Chapter 7

Expressiveness in music performance

Giovanni De Poli

7.1 The quest for expressiveness

During the last decade, lot of research effort has been spent to connect two worlds that seemed to be very distant or even antithetic: machines and emotions. Mainly in the framework of human-computer interaction an increasing interest grew up in finding ways to allow machines communicating expressive, emotional content. Such interest has been justified with the objective of an enhanced interaction between humans and machines exploiting communication channels that are typical of human-human communication and that can therefore be easier and less frustrating for users, and in particular for non technically skilled users.

Starting from the findings from psychology and neurosciences, research has been aimed at developing computational models and algorithms for analysis and synthesis of emotional content.

While from the one hand research on emotional communication found its way into more traditional fields of computer science like Artificial Intelligence, on the other hand novel fields developed explicitly focusing on such issues.

Examples are researches on Affective Computing in the United States, KANSEI Information Processing in Japan and Expressive information processing in Europe. In this section [1] Affective Computing and KANSEI Information Processing are shortly described with reference to the work of the two researchers that in a certain way started the two fields: Rosalind Picard and her group at MIT Media Lab for Affective Computing, and Shuji Hashimoto and his group at Waseda University, Tokyo, for KANSEI Information Processing. In the following sections, analysis and synthesis of expressive content in performing arts (a typical European research stream), with a particular reference to music performance, is presented.

[1] adapted from PhD dissertation of Gualtiero Volpe (2003)

7.1.1 Affective Computing: the American way to artificial emotions

The Affective Computing approach is mainly illustrated in the homonymous book (Picard, 1997).

In her book Picard defines Affective Computing as computing that relates to, arises from, or deliberately influences emotions. Affective Computing addresses the design and implementation of machines that are able to

- recognize emotions,
- express emotions,
- have emotions.

These are human-centred machines that observe their users and sensitively interact with them by expressing emotions depending on what they observed and on the current emotional state of the machine.

- Computers that are able to *recognize* emotions are conceived as systems collecting a variety of input signals ranging from face expressions to voice, movement features (e.g., hand gestures, gait, posture), physiologic measures (e.g., respiration, electrocardiogram, blood pressure, temperature). They perform feature extraction and classification on these inputs (e.g., video analysis of movement, audio analysis of speech) and try to classify the emotion the user is communicating through a reasoning process taking into account information about context, situations, personal goals, social display rules, and other emotion related data. Learning techniques can be employed to adapt recognition to a specific user (e.g., a personal computer can learn the habits of its master to improve its performances in the recognition task). If the computer has an emotional state, this can influence the recognition process.

- Computer that are able to *express emotions* (either depending on instructions given by humans or as a result of an internal mechanism for generating emotions) are systems that modulate audio (e.g., synthetic voice, sound, music) and visual signals (e.g., face, posture, gait of animated creatures, colours) in a way suitable for the emotion that has to be communicated. The expressed emotion can be intentional (i.e., deliberated as a result of a reasoning process) or spontaneous (i.e., reactively triggered). It can directly express the affective state of the machine that can in turn be influenced by the expression of the emotion. Expression partially depends on social display rules.

- If computers *can have* emotions is perhaps one of the most controversial issues in Affective Computing. In her book, Picard proposes to consider five components of an emotional system: a computer can be said to have emotions if all five components are present in it.

  The five components are the following:

  i. Emergent emotions and emotional behaviour i.e. the machine is able to express an emotion through its behaviour even if it does not have any emotion. By observing the machines behaviour, humans naturally tend to attribute an emotional state to the machine.

  ii. Fast primary emotions i.e. mechanisms to generate a kind of hard-wired, reactive responses (especially to potentially harmful events). Fast primary emotions are what Damasio calls primary emotions (Damasio, 1994). Studies about the mechanisms triggering such emotions can be found in neurosciences. They are associated with the inner regions of the brain.

iii. Cognitively generated emotions i.e. emotions that are generated as a result of explicit reasoning. Cognitively generated emotions are slower than fast primary emotions and are usually consequence of deliberate thoughts. They are located in the brain cortex. Several cognitive models of emotion have been developed. One of the most famous is the model by Ortony, Clore, and Collins, usually referred as OCC model (Ortony, Clore, and Collins, 1988) that has been also employed in a number of concrete applications. Originally, the OCC model was not developed for building machines that could have emotions; rather it was conceived as a way for reasoning about emotions. The model develops a collection of rules associating emotions to cognitive evaluations about consequences of events, actions of agents, and aspects of objects.

iv. Emotional experience i.e. the system is cognitively aware of its emotional state. Emotional experience consists of cognitive awareness, physiologic awareness and subjective feelings. If it is possible to have such an emotional experience in a machine and, if yes, how it can be implemented is still an open and quite tricky issue. It relates to consciousness and requires the machine to have sensors able to measure its own emotional state.

v. Body-mind interactions i.e. the emotional state can influence other processes simulating similar human physical and cognitive functions like memory, perception, decision making, learning, goals, motivations, interest, planning, etc.

Research on Affective Computing has been applied in a number of application scenarios, ranging from entertainment, to edutainment, to detection of emotional responses (e.g., frustration) in particular relevant tasks (e.g., learning, driving), to the design and implementation of devices for analysis and synthesis of emotions. Detailed descriptions of ongoing and past research projects can be found in the website of the Affective Computing group at MIT media lab (`http://affect.media.mit.edu/`).

With respect to the three issues mentioned above (i.e., machines recognizing, expressing, and having emotions), we will mainly address the first two aspects, i.e. the design and implementation of algorithms for recognizing and communicating expressive content, rather than with machines that *have* a their own emotional state. In fact, if the goal is to open novel perspective to artistic performances by introducing new tools allowing an extension of the artistic languages by acting on the communicated expressive content through technology, what is mainly needed is

- the possibility to classify and encode in digital format the communicated expressive content in order to process it,
- the ability to produce suitable output to induce emotional reactions in spectators.

In other words, we believe that humans only have emotions. Machines do not need to have them, but they can give more and better support to human activities if they are able to process information not only related to the rational aspects of human behaviour, but also to the emotional ones.

7.1.2 The eastern approach: KANSEI Information Processing

In the same period the Affective Computing research started in the United States, another approach to understanding expressive content communication was developed in Japan: KANSEI Information Processing.

According the Japanese view (Hashimoto, 1997) information processing has three phases:

Physical information processing: physical signals capturing data from the real world (e.g., sound, light, force) are identified as the first target of information processing. Signal processing is the technology field that is mainly responsible of processing such kind of information.

Semantic information processing: The second phase is the semantic information processing to deal with knowledge and rule, that is the field of logic and symbolic knowledge. Artificial Intelligence is the discipline that mainly covers such aspects.

KANSEI information processing: The third target is KANSEI (a Japanese word) that refers to feelings, intuition, and sympathy and according to Hashimoto we are just entering in an historical period in which technology will start to deal with KANSEI, an issue that in the past was often left as a research field for only humanistic or humanistic related disciplines.

The exact meaning of the Japanese word KANSEI is something controversial for western people: it does not have a univocal correspondent in western languages and culture, but is rather associated to a collection of words related to the emotional sphere (e.g., emotion, sensibility, sensuality, sense, feeling). In his paper Hashimoto gives some examples of common uses of the word in Japanese language such as for example Her KANSEI is excellent, He is a man of rich KANSEI, He has no KANSEI, Her KANSEI seems well suited to me, etc. It should be noticed that KANSEI refers to a dynamic process rather than to emotional labels or categories to be applied to expressive contents.

KANSEI Information Processing can be regarded as a coding and decoding process. In other words, KANSEI Information Processing supposes an underlying model in which expressive content is conceived as a kind of high-level information that, in the framework of a human-human communication process, *modulates* the physical signals carrying some usually symbolic message. That is, when a (human) sender sends a message to a (human) receiver he/she encodes in the message some expressive emotional information. Such information together with the symbolic content is embedded in the physical signal carrying the message. When the receiver receives the signal he/she decodes it and extracts both the symbolic message and the additional expressive information the sender encoded into it. Notice that it is not required that the sender deliberately add the expressive information to the message: such additional expressive information can be included unconsciously and can refer to aspects such as personality traits or personal dispositions toward objects, actions, and other people.

By making a comparison with the Affective Computing approach, it can be noticed that all the three aspects of recognizing, expressing, and having emotions are included in the KANSEI process: in fact,

- the sender expresses his/her emotions by encoding them in the physical signals carrying a message,
- the receiver recognizes the emotions expressed by the sender while decoding the message carried by the physical signals,
- sender and receiver have an emotional state that can both influence the encoding/decoding process and be itself the high-level additional expressive information encoded in a message.

KANSEI Information Processing seems therefore to adopt an holistic approach, broader with respect to the Affective Computing perspective because it includes in the same model of encoding/decoding process all the three aspect Affective Computing separately deals with, and because, while Affective Computing is more concerned with emotions, KANSEI rather refers to a wide collection of emotion related aspects (e.g., moods, feelings, personality traits etc.). This difference may reflect a cultural

difference between western and eastern approaches to problem solving: while western people usually tend to divide a problem in sub-problems following a top-down approach and sometime losing the global perspective, eastern people often continue to keep an overall view of the problem even when they are focusing on a specific aspect of it.

7.2 Models, expressiveness and music performance

Music is an important mean of communication where three actors participate: the composer, the performer and the listener. The composer instils into his works his own emotions, feelings and sensations, and the performer communicates them to the listeners. The composer describes his/her musical ideas by a score or a process. The information contained in the score (or produced by the process) has a double function: a descriptive one, as a symbolic representation of the cognitive elements constituting the composition, and a functional one, as a mean to convey instructions to the performer. Other information is implicit in the score and regards performance style and interpretative conventions. The performer interprets these symbols, taking into account the implicit information and his/her personal artistic feeling and aim, and produces the sounds by using a musical instrument. Music performance includes all the human activity that lies between the symbolic score and the music instrument

Music performance is an interesting topic to study for its multidisciplinary valence. In this chapter paradigms and issues emerged in research on modelling expressiveness in music performance will be reviewed and future research perspectives will be discussed. In the following we will discuss performance modelling approaches mainly from a information processing point of view. In this section we will present the basic issue on what models, and computational models, are for and we will discuss expression communication in music performance. In the next section 7.3 we will introduce the aspect of how musical information is represented for modelling purposes. Finally in section 7.4 the main strategies used in model development will be presented in detail. Models for understanding, performance synthesis and artistic creation will be discussed.

7.2.1 Models

Frequently in science, models are employed to evidence and abstract some relations that can be hypothesized, discarding details that are felt to be irrelevant for what is being observed and described. Models can be used to predict the behaviour in certain conditions and compare these results with observations. In this sense, they serve to generalize the findings and have both a descriptive and predictive value.

In the study of music performance, scientists have been developing models for the past few decades. The possibility, offered by advancing technology, of implementing the models and to experiment with their behaviour by simulation gave rise to an increased use of technology in music research. Moreover, computer science and music technology developed many conceptual frameworks and practical tools in the last few decades that are very useful for music performance investigation. For example artificial intelligence, knowledge engineering, soft computing methodologies, physics based models, MIDI instruments, signal processing analysis methods, computer controlled performance, motion capture devices, constitute paradigms and tools that are at the base of many performance models.

The idea of developing computational models of music performance dates back to the first music application of computers.

- The first models were mainly dedicated to music production and experimentation, and were

embedded in computer programs for music synthesis or representation and for interactive performance. Their theoretical assumptions and conceptual foundation were often not explicit. One such application is the Groove system that allowed real time control and editing of performer actions described (graphically or symbolically) by time functions.

- Later models for performance understanding started to be developed (e.g. KTH performance rule system, presented in section 7.6). Their aim is analytical, trying to explain why a performer acts in a certain way and which relation exists between a gesture and its musical effect.

Both kinds of models are based on theoretical concepts and share the idea that an artistic activity can be, at least partially, formalized. We can expect a convergence of efforts toward models that are oriented toward both performance understanding and production.

7.2.2 From mathematical models to information processing models

The classic approach to describe relations in models is by using mathematical expressions among observable (and often measurable) facts called variables or parameters. Developing and then validating mathematical models is the typical way to proceed in science and engineering. Often the variables are distinguished in input variables, supposedly known, and output variables, which are deduced by the model. In this case, inputs can be considered as the causes and output the effect of the phenomenon. A mathematical model can be implemented on a computer allowing to compute the values of output variables corresponding to the provided values of inputs. This process is called simulation and it is widely used to predict the behaviour of the phenomenon in different circumstances.

However, a computer does not only deal with numerical values. More generally, it can be considered as information processing engine. From this perspective, models describe relations among different kinds of information about the phenomenon. Thus, a fundamental problem in developing information processing models is to define which kind of information we want to deal with and how we may represent it on a computer.

The case of music performance is quite interesting; in fact, the information that can be considered regards many aspects. We can distinguish three layers.

- The first is the *physical information* that can be measured, as timing or performer's movements. This information can be represented as numbers and is typically used and processed by mathematical tools.

- The second layer is the *symbolic information* as the score, where the notes are represented by symbols in the common music notation. These symbols refer more to a cognitive organization of the music than to an exact physical value. For example, the duration symbol indicates a division of the meter, while the actual duration of a performed note can vary. Processing at this level uses typical symbolic and logic representations of computer science.

- At a higher level, we have the *expressive information* more related to the affective and emotional content of the music. Recently computer science and engineering started paying attention to this level of information and developing suitable theories and processing tools.

Music and music performance in particular, attracted the interest of researchers for developing and testing such tools. Moreover in performance modelling, all the information levels should be taken into account in a coordinated way. As a consequence, information representation and model structure are crucial topics in model design and will be discussed in section 7.3.

7.2.3 Expressiveness in music performance

The communication of expressive content by music can be studied at three different levels: considering the composer's message, the expressive intentions of the performer, and the listener's perceptual experience. Studies of the first kind are historically more developed. Generally, they analyze the elements of the musical structure and the musical phrasing that are critical for a correct interpretation of composer's message.

The contribution of the performer to expression communication has two facets: to clarify the composer's message enlightening the musical structure and to add his personal interpretation of the piece. A mechanical performance of a score is perceived as lacking of musical meaning and is considered dull and inexpressive as a text read without any prosodic inflexion. Indeed, human performers never respect tempo, timing and loudness notations in a mechanical way when they play a score: some deviations are always introduced, even if the performer explicitly wants to play mechanically.

Thus in general expressiveness refers both to the means used by the performer to convey the composer's message and to his own contribution to enrich the musical message. However, many music performance studies concentrate on the first aspect, trying to understand the performer actions to better convey the musical structure. Simulation models are often evaluated by the musical acceptability of their results, or in other words how well a supposed ideal interpretation of that particular piece is approached. Expressiveness related to the musical structure may depend on the dramatic narrative developed by the performer, on physical and motor constraints or problems (e.g. fingering), on stylistic expectation based on cultural norm (e.g. jazz vs. classic music) and actual performance situation (e.g. audience engagement). Figure 7.1 shows the relation between dynamics profiles and the main elements of music structure of the first measures of a piano performance of Mozart sonata K 545 (figure 7.2). It is particularly evident that the musician emphasized with a decrescendo the end of the first melodic unit (bar 2), the first semi-phrase (bar 4), the first phrase (bar 8) and the period (bar 16).

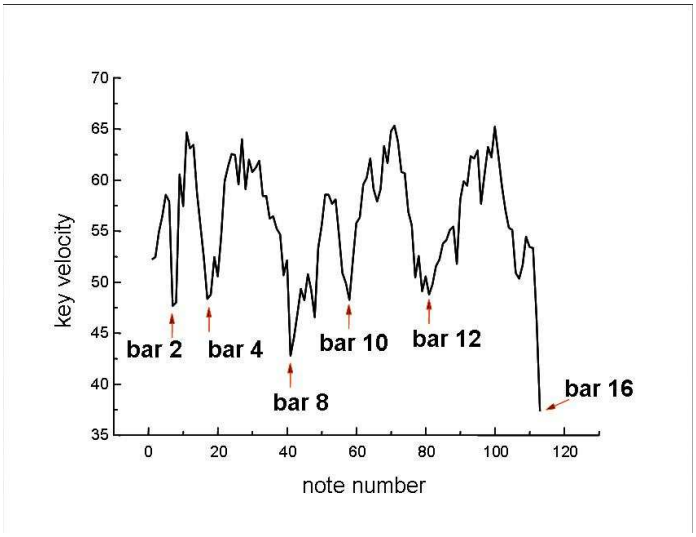


Figure 7.1: Dynamics profiles and the main elements of music structure of the first measures of a piano performance of Mozart sonata K 545 (figure 7.2). It is particularly evident that the musician emphasized with a decrescendo the end of the first inciso (bar 2), the first semi-phrase (bar 4), the first phrase (bar 8) and the period (bar 16).

Recently interest is also growing in taking into account the expression component added by the

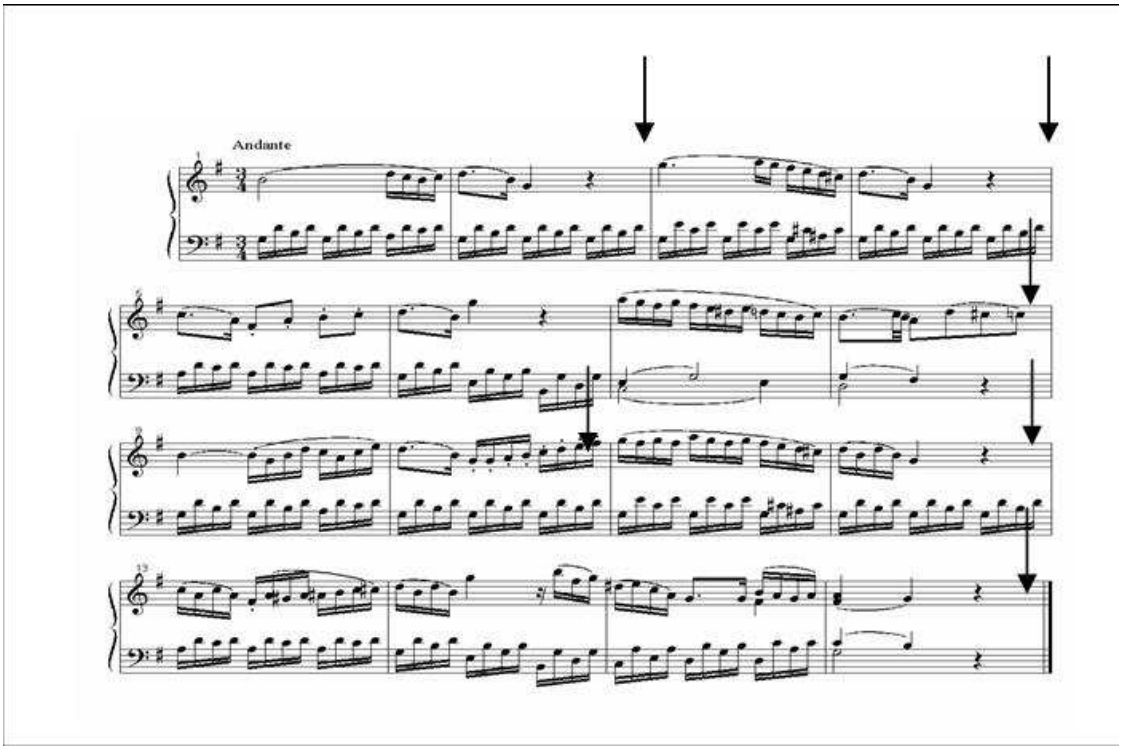


Figure 7.2: Score of the first 16 measures Mozart sonata K 545. The arrows indicate the end of the first inciso (bar 2), the first semi-phrase (bar 4), the first phrase (bar 8) and the period (bar 16).

performer. Some aspects are still strongly related to the musical piece, as performer specific style, and influences of stylistic expectation based on cultural norm (e.g. jazz vs. classic music) or actual performance situation (e.g. audience engagement). Nevertheless, other communicative aspects can be taken into account. Experiments are carried out by asking performers to play the same piece according diverse specific adjectives or nuances or trying to convey different content. The researcher then seeks to understand and model the strategies used in these performances. Often basic emotions are chosen as possible expressions (see section 7.8). and in this case the term expressive performance refers to emotional performance. Notice that sometimes emotions the performer tries to convey can be in contrast with the character of the musical piece. A slightly broader interpretation of expression as KANSEI (Japanese term indicating sensibility, feeling, sensitivity, see sect. 7.1.2) or affective communication see sect. 7.1.1) is proposed in some Japanese or American studies. We prefer the broader term ***expressive intention*** that include emotion, affect as well other sensorial and descriptive adjectives or actions. Furthermore, this term evidences the explicit intent of the performer in communicating expression.

Understanding of specific artistic intentions of top-level performers is more challenging. While artists aim to express aesthetic value, we feel that these qualities are probably impossible to model, without losing their real essence.

7.3 Information and music performance

7.3.1 Expressive performance information

When we want to develop an information processing model, it is important to define which is the relevant information we will use. This choice depends on the phenomenon we are observing and on the available detection techniques. In our case, we want to describe music performance and we can observe the variations a music performer is doing when he plays. This kind of information is often called expressive performance information. The most relevant information used in performance models are discussed in this section.

7.3.1.1 Representation levels

Physical information level At a physical information level, the main expressive parameters, considered in the models, are related to timing of musical events and tempo, dynamics (loudness variation), and articulation (the way successive notes are connected). These parameters are particularly relevant for keyboard instruments. Moreover, they are the basic parameters of the MIDI protocol and thus are easily measurable on electronic music instruments or employable for obtaining a music performance. In some instruments and in the singing voice other acoustic parameters are taken into account such as vibrato and micro-intonation or pedalling at the piano. In contemporary music, timbre is often an essential expressive parameter; sometimes also virtual space location or movement of the sound source is used as expression feature.

These parameters can be measured directly by a MIDI musical instrument or (with more effort) by detecting the performer movements. However, it should be noticed that these measurements depends on an accurate instrument calibration. In fact, the relation of MIDI command with their sonic realization depends greatly on the instrument. Moreover, the Note-off command indicates the beginning of the sound decay and not the ending of the note, as often it would be desired.

Physical information can also be gathered from audio recordings. Additional expressive parameters can be taken into account, such as timbre. However, parameters are more difficult to collect automatically, especially for multi voice music, and depend on the recording conditions. Different methods are often used and thus the measures, reported in the literature, may be not directly comparable. This fact contributes to make the accumulation of knowledge hard. For instance, it is not always clear when exactly a tone exactly starts nor when the attack phase can be considered completed. The amplitude envelope inspection is not sufficient. Therefore, the attack duration of a note can be measured in different ways, leading to dissimilar values. On the other hand, in real time applications we need effective, but not too complex feature-analysis algorithms. It is advisable that the progress of computational analysis techniques should provide useful and standardized tools for performance parameter detection.

The interrelation of these physical parameters is not well understood. Therefore, models often try to separate the parameters and to model their effect separately or to deal with a combination of very few of them. The problem is particularly evident when we want to model some effects that can be rendered in different ways. For example, the performer can emphasize a note by increasing its loudness, or by lengthening its duration or by a slight time shift, or by a particular articulation or timbre modification.

The use of more abstract representations could probably help in separating the low-level features from higher-level ones. This approach would call for multilevel models or a combination of models acting at different abstraction levels. For instance, in the previous example, a model can decide that a

note should be emphasized because of its structural importance and a second model will decide how to realize the emphasis taking into account the context, the expressive resources of the instrument, stylistic expectations etc.. While the performer probably uses such multilevel strategies intuitively in his/her musical practice, a precise definition of intermediate parameters, effective for modelling purpose, is still partial. More research is needed for the selection of these intermediate parameters, for finding a possible quantification, and for assessing their effectiveness.

Symbolic information level As regards symbolic information, the score is a typical reference and it is usually represented as a list of time events. More difficult is the representation of the musical structure. The knowledge is only partially formalized, especially toward classical music. Very few computational models were proposed for automatic (or semiautomatic) structure extraction from the score and their results are not very reliable. Thus the segmentation and the structure is often introduced by hand. The classic paradigm derives from early language modelling and consists in musical grammars represented as a hierarchical tree structure (e.g. phrase, sub phrase, melodic gesture, note). For example the musical period of 16 bars extracted from Mozart sonata K 545 (figure 7.2) can be subdivided into two phrases eight bar long, each phrase into two sub-phrases 4 bar long, sub-phrase 1 and 3 are composed by two incisos two bar long. This structure is shown in figure 7.3, where structural elements are representend as rectangles and the notes as circles. This paradigm is much less appropriate for contemporary music, where other musical parameters and constructs are more pertinent. Music performance research will greatly benefit from theoretic advancements on contemporary music analysis.

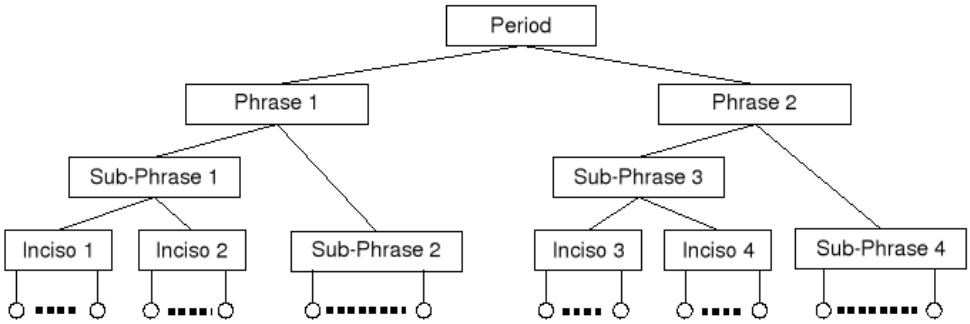


Figure 7.3: Tree representing the hierarchical structure of the first 16 bars extracted from Mozart sonata K 545, shown in figure 7.2: rectangles represent structural elements, circles represent notes.

Expressive information level The understanding of the expressive information is still vague. While its importance is generally acknowledged, the basic constituents are less clear. Often the simple range expressive-inexpressive is used. The most frequently used paradigms, for representing emotions in music performance modelling, are the categorical and the dimensional approach (e.g. valence-arousal space), see section 7.8. In the categorical approach, affect semantics in music has be studied by allowing a large number of listeners to use adjectives (either on a completely free basis, or taken from an elaborate list) to specify the affective content of musical excerpts. Afterwards, the data is analysed and clustered into categories. In this case expression information is represented by label indicating the expression category and eventually a number or adjective indicating the degree of that expression.

In the dimensional approach, there seems to be a considerable agreement about two fundamental dimensions of musical affect processing, namely Valence and Arousal. Valence is about positively or negatively valued affects, while Arousal is about the force of these affects. A third dimension is often noticed, but its meaning is less clearly specified. The dimensional approach was also used with success for other kinds of expressive intentions (see e.g. sect. 7.7. The expressive information is thus represented as a point on a two or three dimensional space, where the interpretation of the coordinates depends on the space used.

In this field too, more research and experimental insight will be very fruitful. On the other hand continuous measurements of subject reactions during a performance, recently used in psychological research, may provide useful data and parameters for performance research.

7.3.1.2 Expressive deviations

Most studies of performance expressiveness aim at understanding the systematic presence of deviations from the musical notation as a communication means between musician and listener. Deviations introduced by technical constraints (such as fingering) or by imperfect performer skill, are not normally considered part of expression communication and thus are often filtered out as noise. Deviations considered in models normally refers to the expressive performance information as discussed above.

The analysis of these systematic deviations has led to the formulation of several models that try to describe their structure, with the aim to explain where, how and why a performer modifies, sometimes unconsciously, what is indicated by the notation in the score. It should be noticed that, although deviations are only the external surface of something deeper and often not directly accessible, they are quite easily measurable, and thus widely used to develop computational models in scientific research and generative models for musical applications.

Reference for computing deviations When we talk of deviation, it is important to define which is the reference used for computing deviation. Different solutions were proposed and the choice depends on the problem we are dealing with.

- Very often the score is taken as reference, both for theoretical (the score represents the music structure) and practical (it is easily available) purposes (see e.g. the KTH model in sect. 7.6. However, the use of a score as reference has some drawbacks for the interpretation of how listeners judge expressiveness.

- Alternative approaches are the intrinsic definitions of expression (expressive deviations defined in terms of the performance itself) or non-structural approaches relating expression to motion, emotion, etc.. The idea is that, from the structural description of a music piece, we can individuate units which can act as a reference at that level. Its subunits will act as atomic parts whose internal detail will be ignored. Then expression is intended as the deviation from the norm as given by a higher level unit. For example, the expressive variations of the durations of beats are expressed in reference (as ratio) of the bar duration. An example of this approach is the hierachical phrasing model of sect. 7.5. Using this intrinsic definition, expression can be extracted from the performance data itself, taking more global measurements as reference for local ones.

- When we studied how a performer plays a piece according to different expressive intentions, we found that a clearer interpretation and best results in simulation are obtainable by using a

neutral performance as reference (see section 7.7). We intend neutral in the sense of a human performance without any specific expressive intention.

- In other cases the mean performance (i.e. the mathematical mean among different performances, by the same or many performers) was taken as reference, when stylistic choices and preferences were investigated.

7.3.2 Event information representation

A key issue is how to represent the musical information. A common way to describe music is as a collection of separate events $EV[n]$, with the event index $n = 1, \cdots, N$. For instance a score can be described as a collection of notes and rests. Each event is characterised by a time reference and by a list of attributes (as pitch, loudness, duration, etc.). Frequently in music performance studies, the collection $EV[\cdot]$ is considered as a sequence of events, i.e. there is no overlap and the end of an event coincides with the beginning of the next event. In this case $n + 1$ is considered as the event index of the successive event, adjacent to n-th event.

A similar approach can be used to represent the hierarchical structure of music. In this case a piece is considered as a macro-event. The macro-event represents the abstract element of the musical structure (e.g. phrase etc.) and can be considered as composed (recursively) by a collection (or sequence) of macro-events or of events (notes and rests). Each macro-event is characterized by a time reference and by a set of attributes. For example in the tree representation of fig. 7.3 macro-events are represented by rectangles and events by circles.

Notice that while the note abstraction is the most common way to think of musical events, it is not the only one. Often sound is continuously varying and not easily sliceable in different notes. The concept of note typically refers to pitched sounds. In general a sound event is characterized by a certain acoustic or perceptual unit and by a beginning and an end.

7.3.2.1 Time information representation

We should distinguish the time reference at symbolic level from the one at the physical level.

- At symbolic level, the space metaphor is often used. Thus the symbolic time reference of the n-th event is called **score position** $x[n]$ and is measured in musical units (where one musical unit is equal to the whole note value) or in metrical units (beats or bars). The symbolic duration of an event is called **length**, while the interval between events is called **distance**. Both can be measured in musical or metrical units. For example in a 3/4 rhythm, the length of 1 bar is 3 beats (quarter note) long or 0.75 musical units (whole notes) long. The corresponding distance of the event at the beginning of a bar from the event at the beginning of the next bar is 3 beats (quarter note) wide or 0.75 musical units (whole notes) wide.

  The symbolic time reference can be absolute (as in the so called piano roll notation) or relative. In the last case the distance from a previous event is specified. For example in the staff notation the score position is not explicitly indicated and the event distance between two adjacent events is represented symbolically as note or rest values $NV[n]$. We can compute the score position of a sequence of symbolic events by

$$x[n + 1] = x[n] + NV[n].$$

Also for macro-event we can have an absolute or relative time reference. The macro-event total length can be computed from the length of the elements of the collection. In the case of a structure composed by sequences, it is just the sum of all the composing elements lengths.

It is possible to express the symbolic time in seconds taking the tempo marking in the score, or normalising the total score length to the performance duration. This representation is called **nominal time** t_{nom} or score time. In this case, the reference time is called nominal onset time $O_{nom}[n]$, event length is called nominal duration $DR_{nom}[n]$, and distance between two adjacent events is called nominal inter onset interval

$$IOI_{nom}[n] = O_{nom}[n+1] - O_{nom}[n]$$

If we call v the score length divided by the the performance duration, we have $t_{nom} = x/v$ and thus

$$O_{nom}[n] = x[n]/v$$

For example with a (allegro) metronome of $MM = 120$ quarters per minute, i.e. 2 quarters per second, a beat will lasts 0.5 seconds. A piece of 16 bars with rhythm 3/4 (e.g. a waltz) will have total length of 12 musical units and 48 beats. The performance played exactly at $MM = 100$ quarters per minute will lasts 28.8 seconds and $v = 12/28.8 = 0.417$ musical units per second or $v = 48/28.8 = 1.67$ beat per second.

- At physical level time is called **performance time** and it is measured in seconds. Performance time is simply represented by the symbol t or by t_p when there is the need to clearly distinguish it from nominal time. The reference time is called onset time $O[n]$. It represents the time of the actual event onset elapsed since the beginning of the piece. The event duration is represented by $DR[n]$, and the distance between two adjacent events is called inter onset interval

$$IOI[n] = O[n+1] - O[n]$$

Notice the difference between inter onset interval $IOI[n]$ and duration $DR[n]$. Inter onset interval refers to music as a sequence of notes and is in relation with the note values indicated in the score, while duration refers to the actual duration of the event and can be shorter or longer than its corresponding IOI. Duration is more related to the sound quality than to the music melodic and metric structure: in fact, notes can be played staccato or legato, greatly affecting their expressive character. Moreover this difference allows to not represent rests, by including their duration into the IOI of the preceding note event.

7.3.2.2 Event attributes

The physical level represents the knowledge on the musical performance as *physical events*. Every event is described by a time reference, called Onset time $O[n]$ normally measured in seconds, and a set of sound attributes. This level corresponds to the same level of abstraction of the MIDI representation of the performance, e.g. as obtained from a sequencer (MIDI list events). A similar event description can be obtained from an audio performance. Often considered sound attributes are: pitch value expressed as frequency $FR[n]$ or MIDI pitch $MP[n]$, Duration $DR[n]$, Intensity $I[n]$ or KeyVelocity $KV[n]$ for MIDI event description and a set of timbre-related parameters. Frequently used timbre parameters are: Brightness $BR[n]$ (measured as the centroid of the spectral envelope and

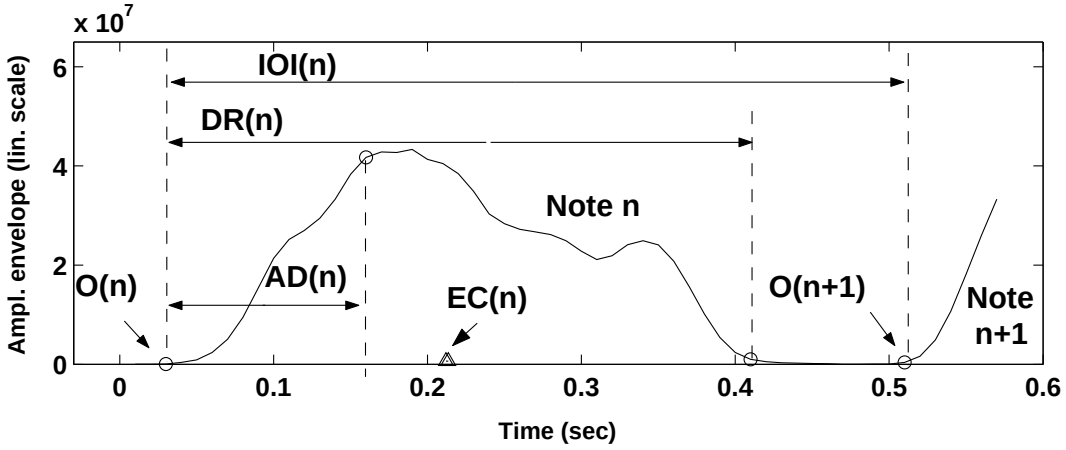


Figure 7.4: Musical parameters involved at event level.

energy envelope, described by Attack Duration $AD[n]$ and Envelope Centroid $EC[n]$ (i.e., the temporal centroid of the dynamic profile of the note). The difference between duration and inter onset interval is expressed by the Legato factor

$$L[n] = DR[n]/IOI[n]$$

which is often used in performance modelling. Figure 7.4 shows the principal parameters introduced. This representation can be obtained from the signal or the sound model representation by a semi-automatic segmentation.

We can figure out a similar representation of the macro-events, each one characterized by a time reference. In music performance modelling it is often convenient to distinguish, also for macro-events, the IOI from the actual duration. This distinction allows for a sort of legato modelling for the different sections of a musical piece, by inserting short micro-pauses to separate the different elements, analogous to punctuation in spoken speech.

7.3.2.3 Expressive timing information representation

The most important aspect is the representation of time. Time can be considered from both a physical and a symbolic point of view. The first one, performance time t, refers to the actual time that can be measured during a performance, while the second refers to the position in the score (e.g. phrase or measure) and it is often called score-time or score position x.

Models of expressive timing normally aim to describe the relation between performance time and score position expressed as $x = x(t)$ or $t = t(x)$. Performers adapt performance time of musical events in subtle way. Understanding models try to explain these variations, while synthesis models compute these variations.

Tempo An important aspect of time representation is tempo, often denoted as v, that is the speed of occurrence of the beats for a given metric structure. Tempo is the ratio between a distance (in beats) and its corresponding duration. Traditionally tempo is measured by a metronome (M.M.) number indicating the number of beats per minute (bpm) of performance time. We can derive tempo as beat per seconds (bps) by dividing the metronome indication by 60. E.g. an allegro tempo played at

MM=144 bpm corresponds to 2.4 bps. Notice that tempo refers to the slope of the representation of position in function of time. A distinction can be made among

- the mean tempo (i.e. the average tempo across the whole piece disregarding possible variations). It can be derived by
$$v_{mean} = \frac{x[N] - x[1]}{O[N] - O[1]}$$

- the main tempo (i.e. the prevailing tempo when passages with momentary variations such as slow start, final retard, fermatas, and amorphous caesuras are deleted);

- the basic tempo (i.e. the central tendency of the tempo over a complete musical excerpt, which is the implied tempo around which the local tempo varies, not necessarily symmetrically). It can be derived as the slope of the least square line approximating of the set of points $(O[i], x[i])$;

- the local tempo v_{loc}, which is maintained only for a short time;

- the event tempo, which is the scaling factor of the single event. It is defined at event level as the nominal length in the score divided by the inter-onset interval
$$v_{ev}[n] = \frac{x[n+1] - x[n]}{O[n+1] - O[n]} = \frac{NV[n]}{IOI[n]}$$
where $NV[n]$ is the symbolic event distance in beats.

- instantaneous tempo defined as
$$v(t) = \frac{d\,x(t)}{dt}$$
We can consider tempo as function of continue time or position, i.e. $v(t)$ or $v(x)$. Notice that particular attention should be payed when we apply it at event representation, which are intrinsically discrete.

When we study music performances, it is convenient to express performance time as function of score position, i.e. $t_{pf} = t_{pf}(x)$. In this way we can easily compare the timing of different performers for the same musical passage. The slope of this function represents durations over musical units and corresponds to the reciprocal of the velocity (tempo), as normally intended. The reciprocal of tempo, i.e. the time between two beats onsets, is called beat duration or beat period and indicates how long a beat lasts. It is measured in seconds per metrical or score unit. Related representations of reciprocal tempo are the measure duration (measured in seconds per measure) and relative inter onset interval $IOI_{rel}[n] = IOI[n]/NV[n]$, i.e. time difference between the next event and the actual event in performance time divided by the symbolic (score) note value. Notice that $IOI_{rel}[n] = 1/v_{ev}[n]$; i.e. it is the reciprocal of event tempo and is measured in seconds per musical or metrical unit. It is frequently used in time performance description.

Although it is still unclear what exactly constitutes the perception of tempo, it seems to be related - at least in metrical music - to the notion of beat or tactus: the speed at which the pulse of the music passes at a moderate rate (i.e. the metrical level at which one counts the beat).

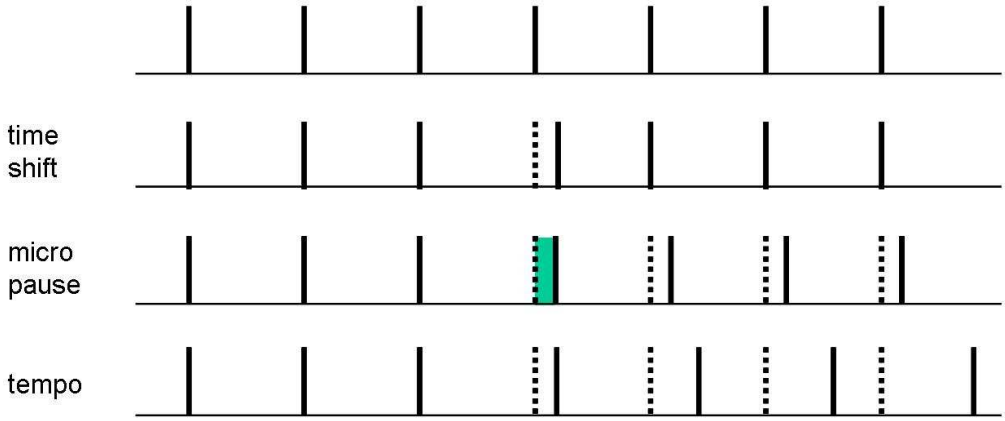


Figure 7.5: Comparison of time shift, micropause insertion and tempo variation.

Time shift In some musical situations or styles of music, where the global tempo is mostly constant, *time-shift* (or event-shift) $TS[n]$, defined as the time deviation of the observed performance onset time $O[n]$ in respect to the expected onset time $O_{exp}[n]$, offers a more natural way of representing timing. It results

$$TS[n] = O[n] - O_{exp}[n]$$

For example in a music played with a fixed beat or pulse in a constant tempo, the expected onset will be $O_{exp}[n] = O_{nom}[n]$. When the tempo is slowly varying, we can compute expected onset time as

$$O_{exp}[n] = O[n-1] + IOI_{exp}[n-1]$$

where $IOI_{exp}[n-1]$ is computed from the event distance $NV[n-1]$ by using the available local estimate of tempo v_{loc} as $IOI_{exp}[n-1] = NV[n-1]/v_{loc}$.

When the n-th event is time shifted by $TS[n]$, we have a characteristic pattern for IOIs. In fact $IOI[n-1]$ is lengthened by $TS[n]$, and $IOI[n]$ is shortened by $TS[n]$. The other IOIs are not affected.

Micro-pauses Sometimes we can observe in music performance timing, the presence of small breathing pauses, called micro-pauses, required occasionally to mark larger structural details of the piece. Micro-pauses are useful to indicate a temporary cessation of the pulse, such as at the end of a major section of a piece, or near the end of a work. It serves to make clear the large-scale structure of the work. These are generally not notated nor quantized - in more recent music they are indicated sometimes by a "comma". Their aim is analogous to punctuation in written text reading. While the presence of a time shift of the n-th event does not affect the onset time of the other events, the presence of a micro-pause after the n-th event will delay the onset time of all the following events. E.g. the insertion of a micro-pause of mp seconds after the n-th event will delay all $\tilde{O}[i] = O[i] + mp$ with $i > n$. A different representation could be by adding mp to $IOI[n]$ and recomputing the onset times. For this reason it is convenient to employ both time shift and micro-pause to describe local timing behaviour of a performance.

Discussion While tempo and timing (time shift and micro-pause) refers both to time values (fig. 7.5), they tend to be perceived somewhat independently by listeners. Thus, timing models should take

into account both aspects trying to separate them. Often expressive timing is considered as describing the timing deviations in a performance (e.g., accentuating notes by lengthening them for a bit, or playing notes after the beat). In addition, timing might be perceived independently of any changing tempo (tempo rubato). So it could be argued that expressive timing and expressive tempo possibly co-exist as two, relatively independent and perceptible aspects of a performance.

From this point of view, expressive timing is seen as a combination of a tempo component (expressing the change of rate over a fragment of music), and a timing (time-shift) component that describes how events are timed (e.g., early or late) with respect to this tempo description.

Music is an organization of events in time and often a hierarchical time structure can be envisaged. Therefore, models are developed for representing performance aspects at different time scales. We may have models at note scale, e.g. for attack time or vibrato, at local scale considering only few notes, e.g. articulation of a melodic gesture, or at a more global scale, e.g. for phrase crescendo. The most complete models deal with the different time scales by using distinct but coordinated strategies.

7.3.2.4 Information representation

Continue vs. discrete time representation. The musical information used in modelling can be represented as values or attributes of discrete time instants (musical events) such as notes or structural units. Alternatively they can be represented as profiles, i.e. as functions of continuous (performance or score) time. An example of discrete time representation is the articulation of timing of individual notes or the micro-pauses between melodic units; an example for continuous time representation is the vibrato of a note or a crescendo curve. The first representation is more related to the symbolic level, while the second one to the physical level. The choice depends on the aim of a model, on availability of data and on their ability to explain. Sometimes models combine both kinds of representations or are able to transform data from one to the other representation, e.g. by interpolation or sampling. For example a crescendo is a discrete parameter at the piano, but not at other instruments as e.g., the violin. Moreover it can be interpreted as continuous curve sampled at the note onsets.

Granularity: continue vs. discrete values Another aspect of the representation is the granularity. When possible the information is represented as numerical values. Sometimes absolute values, e.g. time interval in ms, sometimes relative values, e.g. relative inter onset interval, are used. In this case the inter onset intervals are represented as normalized to their score duration or at a certain metrical level, most often the beat level or the bar level. In this last way, the timing pattern becomes a local tempo indicator. In other situations, the information is categorical describing one choice among few alternatives, e.g. staccato vs. legato, shortening vs. lengthening. Even for granularity, the effectiveness of the representation depends on the problem we are dealing with and on the musical context. However, in symbolic representation of music often the concepts are not easily expressible as numbers or as precisely defined categories. A possibility of using effectively vague definitions is offered by the techniques of soft computing such as fuzzy sets.

7.4 Models of / for music performance

Models are developed with different aims. A basic difference is between models of music performance, i.e. models for understanding (also called analysis models) and models for music performance,

i.e. models able to produce music performances (also called synthesis models). In the following sections, the main paradigms will be presented and discussed.

7.4.1 Model structures

It is often convenient, in developing and using models, to break the problem into simpler parts, each one described and modelled by a proper strategy, and then combine everything into a larger unit. In the following, the principal way used to combine rules or models will be discussed.

- The first, and frequently used, strategy assumes that the partial results computed by sub-models can be added to obtain the final result. Let $x_1, x_2, ..., x_n$ be the inputs of the models and $y_j = f_j(x_1, x_2, ..., x_n)$ be the j-th sub-model, the additive model composition is given by

$$y = \sum_j f_j(x_1, x_2, ..., x_n)$$

  For example, the deviations computed by the KTH rule system (see section 7.6) are obtained by a weighted sum of the deviations computed by the single. Another application is when the final result is obtained as the sum of profiles at different time scale, e.g. the crescendo and accelerando curves computed for phrases and sub-phrases by Todd (see section 7.5). An application of this strategy in analysis is when the principal component analysis (PCA) of measured deviations on a musical passage is used to highlight differences among performing styles of different pianists [Repp 1992]. In fact PCA involves a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. The original data are thus expressed as a linear combination of (few) significant and independent variations around their mean value.

  The additivity hypothesis is attractive from both a mathematical and a practical point of view: it allows the use of many computational tools and it is easily interpretable. However, it may result in over- simplifying and tends to hide the interrelation of different aspects of performance.

- A partially different strategy for combining numerical values consists in multiplying the partial results.

$$y = \prod_j f_j(x_1, x_2, ..., x_n)$$

  It is often used when relative values are employed. Of course taking the logarithms will transform it in an additive strategy.

- More complex is the non linear combination of the sources $y_j = f_j(x_1, x_2, ..., x_n)$.

$$y = F[f_1(x_1, x_2, ..., x_n), \cdots, f_J(x_1, x_2, ..., x_n)]$$

  In this way the interrelations of inputs can taken into account. An example is the use of feed-forward neural networks as general approximators of observed performance deviations.

- Models are sometimes combined using the output of a model as input to a second one, i.e. by functional composition, as in cascade model that compute $y = f[g(x)]$. A typical example is timing function composition as discussed by Honing.

- A more general approach is in hierarchical models when they operate at different abstraction level. The information is processed and combined at the proper level. An example is the distinction of rules and metarules in the KTH system, where the metarules choose the proper setting of basic rules to express, for example, different emotion (see section 7.6).

From another point of view, we may distinguish local models, that acts at note level and try to explain the observed facts in a local context. A different perspective is assumed by phrasing models (see section 7.5) that take into account the higher level of the musical structure or more abstract expression pattern. The two approaches often require different modelling strategies and structures. In certain cases, it is possible to devise a combination of both approaches with the purpose being to obtain better results. The composed models are built by several components, each one aiming to represent the different sources of expression. However, a good combination of the different parts is still quite challenging.

Moreover we can distinguish two kinds of models, according their explanation aims.

- The complete model tries to explain all of the observed performance deviations on the basis of the given data. This approach tends to give very complex models and thus poor insight on the relevant relations. In fact, note level analysis cannot explain all the observed deviations.

- The partial model aims only to explain what can be explained at note level, giving a small and robust set of rules. Moreover when rules for categorical decisions (e.g. play faster of slower) rather than for computing an exact value are used, more understandable results can be obtained.

7.4.2 Comparing performances

A problem that normally arises in performance research is how performances can be compared. In subjective comparison often a supposed ideal performance is taken as reference by the evaluator. In other cases, an actual reference performance can be assumed. Of course subjects with different background can have dissimilar preferences that are not easily made explicit.

However when we consider computational models, objective numerical comparisons would be very appealing. In this case, performances are represented by a set of values. Sometimes the adopted strategies compare absolute or relative values. As measure of distance the mean of the absolute differences can be considered, or the Euclidean distance (square root of difference squares) or maximum distance (i.e. take the maximal difference component). It is not clear how to weight the components, nor which distance formulation is more effective. Different researchers employ different measures.

More basically it is not clear how to combine time and loudness distances for a comprehensive performance comparison. For instance as already discussed, the emphasis of a note can be obtained by lengthening, dynamic accent, time shift, timbre variation. Moreover, it is not clear how perception can be taken into account, nor how to model subjective preferences. How are subjective and objective comparisons related? The availability of good and agreed methods for performance comparison would be very welcome in performance research. A subjective assessment of objective comparison is needed. More research effort on this direction is advisable.

As an example of a possible definition of a norm for comparing performances, if we make reference to MIDI-like representation, we can assume that relevant parameters for comparison are inter onset interval $IOI[n]$, note duration $DR[n]$ and intensity $I[n]$. Given two performances of a piece composed by a sequence of N notes, we can define the differences of the three parameter as $\Delta IOI[n] = IOI_1[n] - IOI_2[n]$, $\Delta DR[n] = DR_1[n] - DR_2[n]$ and $\Delta I[n] = I_1[n] - I_2[n]$. The

performance distance can be expressed by a weighted Euclidean norm as

$$dist = \sqrt{\sum_{n=1}^{N-1} w_{IOI,n} \Delta IOI^2[n] + \sum_{n=1}^{N} w_{DR,n} \Delta DR^2[n] + \sum_{n=1}^{N} w_{I,n} \Delta I^2[n]}$$

where $w_{IOI,n}$, $w_{DR,n}$, $w_{I,n}$ are the weights for each n-th note, depending on the score and on psychoacoustic principles. A simple and good choice is by weighting time and intensity variations so as to balance the effects of the just noticeable differences (JNDs) for the human ear. Thus $w_{IOI} = w_{DR} = 1/JND_{dur}$ and $w_L = 1/JND_{loudness}$. These values of JND are quite variable with many other parameters (i.e. frequency, timbre and position in musical structure). We used two mean values for JNDs: 0.5 dB for sound levels ($JND_{loudness}$) and 5% for time durations (JND_{dur}). This choice means that a timing deviation will be considered of the same importance of a intensity deviation if their values are equal to their respective JNDs.

The latter point deals with the context perception, and interacts with the former point. Many studies have been carried out on this aspect: the musical structure of the piece make some notes particularly important for listeners, and a non accurate fit for these notes can compromise the synthesis. In fact, using a constant weight for every note can produce non-musical results, because taking into account only the JNDs leads to a simplified model of human perception. Moreover, even if we are able to take into account the structural context, in general, it is difficult to assign the correct weights for each kind of piece. Thus, it would be advisable to fine tune the distance definition by increasing the weights of the most relevant notes.

7.4.3 Models for understanding

We may distinguish some strategies in developing the structure of the model and in finding its parameters. The most prevalent ones are analysis-by-measurement and analysis-by-synthesis. Recently some methods from artificial intelligence started being developed: machine learning and case based reasoning.

7.4.3.1 Analysis by measurements

The first strategy, analysis-by-measurement, is based on the analysis of deviations measured in recorded human performances. The analysis aims at recognizing regularities in the deviation patterns and to describe them by means of a mathematical model, relating score to expressive values [see Gabrielsson 1999 for an overview of the main results]. The method consists in different stages:

1. Selection of performances. The choice of good and/or typical performances of the musical excerpt to study is important. Often rather small set of carefully selected performances are used. While normally the performer is left free to play according to his own taste, sometimes for experimental purpose he is asked to play according to specific instructions, e.g. to convey a specific emotion.

2. Measurement of the physical properties of every note. The physical variations of the performance are many: duration, intensity, frequency, envelope, note vibrato; which and how many variables to study depends on the aims and working hypothesis, on the technical possibility of the instrument and on the considered instruments.

3. Reliability control and classification of performances. It is necessary to verify the reliability and consistency of the data obtained from the physical variable measurement, classifying the performance in different categories, with different characteristics, taking into account the collected data.

4. Selection and analysis of the most relevant variables. This stage depends on the two previous ones and it ends temporarily the analytical part of the scheme to give space to the judgment of the listeners, in the following stages.

5. Statistical analysis and development of mathematical interpretation model of the data. The analysis of the selected variables is often carried out on different time scale representations.

The most frequently used approaches are statistical models and mathematical models (see e.g. sect. 7.5). Sometimes multidimensional analysis is applied to performance profiles in order to extract independent patterns. Often the hypothesis that deviations deriving from different patterns or hierarchical levels can be separated and then added is implicitly assumed. This hypothesis helps the modelling phase, but may be oversimplified.

Several methodologies of approximation of human performances were developed using neural network techniques or fuzzy logic approach or using a multiple regression analysis algorithm or linear vector space theory. In these cases, the researcher devises a parametric model and then estimates its parameters that best approximate a set of given performances.

As an alternative to this method that analyses actual music performances, some researchers are performing controlled experiments in collecting and studying performances. The idea is that by manipulating one parameter in a performance (e.g. the instruction to play at a different tempo), the measurements may reveal something of the underlying mechanisms.

7.4.3.2 Analysis by synthesis

The analysis-by-synthesis paradigm takes into account the performance-perception and it starts from the results of the previous stages (steps 1-5 of the previous section) continuing with the following stages.

6. Synthesis of performances with systematic variations. At this stage the researcher produces different versions of the piece in order to have performances in which the physical variables to be studied (duration, intensity, etc.) systematically vary.

7. Judgment on synthesized versions, paying particular attention to the different experimental aspects selected. Knowledge of relevant experimental variables and the designation of useful evaluations scales are required.

8. Control of the reliability judgments and classifications of the listeners. We need to use adequate methods to control the listeners? reliability and their judgments, possibly classifying them in different class.

9. Study of relation between performance and experimental variables. At this point, it is possible to observe the relations between performances with manipulated physical variations and the selected variables asking questions such as: are the listeners sensitive to the manipulations made? If yes, in which way? Are there general effects or interactions among different variables? Which are the most important variables? Can we eliminate some of them?

10. Repetition of the procedure (steps 3-9) until the results converge. In relation to the results of stages 3-9, the process should be continued in an interactive manner until the relations of the selected variables of the performance converge to the experimental variables.

The scheme here described can be modified and extended, but the main concept remains the following: the analysis of the real performances produces hypotheses to be tested through the systematic variations introduced in the synthetic versions. With regard to such variations, it should be noticed that factors must be modified one by one keeping the rest constant. The best method to generate them should be, for instance, to produce simplified versions where only one variable is modified, while imposing constant values to the others. The product will sound rather different from a real performance where all the physical variables change continuously. In order to obtain data about the effect of the other variables and their interaction, we must proceed to further experiments, in a long series of working sessions.

This strategy derives models, which are described with a collection of rules, using an analysis-by-synthesis method. The most important is the KTH rule system presented in section 7.6. In the KTH system, the rules describe quantitatively the deviations to be applied to a musical score, in order to produce a more attractive and human-like performance than the mechanical one that results from a literal playing of the score. Every rule tries to predict (and to explain with musical or psychoacoustic principles) some deviations that a human performer is likely to insert. At first, rules are obtained based on the indications of professional musicians, using knowledge engineering paradigms. Then, the performances, produced by applying the rules, are evaluated by listeners, allowing further tuning and development of the rules. The rules can be grouped according to the purposes that they apparently have in music communication. Differentiation rules appear to facilitate categorization of pitch and duration, whereas grouping rules appear to facilitate grouping of notes, both at micro and macro level. As an example of such rules, let us consider the Duration Contrast rule: it shortens and decreases in amplitude the notes with duration between 30 and 600 ms, depending on their duration according to a suitable function. The value computed by the rule is then weighted by a quantity parameter k.

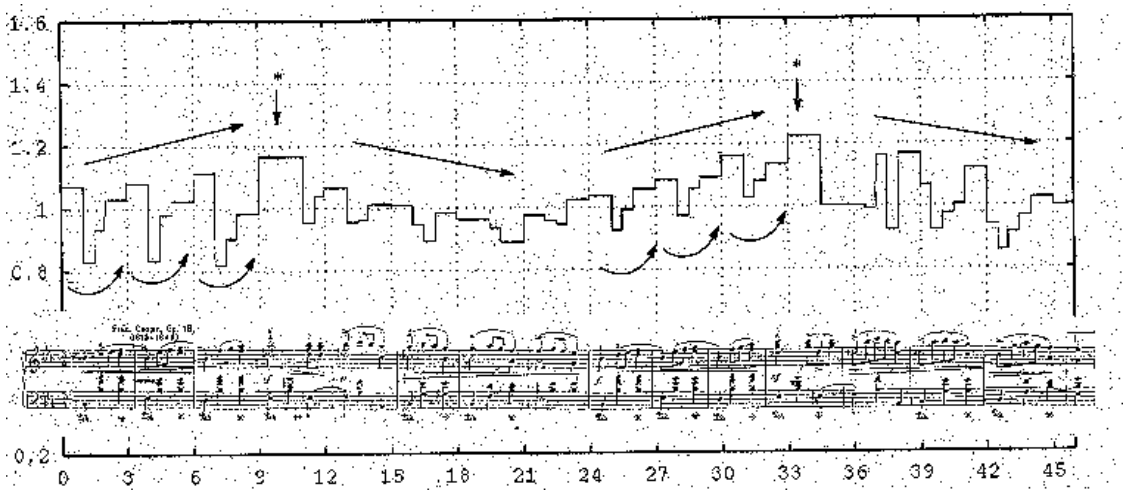


Figure 7.6: Dynamics deviation learned from the training pieces applied to Chopin Waltz Op.18, Op.64 no.2

7.4.3.3 Machine learning

In the traditional way of developing models, the researcher normally makes some hypothesis on the performance aspects s/he want to model and then s/he tries to establish the empirical validity of the model by testing it on real data or on synthetic performances. A different approach, pursued by Widmer and coworkers, instead tries to extract new and potentially interesting regularities and performance principles from many performance examples, by using machine learning and data mining algorithms. The aim of these methods is to search for and discover complex dependencies on very large data sets, without any preliminary hypothesis. The advantage is the possibility of discover new (and possibly interesting) knowledge, avoiding any musical expectation or assumption. Moreover, these algorithms normally allow describing discoveries in intelligible terms. The main criteria for acceptance of the results are generality, accuracy, and simplicity. It can be noticed that when rules for categorical decisions (e.g. play faster of slower) rather than for computing an exact value are used, more understandable results can be obtained. An example is shown in figures 7.6 and 7.7.

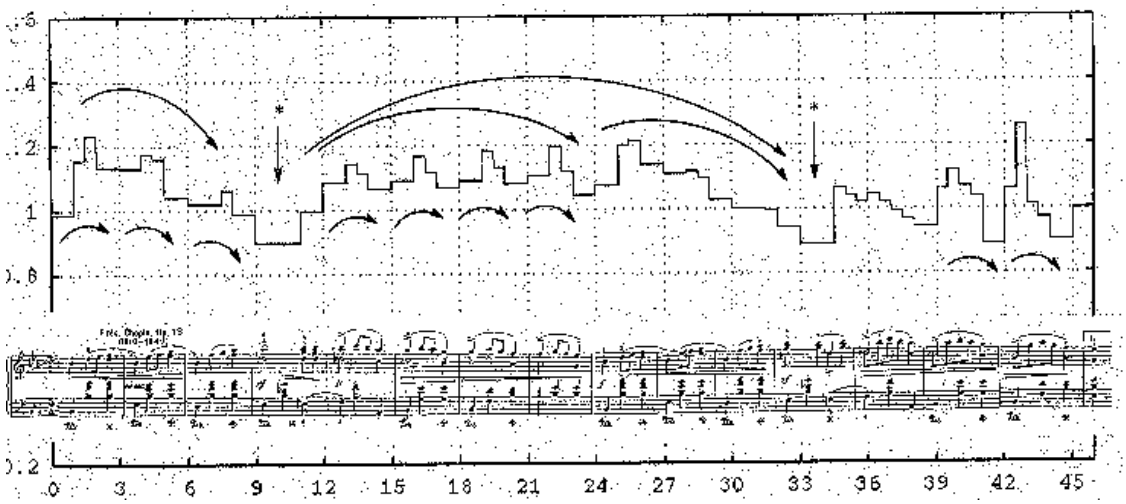


Figure 7.7: Tempo deviation learned from the training pieces applied to Chopin Waltz Op.18, Op.64 no.2

7.4.3.4 Case based reasoning

An alternative approach, much closer to the observation-imitation-experimentation process observed in humans, is that of directly using the knowledge implicit in human performances samples. ***Case based reasoning*** (CBR) [2] is based on the idea of solving new problems by using (often with some kind of adaptation) similar previously solved problems. Two basic mechanisms are used: retrieval of solved problems (called cases) using suitable criteria and adaptation of solutions used in previous cases to the actual problem. The assumption is that similar problems have similar solutions.

The CBR paradigm covers a family of methods that may be described in a common subtask decomposition: the retrieve task, the reuse task, the revise task, and the retain task. Different CBR methods differ in the way of achieving these four tasks.

- The goal of the retrieve task is to recover a set of previously solved problems similar to the

[2] adapted from Arcos (1997)

current problem. The retrieval task is usually performed using, in turn, three subtasks: identify, search, and select tasks.

  - The identify subtask determines, using domain knowledge, the set of relevant aspects of the current problem.
  - The search subtask retrieves a set of precedent cases, using these relevant aspects as similarity criterion,
  - The select subtask has as a goal to rank the set of precedents using domain knowledge.

- Given a set of ordered precedent cases, the reuse task constructs a solution for the current problem adapting the solutions taken in precedent cases. The ranking over cases is interpreted as preference criterion. An usual policy is to consider only the maximal precedent determined by the select subtask.

- When the solution generated by the reuse task is not correct, an opportunity for learning arises. The revision task involves detecting the errors of the current solution and modifying the solution using repair techniques. This phase, that is not present in all CBR methods, takes the result from applying the solution in the real world (or by asking a teacher).

- Finally, the new solved problem is incorporated into the system by the retain task in order to help the resolution of future problems. This task involves selecting which information of the case retain and how to integrate the new case in the memory structure.

CBR is appropriate for problems where many examples of solved problems can be obtained and a large part of the knowledge involved in the solution of problems is tacit, difficult to verbalize and generalize. Moreover new problem solution can be checked by the user and then memorized. Thus, the system learns from experience. The success of this approach greatly depends on the availability of a large amount of well-distributed previously solved problems. These are not easy to collect.

7.4.3.5 Expression recognition models

The methods seen in the previous sections aim at explaining how expression is conveyed by the performer and how it is related to the musical structure. Recently these accumulated research results started giving rise to models that aim to extract and recognize expression from a performance.

In particular, Dannenberg [1997] proposed a style classifier for interactive performance systems, employing a machine learning approach. The features he used to classify are simple parameters that can be extracted from trumpet performances played by one performer and recorded as MIDI data. The classified styles consist of a range of performance intentions: frantic, lyrical, pointillistic, syncopated, high, low, quote and blues.

Friberg (2002) developed a system that combines a low-level cue extraction algorithm with a listener model to predict what emotion the performer is trying to convey in his or her performance. One or several types of listener panels can be stored as models which are used to simulate judgments of new performances based on results from previous listening experiments. From audio input data the following parameters are computed for each tone: interonset duration, relative articulation, peak sound level, attack velocity, and spectral ratio. The spectral ratio is simply defined as the difference in sound level below and above 1000 Hz. The acoustic cues are obtained by computing running averages and standard deviations of the parameters. An estimation of the strength of each intended emotion (happy, angry, sad) is obtained from a regression equation taking the standardized cue values as input variables.

Mion (2003) employed Bayesian Networks for the recognition of expressive content in musical improvisations. From MIDI piano improvisations, the extracted features are: note number, intensity, articulation, inter-onset duration, features pattern. The following expressive intentions described by sensorial adjectives are recognized: slanted, heavy, hopping, vacuous, bold, hollow, fluid, tender.

7.4.4 Models for music production

7.4.4.1 Performance synthesis models

While the models described above were developed mainly for analysis and understanding purpose, often they are used also for synthesis purpose. Starting from expressiveness models, several software systems for the computer automatic generation of musical performances were developed. Moreover, many sequencers now implement functions, called humanizers, that add deviations to the score, computed in a random way or according to specific criteria. The typical scheme is represented in figure 7.8.

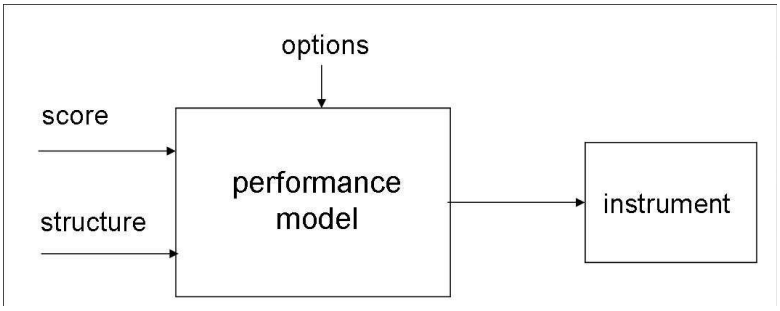


Figure 7.8: Typical structure of a performance synthesis model.

The model defined at Centro di Sonologia Computazionale (CSC), University of Padova, was developed using the results of perceptual and sonological analyses made on professional performances (see also section 7.7). Different applications based on this model were developed. Music performance is an activity that is well suited as a target for multimodal concepts. Music is a nonverbal form of communication that requires both logical precision and intuitive expression. Our research in the creative arts domain has focused on musical mapping of gestural input. In fact, since the control space works at an abstract level, it can be used as an interface between transmodal signals. In particular, we developed an application allowing control of the expressive content of a pre-recorded music performance by means of dancer's movement as captured by a camera. Then expressive features extracted by dancer's movements are used as input for the abstract space. In the entertainment area, we built the application Once upon the time, released as an applet, for the enjoyment of fairy-tales in a remote multimedia environment. In this software, an expressive identity can be assigned to each character in the tale and to the different multimedia objects of the virtual environment. Starting from the storyboard of the tale, the different expressive intentions are located in synthetic control spaces defined for the specific contexts of the tale. The expressive content of audio is gradually modified with respect to the position and movements of the mouse pointer, using the abstract control space described above.

7.4.4.2 Discussion on synthesis models

The idea of automatic expressive music performance, especially when it is applied to the performance of classical music is questionable. We can remark that classical music was not written for this purpose. Even if the models could be very accurate (and they still are not), some very important artistic

aspects of this kind of music will be omitted. When we listen to a recording of a classic music performance, we are aware that it is just a reproduction of an event and not an experience of the music as it was conceived at its time. On the other hand, the possibility to fully model and render the artistic creativity implied in the performance is still to be demonstrated. For the moment, at best, we can expect a reproduction of a specific performance, without a real new creative contribution, that would make listening interesting. Or we can expect the rendering of some, hopefully relevant, aspects of a musically acceptable performance, but not sufficient for a full artistic appreciation.

Performers are particularly sensitive to these aspects and usually look at performance synthesis in a very suspicious manner. An instinctive fear of a possible danger for their competence and even their job can be guessed to contribute, but the cultural motivations are definitely true. On the other hand if we think to music applications, where a real artistic value is not necessary (even if useful as in many multimedia applications), and where the alternative is a mechanic performance of the score (as in many sequencers), automatic performance can be acceptable. From this point of view such models can be used for entertainment application or when it is not necessary to preserve the exact artistic environment of the composition, as in popular music. However, in many occasions a human performer is not available and should be substituted in a certain way. Performance models or processing MIDI recorded performance could be a solution. Notice that the quality of performance processing is much higher when it is based on performance models and knowledge.

Another important application of performance models, even of classical music, is in education. The knowledge embodied in performance models may help teachers to increase their students' awareness for certain performance strategies and to better convey their teaching goals.

7.4.4.3 Models for multimedia application

Representing, modelling and processing expressive information is useful not only for automatic music performance. In fact a user can interact with the model during the performance. We can thus consider interactive performance models where expression is conveyed by a joint action of the user and of the model. This paradigm of human machine interaction for expression communication is not only fruitful in music applications, but it can be extended to many other fields where non-verbal content can be very relevant. We may distinguish two main classes of possible interfaces for the human-machine communication:

- Graphic panel dedicated to the control, where the control variables are directly displayed on the panel and the user should learn how to use it.
- Multimodal, where the user interacts freely through movements and non-verbal communication. Task of the interface is to analyze and to identify human intention correctly.

Expressiveness control is a relevant aspect in multimodal systems. The current state-of-the-art allows for a growing number of applications, from advanced human-computer interfaces in multimedia systems to new kinds of interactive multimodal systems. An explosion of human interface technologies involving ecological interface design, agents, virtual immersive workspaces, decision support systems, avatars, distributed architectures, and computer-supported cooperative work, are appearing into the scene as means to address these complex problems.

Multimodal interfaces have the potential to offer users more expressive power and flexibility, as well as better tools for controlling sophisticated visualization and multimedia output capabilities. As these interfaces develop, research will be needed on how to design complete multimodal-multimedia systems that are capable of highly robust functioning. To achieve this goal, a better expressive content

analysis and processing ability will be essential. The computer science community is just beginning to understand how to design innovative, well integrated, and robust multimodal systems. Most multimodal systems remain bimodal, and recognition technologies related to several human senses (e.g., haptics, smell, taste) have yet to be well represented or included at all within multimodal interfaces. This means that it is very important, for a successful design of multimodal systems, to consider performance models for non-verbal communication.

7.4.5 Models for artistic creation

The situation is different when music is expressively created bearing in mind the use of technology. We are in the era of information society and artists are always more frequently using technology in their artworks. Since the beginning of last century, some musicians started to think how to enlarge the sound palette by using un-conventional instruments. The availability of new electronic and computer generated sounds gave rise to a new kind of music. Artists exploited and innovated greatly the methods of producing and performing music. In the first period of computer music, a lot of research effort was dedicated to sound synthesis and modelling. New synthesis algorithms were discovered, such as frequency modulation, and new paradigms were developed for musical sound generation, such as spectral and physical models. On the other side, models for music representation and algorithmic composition were developed.

Less attention was being paid to the performance aspects. The music was automatically generated from the score as it was written by the composer or generated by the composition program. The composer had to take into account all the nuances often implicit in the score to communicate the expressive content of the music. In this situation, the composer must explicitly preview what the performer normally handles. The composer is also a performer and needs to formalize the performance process. A different approach, to overcome the limitations of computer generated music, was followed by music for live electronics where the performer interacts with technology on the stage transforming in real time the sound produced by traditional or synthetic instruments.

In both cases, a central challenge is the control of the sound synthesis or processing engines (systems, algorithms, etc.). This problem is a typical performance topic and it refers to the need of establishing and computing the relation of musical and compositional aspects with sound parameters, according to the expressive aim of the musician. The inputs are discrete events, as described in the score or generated by computer, and continuous signals, e.g. performer gestures. These inputs should be coordinated and merged to produce and process sound events. In music technology the concept of mapping strategies, which describe these relations, is of great importance. The conventional (and simplest) aspect refers to specific relation; for example how to convert a pitch and loudness information into proper spectral and micro-timing values of a synthetic note. Nevertheless, the word strategies tends to refer to other possible choices and source of information as phrasing, musical character, mood of the performer, stylistic alternatives.

All these aspects are typical music performance issues. Suitable music performance models are very desirable.

Figure 7.9 shows the typical situation of music performance with digital instruments where the electronic instrument performer controls the sound synthesis with gestures and suitable processes. A performance model lies between the symbolic and the audio control level. The performer receives an audio feedback from the instrument as with traditional instruments. In live electronics, the scheme is different (fig. 7.10). Here the live electronics performer processes the sound produced by the instrument performer, acting on his computer. In the live electronic box, we still have score processes and gestures controlling, via a performance model, the sound processing devices. However, in this

case the input is a music sound, already performed. In a certain sense, we have a combined effect of performances (e.g. deviations of deviations) that the models should take into account. The performer receives an audio feedback from both the instrument and the sound processing.

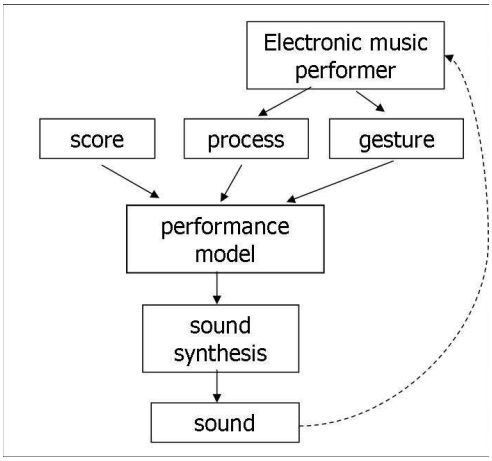


Figure 7.9: Scheme of music performance with digital instruments where the electronic instrument performer controls the sound synthesis with gestures and suitable processes. A performance model lies between the symbolic and the audio control level. The performer receives an audio feedback from the instrument as with traditional instruments.

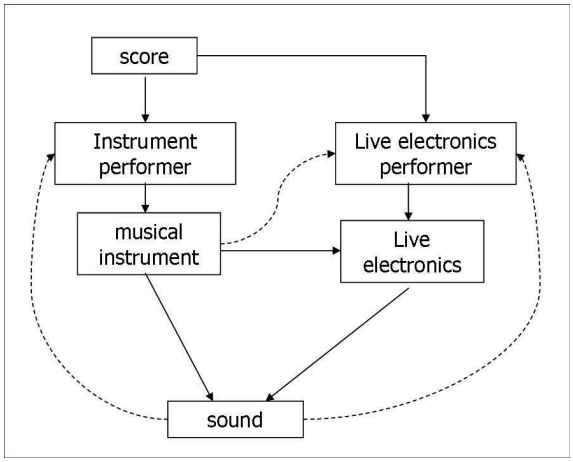


Figure 7.10: Scheme of live electronic music performance. The live electronics performer processes the sound produced by the instrument performer, acting on his computer. In the live electronic box, we still have score processes and gestures controlling, via a performance model, the sound processing devices. The performer receives an audio feedback from both the instrument and the sound processing.

7.4.6 Perspectives

Recently music performance researchers are becoming more aware of the need of a well-founded approach based on strong scientific knowledge. This aim can be faced from two complementary directions. One way is to start from the knowledge gained in classical music performance studies and

formalized in performance models; then generalize their results and apply them to the performance of new music creation. The other direction starts from the practical knowledge of new music creators (often embodied in their music performance systems) in order to extract possible suggestions and proposals of new performance models. From the joint effort of scientists and musicians valid results can be expected and real new tools can be developed, not only inspired to problems and solutions of the past times.

It can be noticed that music performance is an interesting topic for scientific investigation and for technology research: it involves human non-verbal communication, has artistic-creative finality, and requires strong cooperation between art and science - technology. Probably still more important is the fact that music is an immaterial art that has a strong tradition of symbolic representation and abstract thinking. This attitude may explain why musicians were the most enthusiastic and successful in promoting and contributing to the joint development of art and science since the beginning of computer science. In other arts, this collaboration started much later and very often it is restricted to the use of technology rather than a real contribution to joint development of knowledge and tools.

7.5 A dynamic model of phrasing

The idea that there is an intimate relationship between musical motion and physical movement is an old one and can be traced back to antiquity. Classical Greek musical writings can be broadly classified in two distinct schools, a Pythagorean an Aristoxènian school. It is intersting yo notice that whereas for Pythagoreans pitch intervals between notes should be expressed as ratios of numbers, for the Aristoxènians notes are geometrical points in a space (defined by ratios) and intervals the distance between them. It is this concept of a space that enabled Aristoxènus to think in term of melodic motion. The concept of melodic motion relative to an abstract space is central in his thinking. Moreover he makes clear reference to rhythmic movement and its analogy to physical movement.

The idea of a connection between music and motion is a recurrent one. In general the following has been suggested:

- musical movement has two degree of freedom, tonal movement and rhythmic movement;
- this movement is similar to and imitates motion in physical space;
- the object of motion in physical space, to which musical movement alludes, is that of a body or limb.

In this section a dynamic model of phrasing, based on the analogy of physical movement, proposed by Todd, is presented. He consider the score as a trajectory in a 2-D space. The vertical axis is describe a 1-D *pitch space* p while the horizontal axis describe a space like dimension, measured in unit of beats or bars, called *metrical position* x. Thus he distinguishes two kind of motion in music, tonal motion, i.e. pitch as function of time $p(t)$, and rhythmic motion, i.e. metrical position as a function of time $x(t)$. Its model deals with metrical motion.

7.5.1 Definition of the basic terms

A model based on physical analogy has been proposed by Todd(1992,1995). Every note event is described by its onset time $o[n]$, intensity $I[n]$. Let a be the acceleration, u the initial tempo, x score position (measured in units of beats or bars), and t the performance time. Given some analytical

function for acceleration or tempo (velocity), we may obtain either $t = t(x)$ or $x = x(t)$ by integration so that these variables are related by the following system of equations

$$a = a(t) \tag{7.1}$$
$$v = v(t) = \int a(t)dt \tag{7.2}$$
$$x = x(t) = \int v(t)dt \tag{7.3}$$

and

$$a = a(x) \tag{7.4}$$
$$v = v(x) \tag{7.5}$$
$$t = t(x) = \int \frac{1}{v(t)}dx \tag{7.6}$$

where $a(x)$ and $v(x)$ are obtained by solving for $t = t(x)$ and substituting in $a(t)$ or $v(t)$. Conversely, if given a function for position, then the tempo and acceleration may be obtained by differentiation.

7.5.2 The linear tempo model

For instance the classic linear tempo model, i.e. when tempo is supposed to vary linearly in time on a performance segment, assumes that the acceleration (or deceleration as in the final retard) is constant in that segment. The corresponding equations relative to performance time t are

$$a(t) = a \tag{7.7}$$
$$v(t) = \int a(t)dt = u + at \tag{7.8}$$
$$x(t) = \int v(t)dt = ut + \frac{at^2}{2} \tag{7.9}$$

where u is the initial tempo. The equations relative to score position x are

$$a(x) = a \tag{7.10}$$
$$v(x) = \sqrt{u^2 + 2ax} \tag{7.11}$$
$$t(x) = \frac{\sqrt{u^2 + 2ax} - u}{a} \tag{7.12}$$

7.5.3 Energy, tempo and intensity

The model assumes that a piece can be decomposed in a hierarchical sequence of segments, where each segment is on its turn decomposed in a sequence of segments. It is similar as a musical phrase can be decomposed in a sequence of sub-phrases, a sub-phrase on a sequence of melodic gestures, etc.. Every segment is characterized by an accelerando-ritardando pattern and by a crescendo-decrescendo pattern. The models further assumes a linear tempo model, i.e. a constant acceleration in the first phase, followed by a constant deceleration in the second phase. The analogy is the movement of a particle of mass m in a V-shaped potential well of length L, where the depth of teh well and the position of the lower point are parameters of the model. Outside the well the particle moves with

constant velocity. let us assume also that the total energy E of the system is constant and given by $E = T + V$ where $T = mv^2/2$ is the kinetic energy and v is the potential energy linearly varying from zero to a minimum and then linearly returning to zero. Thus the velocity (tempo) is given by $v(x) = \sqrt{2(E - V(x))/m}$. A similar expression is used for the intensity $I(x)$. Notice that this expression corresponds to a parabolic mapping $x(t)$ in the first and in the second phase. The hierarchical structure that a piece is composed by a number of components of this type describing from the global variation over the whole to local fluctuations at the note level. These components are superimposed (summed) onto each other. Thus the complete function is given by

$$v(x) = \sum_j \sqrt{2\frac{E - V_j(x)}{m_j}}.$$

The complete $x(t)$ mapping results shaped as piece-wise parabolas. Some authors try to estimate the parameters of the parabolas from measurements of onset time in real performances. An example of phrasing computed for the theme, from the six variations composed by Beethoven over the duet *Nel cor più non mi sento* (fig. 7.11), is shown in figure 7.12.

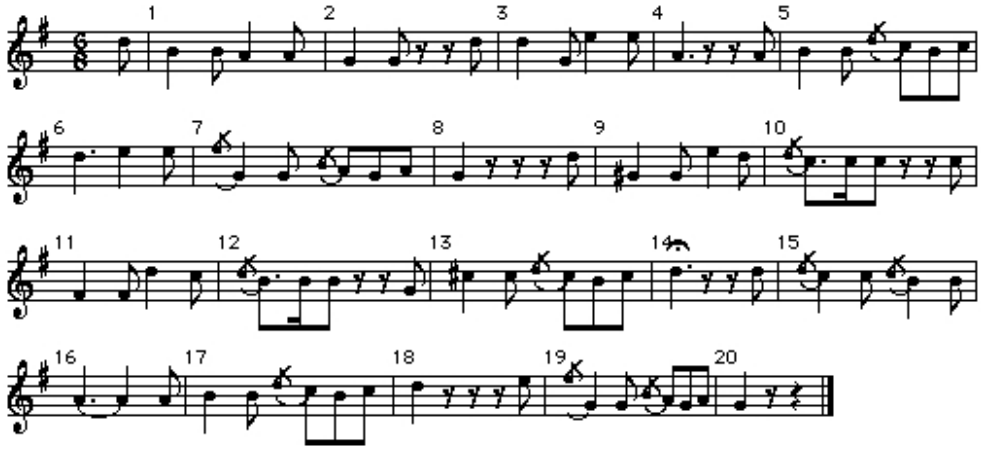


Figure 7.11: Theme from the six variations composed by Beethoven over the duet *Nel cor più non mi sento*.

This model is interesting for describing the typical acceleration-rallentando patterns used in most romantic music performances to communicate the phrasing structure and is quite effective in performance synthesis. It is in alternative with the idea of punctuation (see sect. 7.6), where the boundary of segments are marked by a micro-pause inserted between them. Probably the best way of modelling the phrasing of a piece is using a combination of both methods.

7.5.4 Other parabolic models

The idea of using parabolas as (least square) approximators of observed data is quite common in music performance analysis. But different authors approximate different kind of data. A quite widespread model is parabolic inter onset interval $IOI(x)$ or relative inter onset interval $IOI_{rel}(x)$ as function of score position x or event number n. It is important to notice that different representation of the data to be analysed and modelled tends to evidence different aspects. However the use of different representations makes the comparison of the analyses problematic.

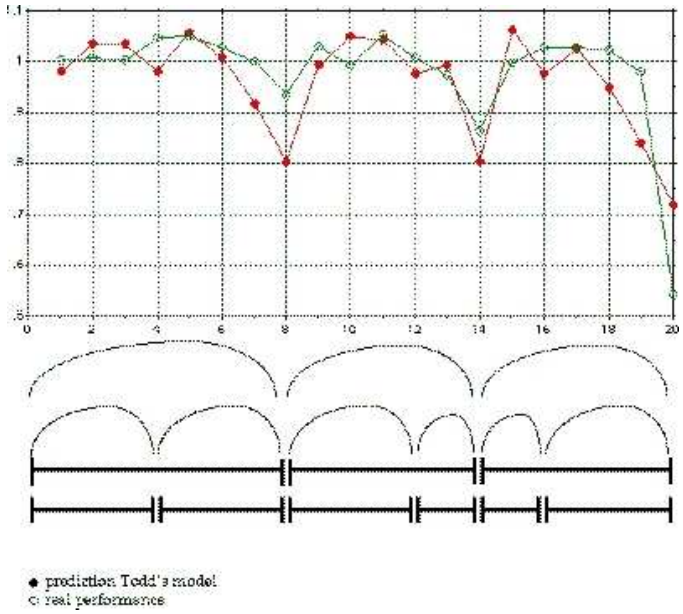


Figure 7.12: Example of phrasing model (score shown in fig 7.11). Each phrase and sub-phrase contributes its own curve. The combination of which is shown here (with black dots, as is a real performance (with white dots).

7.6 KTH music performance rules

adapted from PhD dissertation of Anders Friberg

In this section the most important model, developed using the analysis by synthesis paradigm (see sect. 7.4.3.2), is presented. In the KTH system, the rules describe quantitatively the deviations to be applied to a musical score, in order to produce a more attractive and human-like performance than the mechanical one that results from a literal playing of the score.

7.6.1 From composer to listener: A closer look

The communication of a composer's mental representation of a piece of music to a listener can be assumed to contain three major transformations, as illustrated in Fig. 7.13. : (1) from composer to score (TCS), (2) from score to performance (TSP), and (3) from performance to listener (TPL). The music appears in four different representations in the figure. In addition, the performer has also a mental representation. Of these, only two are easily accessible to a scientific analysis: the score and the performance. The performance is assumed to be the sound signal, i. e. a recording that can be analyzed in terms of physical parameters. The transformation TSP is done by the performer and is the main focus of this study.

It is advantageous to compare a performance to a nominal performance in which the score is simply translated to nominal values of performance parameters; in such a translation simple integer ratios, for instance, are used for converting note values to tone durations. The difference between the actual and the nominal performances constitutes the expressive deviations.

Why do these deviations from the score exist? There are many possible reasons. First, the score serves primary as an aid for the memorization and conservation, as well as for the communication from the composer to the performer. Scores were never intended as exact descriptions of sounding music.

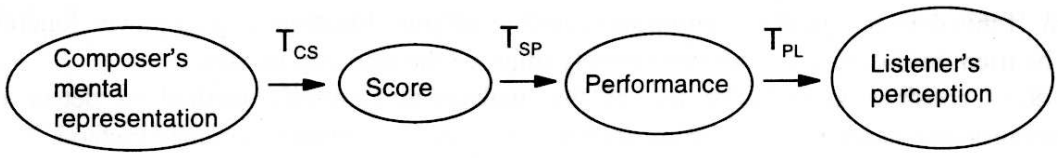


Figure 7.13: From the composer to the listener: the four different music representations and the three corresponding transformations.

Second, as the composer and the performer are unaware of the measured physical quantities, the score may serve as a representation of the cognitive parameters rather than the physical parameters. There is no need to notate cognitive representations that both the composer and the performer agree upon. In this sense the score may be more accurate with respect to cognitive than to physical parameters. Third, over the centuries the liberty of the musicians to exhibit their own, personal interpretation of the composer's piece of art has varied, but has rarely been completely denied by composers. In cases where this liberty was ample, great deviations from a nominal performance can be expected.

7.6.2 Analysis-by-synthesis

As mentioned above, the main method used for developing the rules was analysis-by-synthesis. It was adapted from speech synthesis research where it is considered as a standard method. It was a natural choice since the system developed for text-to-speech translation could be adapted to a score-to-musical-expression system. Here some aspects of this method will be discussed.

The typical start is an idea which is formulated as a tentative rule in the computer. Then this rule is applied to a music example so that the result can be evaluated by listening. This offers an immediate feedback, often suggesting further modifications. The process is then repeated until a satisfactory performance is obtained. Thus in a sense the system acts as a student acquiring some basic knowledge of music interpretation from an expert teacher.

One requirement of this method is that everything must be quantified. A typical observation has been that the exact quantity of each parameter is crucial for a good performance. In determining the dependence of a rule on a certain parameter, such as note duration, it is generally helpful to find two extremes and then to interpolate linearly between them. If this does not yield an appropriate result a different function, e. g., a power function can be tried. In this way we can successively improve the rule step by step.

Let us consider an exclusive use of the analysis-by-synthesis method to detect its advantages and disadvantages as compared to a strict analysis-by-measurement method.

One advantage is that the perception of the music is directly used in the development of the rules, similar to how a musician also act as listener while playing, and use this information as a feedback, see Fig. 7.14. In analysis-by-measurement, the listener's viewpoint, or rather the perception of the music, is not incorporated in the same direct sense.

Another advantage is that the general validity of the hypotheses can directly be tested by applying it to other music examples and that the feedback loop is very short between stating the hypothesis and evaluating the results.

A disadvantage is that conclusions are based on the expertise of just a few people. It raises very high demands on the experts that they are competent, consistent, able to focus on a certain aspect of the performance and that they are sensitive also to small deviations. Another disadvantage is that the

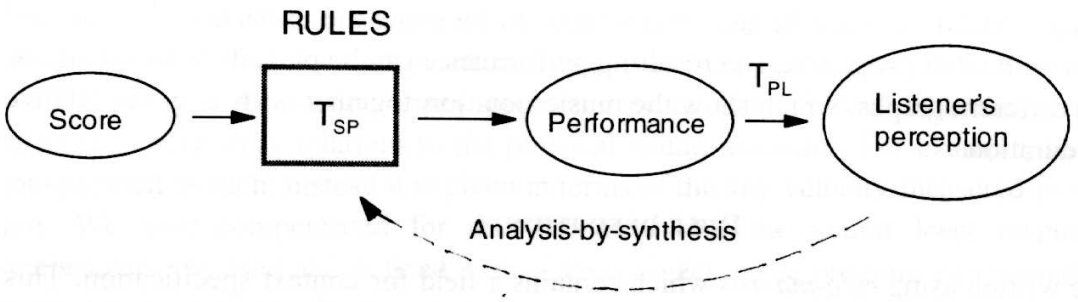


Figure 7.14: From score to listener: the rule transformation and the analysis-by-synthesis loop.

parameters in the rules can in some cases be chosen rather arbitrarily.

For these reasons the current system was not based solely on the analysis-by-synthesis method but also on analysis-by-measurement. This is probably quite essential in performance research. Conversely it is quite important to complement the analysis-by-measurement method by listening tests where the deduced principles are applied to synthetic performances.

7.6.3 Generative rules

The purpose of the rules is to convert the written score complemented with cords symbols and phrase markers, to a musically acceptable performance.

Whenever possible, the resulting deviations computed from the rules are additive. This means that each tone may be processed by several rules, and the deviations made from each rule will be added successively to the parameters of that tone. Most of the rules include the quantity parameter k. This parameter is used to alter the quantity of the manipulation induced by the rule. The default value is $k = 1$. This value is appropriate when all rules are applied. when a rule is used in isolation, slightly higher settings of k can be necessary to produce audible changes. Different settings of k can be used to generate different performances of the same melody.

A rule is expressed as

```
if (condition)
    then (action)
```

where action normally compute a deviation of some parameter.

The rules can be grouped according to the purposes which they apparently have in music communication. Three major principles can be identified: differentiation of categories and grouping. The grouping rules mark which tones belong together and where the structural boundaries are. The differentiation rules increase the differences between tone categories such as pitch classes, intervals, and note values. The emphasis rules emphasise unexpected notes.

All rules are not intended to be used simultaneously. Some of the rules are partly overlapping, as explained below where each rule is discussed. The concept is that the user of the rules may act as a meta-performer where different performances can be realized by selecting rules and rule quantities. The default value of the quantity is $k = 1$. This was developed when many rules were applied simultaneously. When fewer rules are applied higher quantities may be used.

Examples of differentiation rules are:

Double-duration Decreases the IOI contrast for two adjacent notes having the nominal IOI ratio 2 : 1, e.g., a quarter note followed by an eighth note.

Duration-contrast Long notes are lengthened and short note shortened; i.e., comparatively short notes are shortened and softened, while comparatively long notes are lengthened and made louder (see fig. 7.15).

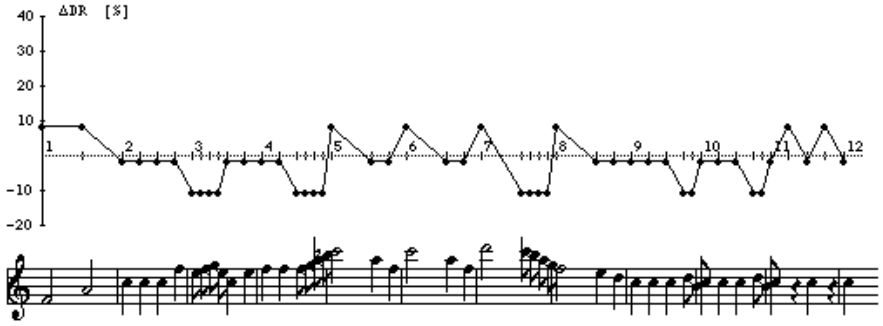


Figure 7.15: Example of Duration Contrast rule $k = 2.2$: Theme from First movement of Quartet in F major for strings, Op 74:2.

Example of emphasis rules is:

Melodic-charge Increases loudness and IOI of notes far away from the root of the current chord along the circle of fifths. The rule is not applicable in atonal music. An analysis of harmony must be provided in the score. Melodic and harmonic charge, as defined below, belong to the same category but are applied on different levels. The idea is to put emphasis on unusual events on the assumption that these events are less obvious, have more tension and are more unstable. The melodic charge $Cmel$ value is defined as a value reflecting the note's distance on the circle of fifths to the root of the current underlying chord. The value of $Cmel$ is largely a distance measure on the circle of fifths with the exception that there is more weight on the subdominant side (see table 7.1). Note that melodic charge is not associated with any particular scale since it is the same in both major and minor tonality.

Table 7.1: Melodic charge $Cmel$ for the various scale tones in a C major or minor scale.

Tone	C	G	D	A	E	B	F#	D	Ab	Eb	Bb	F
$Cmel$	0	1	2	3	4	5	6	6.5	5.5	4.5	3.5	2.5

Examples of grouping rules at microlevel are:

Faster-uphill Decreases IOI of notes in uphill motion of melody. This rule makes the notes "aim" towards the target note, that is, the top note.

Leap-tone-duration The first note in an ascending melodic leap is shortened and the second note lengthened if the preceding and succeeding intervals are by step (less than a minor third). In a descending leap the first note is lengthened and the second shortened. The amount in ms is only dependent on the interval size of the leap (unaffected by the duration). This rule is typically effective in a romantic context with rather long note values.

Examples of grouping rules at macrolevel are:

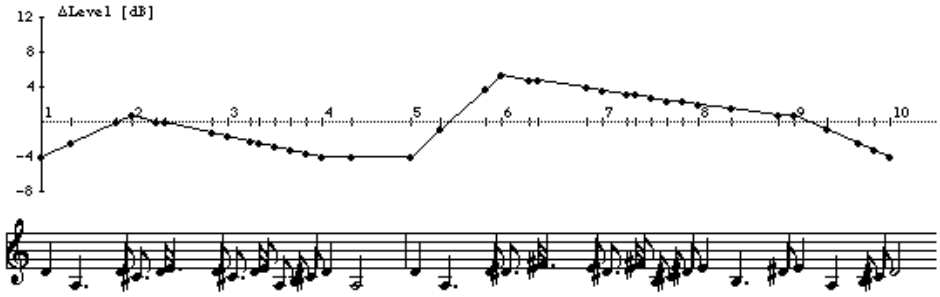


Figure 7.16: Example of harmonic Charge rule $k = 2.5$: F Schubert, Second theme from the First movement of Symphony in b minor, Unfinished

Harmonic-charge Produces rallentando and crescendo when a chord harmonically remote from the current key is approaching and vice versa. Just as with the scale tones, the harmonies in traditional Western tonal music are not equal: there are trivial chord and fantastic chords. Harmonic charge is a concept reflecting the remarkableness of chord in its harmonic context. It is a weighted sum of the chord tones' melodic charges, using the root of the main chord of the key, i. e. the root of the tonic as the reference. This rule marks the distance (related to the distance on the circle of fifths) of the current chord to the root of the current key. Sound level, duration and vibrato frequency are increased in proportion to the harmonic charge value. The increases and decreases of these parameters are gradual with linear interpolation between chord changes (see fig. 7.16). This rule is not applicable in atonal music. An analysis of harmony must be provided in the score. In the case of a temporary new tonal region within a piece, the tonic in the analysis can stay the same since the rule in general works in the intended way in tonal regions close to the original tonic. This also has the advantage that the problem of treating the change of tonic in an overlap region is avoided. One uncertain part of this rule is the chord analysis. In general this can be done on several levels of detail and usually there are also chords which can be analyzed in different ways. The level of the chord analysis in this rule should be on structurally important chords with the exclusion of passing chords.

Phrase-arch Each phrase is performed with an arch-like tempo curve: starting slow, faster in the middle, and ritardando towards the end according to a set of adjustable parameters. The sound level is coupled so that a slow tempo is associated with a low sound level. Phrase boundaries must be marked in the score. The motivation is that music has a hierarchical structure, so that small units, such as melodical gestures, join to form sub-phrases, which join to form phrases etc. When musicians play, they mark the endings of these tone groups.

The way in which it affects the performance can be varied by several additional parameters, for example the hierarchical phrase-level, the amount of lengthening of the last note in each phrase, the position of the turning point. This rule is rather sensitive to musical style and personal taste. In romantic music the amount can be rather large while in Baroque music, for instance, it has to be much lower. There is a large variation seen in measurements of the same piece played by different performers or different pieces played by the same performer. In fig. 7.17 an example is presented of phrase arch applied to F Mendelson, Aria n. 18 from "St. Paul", Op. 36. In this example two other rules are applied: Durational contrast, increasing or decreasing the duration contrasts between note values. In this example the last mentioned alternative have been selected. Punctuation, inserting micropauses after melodic gestures.

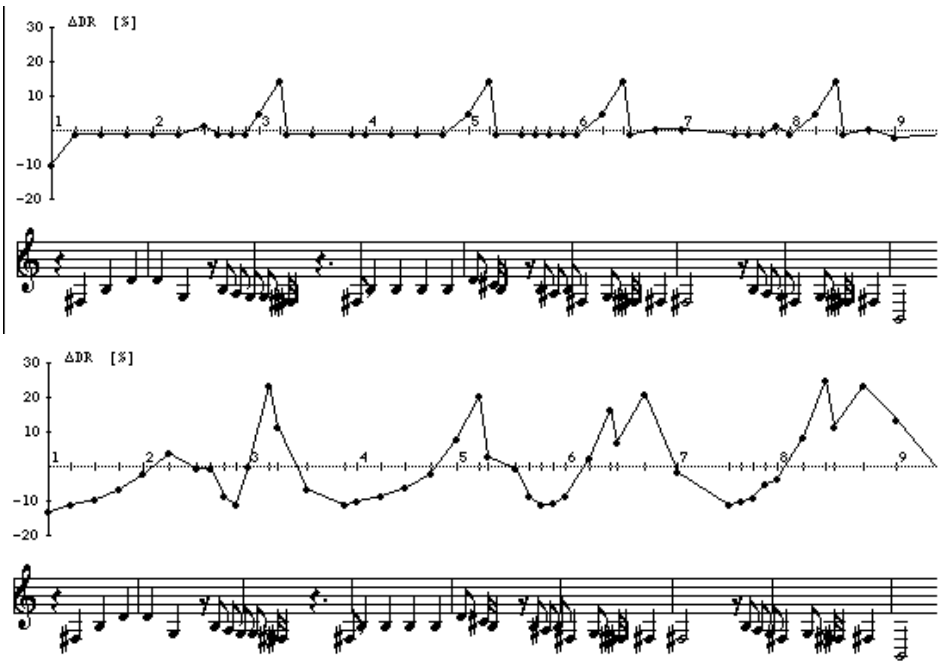


Figure 7.17: Example of Phrase Arch rule: F Mendelsohn, Aria n. 18 from St. Paul, Op. 36. (above) only Duration Contrast and Punctuation; (below) Duration Contrast, Punctuation plus Phrase Arch $k = 1.5$

Punctuation Automatically locates small tone groups and marks them with a lengthening of the last note and a following micropause. This is an attempt to automatically, from the score, identify the musical gestures and transform them to the performance, by inserting a comma realized in term of a micropause at the boundary. These gestures are melodic units consisting of 1 and up to approximately 5-8 tones. The gesture analysis is roughly analogous to grouping analysis at the lowest hierarchical level of the Lerdhal Jackendorf theory (1983), although it includes also the lowest group level which may consist of one single tone. The purpose of Punctuation rule is to find tone units at the end of which it is appropriate to insert a micropause, with the aim of signalling a separation of the different parts of the musical phrase. The punctuation is mostly bottom-up, operating on contexts comprising a maximum of five notes that potentially surround the comma. This rule is composed by a set of 14 finder or eliminator sub-rules. Finder rules mark potential positions of boundary between musical gesture. Eliminator rules indicate positions where boundary markers should not appear. The finder rules use weight values to estimate the importance of the inserted boundary mark. Intervals between adjacent tones will be referred to as steps, and larger intervals as leaps. When a note has received a comma mark, this implies that the comma appears at the end of this note.

The main principles for the finder rules are:

- in melodic leap, with different weights for different contexts,
- after longest of five notes,
- after appoggiatura
- before a note surrounded by longer notes

- after a note followed by two or more shorter notes of equal duration

The eliminator rules remove marks or reduce weights in this cases

- after very short notes
- in a melodic step motion
- when several duration rules interact
- at two adjoining marks in a tone repetition

A real boundary is assumed to exist if the sum of the weights in that position exceeds a certain percentage of the total average of inserted weight values. These boundary marks are introduced in the performance by transforming them in micropauses plus lengthening of the previous tone. The duration of the micropause and of the lengthening are proportional to the preceding note duration. The weight values are not taken into account in this translation.

Final Retard introduces a ritardando at the end of the piece similar to a stopping runner. The tempo at the end of the piece is decreased according to a square-root function of nominal time (or score position)

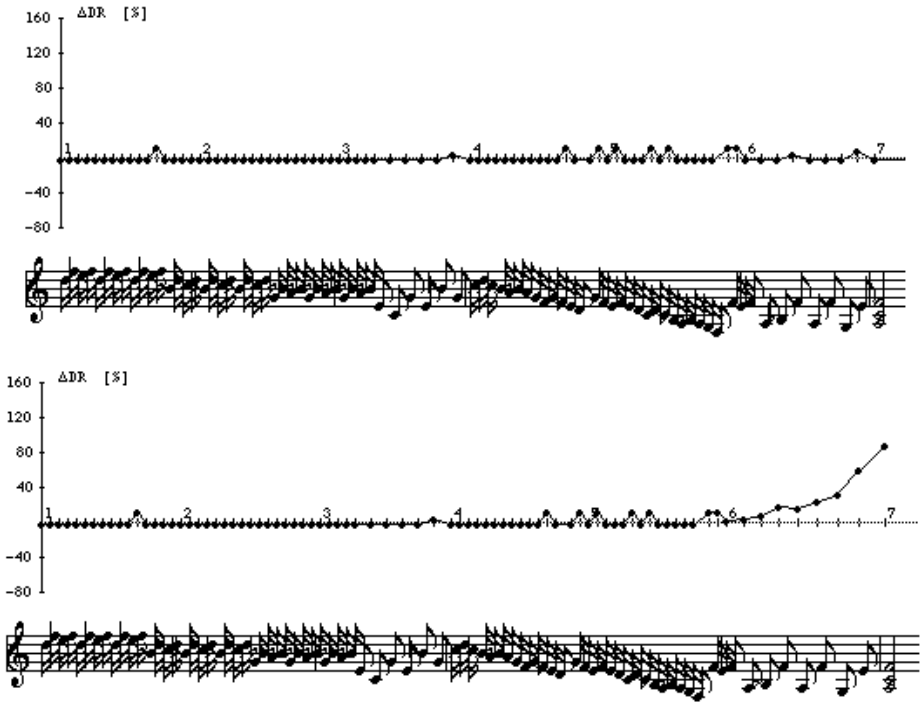


Figure 7.18: Example of Final Retard rule: J S Bach, Invention for two voices, F major, BWV 779. (above) only Punctuation; (below) Punctuation plus Final Retard $k = 1.3$

7.6.4 Macro-Rules for Emotional Expressive Performance

Gabrielsson (1994, 1995) and Juslin (1997a, 1997c) proposed a list of expressive cues that seemed characteristic of each of the emotions fear, anger, happiness, sadness, solemnity, and tenderness (see

Table 1). The cues, described in qualitative terms, concern tempo, sound level, articulation (staccato/ legato), tone onsets and decays, timbre, IOI deviations, vibrato, and final ritardando. These descriptions were used as a starting point for selecting rules and rule parameters that could model each emotion. The cues were restricted to those possible on a keyboard instrument, therefore eliminating the cues of tone onset and decay, timbre, and vibrato, although these do belong to the Gabrielsson and Juslin list of characteristic cues. The method used was analysis by synthesis After trying several musical examples, a consensus was obtained, resulting in a macro-rule (rule palette in DM) consisting of a set of rules and parameters for each emotion. Each macro-rule could be applied with the same parameters to each of the musical examples tried. The rules contained in a macro-rule are automatically applied in sequence, one after the other, to the input music score. The effects produced by each rule are added to the effects produced by previous rules. For example in order to perform a piece with Sadness the Tempo should be Slow, so Tone IOI is lengthened by 30%; Sound Level should be moderate or loud, so Sound Level is decreased by 6 dB; Articulation should be played as Legato; Time Deviations should be moderate, Duration Contrast rule is applied with $k = 2$ and Phrase Arch rule is applied on phrase level and sub-phrase level; and Final Ritardando is applied.

Tables 7.2 to 7.6 show the cue profiles for fear, anger happyness, sadness and tenderness emotion, as outlined by Gabrielsson and Juslin, and the rule setup used for synthesis with Director Musices [from Bresin, 2001].

Table 7.2: Cue profiles and macro rules for fear emotion

Expressive Cue	Gabrielsson and Juslin	Macro-Rule in Director Musices
Tempo	Irregular	Tone IOI is lengthened by 80%
Sound Level	Low Sound	Level is decreased by 6 dB
Articulation	Mostly staccato or non-legato	Duration Contrast Articulation rule
Time Deviations	Large	Duration Contrast rule
	Structural reorganizations	Punctuation rule
	Final acceleration (sometimes)	Phrase Arch rule applied on phrase level
		Phrase Arch rule applied on sub-phrase level
		Final Ritardando

Table 7.3: Cue profiles and macro rules for anger emotion

Expressive Cue	Gabrielsson and Juslin	Macro-Rule in Director Musices
Tempo	Very rapid	Tone IOI is shortened by 15%
Sound Level	Loud	Sound Level is increased by 8 dB
Articulation	Mostly non-legato	Duration Contrast Articulation rule
Time Deviations	Moderate	Duration Contrast rule
	Structural reorganization	Punctuation rule
	Increased contrast between	Phrase Arch rule applied on phrase level
	long and short notes	Phrase Arch rule applied on sub-phrase level

Table 7.4: Cue profiles and macro rules for happyness emotion

Expressive Cue	Gabrielsson and Juslin	Macro-Rule in Director Musices
Tempo	Fast	Tone IOI is shortened by 20%
Sound Level	Moderate or loud	Sound Level is increased by 3 dB
		High Loud rule
Articulation	Airy	Duration Contrast Articulation rule
Time Deviations	Moderate	Duration Contrast rule
		Punctuation rule
		Final Ritardando rule

Table 7.5: Cue profiles and macro rules for sadness emotion

Expressive Cue	Gabrielsson and Juslin	Macro-Rule in Director Musices
Tempo	Fast	Tone IOI is lengthened by 20%
Sound Level	Moderate or loud	Sound Level is decreased by 3 dB
Articulation	Legato	Duration Contrast rule
Time Deviations	Moderate	Duration Contrast rule
		Phrase Arch rule applied on phrase level
		Phrase Arch rule applied on sub-phrase level
Final Ritardando	Yes	Obtained from the Phrase rule with the next parameter

Table 7.6: Cue profiles and macro rules for tenderness emotion

Expressive Cue	Gabrielsson and Juslin	Macro-Rule in Director Musices
Tempo	Slow	Tone IOI is lengthened by 30%
Sound Level	Mostly low	Sound Level is decreased by 6 dB
Articulation	Legato	Duration Contrast rule
Time Deviations	Diminished constrast	Duration Contrast rule
Final Ritardando	Yes	Final Ritardando rule

7.7 Modeling expressive intention in music performance: the Caro system

A musical interpretation is often the result of a wide range of requirements on expressiveness rendering and technical skills. The understanding of why certain choices are, often unconsciously, preferred to others by the musician, is a problem related to cultural aspects and is beyond the scope of this work. However, it is still possible to extrapolate significant relations between some aspects of the musical language and a class of systematic deviations. For our purposes it is sufficient to introduce two sources of expression.

- The first one deals with aspects of musical structures such as phrasing, hierarchical structure of phrase, harmonic structure and so on.

- The second one involves those aspects that are referred to with the term expressive intention, and that relate to the communication of moods and feelings. In order to emphasize some elements of the music structure (i.e. phrases, accents, etc.), the musician changes his performance by means of expressive patterns as crescendo, decrescendo, sforzando, rallentando, etc., otherwise

the performance would not sound musical.

Many works analyzed the relation or, more correctly, the possible relations between music structure and expressive patterns.

Let us call *neutral* performance a human performance played without any specific expressive intention, in a scholastic way and without any artistic aim. Our model is based on the hypothesis that when we ask a musician to play in accordance with a particular expressive intention, he acts on the available freedom degrees, without destroying the relation between music structure and expressive patterns. Already in the neutral performance, the performer introduces a phrasing that translates into time and intensity deviations respecting the music structure. In fact, our studies demonstrate that by suitably modifying the systematic deviations introduced by the musician in the neutral performance, the general characteristics of the phrasing are retained (thus keeping the musical meaning of the piece), and different expressive intentions can be conveyed.

The purpose of the CARO model, developed at CSC-DEI University of Padova, is to control in an automatic way the expressive content of a neutral performance. The CARO model adds an expressive intention to a neutral performance in order to communicate different moods, without destroying the musical structure of the score. The neutral performance can be pre-recorded or derived from the score by a music performance model, as the KTM music performance rule system (see Sect. 7.6). The functional structure of the system used as a test bed is shown in Fig. 7.19.

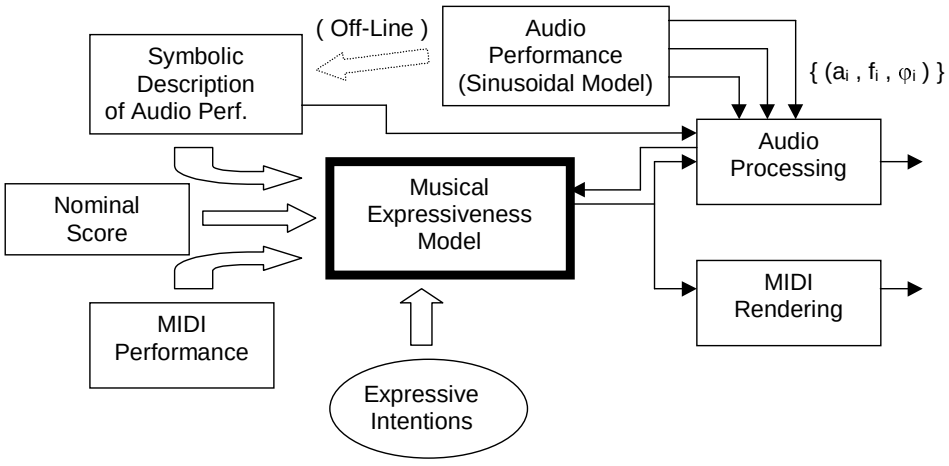
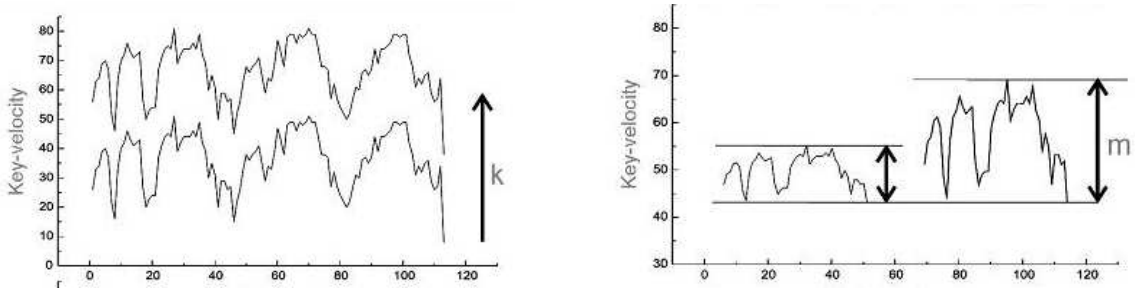


Figure 7.19: Scheme of the CARO model. The input of the expressiveness model is composed by a musical score and a description of a neutral musical performance. Depending on the expressive intention desired by the user, the expressiveness model acts on the symbolic level, computing the deviations of all musical cues involved in the transformation. The rendering can be done by a MIDI synthesizer and/or driving the audio processing engine. The audio processing engine performs the transformations on the pre-recorded audio in order to realize the symbolic variations computed by the model.

The input of the expressiveness model is composed by a description of a neutral musical performance, and a control on the expressive intention desired by the user. The expressiveness model acts on the symbolic level, computing the deviations of all musical cues involved in the transformation. The rendering can be done by a MIDI synthesizer and/or driving the audio processing engine. The audio processing engine performs the transformations on the pre-recorded audio in order to realize the symbolic variations computed by the model. The system allows the user to interactively change the

expressive intention of a performance by specifying its own preferences through a graphical interface.

7.7.1 The expressiveness model



(a) (b)

Figure 7.20: Interpretation of the expressive parameters k and m.

The model is based on the hypothesis, introduced in Section 7.7, that different expressive intentions can be obtained by suitable modifications of a neutral performance. The transformations realized by the model should satisfy some conditions: 1) they have to maintain the relation between structure and expressive patterns, and 2) they should introduce as few parameters as possible to keep the model simple. In order to represent the main characteristics of the performances, we used only two transformations: shift and range expansion/compression. Different strategies were tested. Good results were obtained by a linear instantaneous mapping that, for every P-parameter and a given expressive intention e, is formally represented by the equation:

$$P_e[n] = k_e \overline{P}_0 + m_e \left(P_0[n] - \overline{P}_0 \right) \tag{7.13}$$

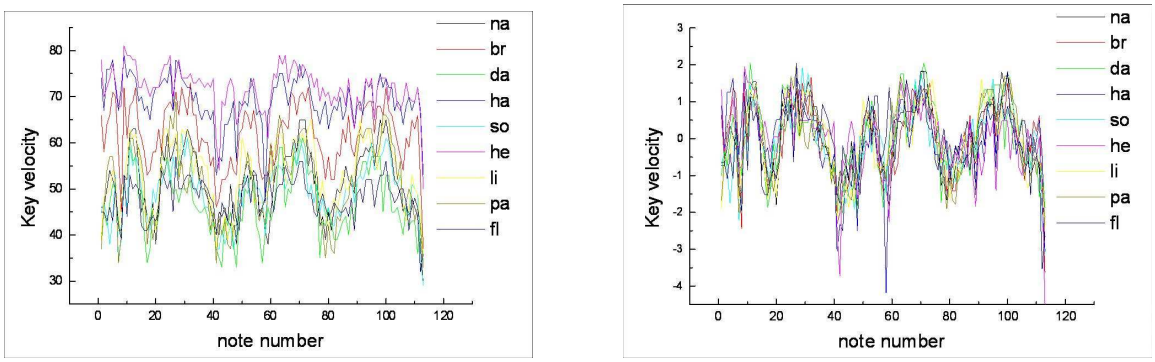
where $P_e[n]$ is the estimated profile of the performance related to expressive intention e, $P_0[n]$ is the value of the P-parameter of the n-th note of the neutral performance, $\overline{P}_0$ is the mean of the profile $P_0[n]$ computed over the entire music sequence, k_e and m_e are respectively the coefficients of shift and expansion/compression related to expressive intention (Fig. 7.20). For example the intensity $I[n]$ parameter is modified according to

$$I[n] = k_I \overline{I}_0 + m_I \left(I_0[n] - \overline{I}_0 \right) \tag{7.14}$$

where $I_0[n]$ is the intensity of a neutral performance and its average value is given by

$$\overline{I}_0 = \frac{1}{N} \sum_n I[n]$$

This choice was motivated by observing data as the ones plotted in fig. 7.21(a), where the intensity profiles of Mozart sonata K 545 (score in figure 7.2), played according diffrerent expressive intention are shown. In fig 7.21(b) the same profiles are plotted after a normalization on the mean values and variance: it can be noticed a common trend. Moreover it was verified that the expressive parameters k and m (affecting mean value and variance) are very robust for the modification of expressive intentions. I.e. a smooth change on their values correspondingly changes the expressive character of the



(a) (b)

Figure 7.21: Intensity profiles (key velocity) of 9 performances of Mozart sonata K 545, (score in figure 7.2) played according different expressive intentions: natural, bright, dark, hard, soft, heavy, light, passionate and flat (a). The same profiles plotted normalized in average and variance: a common trend emerges (b).

performance, without introducing annoying artifacts. Thus, Eq. (7.13) can be generalized to obtain, for every P-parameter, a morphing among different expressive intentions as:

$$P[n] = k(x,y)\overline{P}_0 + m(x,y)\left(P_0[n] - \overline{P}_0\right) \tag{7.15}$$

This equation relates every P-parameter with a generic expressive intention represented by the expressive parameters k and m that constitute the internal representation of the model and that can be put in relation to the time varying position (x, y) in the control space.

7.7.2 The control space

The control space level controls the expressive content and the interaction between the user and the final audio performance. In order to realize a morphing among different expressive intentions we developed an abstract control space, called perceptual parametric space (PPS), that is a two-dimensional space derived by multidimensional analysis (Principal Component Analysis) of perceptual tests on various professionally performed pieces ranging from western classical to popular music. This space reflects how the musical performances are organized in the listener's mind. It was found that the axes of PPS are correlated to acoustical and musical values perceived by the listeners themselves. To tie the control space with the model, we make the hypothesis that a linear relation exists between the PPS axes and every couple of expressive parameters $\{k, m\}$:

$$\begin{aligned} k(x,y) &= a_{k,0} + a_{k,1}x + a_{k,2}y \\ m(x,y) &= a_{m,0} + a_{m,1}x + a_{m,2}y \end{aligned} \tag{7.16}$$

where x and y are the coordinates of the PPS.

7.7.3 Estimation of the model parameters

Event, expressive and the control levels are related by equations 7.13 and 7.16. We will now get into the estimation process of the model parameters (see Fig. 7.22); more details about the relation between x, y and audio and musical values will be given in Sec. 7.7.4.

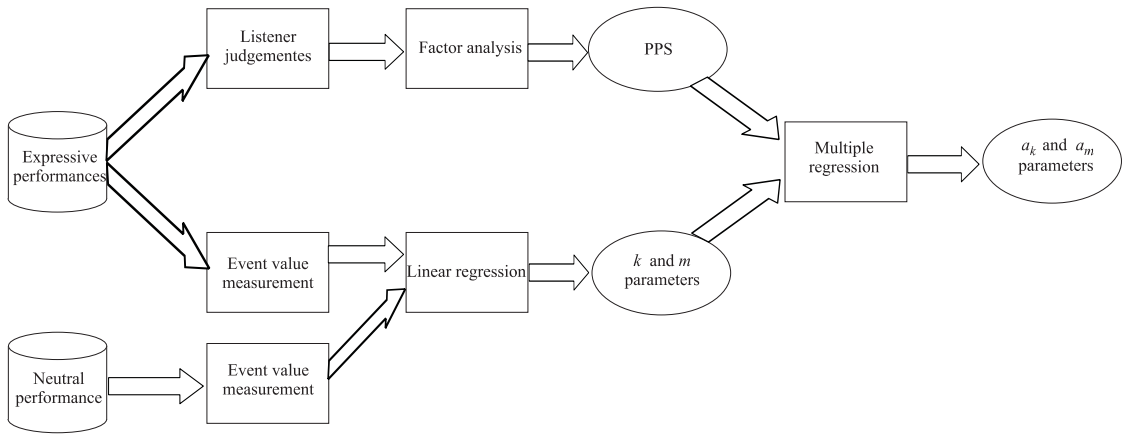


Figure 7.22: Computation of the parameters of the model.

The estimation is based on a set of musical performances, each characterized by a different expressive intention. Such recordings are made by asking a professional musician to perform the same musical piece, each time being inspired by a different expressive intention (see in Sec. 7.7.4 for details). Moreover, a neutral version of the same piece is recorded. Recordings are first judged by a group of listeners, who assign different scores to the performances with respect to a scoring table in which the selectable intentions are reported for more details). Results are then processed by a factor analysis. In our case, this analysis allowed to recognize two principal axes explaining at least the 75% of the total variance. The choice of only two principal factors, instead than three or four, is not mandatory. However, this choice results in a good compromise between the completeness of the model and the compactness of the parameter control space (PPS). The visual interface, being the two-dimensional control space, is effective and easy to realize. Every performance can be projected in the PPS by using its factor loading as x and y coordinates. Let's call (x_e, y_e) the coordinates of the performance e in the PPS. Table 7.8 in Section 7.7.4 shows the factor loadings obtained from factor analysis. These factor loadings are assumed as coordinates of the expressive performances in the PPS.

An acoustical analysis is then carried out on the expressive performances, in order to measure the deviations' profiles of the P-parameters. For each expressive intention, the profiles are used to perform a linear regression with respect to the corresponding profiles evaluated in the neutral performance, in order to obtain k_e and m_e in the model in eq. 7.13. The result is a set of expressive parameters E, for each expressive intention and each of the P-parameters. Given x_e, y_e and k_e, m_e estimated as above, for every P-parameter the corresponding coefficients $a_{k,i}$ and $a_{m,i}$ $(i = 1, 2, 3)$ of equation 7.16 are estimated by multiple linear regression, over expressive intentions.

With this procedue the model parameters can be computed from a set of sample performances. Therefore, it is possible to change the expressiveness of the neutral performance by selecting an arbitrary point in the PPS, and computing the deviations of the low-level acoustical parameters. Let us call x_p and y_p the coordinates of a (possibly time varying) point in the PPS. From eq. 7.16, for every P-parameter, $k(x, y)$ and $m(x, y)$ values are computed. Then, using equation 7.15, the profiles of event-layer cues are obtained. These profiles are used for the MIDI synthesis and as input to the post-processing engine acting at levels one and two, according to the description in the next section.

7.7.4 Results and Applications

We applied the proposed methodology on a variety of digitally recorded monophonic melodies from classic and popular music pieces. Professional musicians were asked to perform excerpts from various musical scores, inspired by the following adjectives: light, heavy, soft, hard, bright, and dark. The neutral performance was also added and used as a reference in the acoustic analysis of the various interpretations. Un-coded adjectives in the musical field were deliberately chosen to give the performer the greatest possible freedom of expression. The recordings were carried out in three sessions, each session consisting of the seven different interpretations. The musician then chose the performances that, in his opinion, best corresponded to the proposed adjectives. This procedure is intended to minimize the influence that the order of execution might have on the performer. The performances were recorded at the CSC-DEI of Padua University in monophonic digital format at 16 bits and 44.1 kHz. In total, twelve score were considered, played with different instruments (violin, clarinet, piano, flute, voice, saxophone) and by various musicians (up to five for each melody). Only short melodies (between 10 and 20 seconds) were selected, allowing us to assume that the underlying process is stationary (the musician doesn't change the expressive content in a so short time window).

Semi-automatic acoustic analyses where then performed in order to estimate the expressive time- and timbre-related cues IOI, L, AD, I, EC, BR. Figure 7.23 shows the time evolution of one of the considered cues, the intensity level I, normalized in respect to maximum Key Velocity, for the neutral performance of an excerpt of Mozart's sonata K545 (piano solo). The score was shown in fig. 7.2.

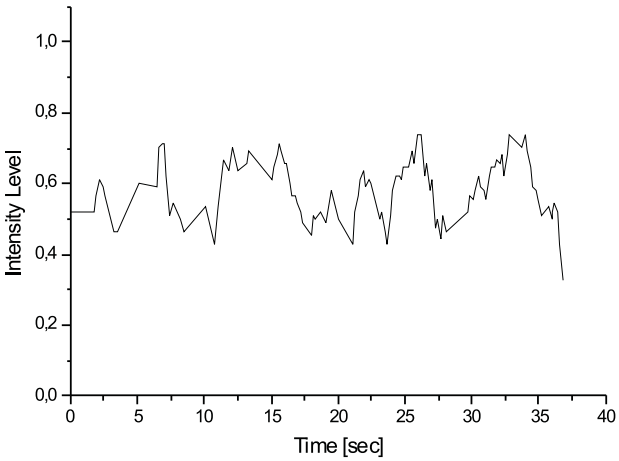


Figure 7.23: Analysis: normalized intensity level of neutral performance of Mozart's sonata K545.

Table 7.7 reports the values of the k and m parameters computed for the Mozart's sonata K545, using the procedure described in Section 7.7.3. For example, it can be noticed that the k-value of the Legato (L) parameter is important for distinguishing *hard* (k = 0,92 means quite staccato) and *soft* (k = 1,43 means very legato) expressive intentions; considering the Intensity (I) parameter, *heavy* and *bright* have a very similar k-value, but a different m-value, that is in *heavy* each note is played with a high Intensity (m = 0,70), on the contrary *bright* is played with a high variance of Intensity (m = 1,06).

The factor loadings obtained from factor analysis carried out on the results of the perceptual test are shown in Table 7.8. These factor loadings are assumed as coordinates of the expressive

Table 7.7: Expressive parameters estimated from performances of Mozart's sonata K545

	IOI		L		AD		I		EC		BR	
	k	m	k	m	k	m	k	m	k	m	k	m
Bright	0.87	0.98	0.68	0.95	0.76	0.96	1.07	1.06	0.90	0.79	1.13	0.80
Dark	1.05	1.01	1.09	1.02	0.93	1.12	0.87	1.05	1.12	1.06	0.67	0.72
Hard	0.95	0.86	0.92	1.06	0.73	0.84	1.06	0.76	0.98	1.04	1.17	0.96
Soft	1.03	1.08	1.43	0.89	1.06	1.02	0.92	1.03	1.18	1.11	0.74	1.05
Heavy	1.16	0.91	1.35	0.98	0.97	1.05	1.06	0.70	0.98	1.06	1.10	0.99
Light	0.90	0.96	0.79	1.12	1.13	1.10	0.97	1.12	0.84	0.84	0.82	1.03

Table 7.8: Factor loadings are assumed as coordinates of the expressive performances in the PPS

	Factor 1	Factor 2
Bright	0.8	0.1
Dark	-0.8	0.28
Hard	-0.4	0.6
Soft	-0.35	-0.7
Heavy	-0.75	0.5
Light	0.6	-0.5

performances in the PPS. It can be noticed that factor 1 distinguishes *bright* (0.8) from *dark* (-0.8) and *heavy* (-0.75), factor 2 differentiates *hard* (0.6) and *heavy* (0.5) from *soft* (-0.7) and *light* (-0.5). From the data such as the ones in Table 7.7 and the positions in the PPS, the parameters of Eq. (7.16) are estimated. Then the model of expressiveness can be used to change interactively the expressive cues of the neutral performance by moving in the two-dimensional control space. The user is allowed to draw any trajectory which fits his own feeling of the changing of expressiveness as time evolves, morphing among expressive intentions (figure 7.24).

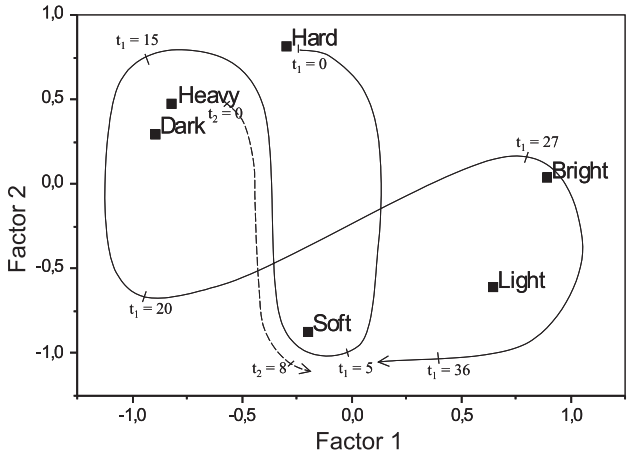


Figure 7.24: Control: trajectories in the PPS space corresponding to different time-evolution of the expressive intention of the performance. Solid line: the trajectory used on the Mozart's theme; dashed line: trajectory used on the Corelli's theme.

As an example, Fig. 7.25 shows the effect of the control action described by the trajectory (solid line) in Fig. 7.24 on the intensity level I (to be compared with the neutral intensity profile show in Fig. 7.23). It can be seen how the intensity level varies according to the trajectory: for instance hard and heavy intentions are played louder than the soft one. In fact, from the Table 7.7, the k values are 1.06 (hard), 1.06 (heavy) and 0.92 (soft). On the other hand, we can observe a much wider range of variation for light performance ($m = 1.12$) than for heavy performance ($m = 0.70$). The new intensity level curve is used, in its turn, to control the audio processing engine in the final rendering step.

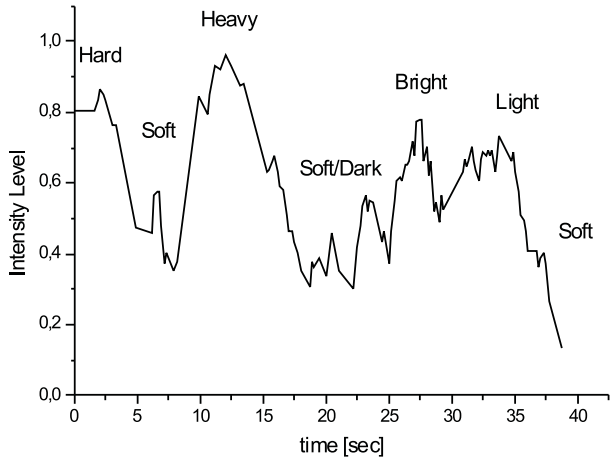


Figure 7.25: Synthesis: normalized intensity level corresponding to the trajectory in Fig. 7.24.

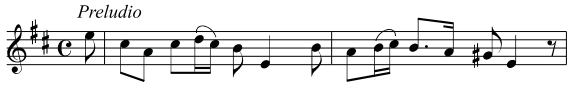
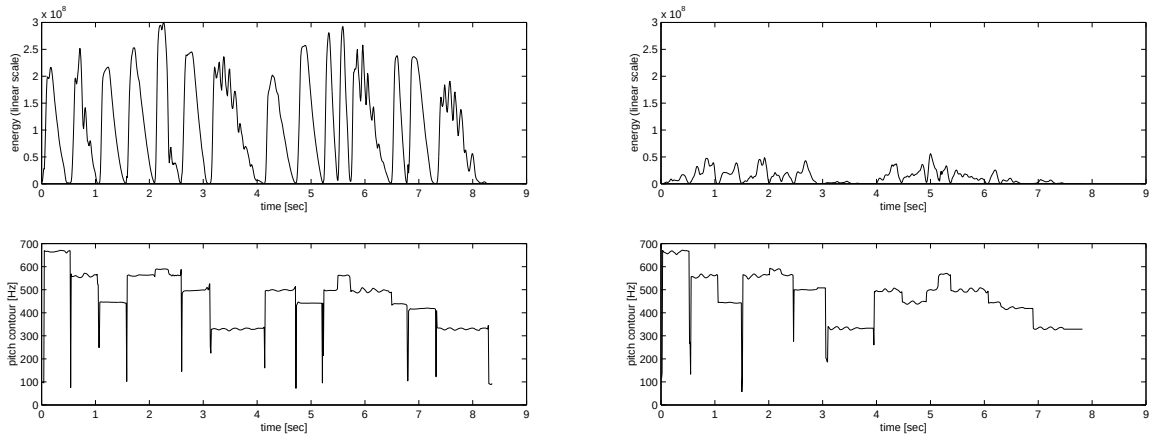


Figure 7.26: Score of the theme of Corelli's sonata op. V.

As a further example, an excerpt from the Corelli's sonata op. V is considered (Fig. 7.26). Figures 7.28(a), 7.27, show the energy envelope and the pitch contour of the original neutral, heavy and soft performances (violin solo). The model is used to obtain a smooth transition from heavy to soft (dashed trajectory in Fig. 7.24) by applying the appropriate transformations on the sinusoidal representation of the neutral version. The result of this transformation is shown in Fig. 7.28(b). It can be noticed that the energy envelope changes from high to low values, according to the original performances (heavy and soft). The pitch contour shows the different behavior of the IOI parameter: the soft performance ($k = 1.03$) is played faster than heavy performance ($k = 1.16$). This behaviour is preserved in our synthesis example.

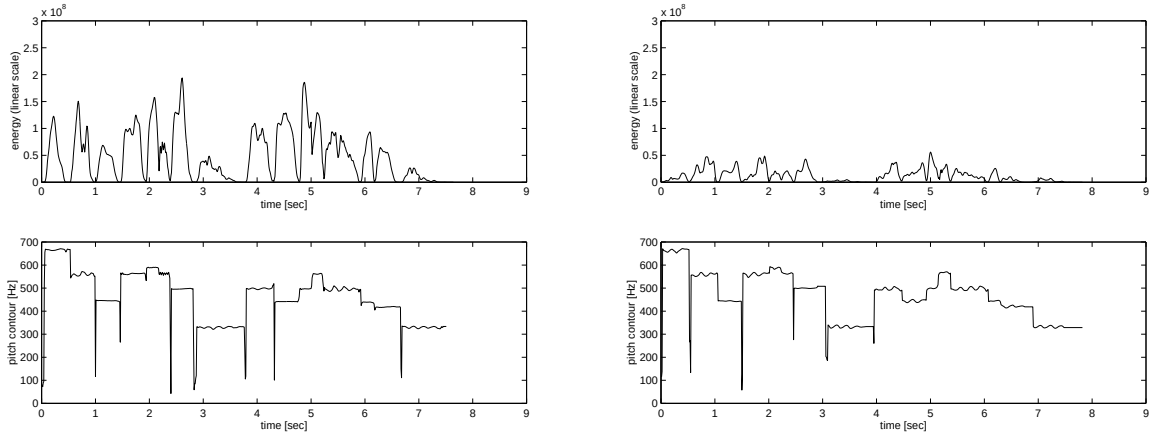
We developed an application *Once upon a time*, released as an applet, for the fruition of fairy-tales in a remote multimedia environment. In these kinds of applications, an expressive identity can be assigned to each character in the tale and to the different multimedia objects of the virtual environment (fig. 7.29). Starting from the storyboard of the tale, the different expressive intentions are located in a control spaces defined for the specific contexts of the tale. By suitable interpolation of the expressive parameters, the expressive content of audio is gradually modified in real time with respect to the position and movements of the mouse pointer, using the model describe above.

This application allows a strong interaction between the user and the audio-visual events. More-



(a) (b)

Figure 7.27: Analysis: Energy envelope and pitch contour of heavy (a) and soft (b) performance of Corelli's sonata op. V.



(a) (b)

Figure 7.28: Energy envelope and pitch contour of Corelli's sonata op. V. (a) Analysis of neutral performance. (b) Synthesis of an expressive morphing. The expressive intention changes smoothly from heavy to soft. The final rendering is the result of the audio transformations controlled by the model and performed on the neutral performance.

over, the possibility to have a smoothly varying musical comment, augments the user emotional involvement, in comparison with the participation reachable using rigid concatenation of different sound comments.

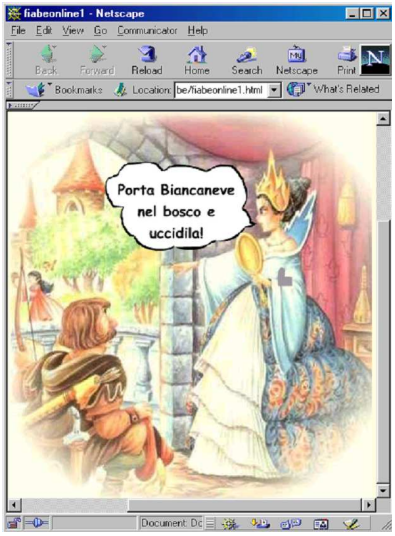


Figure 7.29: *Once upon a time*, an applet for the fruition of fairy-tales in a remote multimedia environment. Different expressive intentions are located in a control spaces defined for the specific contexts of the tale and the expressive content of audio is gradually modified in real time according mouse position.

7.8 Music and emotions

7.8.1 Introduction

In general in human communication, two channels can be distinguished: one transmits explicit messages, which may be about anything or nothing; the other transmits implicit messages about the humans themselves. A lot of research is conducted in understanding the first, explicit channel, but less attention is paid to the second is not as well understood. Understanding the other party emotions is one of the key tasks associated with the second, implicit channel. Aim of the psychology study of music and emotion is understanding the mechanisms that intervene between music reaching a listener and an emotion being perceived, or experienced, by that person as a result of hearing that music. Aim of the scientific and technological study of music and emotion, is to develop models able to describe such phenomena and systems for expression and emotion rendering and recognizing in multimodal communication.

The importance of emotional expression in music communication and its powerful impact on the listener has been recognized throughout history. It is a popular conception, sometimes true, that the composer express his present feelings in his composition. It is more likely that he uses various structural factors for coding and transmitting intended expressions, not directly associated to his present feeling. There is less agreement on how music expresses such affective content and exactly what emotions are most likely to be expressed. The related notion that music induces or produces emotions in listeners also has a venerable history but its validity is still under debate. Moreover the listener usually judge perceived expression of composed music, as realized in performance. Thus both the composed structure and the actual performance jointly contribute to expression communication. In this section, first the concept of emotion will be discussed; then we will briefly review the main topics on emotion and musical structure, followed by a discussion of emotion communication in music performance.

7.8.2 Approaches to conceptualizing emotion in psychology

The two theoretical traditions that have most strongly determined past research in this area are discrete (also called categorical) and dimensional emotion theories.

Categorical approach The assumption of this approach is that people experiences emotions as categories that are distinct from each other. Theorists in this tradition propose the existence of a small number, between 6 and 14, of basic or fundamental emotions that are characterized by very specific response patterns. From these basic emotions, all other emotional states can be derived. The focus is on characteristics that distinguish emotion from one another. There is a reasonable agreement on five basic emotions: happiness, sadness, anger, fear, disgust, surprise.

A problem with this approach is that different researchers propose different sets of basic emotions and the small number of primary basic emotions seems ill adapted to describe the extraordinary richness of the emotional effects of music reported in both fictional and scientific accounts.

Dimensional approach The use of two-dimensional valence-activation models has become very widespread in the affective sciences and is well represented in research on emotional effects of music. This approach has some obvious advantages. It is simple, easily understood by participants in experiments, and highly reliable.

The focus of this approach is on identifying emotions based on their placement on a space with a small number of dimensions. This space is derived from similarity judgments, analyzed using factor analysis or multidimensional scaling. Since the third dimension has been difficult to establish reliably in an empirical fashion via factor analyses, emotions are often defined in terms of a two-dimensional space.

The two major dimensions consist of the valence dimension (pleasant-unpleasant, agreeable-disagreeable) and an activity dimension (active-passive) sometimes also called arousal dimension. If a third dimension is used, it often represents either power or control. The most used representation is the Circumplex model of Russel (see fig. 7.30). It presents a circular structure with activation and valence dimensions. It organize emotions in term of affect appraisal (pleasant - unpleasant) and physiological reaction (high - low arousal).

This approach provides a way of describing emotional states which is more tractable than using words, but which can be translated into and out of verbal descriptions. Translation is possible because emotion-related words can be understood, at least to a first approximation, as referring to positions in activation-emotion space. Moreover this representation is useful for capturing the continuous change in emotional expression during a piece of music.

Identifying the centre as a natural origin has several implications. Emotional strength can be measured as the distance from the origin to a given point in activation-evaluation space. The concept of a full-blown emotion can then be translated roughly as a state where emotional strength has passed a certain limit. An interesting implication is that strong emotions are more sharply distinct from each other than weaker emotions with the same emotional orientation.

A problem with this approach is that specifying the quality of a feeling only in terms of valence and activation does not allow a very high degree of differentiation - qualitatively rather different states can be close neighbours in valence-activation space (e.g., panic fear and hot anger). This is particularly important in research on music, where one may expect a somewhat reduced range of both the unpleasantness and the activation of the states produced by music. In

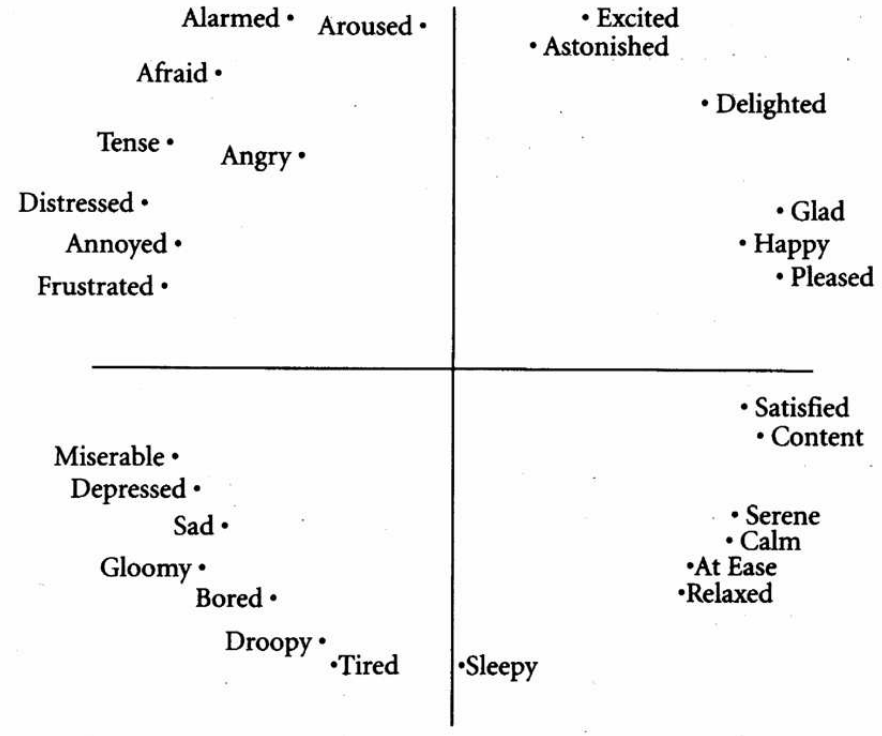


Figure 7.30: Dimensional representation of emotions: the circumplex model of Russel (1980).

consequence, adopting a valence by activation approach, asking listeners to rate their state on these two dimensions, may not allow a very fine-grained separation of the emotional effects of different pieces of music.

Eclectic approaches to describing music-induced emotions The two most popular approaches to studying emotional effects of music, asking listeners to choose between basic emotion labels or rating feeling states on positivenegative/activepassive dimensions are not optimally suited for the task. The third approach regularly encountered in this research area consists of the use of eclectic scales containing verbal affect labels deemed pertinent by a particular researcher for a particular study. Because the whole gamut of emotion labels from natural languages can be used, this approach reflects the primitives of the feeling states produced by music, being more likely to represent the immense richness and complexity of affective reactions to this form of art. In addition, there is an added advantage of flexibility labels can be freely chosen depending on the aim of the study, the nature of the music to be used, the type of listeners recruited, etc. However there is no guarantee that the labels chosen can be reliably judged, that they cover affective phenomena likely to be produced by music, or that they are organized in an economical, non-redundant manner. Most importantly, they render the comparison of data from different studies or a systematic accumulation of findings impossible.

Membership in a particular category is determined by resemblance to prototype exemplars

7.8.3 Sources of emotions in music

It is possible consider two primary sources of emotion in music, namely intrinsic and extrinsic sources.

Intrinsic Intrinsic sources are those that are non-arbitrarily embedded in structural characteristics of the music and that contribute to the creation, maintenance, confirmation, or disruption of schematic expectations. Musical events become more or less expected with reference to other musical events or structures. Expectations may reflect learning or operate at the level of primitive perceptual process, as gestalt laws of perception, where for example the movement in a particular direction creates expectation for further movement in that direction (law of good continuation).

Expectation may be created by the knowledge of (or exposition to) a large body of music, sharing a set of structural regularities, as tonal music. Moreover specific musical style (e.g. AABA song form) may create expectation. For example most compositional styles, as tonal system, have stability points. Approaching these points, tension decrease (e.g. strong beats, tonic); departures increase tension (e.g. weak beats, non-diatonic notes). On the other hand the intrinsic sources do not explain why people experience different emotions listening to the same piece.

Extrinsic. Extrinsic sources refers to causes that are not related to musical facts. Extrinsic sources are of two kinds:

- iconic sources which come about through some formal resemblance between a musical structure and some event or agent carrying emotional tone; E.g. motion, beauty, faith, tension, human character, political conditions. The word iconic evidence that the resemblance of the musical event and the non-musical referent is obvious to anyone familiar with the referent. Notice that this source is embedded in specific structural characteristic of the piece. Changing these characteristics, also the iconic meaning changes.
- associative sources which are premised on arbitrary and contingent relationships between the music being experienced and a range of non-musical factors, which also carry emotional messages of their own. For example the well known effect: *Darling, they are playing our song*.

In general it results that different sources interact in emotion communication.

7.8.4 Emotions and musical structure

There are different factors influencing the perceived emotional expression in music. When we listen to music, we evaluate the expression of a musical piece, as actually realized by the performer. Thus we have a combination of factors deriving by the composer, and coded in the score, and by the performer. In this section the main factors related to the musical structure will be presented; the factors deriving from musical performance will be presented in the next section 7.8.5. Notice that it different the emotional expression perceived, and recognized, by a listener, from his own emotional reaction, produced by the music.

The principal methods used by the researchers in studying expression in music are based on two main paradigms.

- The first one uses real music and study the reported or measured perceived expression by the listener. This approach has good ecological validity, but has the drawback of a difficult separation of possible causes.

- The second approach instead the researcher systematically varies one or more structural factors (e.g. pitch, melodic contour) in short tone sequence without musical context. In this way it is easier to separate the various factors, but ecological validity becomes problematic.
- A third strategy, trying to combine the two approaches, manipulates some structural factors (e.g. tempo, rhythm, mode) in real pieces of music. In this way the musical context is maintained, provided that the alterations do not hamper the musicality of the stimulus.

The listener evaluation of the stimuli is then analysed by multivariate techniques as factor analysis, cluster analysis in order to find a limited number of descriptive dimensions, as seen for the dimensional approach to emotion description (see section 7.8.1). The same different approaches in stimuli selection is used also for the study of expression in music performance.

7.8.4.1 Effects of structural factor

The principal structural factors affecting emotional communication are here briefly summarized.

Tempo Tempo is considered the most important. Fast tempo is associated to activity/excitement, happiness/joy, potency, surprise, anger, fear, while slow tempo to calmness, solemness, sadness, tenderness, boredom, disgust. Tempo usually refers to perceived pulse rate. Perceived speed is affected by note density, density of melodic or harmonic changes.

Mode Major mode is often associated to happiness/joy, graceful, serene, solemn, while minor mode to sadness, dreamy, dignified, tension, disgust, anger.

Loudness Loud music is associated to intensity, power, tension, anger, joy, while soft music to softness, tenderness, sadness, solemnity, fear. Large variations are associated to fear, while small variations to happiness, activity; rapid changes to playful, pleading, while few changes: sadness, peace, dignity

Pitch High pitch: happy, serene, dreamy, exciting, surprise, potency, anger, fear, activity; while low pitch: sad, dignity, vigorous; boredom pleasantness. Large variations: happiness, pleasantness, activity, surprise; while small variations: disgust, anger, fear, boredom.

Melody
- Melody range: Wide range: joy, whimsicality, uneasiness. Narrow range: das, dignified, sentimental, tranquil, delicate, triumphant.
- Melodic direction (pitch contour): Ascending melody: dignity, serenity, tension, happiness; fear, surprise, anger, potency. Descending melody: exciting, graceful, vigorous, sadness.
- Melodic motion: Stepwise motion: dullness. Intervallic leaps: excitement.

Harmony Simple, consonant harmony: happy, relaxed, graceful, serene, dreamy. Complex, dissonant harmony: excitement, tension, vigour, anger, sadness, unpleasantness.

Tonality Chromatic harmony: sad, angry. Tonal harmony: joyful, dull, peaceful. Atonal harmony: angry.

Rhythm Regular/smooth rhythm: happiness, dignity, majesty, peace. Irregular/rough rhythm: amusement, uneasiness, anger. Varied rhythm: joy. Firm rhythm: sadness, dignity, vigour. Flowing/fluent rhythm: happy, graceful, dreamy, serene

Timbre Tones with many harmonics: potency, anger, disgust, fear, activity, surprise. Tones with amplified higher harmonics: anger. Tones with few harmonics: pleasantness, boredom, happiness, sadness. Tones with suppressed higher harmonics: tenderness, sadness. String instruments: anger. Flute: peace.

Articulation Staccato: gaiety, energy, activity, fear, anger. Legato: sadness, tenderness, solemnity, softness. Amplitude envelope: Sharp envelope: anger, happiness, surprise, activity; Round envelope: tenderness, sadness, fear, disgust, boredom.

As a conclusion it can be noticed that Perceived emotional expression is a function of many factors, often interacting among them, and it not easy to separate, in analysis their effect. Actually music abounds of such kinds of interactions that make it so interesting to listen to. Expressive qualities are also related to structural changes, and transitions are seldom investigated. Moreover most studies are on western (classical or popular) music. Not much is known of non western or of contemporary music. Nor the expression appreciation by different cultures. Connections with studies of emotional expression in other areas, e.g. speech, body language would be very useful. Finally we can observe that good composers succeed in communicating expression, without an explicit, rational knowledge of how to code their message. Often their art is based on intuition and implicit knowledge. Thus, may we learn from implicit or explicit knowledge of the composer?

7.8.5 Emotions and musical performance

The musical message is partially coded in the musical structure, as codified in the score. When we listen to a piece of music, we are exposed to the contribution of both the composer and the performer. When a performer plays a piece of music, he aims to communicate the musical structure and his interpretation of the musical message. But he can add his own interpretation of the musical message as discussed in section 7.2.3. Thus expression is communicated not only by the structure of the music, but also by the way it is performed. The process can be schematized as in fig. 7.31. The performer encodes his expressive intention in musical sound, by controlling some specific cues as tempo, loudness, timbre, articulation etc.. The listener combines these cues to decode the expressive intention and arrive to a reliable judgement. While some cues are related to how the sound is produced on the instrument, they are also related to how the performer uses cues to communicate his expressive intention. The redundancy of the cues reduces the uncertainty of the judgement, partially compensating for the fact that a single cue is used for different purposes.

Figure 7.31: Communication of the expressive intention of the performer.

In the following the main expressive cues used by performers to convey some specific emotional (basic) expression are here summarized. The cues are related to tempo and timing, intensity, articulation, timbre characteristics.

Anger: fast tempo, small tempo variability, no ritardando, high intensity, staccato articulation, sharp duration contrast, sharp timbre, spectral noise, abrupt tone attack, accents on unstable notes, large vibrato extent

Sadness: slow tempo, large timing variations, final ritardando, low intensity, legato articulation, small articulation variability, soft duration contrasts, dull timbre, slow time attacks, slow vibrato

Happyness: fast tempo, small tempo variability, small timing variations, high intensity, little intensity variability, staccato, large articulation variability, bright timbre, fast tone attacks, sharp duration contrasts

Fear: fast tempo, large tempo variability, staccato, very low intensity, large intensity variations, soft spectrum, irregular vibrato

Tenderness: slow tempo, large timing articulations, final ritardando, low intensity, small intensity variability, legato, soft duration contrasts, soft timbre, slow attacks, accents on stable notes.

By interpreting the positions of emotions in the valence activation dimensional space, it is possible to hypothesize that some cues have more influence on one dimension than the other. For example, tempo, intensity and articulation seems to have more influence on activation (arousal) dimension, while timbre (possibly combined with intensity) seems to more influent on valence dimension. Medium intensity and high frequency energy are associated to positive emotions, while extreme (low or high) intensity level are associated to negative emotions.

It was studied also which cue is more effective in communicating the perceived expressiveness of the performance. The single cue combination that was most expressive, according listeners, had the following characteristics (in order of predictive strength): legato articulation, soft spectrum, slow tempo, high intensity, slow attacks. It can be noticed that such combination is similar to the cues that express tenderness and sadness. this seems to suggest that there is a close connection between sadness/tenderness and expressiveness.

7.9 Gestural Control of Sound Synthesis

adapted from Wandeley 2004

7.9.1 Introduction

The evolution of computer music has brought to light a plethora of sound synthesis methods available in general and inexpensive computer platforms, allowing a large community direct access to real-time computer-generated sound. Both signal and physical models have reached a point where they can be used in concert situations, although much research continues to be carried on in the subject, constantly bringing innovative solutions and developments.

On the other hand, input device technology that captures different human movements can also be viewed as in an advanced stage, considering both noncontact movements and manipulation. Specifically regarding manipulation, tactile and force feedback devices for both nonmusical2 and musical contexts have already been proposed. We are then in a stage where such devices and sound synthesis methods can be combined to create new computer-based musical instruments, or digital musical instruments (DMI), producing gesturally controlled real time computer-generated sound. The ultimate

goal is to design new DMIs capable of obtaining similar levels of control subtlety as those available in acoustic instruments, but at the same time extrapolating the capabilities of existing instruments.

In short, we need to devise ways to interact with computers in a musical context, i.e., to control multiple continuous parameters that allow the generation of sound in real time. This topic amounts to a branch of knowledge known as HCI. Various questions need to be addressed, such as the following

- Which are the specific constraints that exist in the musical context with respect to general HCI?
- Given the various contexts related to interaction in sound generation systems, what are the similarities and differences within these contexts (interactive installations, DMI manipulation, dancemusic interfaces)?
- How to design systems for these various musical contexts? Which system characteristics are common and which are context specific?

7.9.1.1 HCI and Music

More specifically, gestural control of computer-generated sound can be seen as a highly specialized branch of HCI involving the simultaneous control of multiple parameters, timing, rhythm, and user training.Hunt and Kirk consider various attributes that are characteristic of real-time multi-parametric control systems.

- There is no fixed ordering to the humancomputer dialogue.
- There is no single permitted set of options (e.g., choices from a menu) but rather a series of continuous controls.
- There is an instant response to the users movements.
- The control mechanism is a physical and multipara-metric device which must be learned by the user until the actions become automatic.
- Further practice develops increased control intimacy and, thus, competence of operation.
- The human operator, once familiar with the system, is free to perform other cognitive activities while operating the system (like talking while driving a car).

7.9.1.2 Interaction in a Musical Context

In order to take into account the specifics of musical interaction, one needs to consider the various existing contexts (sometimes called metaphors for musical control) where gestural control can be applied to computer music.

These different interaction contexts are the result of the evolution of electronic technology allowing, for instance, a same input device to be used in different situations: to generate sounds (notes) or to control the temporal evolution of a set of prerecorded notes. These two contexts traditionally correspond to two separate roles in music, those of the performer and the conductor, respectively. Technology has blurred the difference between traditional roles and allowed novel metaphors derived from other areas, such as HCI.

In this section, we will focus on instrument manipulation, or performer-instrument interaction in the context of real-time sound synthesis control. The approach consists on dividing the subject of gestural control of sound synthesis in four parts]:

- definition and typologies of gesture;
- gesture acquisition and input device design;
- synthesis algorithms;
- mapping of gestural variables to synthesis variables.

The goal is to analyze all four parts, which are equally important to the design of new DMIs.

7.9.2 Control of digital musical instrument

The term digital musical instrument is used to represent an instrument that includes a separate gestural interface (or gestural controller unit) from a sound generation unit. Both units are independent and related by mapping strategies.This is shown in Fig. 7.32.

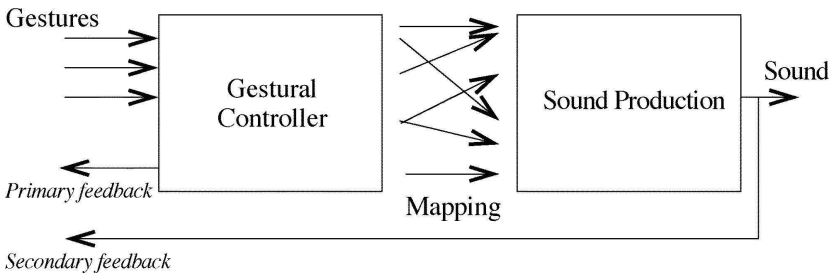


Figure 7.32: A symbolic representation of a DMI.

The term gestural controller4 can be defined here as the input part of the DMI, where physical interaction with the player takes place. Conversely, the sound generation unit can be seen as the synthesis algorithm and its input parameters. The mapping layer refers to the liaison strategies between the outputs of the gestural controller and the input controls of the synthesis algorithm.

This separation is most of the time impossible in the case of acoustic instruments, where the gestural interface is also part of the sound generation unit. If one considers, for instance, a clarinet, the reed, keys, holes, etc., are at the same time both the gestural interface (where the performer interacts with the instrument) and the elements responsible for sound generation. The idea of a DMI is analogous to splitting the clarinet in a way where one could separate these two functions (gestural interface and sound generator) and use them independently.

Clearly, this separation of the DMI into two independent units is potentially capable of extrapolating the functionalities of a conventional musical instrument, the latter tied to physical constraints. On the other hand, basic interaction characteristics of existing instruments may be lost and/or difficult to reproduce, such as tactile/force feedback.

In order to devise strategies concerning the design of new DMIs for gestural control of sound synthesis, it is essential to analyze the characteristics of actions produced by expert instrumentalists during performance. These actions are commonly referred to as gestures in the musical domain. In order to avoid discussing all nuances of the meaning of gesture, let us initially consider performer gestures as performer actions produced by the instrumentalist during a performance, meaning both actions such as prehension and manipulation, and noncontact movements. A detailed discussion is presented.

The importance of the study of gestures in DMI design can be justified by the need to better understand physical actions and reactions that take place during expert performance. Furthermore, gesture

information can also be considered as a form of signal, i.e., they can be processed, transformed, and stored using gesture editors. Gestures can also be synthesized using various models of movement or using rules in a similar way to speech synthesis.

In fact, instrumentalists simultaneously execute various types of gestures during performance. Some of them are necessary for the generation of sound, while others are not although the later are also present in most highly skilled instrumentalists performances.

One can approach the study of gestures in a musical context by either analyzing the possible functions of a gesture during performance or by analyzing the physical properties of the gestures taking place. By identifying gestural characteristics functional, in a specific context, or physiological one can ultimately gain insight into the design of gestural acquisition systems.

Regarding both approaches, one fundamental aspect is the existing feedback available to the performer, be it visual, auditory, or tactile-kinesthetic. Feedback can also be considered, depending on its characteristics, as follows.

- Primary/secondary, where primary feedback encompasses visual, auditory (clarinet key noise, for instance), and tactile-kinesthetic feedback, and secondary feedback relates to the sound produced by the instrument.
- Passive/active, where passive feedback relates to feedback provided through physical characteristics of the system (a switch noise, for instance) and active feedback is the one produced by the system in response to a certain user action (sound produced by the instrument).

7.9.2.1 Gestural Acquisition

Once the characteristics of gestures are known, it is essential to devise an acquisition system that will capture these characteristics. In the case of performerinstrument interaction, this acquisition may be performed in three ways.

Direct acquisition, where one or various sensors are used to monitor performers actions. The signals from these sensors present isolated basic physical features of a gesture: pressure, linear or angular displacement, speed, or acceleration. Each physical variable of the gesture to be captured will normally require a different sensor.

Indirect acquisition, where gestures are extracted from the structural properties of the sound produced by the instrument [33][38].Signal processing techniques can then be used in order to derive performers actions by the analysis of the fundamental frequency of the sound, its spectral envelope, its temporal envelope, etc.

Physiological signal acquisition, the analysis of physiological signals, such as EMG. Commercial systems have been developed based on the analysis of muscle tension and used in musical contexts.

Direct Acquisition. Direct acquisition is performed by the use of different sensors to capture performer actions. Depending on the type of sensors and on the combination of different technologies in various systems, different movements may be tracked.

According to B. Bongers: Sensors are the sense organs of a machine. Sensors convert physical energy (from the outside world) into electricity (into the machine world). There are sensors available for all known physical quantities, including the ones humans use and often with a greater range.

For instance, ultrasound frequencies (typically 40 kHz used for motion tracking) or light waves in the infrared frequency range. Tactile-kinesthetic, or tactual,feedback is composed of the tactile and proprioceptive senses.

Direct acquisition has the advantage of simplicity when compared to indirect acquisition, i.e., one can obtain independent streams of data representing individual control parameters. On the other hand, due to the independence of the variables captured, direct acquisition techniques may underestimate the interdependency of the various variables obtained.

Sensor Characteristics and Musical Applications. Some authors consider that most important sensor characteristics are sensitivity, stability, and repeatability. Other important characteristic relates to the linearity and selectivity of the sensors output, its sensitivity to ambient conditions, etc. A more complete analysis proposes six descriptive parameters applicable to sensors: accuracy, error, precision, resolution, span, and range.

In general instrumentation circuits, sensors typically need to be both precise and accurate, and present a reasonable resolution. In the musical domain, it is often stressed that the choice of a transducer technology matching a specific musical characteristic relates to human performance and perception: for instance, mapping of the output of a sensor that is precise but not accurate to a variable controlling loudness may be satisfactory, but if it is used to control pitch, its inaccuracy will probably be more noticeable.

In music, the use of commercially available sensors developed for other uses is the rule. Only a few researchers have proposed sensors specifically designed for musical use, for instance. Various texts describe different sensors and transducer technologies for general and musical applications.

Analog-to-Digital Conversion: For the case of gesture acquisition with the use of various sensors, the signals obtained at the sensors outputs are usually available in an analog format, basically in the form of voltage or current signals. In order to be able to use these signals as computer inputs, they need to be sampled and converted in a suitable format, usually Musical Instrument Digital Interface (MIDI) or more advanced protocols such as Open Sound Control (OSC). Various analog-to-MIDI converters have been proposed and are widely available commercially. The first examples had already been developed in the 1980s.

Concerning the various discussions on the advantages and drawbacks of the MIDI protocol and its use, strictly speaking, nothing forces someone to use MIDI or prevents the use of faster or different protocols. As already pointed out, the limiting factor regarding speed and resolution is basically the specifications of the MIDI protocol, not the electronics involved in the design.

It is interesting to notice that many existing systems have used communication protocols other than MIDI in order to avoid speed and resolution limitations. One such system is the *transducteur gestuel rtroactif (TGR)* from ACROE. Other papers have proposed different options to implement gesture acquisition interfaces, such as using hardware initially designed for audio processing.

Indirect Acquisition: As opposed to direct acquisition, indirect acquisition provides information about performer actions from the evolution of structural properties of the sound being produced by an instrument. In this case, the only sensor is a microphone, i.e., a sensor measuring pressure or gradient of pressure. Due to the complexity of the information available in the instruments sound captured by a microphone, various real-time signal processing techniques are used in order to distinguish the effect of a performers action from other factors, e.g., the influence of the acoustical properties of the room or the intrinsic properties of the instrument.

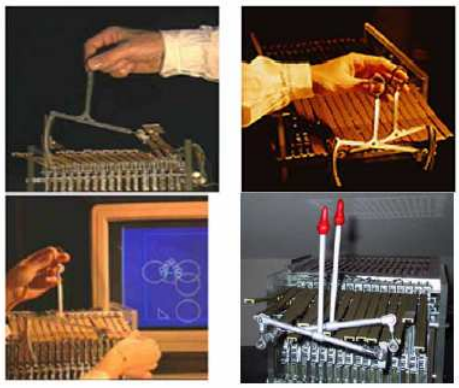


Figure 7.33: *Transducteur gestuel retroactif (TGR)* from ACROE.

Generically, one could identify basic sound parameters to be extracted in real-time as follows.

- Short-time energy, related to the dynamic profile of the signal, indicates the dynamic level of the sound but also possible differences of the instrument position with respect to the microphone.
- Fundamental frequency, related to the sounds melodic profile, gives information about fingering, for instance.
- Spectral envelope, representing the distribution of sound partial amplitudes, may give information about the resonating body of the instrument.
- Amplitudes, frequencies, and phases of sound partials that can alone provide much of the information obtained by the previous parameters.

Sampling Gestural Signals: Obviously, in order to perform the analysis of the above or other parameters during direct or indirect acquisition, it is important to consider the correct sampling of the signal. According to the Nyquist theorem, this frequency needs to be at least twice as high as the maximum frequency of the signal to be sampled.

Although one could reasonably consider that frequencies of performer actions can be limited to a few hertz, fast actions can potentially present higher frequencies. A typical sampling frequency for gestural acquisition is 200 Hz. Some systems may use higher values, up to 1 kHz and other researchers considered the ideal sampling frequency to be around 4 kHz.

7.9.2.2 Gestural Controllers

Once one or several sensors are assembled as part of a unique device, this device is called a gestural controller.

As cited above, the gestural controller is the part of the DMI where physical interaction takes place. Physical interaction here means the actions of the performer, be they body movements, empty-handed gestures, or object manipulation, and the perception by the performer of the instruments status and response by means of tactile-kinesthetic, visual, and auditory senses.

Due to the large range of human actions to be captured by the controller7 and depending on the interaction context where it will be used, its design may vary from case to case. Existing controller designs can be classified as follows:

- Instrument-like controllers, where the input device design tends to reproduce each feature of an existing (acoustic) instrument in detail. Many examples can be cited, such as electronic keyboards, guitars, saxophones, marimbas, and so on.
- Instrument-inspired controllers that although largely inspired by the existing instruments design, are conceived for another use. One example of such controller is the SuperPolm violin developed by S. Goto, A. Terrier, and P. Pierrot, where the input device is loosely based on a violin shape, but is used as a general device to control granular synthesis.
- Extended instruments are instruments augmented by the addition of extra sensors.Commercial augmented instruments included the Yamaha Disklavier, used, for instance, in pieces by J.-C. Risset. Other examples include the flute and the trumpet ,but any existing acoustic instrument may be extended to different degrees by the addition of sensors.
- Alternate controllers, whose design does not follow that of an established instrument. Some examples include the Hands, graphic drawing tablets, etc. For instance, an unorthodox gestural controller using the shape of the oral cavity has been proposed.

For instrument-like controllers, although mostly representing a simplified (first-order) model of the acoustic instrument, many of the gestural skills developed by the performer on the acoustic counterparts can be readily applied to the controller. Conversely, for a nonexpert performer, these controllers present roughly the same constraints as those of an acoustic instrument, i.e., technical difficulties inherent to the former will have to be overcome by the nonexpert performer.

Alternate controllers, on the other hand, allow the use of other gesture vocabularies than those of traditional instrument manipulation, thus being in principle less demanding for nonexpert performers. Even so, performers still have to develop specific skills for mastering these new gestural vocabularies.

These controllers can furthermore be classified into different categories. This fact can be modified by the use of different mapping strategies.

- Touch, expanded range,or immersive controllers [76],depending on the amount of physical contact required from the performer. Mulder also separates immersive controllers into internal, external, and symbolic controllers according to the possibilities of visualization of the control surface. In a different approach, Piringer classifies immmersive controllers into partial or completely immersive controllers.
- Individual or collaborative controllers, depending on whether the instrument is performed by one or multiple performers at one time.
- Metaphorical or ad hoc controllers, and so on.

7.9.3 Mapping of gestural variables to synthesis parameters

Once gesture variables are available either from independent sensors or as a result of signal analysis techniques in the case of indirect acquisition, one then needs to relate these output variables to the available synthesis input variables.

Depending on the sound synthesis method to be used, the number and characteristics of these input variables may vary. For signal model methods, one may have: 1) amplitudes, frequencies, and phases of sinusoidal sound partials for additive synthesis; 2) an excitation frequency plus each formants center frequency, bandwidth, amplitude, and skew for formant synthesis; 3) carrier and modulation

coefficients (f_c : f_m ratio) for FM synthesis, etc. It is clear that the relationship between the gestural variables and the synthesis inputs available is far from obvious.

For the case of physical models, the available input variables are usually the physical parameters of an instrument, such as blow pressure, bow velocity, etc. In this context, the mapping of gestures to the synthesis inputs seems to be more evident, since the relation of these inputs to the synthesis algorithm are directly mapped by the multiple dependencies based on the physics of the particular instrument.

7.9.3.1 Mapping for General Musical Performance

Although simple one-to-one or direct mappings are by far the most commonly used, other mapping strategies can be used. For instance, through the use of several mapping strategies, it has been shown that for the same gestural controller and synthesis algorithm, the choice of mapping strategy became the determinant factor concerning the expressivity of the instrument.

The definition of mapping strategies using instrument-like and perhaps instrument-inspired controllers can benefit from our knowledge of the physics of acoustic instrument. But in the case of an alternate controllers, the possible mapping strategies to be applied are far from obvious, since no model of the mappings strategies to be used is available. Even so, it can be demonstrated that the choice of mappings influences user performance for the manipulation of general input devices in a musical context.

7.9.3.2 Mapping as Multiple Layer

Mapping can be implemented as a single layer between controller outputs and synthesis inputs. In this case, a change of either the gestural controller or synthesis algorithm would imply the definition of a different mapping.

One way to overcome this situation is the definition of mapping as two (or possibly more) independent layers: a mapping of control variables to intermediate parameters and a mapping of intermediate parameters to synthesis variables.

This means that the use of different gestural controllers would necessitate the use of different mappings in the first layer, but the second layer, between intermediate parameters and synthesis parameters, would remain unchanged. Conversely, changing the synthesis method involves the adaptation of the second layer, considering that the same abstract parameters can be used, but does not interfere with the first layer, therefore being transparent to the performer.

The definition of those intermediate parameters or an intermediate abstract parameter layer can be based on perceptual variables such as timbre, loudness, and pitch, but can be based on other perceptual characteristics of sounds, or have no relationship to perception, being then arbitrarily chosen by the composer or performer.

7.10 Commented bibliography

The affective computing approach is presented in Picard [1997] The KANSEI information processing approach is presented in Hashimoto [1997]

A good overview on musc performance research from a psychological point of view is presented in Gabrielsson [1999] and Palmer [1997]. A good example of analysis by measurement is Repp [1992]. Models based on machine learning approach are used by Widmer (see e.g. Widmer [1996] and later

literature of OFAI group). Models based on Case based reasoning were first proposed by Arcos and Mantaras (see e.g. Arcos and de Mantaras. [2001]).

The most widespread computer environment for realizing live electronic music are the Max paradigm based languages Max/MSP, jmax, and Pd (see e.g. Puckette [1991]) (Puckette 1997, 2002) and Eyesweb platform (Camurri et al. [2000]). Many new input devices for music expression have been proposed (see e.g. Paradiso [2000], Wanderley and Orio 2002).

The dynamic model of musical expression was presented in Todd [1992]. KTH music performance rules are described in detail in Friberg [1991]. The Caro system is presented in Canazza et al. [2004]. The reference book on music and emotion is Juslin and Sloboda [2001].

References

Josep Lluis Arcos and Ramon Lopez de Mantaras. An interactive case-based reasoning approach for generating expressive music. *Applied Intelligence*, 14(1):115–119, 2001.

A. Camurri, Hashimoto S., Ricchetti M., Trocca R., Suzuki K., and Volpe G. Eyesweb: Toward gesture and affect recognition in interactive dance and music systems. *Computer Music Journal*, 24(1):57–69, 2000.

S. Canazza, G. De Poli, C. Drioli, A. Roda, and A. Vidolin. Modeling and control of expressiveness in music performance. *The Proceedings of the IEEE*, 96(4):286–301, 2004.

A. Friberg. Generative rules for music performance. *Computer Music J.*, 15(2):56–71, 1991.

A. Gabrielsson. The performance of music. In D. Deutsch, editor, *The psychology of music*, pages 501–602. Academic Press, 1999.

S. Hashimoto. Kansei as the third target of information processing and related topics in japan. In *Proceedings of the International Workshop on KANSEI: The Technology of Emotion.*, pages 101–104, 1997.

P. Juslin and J. Sloboda. *Music and EmotionTheory and Research*. Oxford Univ. Press, 2001.

C. Palmer. Music performance. *Annual Review of Psychology*, 48:115–138, . 1997.

J. Paradiso. New ways to play: Electronic music interfaces. *IEEE Spectrum*, 34(12):18–30, 2000.

R. Picard. *Affective Computing*. MIT Press, 1997.

M. Puckette. Combining event and signal processing in the max graphical programming environment. *Computer Music J.*, 15(3):68–77, 1991.

B. H. Repp. Diversity and commonality in music performance: An analysis of timing microstructure in schumanns rumerei. *J. Acoust. Soc. Am.*, 92(5):2546–2568, 1992.

N.P. Todd. The dynamics of dynamics: A model of musical expression. *Journal of the Acoustical Society of America*, 91 (6):3540–3550, 1992.

G. Widmer. Learning expressive performance: The structure-level approach. *Journal of New Music Research*, 25(2): 179–205, . 1996.

Contents