in terms of BPM than a piece consisting mainly of semi-quavers (sixteenth notes) but still be described with the same word.

6.1.4 Rhythm

Rhythm is the arrangement of events in time. In music, where rhythm has probably reached its highest conscious systematization, a regular pulse or *beat*, appears in groups of two, three and their compound combinations. The first beat of each group is accented. The metrical unit from one accent to the next is called a *bar* or *measure*. This unit is marked out in written scores by vertical lines (*bar lines*) through the staff in front of each accented beat.

Notice that tempo is often defined referring to rhythm and metre. The time signature (also known as ”meter signature”) is used to specify how many beats are in each bar and what note value constitutes one beat. Most time signatures comprise two numbers, one above the other. In text (as in this chapter), time signatures may be written in the manner of a fraction, e.g. 3/4. The first number indicates how many beats there are in a bar or measure; the second number indicates the note value which represents one beat (the ”beat unit”). For example 3/4 indicates three quarter note beats per measure (Fig. 6.8(a)). In this case a metronome indication of 120 BPM (Fig. 6.8(b)) corresponds to $120/60$ beats per second: each quarter lasts $60/120 = 0.5$ sec and the measure lasts $3 \times 0.5 = 1.5$ sec. The duration of a musical unit, i.e. a semibreve, is $4 \times 0.5 = 2$ sec. In general given a time signature n_1/n_2 and a metronome marking m BPM, we have that the beat duration is $d_{beat} = 60/m$ sec, the bar duration $d_{bar} = n_1 \times 60/m$ sec, and the musical unit duration $d_{bar} = n_2 \times 60/m$ sec.

6.1.5 Dynamics

In music, dynamics refers to the volume or loudness of the sound or note. The full terms for dynamics are sometimes written out, but mostly are expressed in symbols and abbreviations (see Table 6.2). There are also traditionally in Italian and will be found between the staves in piano music. In an orchestral score, they will usually be found next to the part to which they apply.

SYMBOL	TERM	MEANING
pp	pianissimo	very soft
p	piano	soft
mp	mezzopiano	medium soft
mf	mezzoforte	medium loud
f	forte	loud
ff	fortissimo	very loud

Table 6.2: Symbols for dymanics notation.

In addition, there are words used to indicate gradual changes in volume. The two most common are *crescendo*, sometimes abbreviated to cresc, meaning ”get gradually louder”; and *decrescendo* or *diminuendo*, sometimes abbreviated to decresc and dim respectively, meaning ”get gradually softer”. These transitions are also indicated by wedge-shaped marks. For example, the notation in Fig. 6.9 indicates music starting moderately loud, then becoming gradually louder and then gradually quieter:

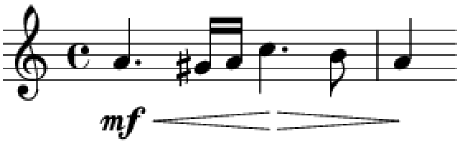


Figure 6.9: Dynamics notation indicating music starting moderately loud (*mezzo forte*), then becoming gradually louder (*crescendo*) and then gradually quieter (*diminuendo*).

6.1.6 Harmony

In music theory, harmony is the use and study of the relationship of tones as they sound simultaneously and the way such relationships are organized in time. It is sometimes referred to as the ”vertical” aspect of music, with melody being the ”horizontal” aspect. Very often, harmony is a result of counterpoint or polyphony, several melodic lines or motifs being played at once, though harmony may control the counterpoint. The term ”chord” refers to three or more different notes or pitches sounding simultaneously, or nearly simultaneously, over a period of time.

Within a given key, chords can be constructed on each note of the scale by superimposing intervals of a major or minor third (four and three semitones, respectively), such as C-E-G giving the C major triad, or A-C-E giving the A minor triad. A harmonic hierarchy similar to the tonal hierarchy has been demonstrated for chords and cadences. The harmonic hierarchy orders the function of chords within a given key according to a hierarchy of structural importance. This gives rise to one of the particularly rich aspects of Western tonal music: harmonic progression. In the harmonic hierarchy, the tonic chord (built on the first degree of the scale) is the most important, followed by the dominant (built on the fifth degree) and the sub-dominant (built on the fourth degree). These are followed by the chords built on the other scale degrees. Less stable chords, that is those that have a lesser structural importance, have a tendency in music to resolve to chords that are more stable. These movements are the basis of harmonic progression in tonal music and also create patterns of musical tension and relaxation. Moving to a less stable chord creates tension, while resolving toward a more stable chord relaxes that tension. Krumhansl has shown that the harmonic hierarchy can be predicted by the position in the tonal hierarchies of the notes that compose the chords.

6.2 Organization of musical events

6.2.1 Musical form

We can compare a single sound, chord, cadence to a letter, a word, or a punctiation mark in language. In this section we will see how all these materials take formal shape and are used within the framework of a musical structure.

6.2.1.1 Low level musical structure

The bricks of music are its *motives*, the smallest unit of a musical composition. To be intelligible, a motive has to consists of at least two notes, and have a clearly recognizable rhythmic pattern, which gives it live. Usually a motive consists of few notes as for example the four notes at the beginning of Beethovens Fifth Symphony. If you recall the continuation of the symphony, you realize that this motive is the foundation of the whole musical building. It is by mean of motive and its development (e.g. repetition, transposition, modification, contrapuntal use, et.) that a composer state, and subsequently explain his idea.

A *figure* figure is a recurring fragment or succession of notes that may be used to construct the accompaniment. A figure is distinguished from a motif in that a figure is background while a motif is foreground

6.2.1.2 Mid and high level musical structure

A musical phrase can consist of one or more motives. The end is marked by a punctuation, e.g. a cadence. Phrases can be combined to form a period or sentence: i.e. a section of music that is relatively self contained and coherent over a medium time scale. In common practice phrases are often four and most often eight bars, or measures, long.

The mid-level of musical structure is made up of sections of music. Periods combine to form larger sections of musical structure. The length of a section may vary from sixteen to thirty-two measures in length - often, sections are much longer. At the macro-level of musical structure exists the complete work formed of motives, phrases and sections.

6.2.1.3 Basic patterns

Repetition, variation and contrast may be seen as basic patterns. These patterns have been found to be effective at all levels of music structure, whether it be shorter melodic motives or extended musical compositions. These basic patterns may be found not only in all world musics, but also in the other arts and in the basic patterns of nature.

Repetition of the material of music plays a very important role in the composing of music and somewhat more than in other artistic media. If one looks at the component motives of any melody, the successive repetition of the motives becomes apparent. A melody tends to ”wander” without repetition of its rhythmic and pitch components and repetition gives ”identity” to musical materials and ideas. Whole phrases and sections of music often repeat. Musical repetition has the form A A A A A A A A A etc..

Variation means change of material and may be slight or extensive. Variation is used to extend melodic, harmonic, dynamic and timbral material. Complete musical phrases are often varied. Musical variation has the form A A1 A2 A3 A4 A5 A6 etc..

Contrast is the introduction of new material in the structure or pattern of a composition of music that contrasts with the original material. Contrast extends the listeners interest in the musical ”ideas” in a phrase or section of music. It is most often used in the latter areas of phrases or sections and becomes ineffective if introduced earlier. Musical contrast has the form A B C D E F G etc..

The patterns of repetition, variation, and contrast form the basis for the structural design of melodic material, the accompaniment to melodic material, and the structural relationships of phrases and sections of music. When these basic patterns are reflected in the larger sectional structure of complete works of music, this level of musical structure defines the larger sectional patterns of music.

6.2.1.4 Basic musical forms

Form in music refers to large and small sectional patterns resulting from a basic model. There are basic approaches to form in music found in cultures around the world. In most cases, the form of a piece should produce a balance between statement and restatement, unity and variety, contrast and connection. Throughout a given composition a composer may:

1. Present a melody and continually repeat it (A-A-A-A-A-A etc.),
2. Present a melody and continually vary it (A A1 A2 A3 A4 A5 etc.),
3. Present a series of different melodies (A-B-C-D-E-F-G etc.),
4. Alternate a repeating melody with other melodies (A-B-A-C-A-D-A-E-A etc.),
5. Present a melody and expand and/or modify it.

Binary form is a way of structuring a piece of music into two related sections, both of which are usually repeated. Binary form is usually characterized as having the form AB. When both sections repeat, a more accurate description would be AABB. Ternary form is a three part structure. The first and third parts are identical, or very nearly so, while the second part is sharply contrasting. For this reason, ternary form is often represented as ABA. Arch form is a sectional way of structuring a piece of music based on the repetition, in reverse order, of all or most musical sections such that the overall form is symmetrical, most often around a central movement. The sections need not be repeated verbatim but at least must share thematic material. It creates interest through an interplay among memory, variation, and progression. An example is A-B-C-D-C-B-A.

6.2.2 Cognitive processing of music information

Adapted from: Mc Adams, Audition: Cognitive Psychology of Music 1996

When we consider the perception of large scale structures like music, we need to call into play all kinds of relationships over very large time scales on the order of tens of minutes or even hours. It is thus of great interest to try to understand how larger scale temporal structures, such as music, are represented and processed by human listeners. These psychological mechanisms are necessary for the sense of global form that gives rise to expectancies that in turn may be the basis for affective and emotional responses to musical works. One of the main goals of auditory cognitive psychology is to understand how humans can ”think in sound” outside the verbal domain. The cognitive point of view postulates internal (or mental) representations of abstract and specific properties of the musical sound environment, as well as processes that operate on these representations. For example, sensory information related to frequency is transformed into pitch, is then categorized into a note value in a musical scale and then ultimately is transformed into a musical function within a given context.

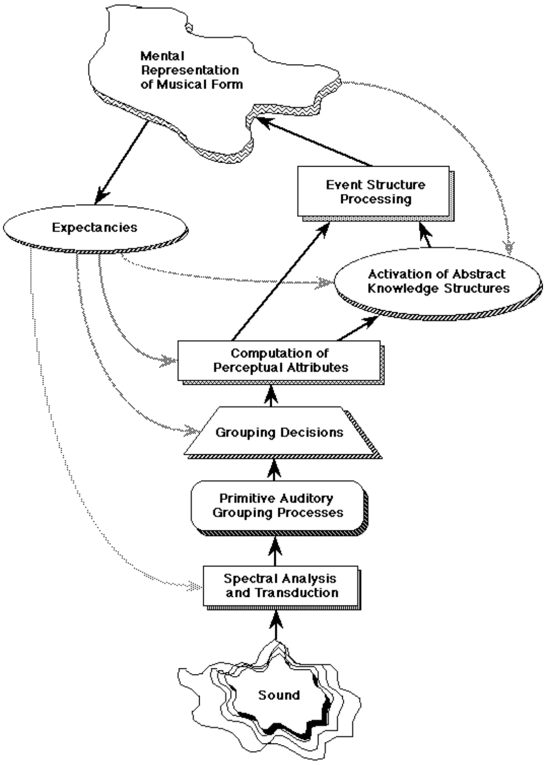

Mental Representation of Musical Form
Expectancies
Event Structure Processing
Activation of Abstract Knowledge Structures
Computation of Perceptual Attributes
Grouping Decisions
Primitive Auditory Grouping Processes
Spectral Analysis and Transduction
Sound

Figure 6.10: Schema illustrating the various aspects of musical information processing [from McAdams 1996].

The processing of musical information may be conceived globally as involving a number of different ”stages” (Fig. 6.10). Following the spectral analysis and transduction of acoustic vibrations in the auditory nerve, the auditory system appears to employ a number of mechanisms (*primitive auditory grouping processes*) that organize the acoustic mixture arriving at the ears into mental ”descriptions”. These descriptions represent events produced by sound sources and their behaviour through time. Research has shown that the building of these descriptions is based on a limited number of acoustic cues that may reinforce one another or give conflicting evidence. This state of affairs suggests the existence of some kind of process (*g*rouping decisions) that sorts out all of the available information and arrives at a representation of the events and sound sources that are present in the environment that is as unambiguous as possible. According to theory of auditory scene analysis, the *computation of perceptual attributes* of events and event sequences depends on how the acoustic information has been organized at an earlier stage. Attributes of individual musical events include pitch, loudness, and timbre, while those of musical event sequences include melodic contour, pitch intervals, and rhythmic pattern. Thus a composer's control of auditory organization by a judicious arrangement of notes can affect the perceptual result.

Once the information is organized into events and event streams, complete with their derived perceptual attributes, what is conventionally considered to be music perception begins.

- The auditory attributes activate *abstract knowledge structures* that represent in long-term memory the relations between events that have been encountered repeatedly through experience in a given cultural environment. That is, they encode various kinds of regularities experienced in the world. Bregman (1993) has described regularities in the physical world and believes that their processing at the level of primitive auditory organization is probably to a large extent innate. There are, however, different kinds of relations that can be perceived among events: at the level of pitches, durations, timbres, and so on. These structures would therefore include knowledge of systems of pitch relations (such as scales and harmonies), temporal relations (such as rhythm and meter), and perhaps even timbre relations (derived from the kinds of instruments usually encountered, as well as their combinations). The sound structures to be found in various occidental cultures are not the same as those found in Korea, Central Africa or Indonesia, for example. Many of the relational systems have been shown to be hierarchical in nature.

- A further stage of processing (*event structure processing*) assembles the events into a structured mental representation of the musical form as understood up to that point by the listener. Particularly in Western tonal/metric music, hierarchical organization plays a strong role in the accumulation of a mental representation of musical form. At this point there is a strong convergence of rhythmic-metric and pitch structures in the elaboration of an event hierarchy in which certain events are perceived to be stronger, more important structurally, and more stable. The functional values that events and groups of events acquire within an event hierarchy generate perceptions of musical tension and relaxation or, in other words, musical movement. They also generate expectancies about where the music should be going in the near future based both on what has already happened and on abstract knowledge of habitual musical forms of the culture– even for pieces that one has never heard before. In a sense, we are oriented–by what has been heard and by what we ”know” about the musical style–to expect a certain type of event to follow at certain pitches and at certain points in time.

- The *expectancies* drive and influence the activation of knowledge structures that affect the way we interpret subsequent sensory information. For example, we start to hear a certain number of pitches, a system of relations is evoked and we infer a certain key; we then expect that future information that comes in is going to conform to that key. A kind of loop of activity is set up, slowly building a mental representation that is limited in its detail by how much knowledge one actually has of the music being heard. It is also limited by one's ability to represent things over the long term, which itself depends on the kind of acculturation and training one has had. It does not seem too extreme to imagine that a Western musician could build up a mental structure of much larger scale and greater detail when listening to a Mahler symphony that lasts one and half hours, than could a person who just walked out of the bush in Central Africa. The reverse would be true for the perception of complex Pygmy polyphonic forms. However, on the one hand we are capable of hearing and enjoying something new, suggesting that there may be inborn precursors to musical comprehension in all human beings that makes this possible. On the other hand, what we do hear and understand the first time we encounter a new musical culture is most likely not what a native of that culture experiences.

The expectancies generated by this accumulating representation can also affect the grouping decisions at the basic level of auditory information processing. This is very important because in music composition, by playing around with some of these processes, one can set up perceptual contexts that affect the way the listener will tend to organize new sensory information. This process involves what Bregman (1990) has called schema-driven processes of auditory organization.

While the nature and organization of these stages are probably similar across cultures in terms of the underlying perceptual and cognitive processing mechanisms involved, the ”higher level” processes beyond computation of perceptual attributes depend quite strongly on experience and accumulated knowledge that is necessarily culture-specific.

6.2.3 Auditory grouping

Sounds and sound changes representing information must be capable of being detected by the listener. A particular configuration of sound parameters should convey consistent percept to the user. *Auditory grouping* studies the perceptual process by which the listener separates out the information from an acoustic signal into individual meaningful sounds (fig. 6.11).

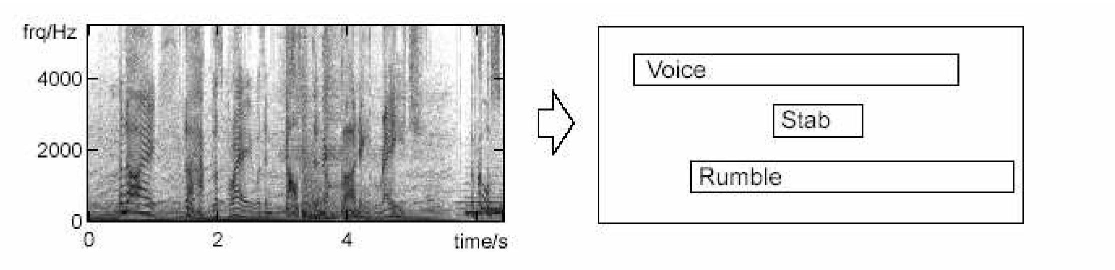


Figure 6.11: Auditory organization

The sounds entering our ears may come from a variety of sources. The auditory system is faced with the complex tasks of:

- Segregating those components of the combined sound that come from different sources.
- Grouping those components of the combined sound that come from the same source.

In hearing, we tend to organise sounds into auditory objects or streams. Bregman (1990) has termed this process Auditory Scene Analysis (fig. 6.12). It includes all the sequential and cross-spectral process which operate to assign relevant components of the signal to perceptual objects denoted *auditory streams*.

The brain needs to group simultaneously (separating out which frequency components that are present at a particular time have come from the same sound source) and also successively(deciding which group of components at one time is a continuation of a previous group). Some processes exclude part of the signal from a particular stream. Others help to bind each stream together.

A stream is

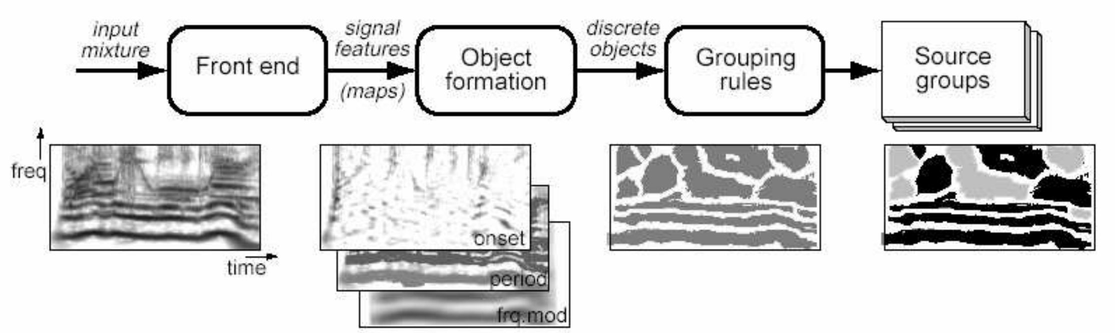


Figure 6.12: Auditory scene analysis

- a psychological organization with perceptual attributes that are not just the sum of the percept of its component but are dependent upon the configuration of the stream.
- a sequence of auditory events whose elements are related perceptually to one another, the stream being segregated from other co-occurring auditory events.
- A psychological organization whose function is to mentally represent the acoustic activity of a single source.

Auditory streaming is the formation of perceptually distinct apparent sound sources. Temporal order judgment is good within a stream but bad between steams. Examples include:

- implied polyphony,
- noise burst replacing a consonant in a sentence,
- click superimposed on a sentence or melody.

An auditory scene is the acoustic pressure wave carrying the combined evidence from all the sound sources present. Auditory scene analysis is the process of decoding the auditory scene, which occurs in auditory perception. Auditory Scene Analysis is a non-conscious process of guessing about ”what's making the noise out there”, but guessing in a way that fits consistently with the facts of the world. For example if a sound has a particular pitch, a listener will probably infer that any other sounds made by that sound source will be similar in pitch to the first sound, as well as similar in intensity, waveform, etc., and further infer that any sounds similar to the first are likely to come from the same location as the first sound. This fact can explain why we experience the sequence of pitches of a tune (Fig. 6.13) as a melody, pitch moving in time. Consecutive pitches in this melody are very close to each other in pitch-space, so on hearing the second pitch a listener will activate our Auditory Scene Analysis inference mechanisms, and assign it to the same source as the first pitch.

If the distance in pitch space had been large, they might have inferred that a second sound source existed, even although they knew that it's the same instrument that's making the sound - this inferred sound source would be a virtual rather than a real source. Hence a pattern such as shown in Figure 6.14(a), where successive notes are separated by large pitch jumps but alternate notes are close together in pitch, is probably heard as two separate and simultaneous melodies rather than one melody leaping around. This tendency to group together, to linearise, pitches that are close together in pitch-space and in time provides us with the basis for hearing a melody as a shape, as pitch moving in time, emanating from a single - real or virtual - source.

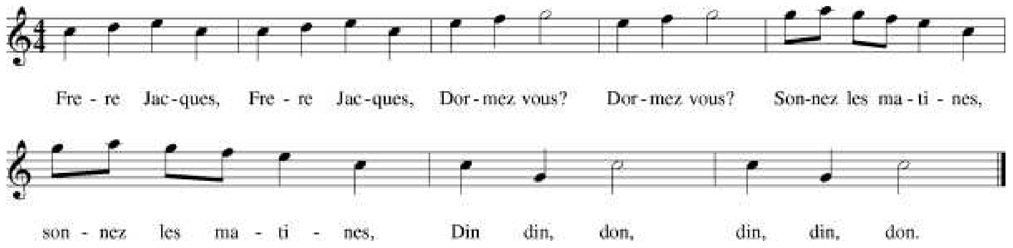

Fre - re Jac-ques, Fre - re Jac-ques, Dor-mez vous? Dor-mez vous? Son-nez les ma - ti - nes,
son - nez les ma - ti - nes, Din din, don, din, din, don.

Figure 6.13: Score of Frere Jacques.

J. S. Bach used them frequently to conjure up the impression of compound, seemingly simultaneous, melodies even though only one single stream of notes is presented. For example, the pattern given in Figure 6.14(b) (from the Courante of Bach's First 'Cello Suite) can be performed on guitar on one string, yet at least two concurrent pitch patterns or streams will be heard - two auditory streams will be segregated (to use Bregman's terminology). We may distinguish analytic vs. synthetic listening. In synthetic perception the information is interpreted as generally as possible, e.g. hearing a room full of voices. In analytic perception, the information is used to to identify the components of the scene to finer levels, e.g. listening to a particular utterance in the crowded room. Interpretation of environmental sounds involves combining analytic and synthetic listening, e.g. hearing the message of a particular speaker.

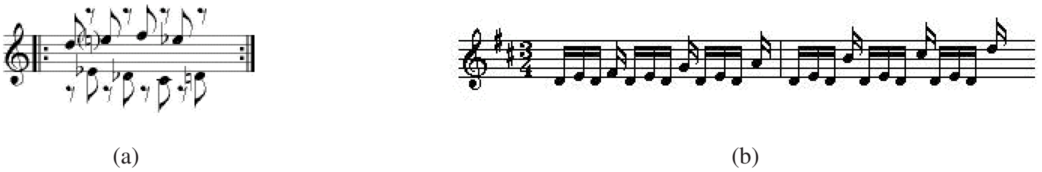


Figure 6.14: (a) Pattern where successive notes are separated by large pitch jumps but alternate notes are close together in pitch, is probably heard as two separate and simultaneous melodies. (b) Excerpt from the Courante of Bach's First 'Cello Suite: two concurrent pitch patterns are heard.

Gestalt psychology theory offers an useful perspective for interpreting the auditory scene analysis beaviour.

6.2.4 Gestalt perception

Gestalt (pronounced G - e - sh - talt) psychology is a movement in experimental psychology that began just prior to World War I. It made important contributions to the study of visual perception and problem solving. The approach of Gestalt psychology has been extended to research in areas such as thinking, memory, and the nature of aesthetics. The word 'Gestalt' means 'form' or 'shape'.

The Gestalt approach emphasizes that we perceive objects as well-organized patterns rather than separate component parts. According to this approach, when we open our eyes we do not see fractional particles in disorder. Instead, we notice larger areas with defined shapes and patterns. The "whole" that we see is something that is more structured and cohesive than a group of separate particles. Gestalt theory states that perceptual elements are (in the process of perception) grouped together to form a single perceived whole (a *g*estalt).

The focal point of Gestalt theory is the idea of *g*rouping, or how we tend to interpret a visual field or problem in a certain way. According to the Gestalt psychologists, the way that we perceive objects, both visual and auditory, is determined by certain principles (gestalt principles). These principles function so that our perceptual world is organised into the simplest pattern consistent with the sensory information and with our experience. The things that we see are organised into patterns or figures. In hearing, we tend to organise sounds into auditory objects or streams. Bregman (1990) has termed this process Auditory Scene Analysis.

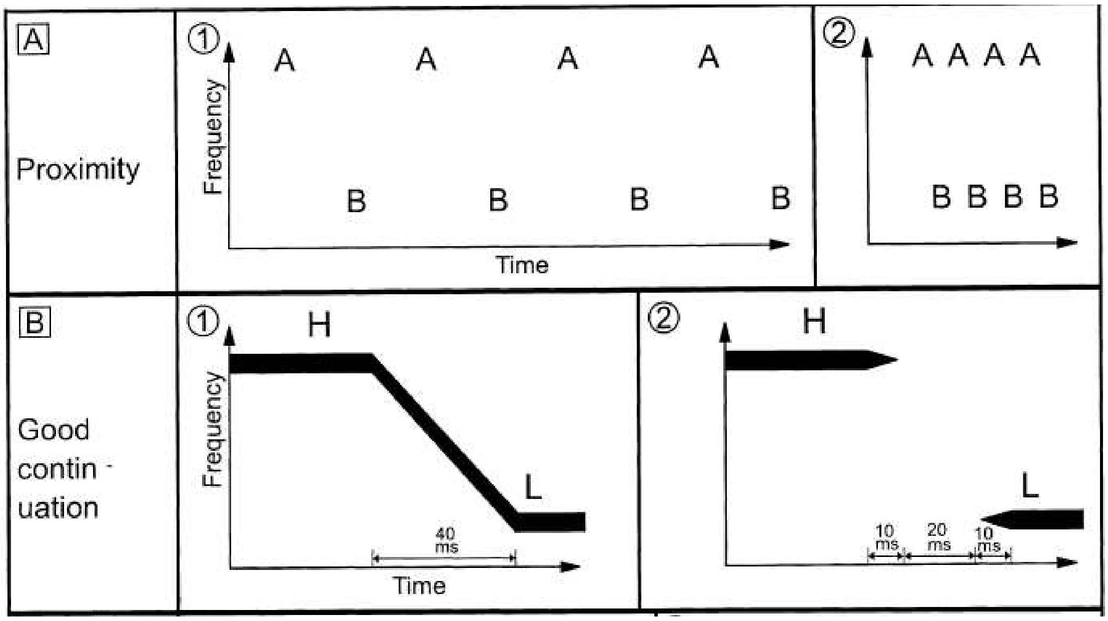


Figure 6.15: Experiments of Proximity and Good Continuation

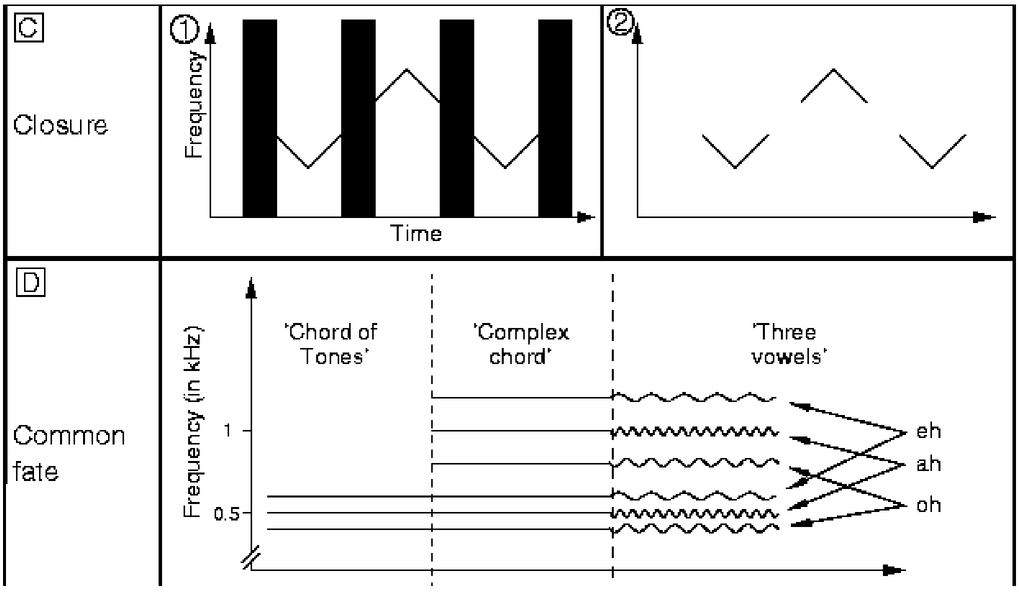


Figure 6.16: Experiments of Closure and Common Fate

The most important principles are

Proximity: components that are perceptually close to each other are more likely to be grouped together. For example temporal proximity or frequency proximity. The principle of proximity refers to distances between auditory features with respect to their onsets, pitch, and loudness. Features that are grouped together have a small distance between each other, and a long distance to elements of another group. Tones close in frequency will group together, so as to minimize the extent of frequency jumps and the number of streams. Tones with similar timbre will tend to group together. Speech sounds of similar pitch will tend to be heard from the same speaker. Sounds from different locations are harder to group together across time than those from the same location.

The importance of pitch proximity in audition is reflected in the fact that melodies all over the world use small pitch intervals from note to note. Violations of proximity have been used in various periods and genres of both Western and non-Western music for a variety of effects. For example, fission based on pitch proximity was used to enrich the texture so that out of a single succession of notes, two melodic lines could be heard. Temporal and pitch proximity are competitive criteria, e.g. the slow sequence of notes A B A B . . . (figure 6.15, A1), which contains large pitch jumps, is perceived as one stream. The same sequence of notes played very fast (figure 6.15, A2) produces one perceptual stream consisting of As and another one consisting of Bs. A visual example is given in figure 6.17: the arrangement of points is not seen as a set of rows but rather a set of columns. We tend to perceive items that are near each other as groups.

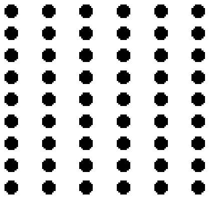
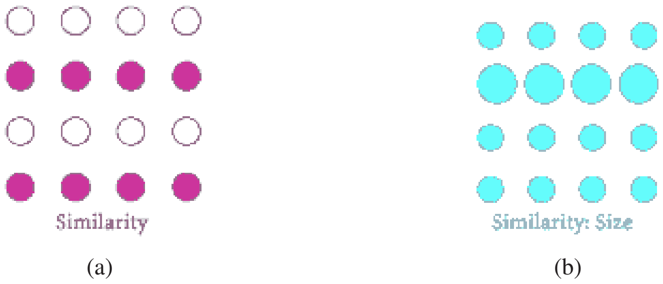


Figure 6.17: Example of proximity gestalt rule

Similarity: components which share the same attributes are perceived as related or as a whole. E.g. colour or form, in visual perception or common onset, common offset, common frequency, common frequency modulation, common amplitude modulation in auditory perception. For example one can follow the piano part in a group of instruments by following the sounds that have the timbre consistent with that of a piano. One can perceptually segregate one speaker's voice from those of others by following the pitch of the voice. Similarity is very similar to proximity, but refers to properties of a sound, which cannot be easily identified with a single physical dimension, like timbre.

A visual example is given in figure 6.18: things which share visual characteristics such as shape, size, color, texture, value or orientation will be seen as belonging together. In the example of 6.18(a), the two filled lines gives our eyes the impression of two horizontal lines, even though all the circles are equidistant from each other. In the example of 6.18(b), the larger circles appear to belong together because of the similiarity in size.



(a)

(b)

Figure 6.18: Example of similarity gestalt grouping principle.

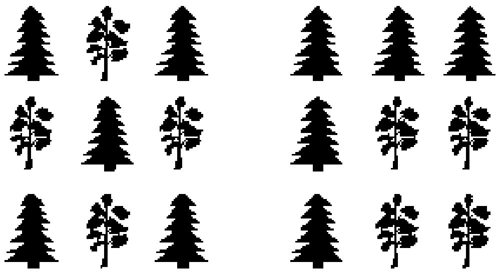


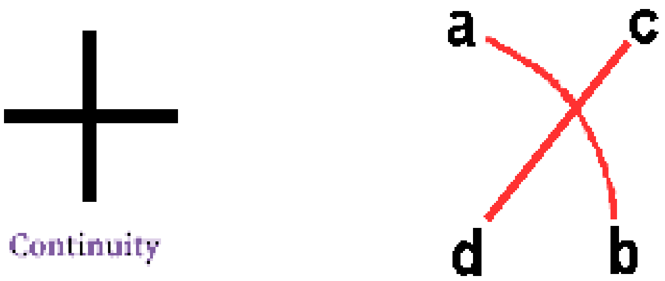
Figure 6.19: Example of similarity gestalt grouping principle.

Another visual example is given in figure 6.19: So in the graphic on the left you probably see an X of fir trees against a background of the others; in the graphic on the right you may see a square of the other trees, partly surrounded by fir trees. The fact that in one we see an X and in the other a square is, incidentally, an example of good form or pragnanz principle, stating that psychological organization will always be as 'good' as prevailing conditions allow. For Gestalt psychologists form is the primitive unit of perception. When we perceive, we will always pick out form.

Good continuation: Components that display smooth transitions from one state to another are perceived as related. Examples of smooth transitions are: proximity in time of offset of one component with onset of another; frequency proximity of consecutive components; constant glide trajectory of consecutive components; smooth transition from one state to another state for the same parameter. For example an abrupt change in the pitch of a voice produces the illusion that a different speaker has interrupted the original. The perception appears to depend on whether or not the intonation contour changes in a natural way. Sound that is interrupted by a noise that masks it, can appear to be continuous. Alternations of sound and mask can give the illusion of continuity with the auditory system interpolating across the mask.

In figure 6.15, B), high (H) and low (L) tones alternate. If the notes are connected by glissandi (figure 6.15, B1), both tones are grouped to a single stream. If high and low notes remain unconnected (figure 1, B2), Hs and Ls each group to a separate stream.

A visual example is given in figure 6.20. The law of good continuation states that objects arranged in either a straight line or a smooth curve tend to be seen as a unit. In figure 6.20(a) we distinguish two lines, one from **a** to **b** and another from **c** to **d**, even though this graphic could represent another set of lines, one from **a** to **d** and the other from **c** to **b**. Nevertheless, we



(a) (b)

Figure 6.20: Examples of good continuation gestalt grouping principle.

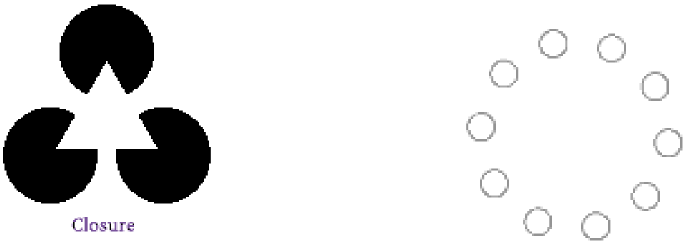
are more likely to identify line **a** to **b**, which has better continuation than the line from **a** to **d**, which has an obvious turn. In figure 6.20(b) we perceive the figure as two crossed lines instead of 4 lines meeting at the centre.

Common Fate Sounds will tend to be grouped together if they vary together over time. Differences in onset and offset in particular are very strong grouping cues. Also, sounds that are modulated together (amplitude or frequency modulation) tend to be grouped together. The principle 'common fate' groups frequency components together, when similar changes occur synchronously, e.g. synchronous onsets, glides, or vibrato.

Chowning (Fig. 6.16, D) made the following experiment: First three pure tones are played. A chord is heard, containing the three pitches. Then the full set of harmonics for three vowels (/oh/, /ah/, and /eh/) is added, with the given frequencies as fundamental frequencies, but without frequency fluctuations. This is not heard as a mixture of voices but as a complex sound in which the three pitches are not clear. Finally, the three sets of harmonics are differentiated from one another by their patterns of fluctuation. We then hear three vocal sounds being sung at three different pitches.

Closure This principle is the tendency to perceive things as continuous even though they may be discontinuous. If the gaps in a sound are filled in with another more intense sound, the original sound may be perceived as being continuous. For example, if part of a sentence is replaced by the sound of a door slam, the speaker's voice may be perceived as being continuous (continuing through the door slam). The principle of closure completes fragmentary features, which already have a 'good Gestalt'. E.g. ascending and descending glissandi are interrupted by rests (Fig. 6.16, C2). Three temporally separated lines are heard one after the other. Then noise is added during the rests (Fig. 6.16 C1). This noise is so loud, that it would mask the glissando, unless it would be interrupted by rests. Amazingly the interrupted glissandi are perceived as being continuous. They have 'good Gestalt': They are proximate in frequency before and after the rests. So they can easily be completed by a perceived good continuation. This completion can be understood as an auditory compensation for masking.

Figure / Ground It is usual to perceive one sound source as the principal sound source to which one is attending, and relegate all other sounds to be background. We may switch our attention from one sound source to another quite easily. What was once figure (the sound to which we were attending) may now become ground (the background sound). An important topics in auditory perception are attention and learning. In a cocktail party environment, we can focus on one



(a) (b)

Figure 6.21: Example of closure.

speaker. Our attention selects this stream. Also, whenever some aspect of a sound changes, while the rest remains relatively unchanging, then that aspect is drawn to the listener's attention ('figure ground phenomenon'). Let us give an example for learning: The perceived illusory continuity (see Fig. 6.16, C) of a tune through an interrupting noise is even stronger, when the tune is more familiar.

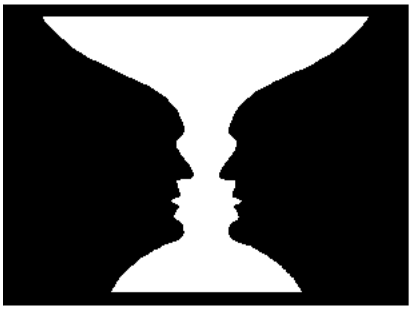


Figure 6.22: Rubin vase: example of figure/ground principle.

The Rubin vase shown in Fig. 6.22 is an example of this tendency to pick out form. We don't simply see black and white shapes - we see two faces and a vase. The problem here is that we see the two forms of equal importance. If the source of this message wants us to perceive a vase, then the vase is the intended figure and the black background is the ground. The problem here is a confusion of figure and ground. A similar everyday example is:

- an attractive presenter appears with a product; she is wearing a 'conservative' dress; eye-tracking studies show substantial attention to the product; three days later, brand-name recall is high;
- an attractive presenter appears with a product; she is wearing a 'revealing' dress; eye-tracking shows most attention on the presenter; brand-name recall is low.

Escher often designed art which played around with figure and ground in interesting ways. Look at how figure and ground interchange in fig. 6.23. Do you see the white horses and riders? Now look for the black horses and riders.

Gestalt grouping laws do not seem to act independently. Instead, they appear to influence each other, so that the final perception is a combination of all of the Gestalt grouping laws acting together. Gestalt theory applies to all aspects of human learning, although it applies most directly to perception and problem-solving.

Figure 6.23: Horses by M. Escher. An artistic example of figure and ground interchange.

6.3 Basic algorithms for melody processing

6.3.1 Melody

Melody may be defined as a series of individual musical events one occurring after another in order so that the composite order constitute a recognizable entity. The essential elements of any melody are duration, pitch, and sound quality (e.g. timbre, texture, and loudness). It represents the linear or horizontal aspect of music and should not be confused with harmony, which is the vertical aspect of music.

6.3.1.1 Melody representations

6.3.1.2 Melodic contour

Contour may be defined as the general shape of an object, often, but not exclusively, associated with elevation or height, as a function of distance, length, or time. In music, contour can be a useful tool for the study of the general shape of a musical passage. A number of theories have been developed that use the rise and fall of pitch level, changes in rhythmic patterns or changes in dynamics as a function of time (or temporal order) to compare or contrast musical passages within a single composition or between compositions of a single composer. One application of the melodic contour is finding out whether the sequence contains repeated melodic phrases. This can be done using computing the autocorrelation.

Parsons showed that encoding a melody by using only the direction of pitch intervals can still provide enough information for distinguishing between a large number of tunes. In Parsons code for melodic contours, each pair of consecutive notes is coded as ”U” (”up”) if the second note is higher than the first note, ”R” (”repeat”) if the pitches are equal, and ”D” (”down”) otherwise. Rhythm is completely ignored. Thus, the first theme from Beethoven’s 8th symphony (Fig. 6.24) would be coded D U U D D D U R D R U U U U. Note that the first note of any tune is used only as a reference point and does not show up explicitly in the Parsons code. Often an asterisk (*) is used in the Parsons code field for the first note. A more precise and effective way of representing contours employs 5-level quantization (++,+,0,-,–) distinguishing between small intervals (steps), which are 1 or 2 semitones wide, from larger intervals (leaps), which are at least 3 semitones wide. The symbols (++,+,0,-,–) are used to code this representation. For example the Beethoven's theme of Fig. 6.24 will be coded as – + + – – – ++ 0 - 0 + + + +.

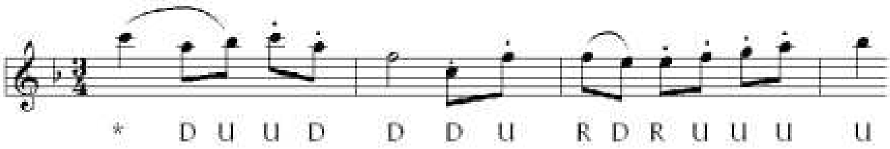


Figure 6.24: Melodic contour and Parson code.

In MPEG-7 the *Melody Contour DS* uses a 5-step contour (representing the interval difference between adjacent notes), in which intervals are quantized. The Melody Contour DS also represents basic rhythmic information by storing the number of the nearest whole beat of each note, which can dramatically increase the accuracy of matches to a query.

For applications requiring greater descriptive precision or reconstruction of a given melody, the *Melody DS* supports an expanded descriptor set and high precision of interval encoding. Rather than quantizing to one of five levels, the precise pitch interval (to cent or greater precision) between notes is kept. Precise rhythmic information is kept by encoding the logarithmic ratio of differences between the onsets of notes in a manner similar to the pitch interval.

6.3.1.3 Similarity measures

When we want to compare melodies, a computable similarity meassure is need. The measures can roughly be classified in three categories: Vector measures, symbolic measures and musical (mixed) measures, according to the computational algorithm used.

- The vector measure treat the transformed melodies as vectors in a suitable real vector space, where methods like scalar products and other means of correlation can be applied to.

- On the contrary the symbolic measures treat the melodies as strings, i.e., sequences of symbols, where well-known measures like Edit Distance or n-gram-related measures can be used.

- The musical or mixed measures typically involve more or less specific musical knowledge and the computation can be from either the vector or the symbolical or even completely different ways like scoring models.

The distance can be computed on different representations of the melodies (e.g. the melody itself, its contour), or some statistic distributions (e.g. pitch classes, pitch class transitions, intervals, interval transitions, note durations, note duration transitions)

6.3.1.4 Edit distance

Approximate string pattern matching is based on the concept of *edit distance*. The edit dsistance $D(A, B)$ between string $A = a_1, \ldots, a_m$ and $B = b_1, \ldots, b_n$ is the minimum number of editing operations required to transform string A into string B, where an operation is an insertion, deletion, or substitution of a single character. The special case in which deletions and insertions are not allowed is called the Hamming distance.

We can define recursively the (edit) distance $d[i,j]$ for going from string $A[1..i]$ to string $B[1..j]$ as

$$d[i,j] = \min \begin{cases} d[i-1,j] + w(a_i,0), & //\text{deletion of } a_i \\ d[i,j-1] + w(0,b_j), & //\text{insertion of } b_j \\ d[i-1,j-1] + w(a_i,b_j) & //\text{match or change} \end{cases} \tag{6.7}$$

where $w(a_i,0)$ is the weight associated with the deletion of a_i, $w(0,b_j)$ is the weight for insertion of a_i, and $w(a_i,b_j)$ is the weight for replacement of element i of sequence A by element j of sequence B. The operation titled ”match/change” sets $w(a_i,b_j) = 0$ if $a_i = b_j$ and a value greater than 0 if $a_i \neq b_j$. Often the weights used are 1 for insertion, deletion and substitution(change) and 0 for match. The initial conditions are given by $d[0,0] = 0$, $d[i,0] = d[i-1,j] + w(a_i,0)$ for $i \geq 1$ and $d[0,j] = d[0,j-1] + w(0,b_j)$ for $j \geq 1$.

The edit distance $D(A,B) = d[n,m]$ can be computed by dynamic programming with running time $O(n \cdot m)$ with the algorithm given in Fig. 6.25.

```
Algorithm EditDistance ($A[1..m], B[1..n,]$, $w_{del}$, $w_{ins}$, $w_{sub}$)
for i from 0 to m
    d[i, 0] := i · $w_{del}$
for j from 0 to n
    d[0, j] := j · $w_{ins}$

for i from 1 to m
    for j from 1 to n
        if A[i] = B[j]
            then cost := 0
            else cost := $w_{sub}$
        d[i,j] := min( d[i-1,j]+$w_{del}$, d[i,j-1]+$w_{ins}$, d[i-1, j-1]+cost )
return d[m,n]
```

Figure 6.25: Dynamic programming algorithm for computing EditDistance.

6.3.2 Melody segmentation

Generally a piece of music can be divided into section and segments at different level. The term grouping describe the general process of segmentation at all levels. Grouping in music is a complex matter. Most computational approaches focused on low-level grouping structure. Grouping events together involves storing them in memory as a larger unit, which is encoded to aid further cognitive processing. Indeed grouping structure plays an important role in recognition of repeated patterns in music. Notice that also the metric structure organize the events on time. However meter involves a framework of level of beats and in itself implies no segmentation; grouping is merely a segmentation without accentural implications.

6.3.2.1 Gestalt based segmentation

Tenny and Polansky proposed a model for small-level grouping in monophonic melodies based on Gestalt rules of proximity (i.e. the preference for grouping boundaries at long intervals between onsets) and similarity (i.e. the preference for grouping boundaries at changes in pitch and dynamics). Moreover the boundary value depends on the context. Thus an interval value in some parameter tends to be a grouping boundary if it is a local maximum, i.e. if it is larger the values immediately preceding and following it. In order to combine the differences of all parameters in a single measure the L_1 norm is proposed, i.e. the absolute values are summed.

The algorithm proceeds in this way:

Algorithm `TenneyLLgrouping`

1. Given a sequence of n tones with pitch $p[i]$ and IOI $ioi[i]$, for $i = 1, \dots, n$
2. for i = 1 to n-1
   Compute the distance $d[i]$ between event i and $i+1$ as
   $$d[i] = ioi[i] + |p[i+1] - p[i]|$$
3. for i = 2 to n-2
   if $d[i-1] < d[i] > d[i+1]$ then i is a low-level boundary point, and $i+1$ is the starting point of a new group.

For higher level grouping the changes perceived at the boundary are taken into account. In order to deal with this, a distinction is made between mean-intervals and boundary-intervals as follows:

- A mean-interval between two groups is the difference between their mean values in that parameter. For the time parameter, the difference of their starting time is considered.
- A boundary-interval is the difference the values of the final component of the first group and the initial component of the second group

The mean-distance between two groups is a weighted sum of the mean-intervals between them, and the boundary-distance is given by a weighted sum of the boundary-intervals between them. Finally the disjunction between two groups is a weighted sum of mean-distance and boundary-distance between them. As a conclusion a group at a higher level will be initiated whenever a group occurs whose disjunction is greater than those immediately preceding and following it.

The algorithm proceeds in the following way:

Algorithm `TenneyHLgrouping`

1. for every group k, the mean pitch is computed by weighting the pitches with the durations
   $$mean_p[k] = \frac{\sum_j p[j] \cdot dur[j]}{\sum_j dur[j]}$$
   where in the summations, j spans all the events in group k.
2. compute the mean-distance
   $$mean\_dist[k] = |mean_p[k+1] - mean_p[k]| + (onset[k+1] - onset[k])$$

3. compute the boundary-distance

$$boundary\_dist[k] = |p[first[k+1]] - p[last[k]]| + (onset[k+1] - on[last[k]])$$

   where $first[k]$ and $last[k]$ are the indexes of the first and last note of group k and $onset[k] = on[first[k]]$.

4. compute the disjunction by

$$disj[k] = w_{md} \cdot mean\_dist[k] + w_{bd} \cdot boundary\_dist[k]$$

5. if $disj[k-1] < disj[k] > disj[k+1]$ then the k-th group is the starting point of a new higher-level segment.

6.3.2.2 Local Boundary Detection Model (LBDM)

In this section, a computational model (developed by Emilios Cambouropoulos 2001), that enables the detection of local melodic boundaries will be described. This model is simpler and more general than other models based on a limited set of rules (e.g. implication realization model seen in sect. 6.6.2) and can be applied both to quantised score and non-quantised performance data.

The Local Boundary Detection Model (LBDM) calculates boundary strength values for each interval of a melodic surface according to the strength of local discontinuities; peaks in the resulting sequence of boundary strengths are taken to be potential local boundaries.

The model is based on two rules: the Change rule and the Proximity rule. The Change rule is more elementary than any of the Gestalt principles as it can be applied to a minimum of two entities (i.e. two entities can be judged to be different by a certain degree) whereas the Proximity rule requires at least three entities (i.e. two entities are closer or more similar than two other entities).

- Change Rule (CR): Boundary strengths proportional to the degree of change between two consecutive intervals are introduced on either of the two intervals (if both intervals are identical no boundary is suggested).

- Proximity Rule (PR): If two consecutive intervals are different, the boundary introduced on the larger interval is proportionally stronger.

The Change Rule assigns boundaries to intervals with strength proportional to a degree of change function S_i (described below) between neighbouring consecutive interval pairs. Then a Proximity Rule scales the previous boundaries proportionally to the size of the interval and can be implemented simply by multiplying the degree-of-change value with the absolute value of each pitch/time/dynamic interval. This way, not only relatively greater neighbouring intervals get proportionally higher values but also greater intervals get higher values in absolute terms - i.e. if in two cases the degree of change is equal, such as sixteenth/eighth and quarter/half note durations, the boundary value on the (longer) half note will be overall greater than the corresponding eighth note.

The aim is to develop a formal theory that may suggest all the possible points for local grouping boundaries on a musical surface with various degrees of prominence attached to them rather than a theory that suggests some prominent boundaries based on a restricted set of heuristic rules. The discovered boundaries are only seen as potential boundaries as one has to bear in mind that musically interesting groups can be defined only in conjunction with higher-level grouping analysis (parallelism, symmetry, etc.). Low-level grouping boundaries may be coupled with higher-level theories so as to produce optimal segmentations (see fig. 6.26).

Figure 6.26: Beginning of Frère Jacques. Higher-level grouping principles override some of the local detail grouping boundaries (note that LBDM gives local values at the boundaries suggested by parallelism - without taking in account articulation.

In the description of the algorithm only the pitch, IOI and rest parametric profiles of a melody are mentioned. It is possible, however, to construct profiles for dynamic intervals (e.g. velocity differences) or for harmonic intervals (distances between successive chords) and any other parameter relevant for the description of melodies. Such distances can also be asymmetric; for instance the dynamic interval between p and f should be greater that between f and p.

Local Boundary Detection algorithm description Given a melodic sequence of n tones, where the i-th tone is represented by pitch $p[i]$, onset $on[i]$, offset $off[i]$.

A melodic sequence is converted into a number of independent parametric interval profiles P_k for the parameters: pitch (pitch intervals), ioi (interonset intervals) and rest (rests - calculated as the interval between current onset with previous offset). Pitch intervals can be measured in semitones, and time intervals (for IOIs and rests) in milliseconds or quantised numerical duration values. Upper thresholds for the maximum allowed intervals should be set, such as the whole note duration for IOIs and rests and the octave for pitch intervals; intervals that exceed the threshold are truncated to the maximum value. Thus we have

Algorithm `LBDM`

```
1. Given: pitch $p[i]$, onset $on[i]$, offset $off[i]$ for $i = 1, \dots, n$.
2. Compute the pitch profile $P_p$ as $P_p[i] = |p[i+1] - p[i]|$ with $i = 1, \dots, n-1$.
3. Compute the IOI profile $P_{IOI}$ as $P_{IOI}[i] = |on[i+1] - on[i]|$ with $i = 1, \dots, n-1$.
4. Compute the rest profile $P_r$ as $P_r[i] = \max(0; on[i+1] - off[i])$ with $i = 1, \dots, n-1$.
5. for each profile $P_k$,
   compute the strength sequence $S_k$ with algorithm ProfileStrength
6. Compute the boundary strength sequence $LB$ as a weighted average of the individual strength
   sequences $S_k$. I.e.
       $LB[i] = w_{pitch} S_p[i] + w_{ioi} S_{IOI}[i] + w_{rest} S_r[i].$
7. Local peaks in this overall strength sequence $LB$ indicate local boundaries.
```

The suggested weights for the three different parameters are $w_{pitch} = w_{rest} = 0.25$ and $w_{ioi} = 0.50$.

In order to compute the profile strength the following algorithm is used.

Algorithm `ProfileStrength`

```
1. Given the parametric profile $P_k = [x[1], \dots x[n-1]]$
```

2. Compute the degree of change $r[i]$ between two successive interval values x_i and $x[i+1]$ by:

$$r[i] = \frac{|x[i] - x[i+1]|}{x[i] + x[i+1]}$$

   if $x[i] + x[i+1] \neq 0$ and $x[i], x[i+1] \geq 0$; otherwise $r[i] = 0$.

3. Compute the strength of the boundary $s[i]$ for interval $x[i]$ which is affected by both the degree of change to the preceding and following intervals, and is given by the function:

$$s[i] = x[i] \cdot (r[i-1] + r[i])$$

4. Normalise the strength sequence in the range $[0, 1]$, by computing $s[i] = s[i]/\max_j(s[j])$

5. Return the sequence $S = \{\, s[2], \ldots, s[n-1] \,\}$

6.3.3 Tonality: Key finding

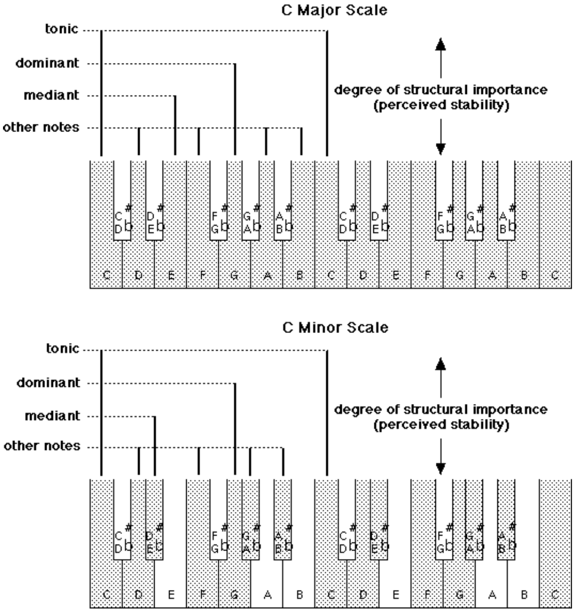


Figure 6.27: Piano keyboard representation of the scales of C major and C minor. Notes in each scale are shaded. The relative importance of the first (tonic - C), fifth (dominant - G) and third (mediant - E) degrees of the scale is illustrated by the length of the vertical bars. The other notes of the scale are more or less equally important followed by the chromatic notes that are not in the scale (unshaded) [from McAdams 1996].

In the Western tonal pitch system, some pitches and chords, such as those related to the first and fifth degrees of the scale (C and G are the tonic and dominant notes of the key of C major, for example) are structurally more important than others (Fig. 6.27). This hierarchization gives rise to a sense of key. In fact when chords are generated by playing several pitches at once, the chord that is considered to be most stable within a key, and in a certain sense to ”represent” the key, comprises the first, third and fifth degrees of the scale. In tonal music, one can establish a sense of key within a given major or minor scale and then move progressively to a new key (a process called modulation) by introducing notes from the new key and no longer playing those from the original key that are not present in the new key.

Factors other than the simple logarithmic distance between pitches affect the degree to which they are perceived as being related within a musical system. The probe tone technique developed by Krumhansl has been quite useful in establishing the psychological reality of the hierarchy of relations among pitches at the level of notes, chords, and keys. In this paradigm, some kind of musical context is established by a scale, chord, melody or chord progression, and then a probe stimulus is presented. Listeners are asked to rate numerically either the degree to which a single probe tone or chord fits with the preceding context or the degree to which two notes or chords seem related within the preceding context. This technique explores the listener's implicit comprehension of the function of the notes, chords, and keys in the context of Western tonal music without requiring them to explicate the nature of the relations.

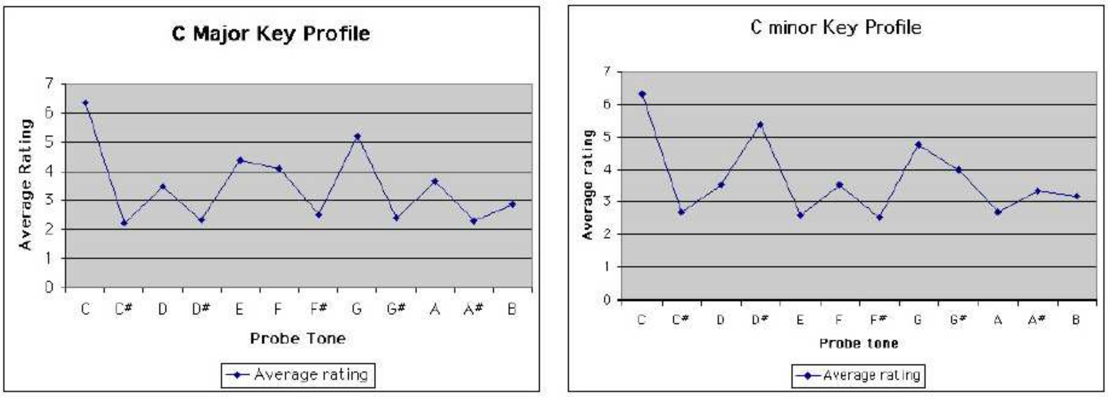


Figure 6.28: C Major and C minor profiles derived with the probe-tone technique from fittingness ratings by musician listeners.

If we present a context, such as a C major or C minor scale, followed by a single probe tone that is varied across the range of chromatic scale notes on a trial-to-trial basis, a rating profile of the degree to which each pitch fits within the context is obtained. This quantitative profile, when derived from ratings by musician listeners, fits very closely to what has been described intuitively and qualitatively by music theorists (Fig. 6.28). Note the importance of the tonic note that gives its name to the scale, followed by the dominant or fifth degree and then the mediant or third degree. These three notes form the principal triad or chord of the diatonic scale. The other notes of the scale are of lesser importance followed by the remaining chromatic notes that are not within the scale. These profiles differ for musicians and non-musicians. In the latter case the hierarchical structure is less rich and can even be reduced to a simple proximity relation between the probe tone and the last note of the context.

Krumhansl has shown (fig. 6.29) that the hierarchy of tonal importance revealed by these profiles is strongly correlated with the frequency of occurrence of notes within a given tonality (the tonic appears more often than the fifth than the third, and so on). It also correlates with various measures of tonal consonance of notes with the tonic, as well as with statistical measures such as the mean duration given these notes in a piece of music (the tonic often having the longest duration).

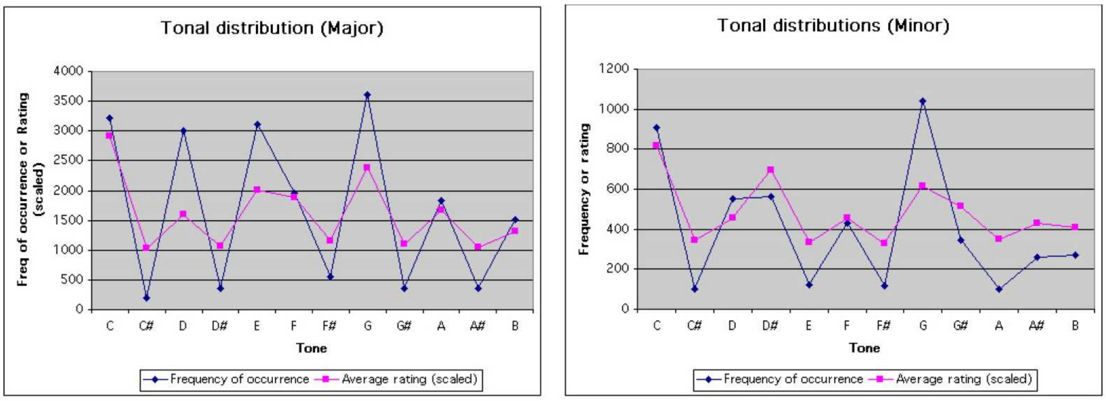


Figure 6.29: Comparison between tonal hierarchies and statistical distribution of tones in tonal works. It is shown the frequency of occurrence of each of the 12 chromatic scale tones in various songs and other vocal works by Schubert, Mendelssohn, Schumann, Mozart, Richard Strauss and J. A. Hasse. and the key profile (scaled).

6.3.3.1 Key finding algorithm

These correlations are the base of the classic key finding algorithm of Krumhansl-Schmuckler (as explained in Krumhansl's book Cognitive Foundations of Musical Pitch [Oxford University Press, 1990]). Each key has a key-profile: a vector representing the optimal distribution of pitch-classes for that key. The `KSkeyFinding` algorithm works as follows.

Algorithm `KSkeyFinding`

1. Given a music segment of n tones, with pitch $p[i]$, duration $dur[i]$, for $i = 1, \dots, n$.
2. Given the key profiles, 12 for major key and 12 for minor key
3. Compute the pitch class distribution vector pcd[0..11], taking into account the tone duration with:

```
for i from 1 to n
    pcd[i] = 0
for i from 1 to n
    pc = $p$[i] mod 12
    pcd[pc] = pcd[pc] +$dur[i]$
```

4. Compute correlations of for all 24 major and minor pitch-class keys
5. Assume that the estimated key for the passage is given by the largest positive correlation.

In this method, the input vector for a segment represents the total duration of each pitch-class in the segment. The match between the input vector and each key-profile is calculated using the standard correlation formula.

For example, if we take opening bar of Yankee Doodle, as shown in fig. 6.30, we find that: the sum of the durations of the G naturals gives .75 of a minim, the durations of the B naturals add up to

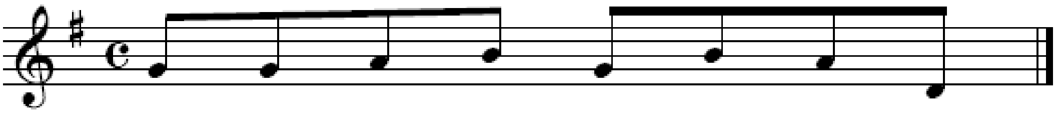


Figure 6.30: Example of Krumhansl-Schmuckler key finding algorithm: opening bar of *Yankee Doodle*.

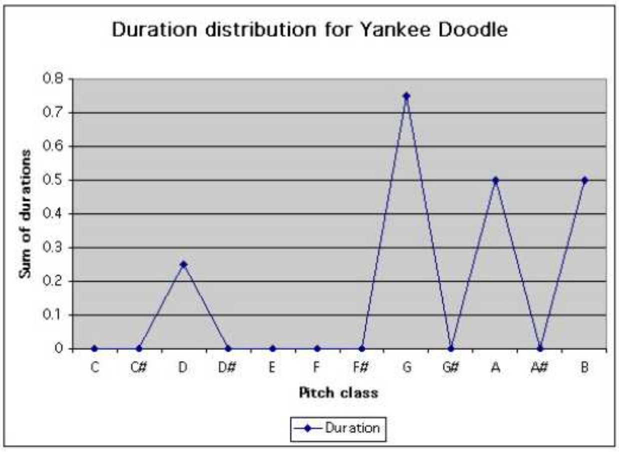


Figure 6.31: Example of Krumhansl-Schmuckler key finding algorithm: duration distribution of *Yankee Doodle*.

Key	Score	Key	Score
C major	0.274	C minor	-0.013
C sharp major	-0.559	C sharp minor	-0.332
D major	0.543	D minor	0.149
E flat major	-0.130	E flat minor	-0.398
E major	-0.001	E minor	0.447
F major	0.003	F minor	-0.431
F sharp major	-0.381	F sharp minor	0.012
G major	**0.777**	G minor	0.443
A flat major	-0.487	A flat minor	-0.106
A major	0.177	A minor	0.251
B flat major	-0.146	B flat minor	-0.513
B major	-0.069	B minor	0.491

Table 6.3: Correlation between the graph showing the durations of the various pitches in the Yankee Doodle excerpt and each of the major and minor key profiles.

half a minim, the durations of the A naturals add up to half a minim and there is one quaver D natural. We can then draw a graph showing the durations of the various pitch classes within the passage being analysed, as shown in fig 6.31. The next step in the algorithm is to calculate the correlation between this graph and each of the 24 major and minor key profiles. This table (tab. 6.3) shows the correlation between this graph showing the durations of the various pitches in the Yankee Doodle excerpt and

C key note names	C	C#	D	D#	E	F	F#	G	G#	A	A#	B
major key	5	2	3.5	2	4.5	4	2	4.5	2	3.5	1.5	4
minor key	5	2	3.5	4.5	2	4	2	4.5	3.5	2	1.5	4

Table 6.4: Temperley key profiles. The note names refer to C major and C minor key.

each of the major and minor key profiles. The algorithm then predicts that the perceived key will be the one whose profile best correlates with the graph showing the distribution of tone durations for the passage. So in this case, the algorithm correctly predicts that the key of Yankee Doodle is G major.

A variation of the key finding algorithm is proposed in Temperley 2001 (`KSTkeyFinding` algorithm). In this method, the input vector for a segment simply has 1 for a pitch-class if it is present at all in the segment (the duration and number of occurrences of the pitch-class are ignored) and 0 if it is not; the score for a key is given by the sum of the products of key-profile values and corresponding input vector values (which amounts to summing the key-profile values for all pitch class present in the segment). Moreover the key profiles were heuristically adjusted and are given in Table 6.4. Notice that given a C major key profile, the other major key profiles can be simply obtained by acyclical shift, and in a similar way all the minor key profiles can be obtained from the Cminor key profile.

The `KSTkeyFinding` algorithm works as follows.

Algorithm `KSTkeyFinding`

1. Given a music segment of n tones, with pitch $p[i]$, for $i = 1, \ldots, n$.
2. Given the (modified) key profiles, 12 for major key and 12 for minor key
3. Compute the pitch class vector pv, where $pv[k] = 1$ if pitch class k is present in the music segment, else $pv[k] = 0$. I.e.

```
**for** $k$ **from** 0 **to** 11
    $pv[k]$ = 0
**for** i **from** 1 **to** n
    $pv[\,p[i]\,]$ = 1
```

4. for all 24 major and minor key profiles,
   Compute the scalar product of pv with the key profile vector kp as

$$\sum_j pv[j] \cdot kp[j]$$

5. Assume that the estimated key for the passage is given by the largest positive scalar product.

6.3.3.2 Modulation

The key finding algorithms produce a single key judgement for a passage of music. However, a vital part of tonal music is the shift of keys from one section to another. In music, modulation is most commonly the act or process of changing from one key (tonic, or tonal center) to another.

The key finding algorithm could easily be run on individual sections of a piece, once these sections were determined. It is possible to handle modulation: in considering a key for a segment, a penalty is assigned if the key differs from the key of the previous segment. In this way, it will prefer to remain in the same key, other things being equal, but will change keys if there is sufficient reason to do so. This task can be dealt with an algorithm similar to Viterbi algorithm, which can be implemented by dynamic programming as the following `KeyModulation` algorithm.

Algorithm `KeyModulation`

```
Given $m$ music segments
for every segment $i = 1, \dots, m$
    compute $q[i, \cdot]$ vector of key weights by a key finding algorithm
Let $d[1, \cdot] = q[1, \cdot]$
for $i$ = 2 to m
    for $j$ = 0 to 23
        $d[i, j] = q[i, j] + \max_k (d[i - i, k] - w(k, j))$
        $pr[i, j] = \arg\max_k (d[i - i, k] - w(k, j))$
$key[m] = \arg\max_j d[m, j]$

for $i$ = $m$-1 downto 1
    $key[i] = pr[key[i + 1]]$
```

In this algorithm, the vector position $pr[i, j]$ contains the best previous key which conducted to the j-th key estimation of the segment i. The function $w(k, j)$ gives the penalty for passing from k to j key. The penalty value is zero if there is no key chance: i.e. $w(j, j) = 0$.

With this strategy, the choice does not depends only on the segment in isolation, but it takes into account also previous evaluations. At each segment each key receives a local score indicating how compatible that key is with the pitches of the segment. Then we compute the best so far analysis ending at that key. The best scoring analysis of the last segment can be traced back to yield the preferred analysis of the entire piece. Notice that some choices can be changes as we proceed in the analysis of the segments. In this way the dynamic programming model gives a nice account of an important phenomenon in music perception: the fact that we sometimes revise our initial analysis of a segment based on what happens later.

6.4 Music Information Retrieval: Issues, Problems, and Methodologies

by Nicola Orio

6.4.1 Introduction

The core problem of Information Retrieval (IR) is to effectively retrieve documents which convey content being relevant to the user's information needs. Effective and efficient techniques have been developed to index, search, and retrieve documents from collections of hundreds of thousands, or millions of textual items.

The most consolidated results have been obtained for collection of documents and user's queries written in textual form and in English language. Statistical and probabilistic techniques have lead to the most effective results for basic system functions and are currently employed to provide advanced information access functions as well. The content description of media being different from text, and the development of different search functions are necessary steps for content-based access to Digital Libraries (DL). This statement mainly applies to cultural heritage domain, where different media and search functions live together.

In order to provide a content-based multimedia access, the development of new techniques for indexing, searching, and retrieving multimedia documents have recently been the focus of many researchers in IR. The research projects in DLs, and specifically those carried out in cultural heritage domain, have shown that the integrated management of diverse media - text, audio, image, video - is necessary.

The problem with content-based access to multimedia data is twofold.

- On the one hand, each media requires specific techniques that cannot be directly employed for other media.
- On the other hand, these specific techniques should be integrated whenever different media are present in a individual item.

The core IR techniques based on statistics and probability theory may be more generally employed outside the textual case and within specific non-textual application domains. This is because the underlying models, such as the vector-space and the probabilistic models, are likely to describe fundamental characteristics being shared by different media, languages, and application domains.

6.4.1.1 Digital Music and Digital Libraries

There is an increasing interest towards music stored in digital format, which is witnessed by the widespread diffusion on the Web of standards for audio like MP3. There are a number of reasons to explain such a diffusion of digital music.

- First of all, music is an art form that can be shared by people with different culture because it crosses the barriers of national languages and cultural backgrounds. For example, tonal Western music has passionate followers also in Japan and many persons in Europe are keen on classical Indian music: all of them can enjoy music without the need of a translation, which is normally required for accessing foreign textual works.
- Another reason is that technology for music recording, digitalization, and playback, allows for an access that is almost comparable to the listening of a live performance, at least at the level of audio quality, and the signal to noise ratio is better for digital formats than for many analog formats. This is not the case of other art forms, like painting, sculpture or even photography, for which the digital format is only an approximate representation of the artwork. The access to digitized paintings can be useful for studying the works of a given artist, but cannot substitute the direct interaction with the real world works.
- Moreover, music is an art form that can be both cultivated and popular, and sometimes it is impossible to draw a line between the two, as for jazz or for most of ethnic music.

These reasons, among others, may explain the increasing number of projects involving the creation of music DLs. A music DL allows for, and benefits from, the access by users from all over the world, it helps the preservation of cultural heritage, and it is not tailored only to scholars' or researchers' needs. More in general, as music is one of the most important means of expression, the organization, the integration with other media, and the access to the digitized version of music documents becomes an important multimedia DL component. Yet, music has some peculiarities that have to be taken into account when developing a music DL. In figure 6.32 the architecture of a music information retrieval system is shown.

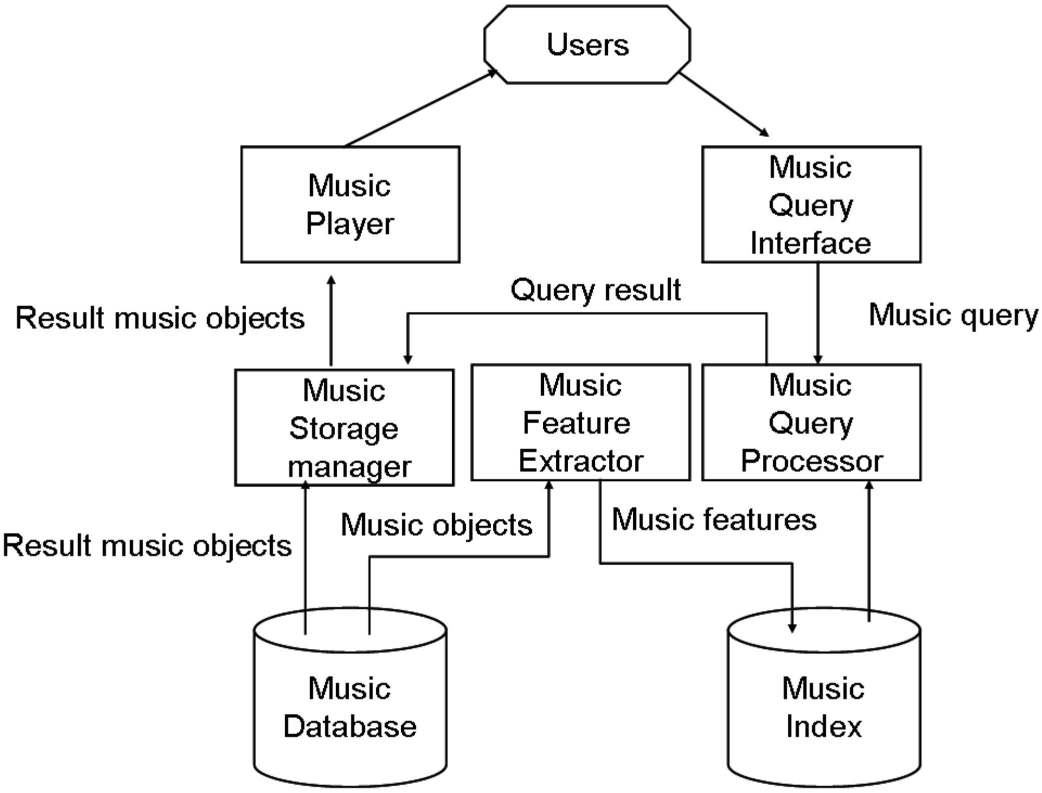


Figure 6.32: Architecture of a music information retrieval system

6.4.1.2 Music Information Retrieval

Specific and effective techniques being capable of indexing and retrieving such multimedia documents as the music ones need to be designed and implemented.

Current approaches to Music Information Retrieval (MIR) are based either on string matching algorithms or textual bibliographic catalogue.

- Sting matching approach makes content-based retrieval very difficult - indeed, retrieving textual files using Unix grep-like commands gives poor results.
- Textual bibliographic catalogue approach makes content-based retrieval impossible since the music content cannot be described by bibliographic catalogue.

The requirement for a content-based MIR has been stressed within the research area of music information systems as well. The developments in the representation of music suggest a need for an information retrieval philosophy directed toward non-text searching and eventual expansion to a system that encompasses the full range of information found in multimedia documents. As IR has dealt with the representation and the disclosure of content from its early days, it is natural to think that IR techniques should be investigated to evaluate their application to music retrieval. According to

McLane “what has been left out of this discussion, and will no doubt be a topic for future study, is the potential for applying some of the standard principles of text information retrieval to music representations”.

- If we follow the hypothesis that the use of standard principles of text information retrieval to index and retrieve music documents is possible, then the design of ad-hoc segmentation algorithms to produce musical ‘lexical units’ like words in textual documents is required.

  The concept of lexical unit may vary depending on the approach. A lexical unit can be: a fixed-length string, the incipit, a complete theme, a melodic phrase, and so on. Music is a continuous flow of events (e.g., notes, chords, and unpitched percussive sounds) without explicit separators, if not those perceived by listeners. Also music representation lacks of separators of lexical units, because it conveys information only about macro-events, like changes in tonality or the presence of repetitions. It is therefore necessary to automatically detect the perceived lexical units of a music document to be used like words in textual documents.

- Moreover, content-based MIR requires the design of normalization algorithms. Once detected, musical lexical units occur in documents with many variants like textual words do within textual documents. For example, a melodic pattern may occur in many music works, perhaps composed by different authors, with small deviations of note intervals or timing. Despite these deviations, different patterns may be perceptually similar, hence conveying the same music perception. It is therefore necessary to detect these variants and conflate all the similar musical lexical units into a common stem expressing the same music perception. This conflation process is analogous to the one performed in the textual case for detecting word stems through, for example, the Porter’s stemming algorithm.

To allow the integration of automatic music processing techniques with automatic IR techniques, segmentation and normalization algorithms are applied also on music queries.

In a content-based music IR system, users may be able to interact with the system by using the same language, that is the music language. This because content-based MIR requires users to be able of expressing the music document content. The most natural way of express music content is singing and playing music. This approach is often referred to as the query by example paradigm. Therefore, users should be provided with interfaces and search functions so that they can play music and send a music query to the system.

To make content-based music retrieval possible, query content and document content have to be matched: Describing query content is then necessary. If we regard music queries as music documents, segmentation and normalization can be performed also on music queries using the same algorithms used for disclosing document content.

6.4.2 Issues of Content-based Music Information Retrieval

Music, in its different representations, can be considered as another medium together with text, image, video, and speech. Nevertheless, there are some issues that make music different from other multimedia IR application domains. The issues we address are form, instantiation, dimension, content, perception, user profile, and formats. The most relevant issues are describes in the following Sections.

6.4.2.1 Peculiarities of the Music Language

The same entity, i.e. a music work, can be represented in two different main forms: the notated and the acoustic form, respectively corresponding to score and performance. Hence the communication in music is performed at two levels:

- the composer translates his intentions in a music structure (music as a composing art),
- the musician translates the written score into sounds (music as a performing art).

Also users may have different needs, in particular the music scholar may look for a given composition, while the melomane may look for a particular performance.

Each music work may have different instantiations. As musicians can interpret scores, the resulting performances may differ and therefore more performances correspond to an individual score. Furthermore, the same music work may be transcribed into different scores, depending on the revisers’ choices. As a consequence, different performances and scores may rely to the same music work.

Different dimensions characterize the information conveyed by music. Melody, harmony, rhythm, and structure are dimensions, carried by the written score, that may be all or in part of interest for the final user. In the case of a performance other dimensions should be added, for instance timbre, articulation, and timing. It is likely that the dimensions of interest vary with the level of user’s expertise and the specific user’s search task. As described in Section 6.4.2.3, different formats are able to capture only a reduced number of dimensions. Therefore, the choice of a representation format has a direct impact on the degree to which a music retrieval system can describe each dimension.

While text, image, video, or speech-based documents in general convey some information that form their content, it is still unclear what type of content, if any, music works do convey. Let us consider an example: the concept of tempest can be described with a textual document, such as the first chapter of Shakespeare’s ‘The Tempest’, a painting, such as the landscape of Giorgione’s ‘The Tempest’, a video or speech, such as broadcasting news about, for instance, a tornado. All these media are able to convey, among all the other information, the concept of tempest. There are up to forty music works of tonal Western music whose title is related to tempests, among those the most famous probably are Beethoven’s Sixth Symphony IV Movement, Rossini’s Overture of ‘William Tell’, and Vivaldi’s Concerto ‘La Tempesta di Mare’. These works differ in music style, form, key and time signature, and above all the user may be not able to recognize that the work is about a tempest and not just pure music.

In principle, music language does not convey information as, for instance, text or video do. Many composers wrote music to stir up emotions, and in general they aimed to communicate no specific information to the listener. The final user feels emotions on listening to the music, and he interprets some information independently from the composer’s and performer’s thought and differently from the other users. There is a particular kind of music works, called musica a programma, in which the title (like Vivaldi’s ‘The Spring’) or a lyric (like Debussy’s ‘Prlude l’aprs-midi d’un faune’) suggests a meaning to the listener; this sort of textual data would be better managed using a database system rather than a IR system. Moreover in sung music, such as Cantatas, the accompanied text gives the work some meaning, yet that sort of text would require ad-hoc IR techniques to be effectively managed. In general the availability of textual material together with music documents is insufficient.

It is then important to consider how music is perceived and processed by listeners, to highlight which kind of content is carried by this medium. A number of different theories was proposed by musicologists, among which the most popular ones are the Generative Theory of Tonal Music (see Sect. 6.6.1) and the Implication-Realization Model (see Sect. 6.6.2). In both cases it is stated that listeners perceive music as structured and consisting of different basic elements. Therefore, even if music notation and performance lack of explicit separators (like blanks or commas in text) musicians and listeners perceive the presence of small elements which constitute the music work: we can consider these elements as the lexical units for a content-based approach to MIR. It is likely that all the dimensions of music language can be segmented in their lexical units and be used to extract a content from a music document.

6.4.2.2 The Role of the User

As always happens in IR, the effectiveness of techniques does strongly depend on the final user. DL systems does indeed interact with final users of very diverse types and with different levels of expertise in the use of the system itself. This is particularly true for music DLs, because there is a great difference in users' expertise depending on the practice of a musical instrument, the ability of reading a score, the knowledge of harmony rules, the familiarity with composition styles, and so on. Users may have different needs, for instance a music scholar may look on how a given cadenza is used by different authors, while a melomane may look for a particular performance of a well-known musician. This is a key aspect in the design of a methodology for content-based MIR, because it affects the choice of the dimension to be used for describing a music work, that is which kind of content has to be extracted from it.

Considering that access to DL is widely spread to users of any type, final users of a music DL may not have a deep knowledge of music language. Therefore, melody seems to be the most suitable dimension. In fact, almost everybody can recognize simple melodies and perform them at least by singing or humming. In this case, lexical units can be considered the musical phrases, which may be defined as short excerpts of the melody which constitute a single musical gesture. Moreover, melody carries also explicit information about rhythm and implicit information about harmony.

Melody can be the most suitable evidence for content-based music retrieval, it may however be the case that only a part of the melody can effectively be exploited as useful evidence for music document and query description. This implies that, if phrases can be detected by means of some segmentation algorithms, then it is likely that some of these phrases are 'good' descriptors of the music content from users' point of view, while others can be dropped since they give little contribution to the music content description and may negatively affect efficiency. This latter consideration leads us to contemplating the possibility of building lists of stop phrases, that may be dropped from the index of phrases similarly to the textual case. However, it is still unclear if stop phrases exist how users perceive them. While one can identify a word as stop word because it has no, little, or less meaning than keywords, one cannot identify a phrase as stop phrase because it is very difficult to say what 'phrase meaning' does mean, and frequency-based stop phrase list construction may be a difficult task because, for instance, users may recall melody excerpts just because they are very frequent in a musical genre.

6.4.2.3 Formats of Music Documents

As previously mentioned, the communication in music is achieved at two levels, corresponding to two forms: the composer translates his intentions into a musical structure, that is represented by a music score, and the musician translates the written score into a performance, that is represented by a flow of acoustic events. A number of different digital formats correspond to each form. It can be noted that, as musicians can interpret scores, the resulting performances differ and therefore more than one performance correspond to a single score. Even if the two forms can be considered as instantiations of the same object, they substantially differ in the information that can be manually or automatically extracted from their respective formats.

The first problem which arises in the automatic processing of music is then that a music work may be digitally stored in different formats. The same music piece can be represented, for example,

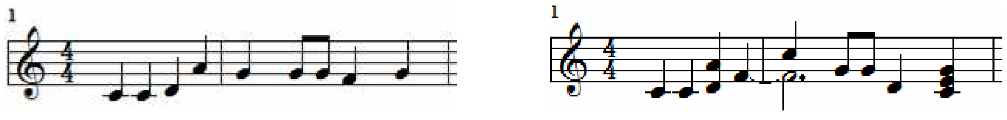
- by a reproduction of the manuscript,

- by a symbolic notation of the score,
- by a sequence of time-stamped events corresponding to pitched and unpitched sounds,
- or by a digital recording of an acoustic performance.

Each format carries different information on the content of the document. For instance, at the state-of-the-art it is impossible to recover informations about the written score from the digital sampling, e.g. stored in a compact disk, of a polyphonic audio signal, and the score carries no information about the timbre, expressive timing and other performing parameters. Hence, the documents format has to be chosen depending on the aims of the DL, which may encompass preservation, displaying, listening, indexing, and retrieval, and so on. As an example, preservation requires high quality audio coding and dissemination over the Internet requires lossy compression.

Formats for digital music documents can be divided in two classes.

- The score is a structured organization of symbols, which correspond to acoustic events; the score is a direct representation of all the dimensions of music (i.e., melody, harmony, and rhythm) and it usually contains all the information that is relevant for classifying and cataloguing: type of movement, time and key signatures, composer's notes, and so on. The symbolic nature of the score allows for an easy representation of its content, and many proposed formats represents score in the form of a textual markup language, for instance ABC and GUIDO.
- The performance is made of a sequence of gestures performed by musicians on their musical instruments; the result is a continuous flow of acoustic waves, which correspond to the vibration induced on musical instruments. Even if all the dimensions of music are embedded in a performance, it requires high-level information processing to recognize them. In particular, only experienced musicians can recognize all the dimensions of music from listening to a performance and, at the state of the art, there is no automatic system that can recognize them from an acoustic recording, apart from trivial cases. The nature of a performance does not allow for an easy representation of its content. The formats adopted to digitally represent performances, such as AIFF (Audio Interchange File Format, proposed by Apple Computers) or MP3 (MPEG1, Layer3), are a plain digital coding of the acoustic sound waves, with a possible data compression.



(a) (b)

Figure 6.33: Example of a melody

We present now an example of different representations of a melody with reference to fig. 6.33(a). we can represent as absolute or relative values.

- Absolute measure:
  - Absolute pitch: `C5 C5 D5 A5 G5 G5 G5 F5 G5`
  - Absolute duration: `1 1 1 1 1 0.5 0.5 1 1`

  - Absolute pitch and duration:
    `(C5,1)(C5,1)(D5,1)(A5,1)(G5,1)(G5,0.5)(G5,0.5)(F5,1)(G5,1)`

- Relative measure:
  - Contour (in semitones): `0 +2 +7 -2 0 0 -2 +2`
  - IOI (Inter onset interval) ratio: `1 1 1 1 0.5 1 2 1`
  - Contour and IOI ratio:
    `(0,1)(+2,1)(+7,1)(-2,1)(0,0.5)(0,1)(-2,2)(+2,1)`

In a polyphonic case (see fig. 6.33(b)) we can represent in different ways.

- Keep all information of absolute pitch and duration (start_time, pitch, duration)
  `(1,C5,1)(2,C5,1)(3,D5,1)(3,A5,1)(4,F5,4)(5,C6,1)(6,G5,0.5)(6.5,G5,0.5)...`
- Relative note representation: Record difference of start times and contour (ignore duration)
  `(1,0)(1,+2)(0,+7)(1,-4) ...`
- Monophonic reduction, e.g. select one note at every time step (main melody selection)
  `(C5,1)(C5,1)(A5,1)(F5,1)(C6,1)...`
- Homophonic reduction (chord reduction), e.g. select every note at every time step
  `(C5)(C5)(D5,A5)(F5)(C6)(G5)(G5) ...`

With the aim of taking into account all the variety in which music information can be represented, it has been proposed the Standard Music Description Language (SMDL), as an application of the Standard ISO/IEC Hyper-media/Time-based Structuring Language. In SMDL, a music work is divided into different domains, each one dealing with different aspects, from visual to gestural, and analytical. SMDL provides a linking mechanism to external, pre-existing formats for visual representation or storage of performances. Hence SMDL may be a useful way for music representation standardization, but the solution is just to collect different formats rather that proposing a new one able to deal with all the aspects of the communication in music.

A Note on MIDI A format that can be considered as a compromise between the score and the performance forms is MIDI (Musical Instrument Digital Interface), which was proposed in 1982 for data exchange among digital instruments. MIDI carries both information about musical events, from which it is possible to reconstruct an approximate representation of the score, and information for driving a synthesizer, from which it is possible to listen to a simplified automatic performance. It seems then that MIDI draws a link between the two different forms for music representation. This characteristics, together with the fortune of MIDI as an exchange format in the early times of the Internet, can explain why many music DLs and most projects regarding music indexing and retrieval refer to it. Some of the research work on music information retrieval take advantage of the availability of MIDI files of about all the different music genres and styles. MIDI files are parsed in order to extract a representation of the music score, and then indexed after different preprocessing.

Nevertheless, MIDI is becoming obsolete and users on the Internet increasingly prefer to exchange digital music stored in other formats such as MP3 or RealAudio, because they allow for a good audio-quality with a considerably small dimension of the documents size. Moreover, if the goal of a music DL is to preserve the cultural heritage, more complete formats for storing both scores and performances are required. Being a compromise between two different needs – i.e., to represent symbols and to be playable – MIDI turns out to fit neither the needs of users who want to access to a complete digital representation of the score, nor to users who want to listen to high-quality audio performances.

6.4.2.4 Dissemination of Music Documents

The effectiveness of a retrieval session depends also on the ability of users to judge whether retrieved documents are relevant to their information needs. The evaluation step, in a classical presentation-evaluation cycle, for an information retrieval session of textual documents usually benefits from tools for browsing the document (e.g., the 'find' function), in particular when the size of documents is large. Moreover, a general overview of the textual content may help users to judge the relevance of most of the retrieved documents.

Users of a music DL cannot take advantage of these shortcuts for the evaluation of documents relevance, when they are retrieving music performances. This is due to the central role played by time in the listening to music. A music performance is characterized by the organization of music events along the time axis, which concatenates the single sounds that form the whole performance. Changing playback speed of more than a small amount may result in a unrecognizable performance. In other words, it requires about 20 minutes to listen to a performance that lasts 20 minutes. It may be argued that many music works are characterized by their incipit, that is by their first notes, and hence a user could be required to listen only to the first seconds of a performance before judging its relevance to his information needs. Anyway, the relevant passage of a music document – e.g., a theme, the refrain – may be at any position in the time axis of the performance.

A tool that is often offered by playback devices is the 'skip' function, that allows for a fast access to a sequence of random excerpts of the audio files, to help listeners looking for given passages. Everyone who tried to find a particular passage in a long music performance, knows that the aid that the skip function gives when accessing to music documents is not even comparable with the find function for textual documents. This is partially due to the fact that auditory information does not allow a snapshot view of the documents as visual information does. The evaluation of relevance of retrieved music documents may then be highly time-consuming, if tools for a faster access to document content are not provided.

6.4.3 Approaches to Music Information Retrieval

There is a variety of approaches to MIR and there are many related disciplines involved. Because of such wide varieties, it is difficult to cite all the relevant work. Current approaches to MIR can broadly be classified into data-based and content-based approaches. For the aims of scientific research on multimedia IR, content-based approaches are more interesting, nevertheless the use of auxiliary textual data structures, or metadata, can frequently be observed in approaches to non-textual, e.g. image or video document indexing. Indeed, textual index terms are often manually assigned to multimedia documents to allow users retrieving documents through textual descriptions.

6.4.3.1 Data-based Music Information Retrieval

Data-based MIR systems allow users for searching databases by specifying exact values for predefined fields, such as composer name, title, date of publication, type of work, etc., in which cases we actually speak about exact match retrieval. Data-based approaches to MIR makes content-based retrieval almost impossible since the music content cannot easily be conveyed simply by bibliographic catalogue only.

Indeed, music works are usually described with generic terms like 'Sonata' or 'Concerto' which are related only to the music form and not the actual content. From an IR point of view, data-based approaches are quite effective if the user can exhaustively and precisely use the available search fields. However, bibliographic values are not always able to describe exhaustively and precisely the content of music works. For example, the term 'Sonata' as value of the type of work cannot sufficiently discriminate all the existing sonatas.

Moreover, many known work titles, such as the Tchaikovskij's 'Pathetic', are insufficient to express a final user's query whenever he would find the title not being a good description of the music work. The use of cataloging number, like K525 for Mozart's 'Eine Kleine Nachtmusic', will be effective only if the user has a complete information on the music work, and in this case a database system will suffice.

Searching by composer name can be very effective. However, some less known composers and their works may not be retrieved if only because the authors are little known. Content-based MIR may allow for the retrieval of these pieces since querying by a known melodic pattern, such as a Mozart's one, may retrieve previously not considered or unknown composers. On the other hand, for a prolific composer, just like Mozart, a simple query by composer's name will retrieve an extremely high number of documents, unbearable for the final user.

6.4.3.2 Content-based Music Information Retrieval

Content-based approaches take into account the music document content, such as notation or performance, and automatically extract some features, such as incipites or other melody fragments, timing or rhythm, instrumentation, to be used as content descriptors. Typical content-based approaches are based on the extraction of note strings from the full-score music document. If arbitrarily extracted, note strings may be meaningless from a musical point of view because no music information is exploited to detect those strings, yet allows for a good coverage of all the possible features to be extracted.

Content-based approaches to MIR can sometimes be oriented to disclosing music document semantic content using some music information, under the hypothesis that music documents can convey some meaning and then some fragments can effectively convey such meaning. In the latter case, some music information is exploited to detect those strings so that the detected strings can musically make sense if, for instance, they were played.

The research work on this area of MIR can be roughly divided in two categories:

- on-line searching techniques, which compute a match between a representation of the query and a representation of the documents each time a new query is submitted to the system;

- indexing techniques, which extract off-line from music documents all the relevant information that is needed at retrieval time and perform the match between query and documents indexes.

Both approaches have positive and negative aspects.

- From the one hand, on-line search allows for a direct modelling of query errors by using, for instance, approximate pattern matching techniques that deal with possible sources of mismatch, e.g. insertion and/or deletion of notes. This high flexibility is balanced by high computational costs, because the complexity is at least proportional to the size of the document collection (and, depending on the technique, to the documents length).

- From the other hand, indexing techniques are more scalable to the document collection, because the index file can be efficiently accessed through hashing and the computational complexity depends only on query length. The high scalability is balanced by a more difficult extraction of document content, with non trivial problems arising in case of query errors that may cause a complete mismatch between query and document indexes.

Both approaches had given interesting and promising results. Yet, indexing approaches need to be investigated in more detail because of the intrinsic higher computational efficiency.

Previous work on on-line search has been carried out following different strategies. A first approach is based on the use of pattern discovery techniques, taken from computational biology, to compute occurrences of a simplified description of the pitch contour of the query inside the collection of documents. Another approach applies pattern matching techniques to documents and queries in GUIDO format, exploiting the advantages of this notation in structuring information. Approximate string matching has been used. Markov chains have been proposed to model a set of themes that has been extracted from music documents, while an extension to hidden Markov models has been presented as a tool to model possible errors in sung queries.

An example of research work on off-line document indexing has been presented in[8]. In that work melodies were indexed through the use of N-grams, each N-gram being a sequence of N pitch intervals. Experimental results on a collection of folk songs were presented, testing the effects of system parameters such as N-gram length, showing good results in terms of retrieval effectiveness, though the approach seemed not be robust to decreases in query length. Another approach to document indexing has been presented in[24], where indexing has been carried out by automatically highlighting music lexical units, or musical phrases. Differently than the previous approach, the length of indexes was not fixed but depended on the musical context. That is musical phrases were computed exploiting knowledge on music perception, in order to highlight only phrases that had a musical meaning. Phrases could undergo a number of different normalization, from the complete information of pitch intervals and duration to the simple melodic profile.

Most of the approaches are based on melody, while other music dimensions, such as harmony, timbre, or structure, are not taken into account. This choice may become a limitation depending on the way the user is allowed to interact with the system and on his personal knowledge on music language. For instance, if the query-by-example paradigm is used, the effectiveness of a system depends on the way a query is matched with documents: If the user may express his information need through a query-by-humming interface, the melody is the most likely dimension that he will use. Moreover, for non expert users, melody and rhythm (and lyrics) are the more simple dimensions for describing their information needs.

Query processing can significantly differ within content-based approaches. After a query has been played, the system can represent it either as a single note string, or as a sequence of smaller note fragments. The latter can be either arbitrary note strings, such as n-grams, or fragments extracted using melody information. Regarding the query as a single note string makes content-based retrieval very difficult since it would be similar to retrieving textual files using Unix grep-like commands which provides very poor results. On the contrary, extracting fragments using melody information can result in a more effective query description. We then speak about partial match retrieval.

6.4.3.3 Music Digital Libraries

Digital library projects have been carried out for designing, implementing, and testing real MIR systems. Some of them implement data-based, content-based, or both approaches to MIR. We cite some of the projects being most relevant to our research aims. The reader can access to the cited papers to have a complete description of methods and systems. The VARIATIONS digital library has been reported in [9], while the MELDEX project is reported in [4]. A project involved the University of Milan and the Teatro alla Scala, Milan [10] to implement a multimedia object-relational database storing the music contents of the archive, as well as catalogue data about the nights at the Teatro alla Scala. The access to the archive is basically based on fragment extraction and approximate string matching. A feasibility study was conducted for the ADMV (Digital Archive for the Venetian Music of the Eighteenth century) digital library project [3]. The feasibility study allowed for defining architecture, technology, and search functions for a data and content-based MIR and database management system. The system complexity is due to the number of inter-relationships of all the aspects being typical of a real effective DL: distributed databases, preservation, wide area networking, protection, data management, content-based access.

6.4.4 Techniques for Music Information Retrieval

Content-based MIR is a quite new research area, at least compared to classical textual IR. For this reason, most of the techniques applied to retrieve music documents derive from IR techniques. In this section, after introducing some terminology typical of content-based description of music documents, techniques for MIR and their relationship with IR techniques are described. A final example is given on how evaluation can be carried out.

6.4.4.1 Terminology

There is a number of terms that have a special meaning for the research community on MIR.

A **feature** is one of the characteristics that describe subsequent notes in a score. A note feature can be: the pitch, the pitch interval with the previous note (PIT), a quantized PIT, the duration, the interonset interval with the subsequent note (IOI), the ratio of IOI with the previous note, and so on. All the features can be normalized or quantized. In the example of sect. 6.4.5.4, features are related to pitch and rhythm that, though usually correlated, can be treated independently. For example, many songs can be guessed only by tapping the rhythm of the melody while other ones can be easily recognized even if played with no tempo or rubato.

A **string** is a sequence of features. Any sequence of notes in a melody can be considered a string. It can be noted that strings can be used as representative of a melody, which is the idea underlying many approaches to MIR, but the effectiveness by which each string represents a document may differ. For instance, it is normally accepted that the first notes of a melody play an important role in recognition, or that strings that are part of the main theme or motif are good descriptors as well. String length is an important issue: Long strings are likely to be effective descriptors, yet they may lead to problems when the user is request to remember long parts of a melody for querying a MIR system. Often, strings shorter than three notes can be discarded, because they can be considered not significant descriptors.

A **pattern** is a string that is repeated at least twice in the score. The repetition can be due to the presence of different choruses in the score or by the use of the same music material (e.g., motifs, rhythmical cells) along the composition. Each pattern is defined by the string of features, by its length n and by the number of times r it is repeated inside the score. All patterns that appear only inside longer patterns have been discarded in the example of sect. 6.4.5.4. The computation of patterns can be done automatically using well known algorithms for pattern discovery. Given a particular feature, patterns can be considered as effective content descriptors of a music document. Depending on the selected feature, patterns carry different information about document content.

It can be noted that a music documents may be directly indexed by its strings. In particular, it can be chosen to describe a document with all its strings of a given length, usually from 3 to 5 notes, that are called *n-grams*. The n-gram approach is a simple, but often effective, alternative to more complex approaches that are based on melodic information. In the following sections, patterns are considered as possible content descriptors, yet the discussion may be generalized to n-grams, musical phrases, and so on. Moreover, in the following discussion, three kinds of features are considered for the pattern selection step – the interonset interval (IOI) normalized to the quarter note, the pitch interval (PIT) in semitones, and both (BTH).

6.4.5 Document Indexing

Document indexing is a mandatory step for textual information retrieval. Through indexing, the relevant information about a collection of documents is computed and stored in a format that allows easy and fast access at retrieval time. Document indexing is carried out only when the collection is created or updated, when users are not yet accessing the documents, and then the problems of computational time and efficiency are usually less restrictive. Indexing speeds up retrieval time because it is faster to search for a match inside the indexes than inside the complete documents.

Following the terminology introduced in the previous section, each document may be indexed by a number of patterns of different length and with different multiplicity. If it is assumed that patterns are effective descriptors for document indexing, the first step of document indexing consists in the automatic computation of the patterns of each document. As previously mentioned, relevant features which are usually taken into account are IOI, PIT, and BTH. Pattern computation can be carried out with a ad-hoc algorithms that compute exhaustively all the possible patterns, and store them in a hash table.

An exhaustive pattern discovery approach highlights a high number of patterns that have little or no musical meaning; for instance, a pattern that is repeated only two or three times in a document is likely to be computed by chance just because the combination of features is repeated in some notes combinations. Moreover, some patterns related to scales, repeated notes, or similar musical gestures, are likely to appear in almost all documents and hence to be poor discriminants among documents. In general, the degree by which a pattern is a good index may vary depending on the pattern and on the document. This is a typical situation of textual information retrieval, where words may describe a document to a different extent. For this reason it is proposed to apply the classical $tf \cdot idf$ weighting scheme.

The extent by which a pattern describes a document is the result of the multiplication of two terms. The **term frequency** is the number of occurrences of a given pattern inside a document. Hence, the term frequency of pattern p for document d can be computed as

$$tf_p^d = \text{\# occurrences of } p \in d$$

The **inverse document frequency** takes into account the number of different documents in which a patters appears. The inverse document frequency of pattern p can be computed as

$$idf_p = -log \frac{\text{\# documents containing } p}{\text{\# documents}}$$

Relevant patterns of a document may have a high tf – they are frequent inside the document – and/or a high idf – they are infrequent across the collection.

For the aims of indexing, a document is described by a sparse array, where each element is associated to a different pattern in the collection. The value of each element is given by the $tf \cdot idf$ value. The index is built as an inverted file, where each term of the vocabulary is a different pattern in a given notation (i.e., a text string). Each entry in the inverted file corresponds to a different pattern, and can efficiently be computed in an expected time $O(1)$ with an hashing function. Given the different sets of features, three inverted files are built, respectively for features IOI, PIT, and BTH. Inverted files can be efficiently stored in memory, eventually using compression, and fast accessed at retrieval time. The size of the inverted file and the implementation of the hashing function depend on the number of different patterns of the complete collection.

It may be useful to fix the maximum allowable pattern length to improve indexing. In fact, it is likely that very long patterns are due to repetitions of complete themes in the score and taking into account also them will give a quite sparse inverted file. Moreover, it is unlikely that a user will query the system singing a complete theme. These considerations suggest that long patterns could be truncated when they are over a given threshold.

6.4.5.1 Query Processing

For the query processing step, it can be assumed that users interact with the system according to a query-by-example paradigm. In particular, users should be able to describe their information needs by singing (humming or whistling), playing, or editing with a simple interface a short excerpt of the melody that they have in mind. Pitch tracking can be applied to the user's query in order to obtain a transcription in a notation format, such as a string of notes. The string representing the translated query needs to undergo further processing, in order to extract a number of descriptors that can be used to match the query with potentially relevant documents. It is normally assumed that a query is likely to contain strings that characterize the searched document, either because they appear very often inside its theme or because they are peculiar of that particular melody. In other words, a query is likely to contain relevant patterns of the searched document, which may have a high tf and/or idf.

The automatic detection of relevant strings cannot be carried out through pattern analysis, because normally queries are too short to have repetitions and hence to contain patterns. A simple approach to extract relevant strings, or potential patterns, from a query consists in computing all its possible substrings. That is, from a query of length q notes are automatically extracted $q - 2$ strings of three notes, plus $q - 3$ strings of four notes, and so on until the maximum allowable length for a pattern is reached. This approach can be considered similar to query expansion in textual information retrieval, which is known to increase recall at the risk of lowering precision. On the other hand, it is expected that most of the arbitrary strings of a query will never form a relevant pattern inside the collection, and then the negative effects on precision could be bounded.

6.4.5.2 Ranking Relevant Documents

At retrieval time, the strings are automatically extracted from the query and matched with the patterns of each document. The computation of potentially relevant documents can be carried out computing the distance between the vector of strings representing the query and the vector of patterns representing each document. Hence, for each document a Retrieval Status Value (RSV) is calculated, the higher the RSV, the closer the document with the query. A rank list of potentially relevant documents is computed from RSVs, obtaining a different rank lists for each of features used.

In general the orderings of documents in the rank lists differ. Differences may be due to many factors, as the diverse importance of rhythm and melodic profile for a the document collection, the

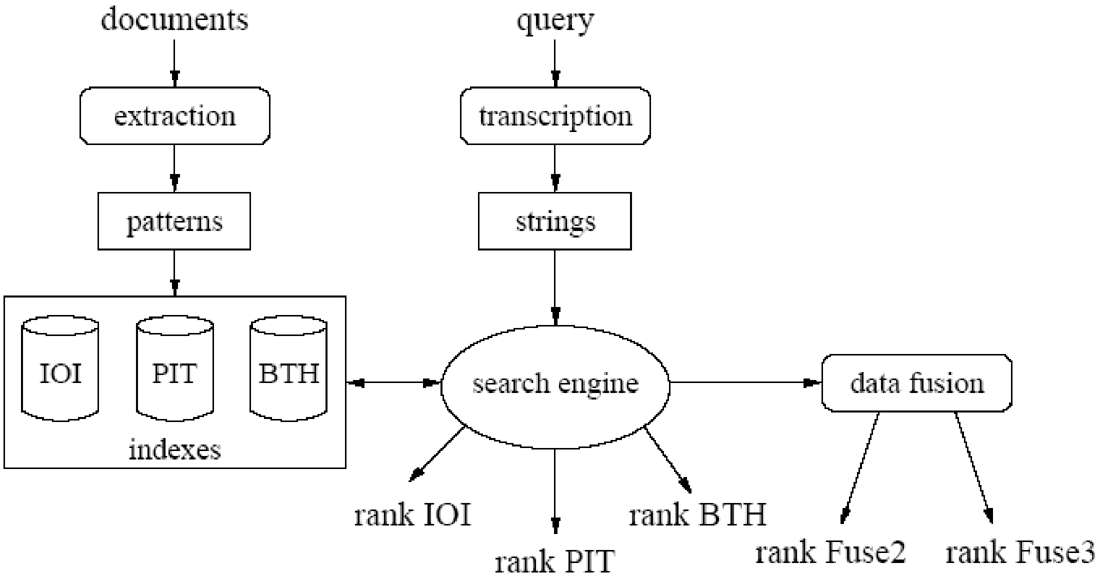


Figure 6.34: The phases of a methodology for MIR: Indexing, retrieval, and data fusion

effect of errors in the query, the kind of melodic excerpt chosen by the user as a representative of his information needs. It is expected that BTH ranking will give high scoring to the relevant documents when the query is sufficiently long and correctly played, because BTH patterns are a closer representation of the original melody. On the other hand, IOI and PIT are robust to query errors in melodic profile and rhythm, respectively. Moreover, simple representations as IOI and PIT are expected to be less sensitive to query length because of the possible presence of subpatterns of relevant motifs.

It is possible to take advantage from the existence of different rank lists by fusing together the results, in order to give the user a single rank list which takes into account the results of the three parallel approaches. This is a typical problem of **data fusion**, an approach that is usually carried out in the research area of Meta Search Engines, where the results obtained by different indexing and retrieval methodologies are combined – or fused – together according to a predefined weighting scheme. Since the RSVs of individual search engines are not known, or not comparable with others, the classical approach to data fusion is based on the information of rank only. In the case of MIR based on parallel features, the fusion can be carried out directly using the RSVs, because they are all based on the same $tf \cdot idf$ scheme. A new RSV can be computed as a weighted sum of RSVs of single features obtaining a new rank list.

A complete methodology for MIR shown in Figure 6.34, where steps undertaken at indexing time are shown on the left, while the operations that are performed at retrieval time are shown on the right. From Figure 6.34 and the above discussion, it is clear that the computational complexity depends on the query length – i.e., the number of strings that are computed from the query – while it is scalable on the number of documents. This is an important characteristic given by indexing techniques, because the time needed to reply to a query can be reasonably low also for large collections of documents.

6.4.5.3 Measures for Performances of MIR Systems

The output of almost any information retrieval system, and this applies also to MIR, is a ranked list of potentially relevant documents. It is clear that only the final user can judge if the retrieved documents are really relevant to his information needs. That is, the user should evaluate system performances in terms of retrieval effectiveness. There are two main reasons why the user may not be satisfied by the result of an information retrieval system.

- the system does not retrieve documents that are relevant for the user information needs – which is usually called **silence effect**;
- the system retrieves documents that are not relevant for the user information needs – which is usually called **noise effect**

All real systems for MIR try to balance these two negative effects. From the one hand, a high silence effect may result in not retrieving all the music documents that are similar to a given query sung by the user. From the other hand, a high noise effect may cause the user to spend great part of a retrieval session in listening to irrelevant documents.

Even if user satisfaction plays a central role in the evaluation of performances of a MIR system, and in general of any IR system, user studies are very expensive and time consuming. For this reason, the IR research community usually carries out automatic evaluation of the proposed systems using commonly accepted measures. In particular, there are two measures that are connected to the concepts of silence and noise effects. The first measure is **recall**, which is related to the ability of a system to retrieve the highest percentage of relevant documents (thus minimizing the silence effect). Recall is defined as

$$\text{recall} = \frac{\text{\# relevant retrieved}}{\text{\# total relevant}}$$

that is the number of relevant documents retrieved by the system divided by the total number of relevant documents in the complete database of documents. The second measure is **precision**, which is related to the ability of the system of retrieving the lowest percentage of irrelevant documents (thus minimizing the noise effect). Precision is defined as

$$\text{precision} = \frac{\text{\# relevant retrieved}}{\text{\# total retrieved}}$$

that is the number of relevant documents retrieved by the system divided by the total number of retrieved documents. An ideal system retrieved only relevant documents, and hence has 100% recall and precision. For real systems, high precision is usually achieved at the cost of low recall and viceversa.

Both precision and recall do not take into account that a MIR system may output a rank list of documents. For this reason it is a common practice to compute these measures also for the first N documents (for $N \in \{5, 10, 20, \dots\}$) and, in particular, to compute the precision at given levels of recall. Another approach is to summarize these measures, and the effect of the documents rank, in a single measure. For instance, the **average precision** is computed as the mean of the different precisions computed each time a new relevant document is observed in the rank list.

The evaluation of MIR systems is usually carried out on a test collection according to the Cranfield model for information retrieval, which is used at the Text REtrieval Conference (TREC). A test collection consists in a set of documents, a set of queries, and a set of relevance judgments that match documents to queries. The creation of a common background for evaluation is still an open issue in the MIR community, hence each research group created its own test collection from scratch. A "good" test collection should be representative of real documents and, in particular, of real user's queries. The size of the document set, as well as the way queries are collected, may deeply influence the evaluation results. Relevance judgments should be normally given by a pool of experts in the music domain, which is an expensive task, but they can also be automatically constructed when queries are in the form of excerpts of a known tune. In this latter case, only the document from which the query derives is considered as relevant.

6.4.5.4 An Example of Experimental Evaluation

In the following paragraphs, the result of an experimental evaluation of a running MIR system are reported. The system is based on pattern analysis, based on three alternative features (IOI, PIT, and BTH) and data fusion techniques applied to the combination of IOI and PIT, called Fuse2, and the combination of all the three features, called Fuse3.

The Test Collection A small test collection of popular music has been created using 107 Beatles' song in MIDI format downloaded from the Web. As for any test collection, documents may contain errors. In a preprocessing step, the channels containing the melody have been extracted automatically and the note durations have been normalized; in case of polyphonic scores, the highest pitch has been chosen as part of the melody. After preprocessing, the collection contained 107 complete melodies with an average length of 244 notes, ranging from 89 of the shortest melody to 564 of the longest. Even if a number of approaches for performing automatic theme extraction has been already proposed in the literature, the methodology relies on indexing of complete melodies, because repetitions of choruses and verses can be taken into account by the $tf \cdot idf$ measure.

A set of 40 queries has been created by randomly selecting 20 themes in the dataset and using the first notes of the chorus and of the refrain. The initial note and the length of each query were chosen to have recognizable motifs that could be considered representative of real users' queries. The queries had an average length of 9.75 notes, ranging from 4 to 21 notes. Only the theme from which the query was taken was considered as relevant. Using this initial set of correct queries, an alternative set has been created by adding errors on pitch, duration, and both, obtaining a new set of 120 queries. A simple error model has been applied, because errors were uniformly distributed along the notes in the queries, with a probability of about 13.3%. As for many approaches to approximate string matching, an error can be considered the result of a deletion and an insertion, thus these alternative sources of errors have not been explicitly modelled. Tests on robustness to query length were carried out by automatically shortening the initial queries by an increasing percentage, disregarding the fact that query would not sound musical. In this way, 160 more queries with decreasing length have been automatically generated. For all the modified queries, only the theme of initial query was considered as relevant. In the following, we will refer to the only relevant document with the term *r-doc* for all the experiments.

Truncation of Patterns All the experimental analyses, whose results are shown in the following sections, have been carried out after truncating patterns longer than a given threshold t. When a pattern $[f_1 \dots f_n]$ had a length of $n > t$, it has been replaced (in the indexing step) by all its subpatterns of exact length t, that is the $n - t + 1$ subpatterns $[f_1 \dots f_t]$, $[f_2 \dots f_{t+1}]$, and so on until $[f_{n-t} \dots f_n]$, where some of the subpatterns may be already extracted, because they were part of other motifs.

With the aim of computing the optimal threshold for the test collection, five different thresholds have been tested, respectively 5, 7, 10, 15, and 20 notes. The retrieval effectiveness decreased with high values of the threshold, meaning that a compact representation of patterns can be more effective than longer ones. The average precision was approximately constant when thresholds higher than $15 - 20$ notes were applied, probably because the number of different patterns longer than 20 notes is less than 8% and with a low value of r. The use of short patterns can be a useful way to control the increase of the index when new documents are added to the collection. Due to simple combinatorial reasons, the number of different patterns is bounded by the pattern length; on the other hand, the use of short patterns has the drawback of a higher number of patterns that are in common among documents, which may lower precision. It is interesting to note that data fusion approaches gave consistently better results than single approaches. This behaviour has been found in all our experiments, which are presented in the following sections, where results are shown only for $t = 5$.

Retrieval Effectiveness The first detailed analysis regarded the retrieval effectiveness with the set of 40 correct queries. Results are shown in Table 6.5, where the average precision (Av.Prec.), the percentage queries that gave the r-doc within the first k positions (with $k \in \{1, 3, 5, 10\}$), and the ones that did not give the r-doc at all ("not found"), are reported as representative measures. As it can be seen, IOI gave the poorest results, even if for 90% of the queries the r-doc were among the first three retrieved. The highest average precision using a single feature was obtained by BTH, with the drawback of an on-off behaviour: either the r-doc is the first retrieved or it is not retrieved at all (2.5% of the queries). PIT gave good results, with all the queries that found the r-doc among the first three documents.

	IOI	PIT	BTH	Fuse2	Fuse3
Av.Prec.	0.74	0.93	0.98	0.96	0.98
$= 1$	57.5	87.5	97.5	92.5	95.0
≤ 3	90.0	100	97.5	100	100
≤ 5	95.0	100	97.5	100	100
≤ 10	97.5	100	97.5	100	100
not found	0	0	2.5	0	0

Table 6.5: Retrieval effectiveness for correct queries

The best results for Fuse2 and Fuse3 have been obtained assigning equal weights to the single ranks. When the $tf \cdot idf$ scores had different weights an improvement was still observed in respect to single rankings, though to a minor extent. For this reason, results for Fuse2 and Fuse3 are presented only when equal weights are assigned.

Robustness to Errors in the Queries Users are likely to express their information needs in an imprecise manner. The query-by-example paradigm is error prone because the example provided by the user is normally an approximation of the real information need. In particular, when the user is asked to sing an excerpt of the searched document, errors can be due to imprecise recall of the melody, problems in tuning, tempo fluctuations, and in general all the problems that untrained singers have. Moreover, transcription algorithms may introduce additional errors in pitch detection and in melody segmentation. The robustness to errors has been tested on an experimental setup. Since indexing is carried out on melodic contour and on rhythm patterns, the errors that may affect the retrieval effectiveness regard the presence of notes with a wrong pitch and a wrong duration. As previously mentioned, a set of queries with automatically added errors has been generated in order to test the robustness of the approach in a controlled environment.

As expected, the performances of IOI dropped for queries with errors in rhythm and the same applied to PIT for queries with errors in pitch. The same considerations apply to BTH in both cases, with an even bigger drop in the performances. It is interesting to note that data fusion allowed for compensating the decreases in performances of single ranks, giving for both Fuse2 and Fuse3 an average precision equal to the one obtained without errors. In the case of errors in both pitch and rhythm, also Fuse2 and Fuse3 had a decrease in performances, even if their average precision was consistently higher than the one of single features.

The experimental results showed that Fuse3 gave a considerable improvement in respect to the single rankings contribution. A query-by-query analysis showed that this behaviour is due to the fact that the sum of $tf \cdot idf$ scores of the single features gave always a new ranking where the r-doc was at the same level of the best of the three separate ranks; that is, if one of the three gave the r-doc as the most relevant document, also Fuse3 had the r-doc in first position. Moreover, for some queries, the fused rank gave the r-doc at first position even if none of the three single ranks had the r-doc as the most relevant document. These improvements can be explained by two factors: First, when the r-doc was retrieved at top position by one of the features, it had a very high $tf \cdot idf$ score that gave an important contribution to the final rank; Second, the r-doc was often retrieved with a high rank by two or three of the features, while in general other documents were not considered as relevant by more than one feature. Similar considerations apply, though at a minor extent, also to Fuse2.

Dependency to Query Length A final analysis has been carried out on the effects of query length to the retrieval effectiveness. It is known that users of search engines do not express their information needs using much information. The community of information retrieval had to face the problems of finding relevant information also with vague or short queries. To some extent, a similar problem applies to MIR because users may not remember long excerpts of the music documents they are looking for. Moreover, untrained singers may not like to sing for a long time a song that they probably do not know very well. The effects of query length on a MIR system should then be investigated.

Tests on the dependency to query length have been carried out on a set of queries that were obtained from the original set of queries by shortening the number of notes from 90% to 60% of their original lengths. With this approach, queries may become very short, for instance a query of two notes cannot retrieve any document because patterns shorter than three notes are not taken into account.

Consistently with previous results, Fuse3 gave the best performances and showed a higher robustness to decrease in query length. Also in this case results showed that the data fusion approach was enough robust to changes in the initial queries. As previously mentioned, each initial query has been created selecting a number of notes that allowed to recognize the theme by a human listener. Moreover, each query was made by one or more musical phrases – or musical gestures or motifs – considering that a user would not stop singing his query at any note, but would end the query in a position that have a “musical sense”. For this reason, tests on query length can give only a general indication on possible changes in retrieval effectiveness.

6.4.6 Conclusions

This section present a short overview on some aspects of music IR. In particular, the issues typical of the music language have been discussed, taking into account the problems of formats and the role of the user. A number of approaches that have been proposed in the literature are presented, in particular the ones related to music Digital Libraries.

There are a number of aspects that are beyond the scope of this overview. In particular, all the research work related to audio processing that, even if not central to music IR, plays an important role in creating tools for classification of audio files and automatic extraction of low level features, that may be useful for expert users.

6.5 Commented bibliography

The reference book for Auditory scene analysis is Bregman [1990]. The Implication realization model is described in Narmour [1990]. The Local Boundary Detection algorithm is presented in Cambouropoulos [2001]. The Generative Theory of Tonal Music is described in Lerdahl and Jackendoff [1983].

Research on automatic metadata extraction for MIR can be classified in two main fields, depending on the two different classes of formats in which a music document can be represented: the automatic extraction of relevant information from a music score, which is typically achieved through melody segmentation and indexing; the automatic categorization of a music recording, which is typically achieved through audio classification. In this chapter we deal with the first field.

In the case of melody segmentation and indexing, the main assumption is that it is not possible to use textual descriptors for music documents, in particular for compositions and for melodies. Since it is not clear what kind of meaning is conveyed by a music document, the common approach is to describe a document using perceptually relevant elements, that may be in the same form of the document itself (that is the only way to describe music is through music). Clearly, the alternative description of a music document should be more compact and summarize the most relevant information, at least from a perceptual point of view. The music language may be characterized by different dimensions, which may regard the score representation ? e.g., melody, harmony, rhythm ? the recording of performances ? e.g., timbre, instrumentation ? and high level information ? e.g., structure, musical form. Among the different dimensions, melody seems to be the most suitable for describing music documents. First of all, users are likely to remember and use, in a query-by-example paradigm, parts of the melody of the song they are looking for. Moreover, most of the dimensions require a good knowledge of music theory to be effectively used, reducing the number of potential users to scholars, composers, and musicians. Finally, melody can benefit from tools for string analysis and processing to extract relevant metadata. For these reasons, most of the research work on metadata extraction focused on melody segmentation and processing. The need for automatic melody processing for extracting relevant information to be used as alternative descriptors, arises from the fact that the melody is a continuous flow of events. Even though listeners perceive the presence of elements in the melodic flow, which may be called lexical units, there is no explicit separator to highlight boundaries between them. Moreover, it is well known that there are parts of the melody ? e.g., the incipit, the theme, the leit-motiv, and so on ? that are more relevant descriptors of a music document than others. Yet, the automatic labelling of these relevant parts needs ad-hoc techniques.

One of the first works, probably the most cited in the early literature on MIR, is Ghias et al. [1995]. In this paper it is proposed the use of a query-by-example paradigm, with the aim of retrieving the documents that are more similar to the melody excerpts sung by the user: both documents and queries are transformed in a different notation that is related to the melodic profile. An alternative approach to MIR is proposed in Blackburn and DeRoure [1998], where metadata is automatic computed and stored in a parallel database. Metadata is in the form of hyperlinks between documents that are judged similar by the system.

Music language is quite different from other media, because it is not clear if music conveys a meaning and how a music document can be effectively described; this mostly because perception plays a crucial role in the way users can describe music. The important issue of perception is faced in Uitdenbogerd and Zobel [1998], where a user study is presented on users? melody representation. The knowledge of music structure is exploited in Melucci and Orio [1999] for extracting relevant information, where music documents and queries are described by surrogates made of a textual description of musical lexical units. Experiments on normalization are also reported, in order to cope with variants in musical lexical units that may describe similar documents. In Bainbridge et al. [1999] is proposed a multimodal description of music documents, which encompasses the audio, a visual representation of the score, the eventual lyrics, and other metadata that are automatically extracted from files in MIDI format.

An alternative approach to automatically compute melodic descriptors of music documents is presented in Bainbridge et al. [1999], which is based on the use of N-grams as musical lexical units. Alternatively, musically relevant phrases are proposed in Melucci and Orio [2000], where an hypertextual structure is automatically created among documents and musical phrases. In this case a document is described by a set of links to similar documents and to its most relevant phrases. Musical structure is exploited in Hoos et al. [2001] for computing a set of relevant features from a music document in a complex notation format.

Alternatively to previous works, in Birmingham et al. [2001] it is proposed that a good descriptor of a music document is its set of main themes, which are units longer than N-grams or musical phrases. Themes are modelled through the use of Markov chains. An extension to hidden Markov models is presented in Shifrin et al. [2002], where possible mismatches between the representation of the query and of the documents are explicitly modelled by emission probabilities of Hidden Markov Models states. An evaluation of different approaches is presented in Hu and Dannenberg [2002], where the problem of efficiency is raised and discussed.

References

D. Bainbridge, C.G. Nevill-Manning, I.H. Witten, L.A. Smith, and McNab R.J. Musical information retrieval using melodic surface. In *Proc. International Symposium on Music Information Retrieval*, pages 161–169, 1999.

W.P. Birmingham, R.B. Dannenberg, G.H. Wakefield, M. Bartsch, D. Bykowski, D. Mazzoni, C. Meek, M. Mellody, and W. Rand. Musart: Music retrieval via aural queries. In *Proc. International Symposium on Music Information Retrieval*, pages 73–82, 2001.

S. Blackburn and D. DeRoure. A tool for content based navigation of music. In *Proc. ACM Multimedia Conference*, pages 361–368, 1998.

A. S. Bregman. *Auditory Scene Analysis*. MIT Press, 1990.

E. Cambouropoulos. The local boundary detection model (lbdm) and its application in the study of expressive timing. In *Proc. Int. Computer Music Conf.*, 2001.

A. Ghias, J. Logan, D. Chamberlin, and B.C. Smith. Query by humming: Musical information retrieval in an audio database. In *Proceedings of ACM Digital Libraries (DL) Conference*, pages 231–236, 1995.

H.H. Hoos, K. Renz, and M. Gorg. GUIDO/MIR - an experimental musical information retrieval system based on guido music notation. In *Proc. International Symposium on Music Information Retrieval*, pages 41–50, 2001.

N. Hu and R.B. Dannenberg. A comparison of melodic database retrieval techniques using sung queries. In *Proc. ACM/IEEE Joint Conference on Digital Libraries*, pages 301–307, 2002.

F. Lerdahl and R. Jackendoff. *A Generative Theory of Tonal Music*. The MIT Press, 1983.

M. Melucci and N. Orio. Musical information retrieval using melodic surface. In *Proc. 4th ACM Conference on Digital Libraries*, pages 152–160, 1999.

M. Melucci and N. Orio. Smile: a system for content-based musical s information retrieval environments. In *Proc. Intelligent Multimedia Information Retrieval Systems and Management (RIAO) Conference*, pages 1246–1260, 2000.

Eugene Narmour. *The Analysis and cognition of basic melodic structures : the implication-realization model*. University of Chicago Press, 1990.

J. Shifrin, B. Pardo, C. Meek, and W. Birmingham. Hmm-based musical query retrieval. In *Proc. ACM/IEEE Joint Conference on Digital Libraries*, pages 295–300, 2002.

A. Uitdenbogerd and J. Zobel. Manipulation of music for melody matching. In *Proc. ACM Multimedia Conference*, pages 235–240, 1998.

6.6 Appendix

6.6.1 Generative Theory of Tonal Music of Lerdahl and Jackendorf

Lerdahl and Jackendoff (1983) developed a model called Generative Theory of Tonal Music (GTTM). This model offers a complementary approach to understanding melodies, based on a hierarchical structure of musical cognition. According to this theory music is built from an inventory of notes and a set of rules. The rules assemble notes into a sequence and organize them into hierarchical structures of music cognition. To understand a piece of music means to assemble these mental structures as we listen to the piece.

It seeks to elucidate a number of perceptual characteristics of tonal music - segmentation, periodicity, differential degrees of importance being accorded to the components of a musical passage or work, the flow of tension and relaxation as a work unfolds - by employing four distinct analytical levels, each with its own more-or-less formal analytical principles, or production rules. These production rules, or Well-Formedness rules, specify which analytical structures may be formed - which analytical structures are possible - in each of the four analytical domains on the basis of a given musical score. Each domain also has a set of Preference Rules, which select between the possible analytical structures so as to achieve a single ”preferred” analysis within each domain.

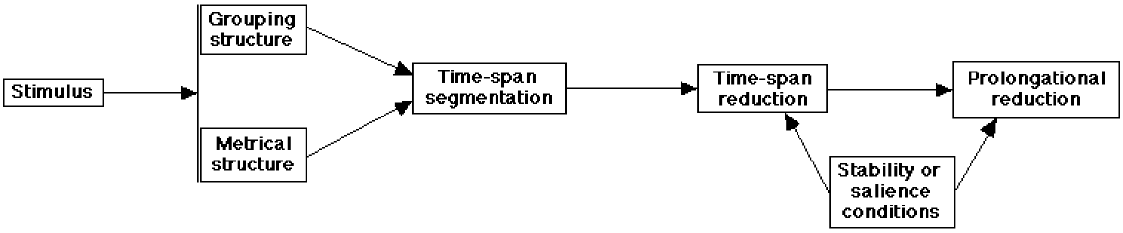


Figure 6.35: Main components of Lerdahl and Jackendoff’s generative theory of tonal music.

GTTM proposes four types of hierarchical structures associated with a piece: the grouping structure, the metrical structure, the time-span reduction structure, and the prolongational reduction structure (fig. 6.35).

The grouping structure describes the segmentation units that listeners can establish when hearing a musical surface: motives, phrases, and sections.

The metrical structure describes the rhythm hierarchy of the piece. It assign a weight to each note depending on the beat in which is played . In this way notes played on strong (down) beats have higher weight than notes played on week (up) beats.

The time-span reduction structure is a hierarchical structure describing the relative structural importance of notes within the audible rhythmic units of a phrase (see Fig. 6.36). It differentiate the essential parts of the melody from the ornaments. The essential parts are further dissected into even more essential parts and ornament on them. The reduction continues until the melody is reduced to a skeleton of the few most prominent notes.

The prolongational reduction structure is a hierarchical structure describing tension-relaxation relationships among groups of notes. This structure captures the sense of musical flow across phrases, i.e. the build-up and release of tension within longer and longer passages of the piece, until a feeling of maximum repose at the end of the piece. tension builds up as the melody departs from more stable notes to less stable ones and is discharged when the melody returns to stable notes. tension and release are also felt as a result of moving from dissonant chords to consonant ones, from non accented notes to accented ones and from higher to lower notes.

The four domains - Metrical, Grouping, Time-Span and Prolongational - are conceived of as partially interdependent and at the same time as modelling different aspects of a listener's musical intuitions.

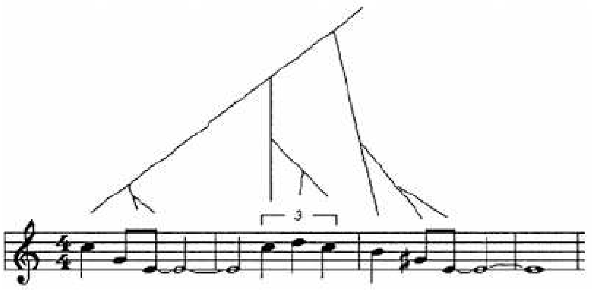


Figure 6.36: Example of a time-span tree for the beginning of the All of me ballad [from Arcos 1997].

Each of these four components consists of three sets of rules:

Well-formedness Rules which state what sort of structural descriptions are possible. These rules define a class of possible structural descriptions.

Preference Rules which try to select from the possible structures the ones that correspond to what an experienced listener would hear. They are designed to work together to isolate those structural descriptions in the set defined by the well-formedness rules that best describe how an expert listener interprets the passage given to the theory as input.

Transformational Rules that allow certain distortions of the strict structures prescribed by the well-formedness rules.

The application of their theory to the first four bars of the second movement of Mozart's K.311 is shown in fig. 6.37 and 6.38. The Metrical analysis (shown in the dots below the piece in Figure 6.37) appears self-evident, deriving from Well-Formedness Rules such as those stating that "Every attack point must be associated with a beat at the smallest metrical level present at that point in the piece"

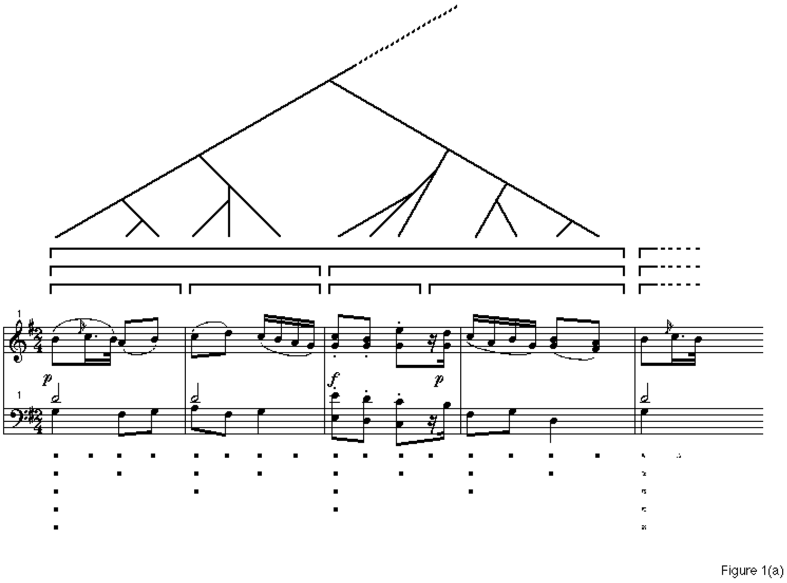


Figure 6.37: Example of GTTM analysis of the first four bars of the second movement of Mozart's K.311: Metrical analysis (dots below the piece) and Time-Span analysis (tree-structure above the piece) [from Cross 1998].

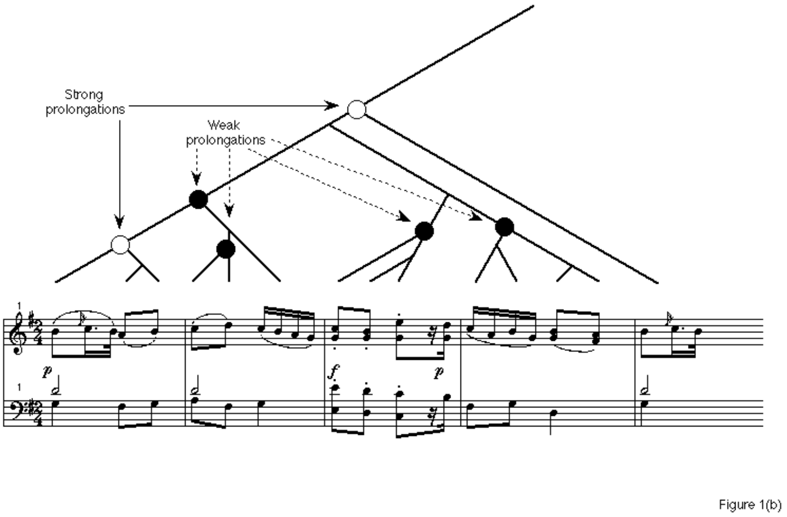


Figure 6.38: Example of GTTM analysis of the first four bars of the second movement of Mozart's K.311: Prolongational analysis [from Cross 1998].

(although the lowest, semiquaver, level is not shown in the figure), "At each metrical level, strong beats are spaced either two or three beats apart", etc. These Well-Formedness rules are supplemented by Preference rules, that suggest preference should be given to e.g., "metrical structures in which the strongest beat in a group appears relatively early in the group", "metrical structures in which strong beats coincide with pitch events”, etc.

The Grouping structure (shown in the brackets above the piece in Figure 6.37) appears similarly self-evident, being based on seemingly truistic Well-Formedness rules such as ”A piece constitutes a group”, ”If a group contains a smaller group it must contain all of that smaller group” (thus ensuring a strictly nested hierarchy), etc. Preference rules here specify such matters as the criteria for determining group boundaries (which should occur at points of disjunction in the domains of pitch and time), conditions for inferring repetition in the grouping structure, etc. Thus a group boundary is formed between the end of bar two and the beginning of bar three both in order to ensure the symmetrical subdivision of the first four bars (themselves specifiable as a group in part because of the repetition of the opening of bar one in bar five) and because the pitch disjunction occurring between the G and the C is the largest pitch interval that has occurred in the upper voice of the piece up to that moment. Perhaps the only point of interest in the Grouping analysis is the boundary between the third quaver of bar three and the last semiquaver of that bar, brought about by the temporal interval between the two events (again, the largest that has occurred in the piece up to that moment). Here, the Grouping structure and the Metrical structure are not congruent, pointing-up a moment of tension at the level of the musical surface that is only resolved by the start of the next group at bar five.

The Time-Span analysis (tree-structure above the piece in Figure 6.37) is intended to depict the relative salience or importance of events within and across groups. The Grouping structure serves as the substrate for the Time-Span analysis, the Well-Formedness rules in this domain being largely concerned with formalising the relations between Groups and Time-Spans. The Preference rules suggest that metrically and harmonically stable events should be selected as the ”heads” of Time-Spans, employment of these criteria resulting in the straightforward structure shown in the Figure. This shows clearly the shift in metrical position of the most significant event in each Group or Time-Span, from downbeat in bar one to upbeat crotchet in bars two and three to upbeat quaver in bar four.

A similar structure is evident in the Prolongational analysis (Figure 6.38), which illustrates the building-up and release of tension as a tonal piece unfolds. The Prolongational analysis derives in part from the Time-Span analysis, but is primarily predicated on harmonic relations, which the Well-Formedness and Preference rules specify as either prolongations (tension-producing or maintaining) or progressions (tension-releasing).

Lerdahl and Jackendoff’s theory however lack of a detailed, formal account of tonal-harmonic relations and tend to neglect of the temporality of musical experience. Moreover it let the analyst to make different choices that are quite difficult to formalize and implement on a computational model. Although the authors attempt to be thorough and formal throughout the theory, they do not resolve much of the ambiguity that exists through the application of the preference rules. There is little or no ranking of these rules to say which should be preferred over others and this detracts from what was presented as a formal theory.

6.6.2 Narmour’s implication realization model

[1]

An intuition shared by many people is that appreciating music has to do with expectation. That is, what we have already heard builds expectations on what is to come. These expectations can be fulfilled or not by what is to come. If fulfilled, the listener feels satisfied. If not, the listener is surprised or even disappointed. Based on this observation, Narmour proposed a theory of perception

[1] adapted from Mantaras AI Magazine 2001

and cognition of melodies based on a set of basic grouping structures, the Implication/Realization model, or I/R model.

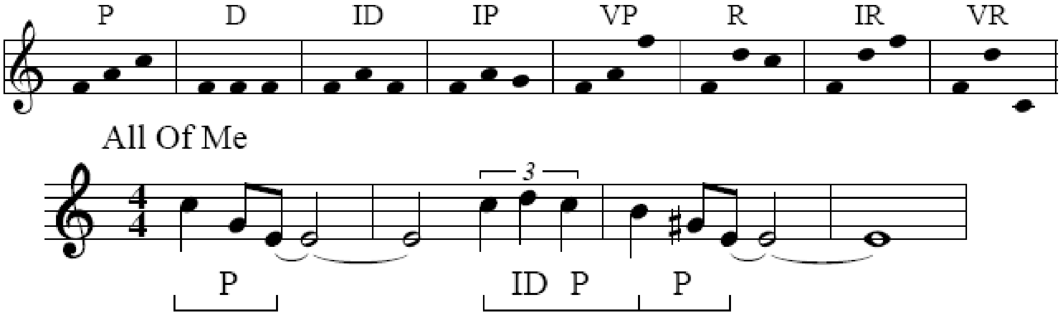


Figure 6.39: Top: Eight of the basic structures of the I/R model. Bottom: First measures of All of Me, annotated with I/R structures.

According to this theory, the perception of a melody continuously causes listeners to generate expectations of how the melody will continue. The sources of those expectations are two-fold: both innate and learned. The innate sources are hard-wired into our brain and peripheral nervous system, according to Narmour, whereas learned factors are due to exposure to music as a cultural phenomenon, and familiarity with musical styles and pieces in particular.

The innate expectation mechanism is closely related to the gestalt theory for visual perception. Narmour claims that similar principles hold for the perception of melodic sequences. In his theory, these principles take the form of implications: Any two consecutively perceived notes constitute a melodic interval, and if this interval is not conceived as complete, or closed, it is an implicative interval, an interval that implies a subsequent interval with certain characteristics. In other words, some notes are more likely to follow the two heard notes than others. Two main principles concern registral direction and intervallic difference.

- The principle of registral direction states that small intervals imply an interval in the same registral direction (a small upward interval implies another upward interval, and analogous for downward intervals), and large intervals imply a change in registral direction (a large upward interval implies another upward interval and analogous for downward intervals).

- The principle of intervallic difference states that a small (five semitones or less) interval implies a similarly-sized interval (plus or minus 2 semitones), and a large intervals (seven semitones or more) implies a smaller interval.

Based on these two principles, melodic patterns can be identified that either satisfy or violate the implication as predicted by the principles. Such patterns are called structures and labelled to denote characteristics in terms of registral direction and intervallic difference. Eight such structures are shown in figure 6.39(top). For example, the P structure (Process) is a small interval followed by another small interval (of similar size), thus satisfying both the registral direction principle and the intervallic difference principle. Similarly the IP (Intervallic Process) structure satisfies intervallic difference, but violates registral direction.

Additional principles are assumed to hold, one of which concerns closure, which states that the implication of an interval is inhibited when a melody changes in direction, or when a small interval is followed by a large interval. Other factors also determine closure, like metrical position (strong metrical positions contribute to closure, rhythm (notes with a long duration contribute to closure), and harmony (resolution of dissonance into consonance contributes to closure).

These structures characterize patterns of melodic implications (or expectation) that constitute the basic units of the listener perception. Other resources such as duration and rhythmic patterns emphasize or inhibit the perception of these melodic implications. The use of the implication-realization model provides a musical analysis of the melodic surface of the piece.

The basic grouping structure are shown in fig. 6.39:

P (process) structure a pattern composed of a sequence of at least three notes with similar intervallic distances and the same registral direction;

ID (intervallic duplication) structure a pattern composed of a sequence of three notes with the same intervallic distances and different registral direction;

D (duplication) structure a repetition of at least three notes;

IP (intervallic process) structure a pattern composed of a sequence of three notes with similar intervallic distances and different registral direction;

R (reversal) structure a pattern composed of a sequence of three notes with different registral direction; the first interval is a leap, and the second is a step;

IR (intervallic reversal) structure a pattern composed of a sequence of three notes with the same registral direction; the first interval is a leap, and the second is a step;

VR (registral reversal) structure a pattern composed of a sequence of three notes with different registral direction; both intervals are leaps.

In fig. 6.39 (bottom) the first three notes form a P structure, the next three notes an ID, and the last three notes another P. The two P structures in the figure have a descending registral direction, and in both cases, there is a duration cumulation (the last note is significantly longer).

Looking at melodic grouping in this way, we can see how each pith interval implies the next. Thus, an interval can be continued with a similar one (such as P or ID or IP or VR) or reversed with a dissimilar one. That is, a step (small interval) is followed by a leap (large interval) between notes in the same direction would be a reversal of the implied interval (another step was expected, but instead, a leap is heard) but not a reversal of direction. Pitch motion can also be continued by moving in the same direction (up or down) or reversed by moving in the opposite direction. The strongest kind of reversal involves both a reversal of intervals and of direction. When several small intervals (steps) move consistently in the same direction, they strongly imply continuation in the same direction with similar intervals. If a leap occurs instead of a step, it creates a continuity gap, which triggers the expectation that the gap should be filled in. To fill it, the next step intervals should move in the opposite direction from the leap, which also tends to limit pitch range and keeps melodies moving back toward a centre.

Basically, continuity (satisfying the expectation) is nonclosural and progressive, whereas reversal of implication (not satisfying the expectation) is closural and segmentative. A long note duration after reversal of implication usually confirm phrase closure.

Any given melody can be described by a sequence of Narmour structures. Fig. 6.40 Narmour's analysis of the first four bars of the second movement of K.311 is shows. Letters (IP, P, etc.) within the ”grouping” brackets identify the patterns involved, while the b's and d's in parentheses above

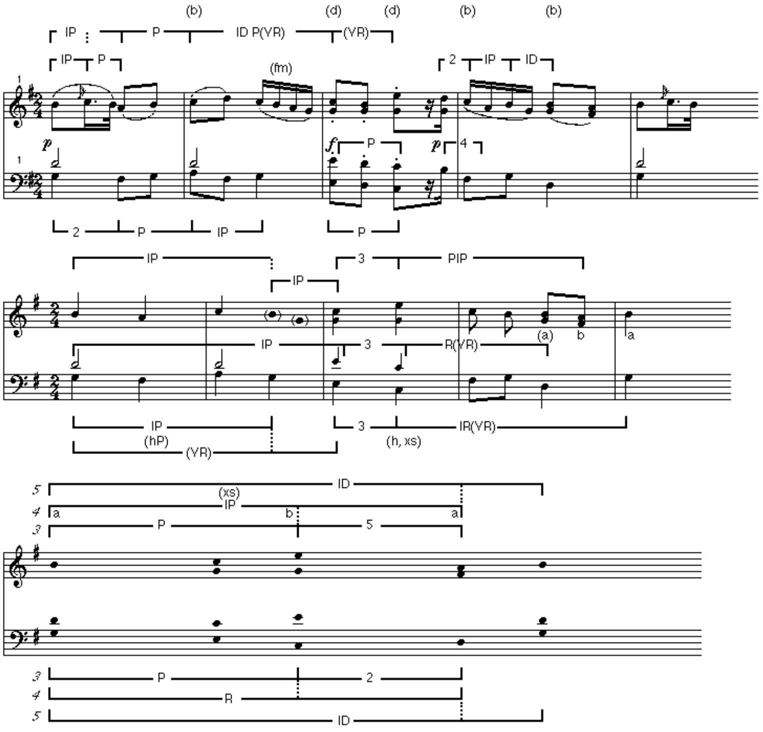


Figure 6.40: Example of Narmour analysis of the first four bars of the second movement of Mozart's K.311 [from Cross 1998].

the top system indicate the influence of, respectively, metre and duration. The three systems show the progressive ”transformation” of pitches to higher hierarchical levels, and it should be noted that the steps involved do not produce a neatly nested hierarchy of the sort that Lerdahl and Jackendoff's theory provides.

Contents