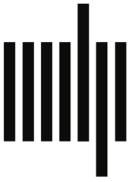
Embodied Music Cognition and Mediation Technology

Marc Leman



The MIT Press

From The MIT Press

MITCogNet

This book was set in Sabon on 3B2 by Asco Typesetters, Hong Kong. Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Leman, Marc, 1958–
Embodied music cognition and mediation technology / Marc Leman.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-262-12293-1 (hardcover : alk. paper)
1. Music—Psychological aspects. 2. Music—Physiological aspects. 3. Musical perception. I. Title.
ML3800.L57 2007
781′.11—dc22 2006035169

10 9 8 7 6 5 4 3 2 1

6 Interaction with Musical Instruments

Corporeal articulation and intentionality, two concepts introduced in the previous chapters, provide a framework for understanding musical involvement. They allow us to understand the direct impact of physical energy on the human body and the subsequent translation of this sensing into action-oriented values, meanings, goals, and mental representations. A central part of the embodied music cognition theory concerns the role of the human body as a mediator between the human mind and forms of the physical environment. It provides an understanding of how the physical energy of music can be related to an ontology of action-oriented behavior and how a subsequent conceptualization of this relationship may lead to verbal descriptions of musical experience and further use in symbolic communication.

Now the question is how the above theory of musical involvement can contribute to practical applications in mediation technology, that is, to tools that can provide access to (technologically encoded) music. Two chapters aim to give a provisional answer to this question. The present chapter is about how human subjects can enhance artistic expressiveness with the help of interactive music systems. Chapter 7 is about how human subjects can search for and retrieve music from a database.

In both chapters, it is assumed that machines may be furnished with tools that become extensions of the human body. In addition, these tools may be conceived of as more or less independent agents with which humans can interact at a level in agreement with the assessment of intentional actions and nonverbal corporeal communication. Flexible human–machine communication is still far beyond the capabilities of any existing system, yet the concept of intentional verbal and nonverbal interaction is an important and useful metaphor in human–machine communication research. It draws upon the idea that a machine could be developed as a *you* (a second person, or social agent) with whom the

subject interacts and communicates. In both cases corporeal articulations and verbal descriptions play an important role. This chapter and the next show how the human body, as a natural mediator between the mental world and the world of physical entities, can be extended with technology, thus allowing a more flexible access to music encoded and stored in machines.

An important aspect of technology-based mediation is concerned with interactivity. Playing a musical instrument is an interactive activity, and the musical instrument can be seen as the technology which mediates between human mind (musical ideas) and physical energy (music as sound). In recent decades, acoustic instruments have evolved to electronic instruments to music environments and multimedia platforms. In this development, interactivity has been a central concern, as well as one of the driving forces for innovation in music research and far beyond.

In this chapter, interactivity is investigated in relation to music systems. It is assumed that the foundations of interaction can be found in acoustic instruments and that the detailed study of the way in which performers handle acoustic instruments may reveal basic components of an embodied interaction and communication pattern that can be exploited for the development of electronic interactive systems.

This chapter has four sections. In the first section, involvement with music is considered in terms of multimodal experiences and multimedia technology. It defines a global background for the second section, in which a model of musical communication is worked out. This model starts from a case study in which the relationship between playing an acoustic musical instrument and corporeal attuning to the resulting sound is studied in detail. The third section is concerned with an analysis of the constraints that define musical communication. These constraints are related to problems encountered in electronic instruments, environments, and multimedia systems. Finally, in the fourth section, the model and its implied constraints of musical communication are applied to a concrete example of an interactive multimedia system.

6.1 Multimodal Experience and Multimedia Technology

The idea that musical communication involves all senses, and therefore is a multimodal experience, is not new. It is a central concept of ethnomusicology, in which music and sound are seen as part of a multitude of energies and events having social and cultural signification (Merriam,

1964). In most cultures, music is integrated with dance and social/cultural functions. Music forms part of visual and tactile events, such as actions, movements, and interactions between performers and listeners. Moreover, even if the music is limited to a single energetic channel such as audio (as in radio, CD, or iPod), then the musical experience can still be said to be a multimodal experience. Music moves the body, evokes emotional responses, and generates associations with spaces and textures. Music as sound involves all senses, but often music is also embedded in other physical energies that have an impact on how music is experienced.

In Western culture, there has been an interesting development toward an integrated use of different media technologies in the production of music. In the early seventeenth century, Florentine humanists introduced the concept of *opere* (works) as a new and strategic attempt to integrate media of expression—such as singing, reciting, movement, and musical accompaniment—in close synergy with each other. This set the stage for an important development which in the nineteenth century culminated in Wagner's concept of *Gesamtkunstwerk*, the unification of all the arts into a single medium of artistic expression (Packer and Jordan, 2001). In the 1950s, with the development of electronic equipment, composers of avant-garde music explored the integration of electronic media technologies in musical spectacles and music theater. This was based on the use of tape players, film projectors, light organs, and other devices that allowed the electronic manipulation of reality. Popular musical culture also became increasingly involved with electronic media. Meanwhile, most pop music concerts became spectacles of sound, drama, dance, light, smells, and tactile modalities, using electrically powered multimedia engines. Recent developments indicate that avant-garde music moves in the direction of multimedia performances and virtual reality. The link between multimodal experiences and multimedia technologies can be seen as an extension of a long tradition that puts the human body (again) at the center of musical activity, as it is in many non-Western music cultures.

The connective thread in these developments seems to be the desire to enhance the expressive power of music, using technology as a means. For that purpose, the integration of sound with other types of energy can be seen as enhancing the effect of peak experiences, of being immersed within the music (see section 1.1).

Of particular interest in this context is the increase in stimuli levels, which runs parallel with the development of multimedia. During the

course of the nineteenth century and the first half of the twentieth century, orchestras grew in size, and musical instruments increasingly produced more energy as manufacturers broadened their sonic and dynamic ranges (Sabbe, 1998). In the second half of the twentieth century in particular, with the advent of modern dance music, the sound intensity at concerts came to have an ever-increasing range.[1]

In fact, both the increase of sound intensity and the integration of multiple media are likely to facilitate peak experiences. They are examples of the human predilection for using music to become totally immersed in energy. Other approaches include phenomena such as trances and drugs. In what follows, however, I restrict my account to technology, and in particular to interactive technologies, whose development can be seen as an extension of the human body to reach peak experiences.

6.1.1 Multimedia Micro-integration

Before going deeper into the interactive aspect, it is of interest to mention that the contribution of modern digital technology to music and multimedia is particular. Compared with previous stages in the history of multimedia and music, the main novelty of modern digital technology is concerned with the encoding, exchange, and integration of energy, using different levels of description. In the modern concept, multimedia are no longer conceived of as a juxtaposition, the placing together, both synchronically and diachronically, of different media related to sound, acting, decor, and lighting, but rather as a micro-integration, or close linkage, of different media. This micro-integration is possible because of computational platforms that allow the processing of different media at different levels of description which are mutually exchangeable, from low-level descriptions of physical energy to high-level content-based descriptions of artistic expressiveness.

Micro-integration is an important concept because it offers new opportunities for artistic exploration. For example, it allows the parameters of musical expressiveness to be extracted from one modality, say sound, and then to be reused in another modality, for example, in computer animation, where it is used to modify the expressive movement of an avatar on a screen (Mancini et al., 2006). All this can be realized in real time, and the computational platform can be configured so that it acts as an independent agent or a virtual environment with which the artist can interact.

Micro-integration allows humans to communicate with machines that extrapolate, enhance, or transfer aspects of our multimodal experience in real and virtual environments. Mixed and virtual realities use machine technology to cope with the multimodal nature of corporeal articulations, intentions, expressions, and expressiveness. Thanks to the availability of digital electronic technology (sensors, computers) and software applications, a whole new area for artistic exploration has become available. In this, the interaction between multimodal experiences and multimedia technology plays a key role. How should we conceive that interaction, and how does it relate to the human body? What are the invariant components in that interaction, and what are the constraints that confine it?

6.1.2 Approach

In this chapter, multimodal experience and multimedia technology are related to the theory of corporeal articulation, thus providing a foundation for practical applications in musical interaction in terms of embodied cognition. This is conceived from two different viewpoints: (1) in terms of a human subject, which stresses the multimodal component in interaction, and (2) in terms of an interactive technology, which stresses the multimedia component in mediation.

- The multimodal aspect of musical interaction draws on the idea that the sensory systems—auditory, visual, haptic, and tactile, as well as movement perception—form a fully integrated part of the way the human subject is involved with music during interactive musical communication. It is hypothesized that through corporeal articulation, multimodal experience of music (through movement, vision, audio) is translated into components of our subjective action-oriented ontology, and vice versa. Corporeal articulation should thus be seen as a unified principle that links mental processing with multiple forms of physical energy.
- It is straightforward to assume that any technology which mediates between mental processing and multiple physical energies should be based on multimedia, that is, on tools that take into account the different ways energy manifests itself as a function of human interaction. These tools can function as an extension of the human body, the natural mediator between musical energy and mental representations. In what follows, musical instruments are conceived of as multimedia mediators. They evolved from acoustic musical instruments to electronic music

instruments to multimedia environments and multimedia platforms.[2] They rely on common principles of human interaction and mind/body/matter transitions. The main problem to be solved is why, and to what extent, multiple media (e.g., audio, visual, haptic, tactile) are necessary for music mediation, and how these media can be designed such that they cope with principles of human interaction.

In what follows, I show that principles of multimodal and multimedia human/technology interaction can be studied in acoustical music instruments and that the implied principles are relevant to the development of electronic musical instruments and environments, and multimedia platforms.

6.2 The Communication of Intended Action

In this section, music interaction is considered in terms of the communication between a musician and a listener, using a musical instrument as mediator. The relevant questions are to what extent this communication can be measured and to what extent we can infer from it a model of musical communication and interaction that can be extended in the electronic domain.

In what follows, I start the discussion with a case study which suggests that musicians may encode gestures in sound, while listeners may decode particular intended aspects of these gestures through corporeal resonance behavior. The study suggests that these encoding/decoding processes may enable the communication of intended actions.

Since the mid-1990s, the study of musical intentions has focused on the study of musical expressiveness and, related to that, the study of gestural control and performer nuances (e.g., Dannenberg and De Poli, 1998; Widmer, 2001; Widmer and Tobudic, 2003; De Poli, 2004; Widmer and Goebl, 2004). In several experiments, it was shown that the musician's intentions are reflected in the sound structure and the cues that can be extracted from it, such as timing, articulation, loudness, and sound color (e.g., Canazza et al., 1997a, 1997b). Research on musical expressiveness has been stimulated by studies that envision the automatic performance of a musical score. A piece of music played without expression sounds dull and boring, but with expressiveness added, the music may acquire a more lively character. The idea is to capture expressiveness in performance rules, and to use these rules to drive the synthesis of musical scores (Friberg, Colombo, et al., 2000; Sundberg et al., 2003;

Zanon and De Poli, 2003). Studies also show that listeners are capable of capturing important aspects of the intended expressive meaning in music. They may recognize what kind of expressive intention the musician wanted to communicate (Gabrielsson and Juslin, 2003). This finding suggests that the musician can encode particular intentions which the listener can decode. More particularly, these message may relate to something as elusive as expressiveness.

What is not clear, however, is how decoding could work as a mechanism, and how the listener may have knowledge of the particular intention that is encoded in the music. In most studies on musical expressiveness, the underlying assumption is that the listener has in mind a representation of the symbolic structure of the music—as a score, say—which is used as a reference frame to compare performance nuances. From that disembodied comparison, the listener would be able to infer the expressive character of music.[3] This approach assumes that listeners hold a cognitive map of deviation patterns which they use for the recognition of expressiveness in the music. A second hypothesis has been proposed by Juslin and Laukka (2003), who claim that the expressive code used in music can be derived from that used in speech.

I propose a third account, based on the idea that the listener is able to decode aspects of the performer's expressive intentions on the basis of corporeal resonances with the implied moving sonic forms. This account assumes that perception and understanding of musical expressiveness is based on corporeal resonance behavior which relates sound energy to the subjective action-oriented ontology. The decoding process can be effective if the encoding process is effective as well, that is, if the composer and performer succeed in translating aspects of the subjective action-oriented ontology into sound energy. Evidence for this theory would consist in showing possible links between the intentions of the performer's gestural control and the intentions of the listener's embodied perception. This idea is studied in more detail in the following case study.

6.2.1 *Guqin* Music

The case study is based on an analysis of *guqin*-playing. The *guqin* is a Chinese plucked string instrument of the zither family, considered to be the oldest such Chinese instrument, with a history of about three thousand years. The *guqin* can be roughly described as an instrument which is played by plucking the string with the right hand and by manipulating the string with the left hand in order to produce different pitches.[4] The

guqin is very suited for this type of study because the sound reflects the player's subtle gestural control in a direct way. In contrast with the violin or guitar, the *guqin* involves no other mediating device, such as a bow or frets.

In *guqin*-playing, as with many other instruments, the short-term goal of a performance is to produce a sequence of tones. All movements that contribute to the formation of a tone can be considered as constituent of a tone gesture. The whole piece can be conceived as being produced by a sequence of such tone gestures. However, when looking at individual tones, in particular tones with pitch-sliding, it becomes clear that the tone gesture itself consists of several more elementary movements (Li and Leman, submitted). For example, a single *guqin* tone may first go up in pitch, then go down in pitch, and end with a vibrato (fast up-and-down of pitch). Gestural control is then accomplished by moving the left thumb on the string from left to right, and from right to left, followed by a rapid repetitive movement from left to right and back. Each movement, from left to right or from right to left, can be considered an elementary movement.

Figure 6.1 shows a short fragment of a piece of *guqin* music in Western notation. In this score, arrows are used to indicate pitch-sliding effects. Sonic entities that define a *guqin* tone are described using one or more notes, and bars should be read as rough indications of the meter. The tones are numbered from 1 to 20. Table 6.1 provides a summary of how these tones were produced: which finger was used, which string was played, whether the string was pressed/stopped by left-hand fingers or

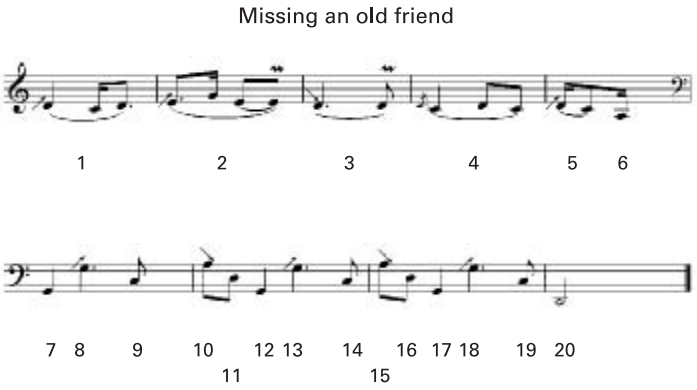


Figure 6.1
Western notation of a short *guqin* piece titled "Missing an Old Friend."

Table 6.1
Description of twenty sound entities (called tones) in terms of control characteristics

Tone	Finger	String	Open/ Stopped	Pitch movement	Finger movement (left hand)	Type of gesture (left hand)
1	1	7	S	UVDUV	RVLRV	E
2	1	7	S	UVUDV	RVRLV	E
3	1	7	S	DVV	LVV	E
4	1	7	S	DVUVDV	LVRVLV	E
5	1	7	S	UDV	RLV	E
6	4	7	S	/	/(press only)	/
7		4	O			AP
8	4	6	S	UV	RV	E
9	4	6	O			EP
10	4	7	S	D	L	E
11	4	7	O			E(A)
12		4	O			AP
13	4	6	S	UV	RV	E
14	4	6	O			EP
15	4	7	S	D	L	E
16	4	7	O			E(A)
17		4	O			AP
18	4	6	S	UV	RV	E
19		6	O			
20		2	O			

Note: The number of each tone corresponds to the number in the score of figure 6.1.
4, left hand ring finger (*Ming*); 1, left thumb; S, stopped string; O, open string; U, pitch up; D, pitch down; V, vibrato (pitch, hand); R, movement to the right; L, movement to the left; E, effective movement; P, preparatory movement; A, ancillary movement.

not (open), if pitch goes up or down, if the finger goes left or right, and if gestural control is executed by the left hand.

The gesture that produces a sliding tone will typically consist of a sequence of coordinated movements of left hand and right hand. For example, to play tone 1, the finger of the right hand will provide energy by plucking the string at onset time. The thumb of the left hand (finger 1) will press the string and move to the right, so that the string becomes shorter and pitch goes up to note d. This will be followed by a rapid alternation from left to right, that is, of shortening and lengthening, which results in a short vibrato. Then the thumb will move to the left, so that the string becomes longer and pitch goes down to note c, and then again to the right (note d), which again involves a vibrato.

6.2.2 Corporeal Articulations and Elementary Movements

In what follows, gestural control of the *guqin* is analyzed in terms of its constituent elementary movements, so that it can be related to the corporeal attuning of listeners.

Elementary Movements

An elementary movement can be defined as the movement of a body part between two points in space (Gibet et al., 2003). This movement is characterized by a bell-shaped velocity pattern (figure 6.2). In a typical pointing task, where a finger moves from point A to point B, the velocity of the finger will increase, reach a maximum, and then decrease in order to arrive at point B. Moving a finger on the string of a *guqin* can be considered a pointing task. In this example, the asymmetric form of the bell-shaped curve of figure 6.2 may reflect the fact that the initial (more or less linear) part of the movement is largely preprogrammed, while the second (nonlinear) part is influenced by sensory feedback after touching the string shortly after the maximum speed is reached. In general, while playing, the movements of the left finger on the string of a *guqin* can be considered as a multipoint movement constrained by sensory feedback. In essence, playing is then reduced to a sequence of elementary movements.

Gesture and Action

Whether an elementary movement can be considered a gesture (that is, a movement with a defined meaning that stands on its own) or a gesture component (that is, a movement which forms part of an action) depends on the level at which one looks at the gesture. Clearly, a displacement of

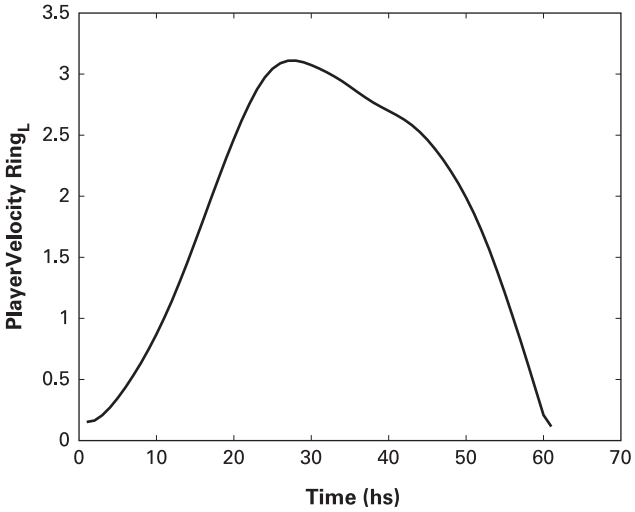


Figure 6.2
Elementary movement. The horizontal axis represents time in hundredths of a second, the vertical axis represents velocity (in relative units). The curve represents the movement of a finger in a more or less straight line from point A to point B.

the finger from point A to point B is often just an element in a sequence of movements which together characterize the gestural control of a *guqin* tone. When a gestural control has a particular goal—to play a *guqin* tone—it is called an action.

Monitoring Gestural Control in *Guqin* Music

The displacement of a marker attached to a particular body part (e.g., finger, wrist, elbow, head) can be monitored with an infrared camera system (figure 6.3).[5] From this displacement in three dimensions, it is possible to derive the velocity (first derivative), and thus to obtain a displacement and velocity curve of each marked body part. The top panel of figure 6.4 shows the displacement curve and the velocity curve of the (left) thumb during the first part of the piece. The top panel of figure 6.5 shows the displacement curve and the velocity curve of the (left) ring finger[6] during the second part of the piece.

Segmentation of Gestural Control

Given that playing music is a continuous activity, the camera will record a continuous displacement from which a continuous velocity pattern will be derived. In view of the concept of elementary movements, this

Figure 6.3
Infrared recording equipment and *guqin* player.

continuous velocity pattern can be interpreted as a sequence of bell-shaped forms. The zero crossings, the points at which the curve goes through the horizontal line, suggest a segmentation of the continuous movement in terms of elementary pointing movements.

Zero crossings can be used as a criterion to segment the measured velocity of a marker into elementary velocity patterns.[7] In figures 6.4 and 6.5, only large movements have been segmented. Smaller movements, which are typical for vibrato, are neglected. In fact, it is straightforward to consider vibrato as a self-contained movement after all.

Sonogram of *Guqin* Music

The bottom panels of figures 6.4 and 6.5 show sonogram representations of the *guqin* music. The vertical axes show frequency in hertz, and the horizontal axes show time (amplitude is in black). The onset of each tone generates a short burst of energy over all frequencies. This can be used as a segmentation marker for tones. The tone numbers are written below each sonogram.[8] The dashed vertical lines indicate segmentation marks of the gestural control (displayed in the top panel). On the sono-

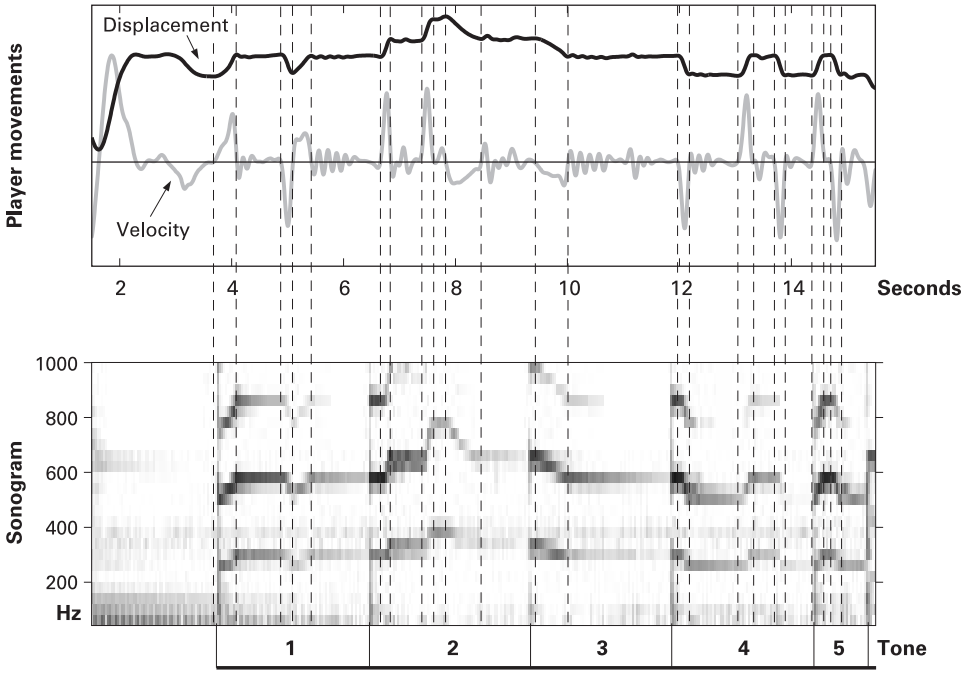


Figure 6.4
Player movements and sonogram of the first part of "Missing an Old Friend" (figure 6.1). The top panel shows the displacement and the velocity of the thumb. The bottom panel shows the sonogram. Vertical lines indicate segments at zero crossings of the velocity curve. To facilitate reading, only large movements have been segmented. Vibrato is not segmented.

gram representation they clearly mark the difference between gestural control and sonic output.

6.2.3 Gestural Control and Its Effect on Sound

A tone can be conceived as the sonic encoding of an action. It starts with the pluck, and all pitch shifts within the tone are considered part of the tone.

In order to have a better view of the relationship between the player's action and the resulting tones, tones have been analyzed according to the descriptions summarized in the legend of table 6.1. The following considerations should be taken into account:

- A string is played open (O) or stopped (S). An open string is often plucked by the right hand (R). In that case, the left hand (L) does not

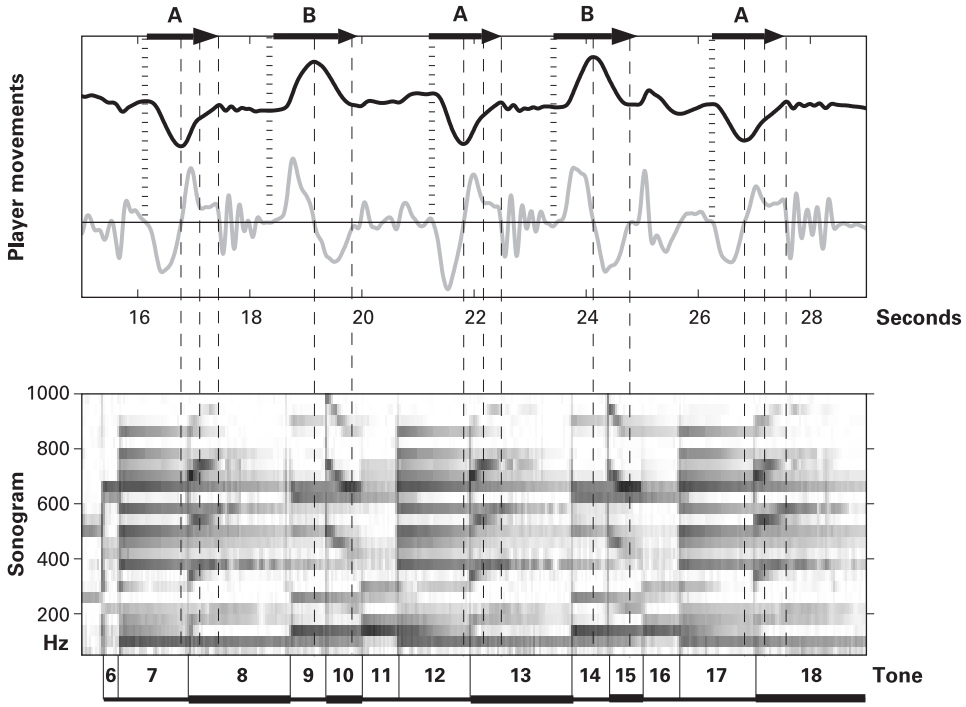


Figure 6.5
Player movements and sonogram of the second part of "Missing an Old Friend" (figure 6.1). The top panel shows the displacement and the velocity of the ring finger. The bottom panel shows the sonogram. Vertical lines indicate segments at zero crossings of the velocity curve. To facilitate reading, only large movements have been segmented. Vibrato is not segmented. The arrows in the top panel indicate control gestures that contain countermovements. (See text for further explanation.)

touch the string. The resulting pitch is then only the pitch of the string. However, there are cases where an open string is plucked by the left hand as well.

- In this piece, the stopped string is played by two fingers of the left hand, the thumb (finger 1) and the ring finger (finger 4). The first five tones are controlled by the thumb and the other tones are mainly controlled by the ring finger. In all stopped tones of the example, the left hand controls the pitch.
- The *guqin* has seven strings, numbered from 1 to 7, which correspond to the notes c, d, f, g, a, c, and d.

- The vibrato (V) is defined as a sequence of fast up-and-down pitch (or hand) movements. It is considered as a self-contained movement (or hand movement).
- The pitch of stopped tones can go up (U) or down (D). This corresponds to the movement of the finger to the right (R) or to the left (L).
- The gesture made by a finger of the left hand is called effective (E) when it actually produces a tone. It is called preparatory (P) when it is preparing the production of a tone. A movement is called ancillary (A) when it is neither effective nor preparatory.

Using these labels, the musical fragment can be described as follows:

- Tone 1 is played by the thumb (finger 1) on string 7. As shown in the sonogram of figure 6.4, the tone first goes up in pitch, followed by a vibrato. Then the pitch goes down and up and ends in a vibrato. The corresponding movement can be seen in the displacement curve (top panel). The thumb moves to the right, makes a vibrato, moves to the left and then to the right, and finally makes a vibrato. The velocity curve shows the change of movement. Displacement to the right (upward on the displacement curve) corresponds to positive bell-like curves, while displacement to the left (downward on the displacement curve) corresponds to negative bell-like curves. The horizontal line corresponds to zero velocity (no displacement). Note the interesting negative curve just before the start of the first tone, right where the arrow indicates that this is the velocity curve. This shape indicates a typical preparatory movement which anticipates the effective gesture. As can be seen on the displacement curve, the anticipation movement is in the opposite direction from the effective movement. Note also that the onset of tone 1 starts a fraction of a second later than the actual start of the upward movement. In *guqin*-playing, this technique is known as hidden head sliding. It is used to avoid the emphasis on the beginning of the tone, and thus produces a light, smooth, and gentle tone quality. The same technique, often more pronounced, is used in tones 5, 8, 10, 13, 15, and 18 (figure 6.5, top panel).
- Tones 2 to 5 can be described as similar to tone 1. These tones have no preparatory movements, which can be explained by the fact that their pitch starts at the pitch level on which the previous tone ended. The player already knows the starting position.
- Tone 6 (figure 6.5) has a stable pitch. This tone is plucked by the left thumb while the left ring finger is pressed 10 cm to the left of the left

thumb on the same string. There is very little movement because the ring finger is about to press on the string, and it is very close to the pressing point.

- Tone 7 has a stable pitch which results from plucking open string 4, using a finger of the right hand. The left ring finger first makes ancillary movements but during the second half of the tone, it anticipates the generation of the next tone.
- Tone 8 goes up in pitch and is followed by a vibrato. This tone is produced by the left ring finger on string 6, by a movement which goes to the right and is followed by a rapid left/right movement. This movement is effective for sound control. Note that the first part of the movement has been anticipated by a fraction of a second. Indeed, the tone starts when the velocity of a movement to the right (upward) is at top speed. The pitch shift consists of two parts, a rapid rise followed by a slower rise. This is reflected in the displacement curve as well as in the velocity curve. In order to be able to perform this subtle gestural control, the movement had been prepared at the end of the previous tone. The anticipatory movement seems to start from the position where the tone will end, and the time to perform this countermovement is about equal to the time to perform the movement that allows the finger to touch the string and thus to generate the sound.
- Tone 9 has a stable pitch which results from plucking string number 6 with a finger of the right hand. Meanwhile, the movement of the left ring finger is anticipating the production of tone 10.
- Tone 10 goes down in pitch. It is controlled by the left ring finger on string 7, through a movement to the left. The gesture is effective for the sound production. Also, the tone starts when the speed of the movement to the left (downward on the displacement curve) reaches its maximum. This gesture is also anticipated during the sounding of the previous note, and the anticipation has the character of a countermovement in space and in time.
- Tone 11 is a stable pitch which results from plucking string number 7 with the left ring finger. This plucking movement is reflected in a sharp peak in velocity.

The sequence of tones 12 to 16 is a repetition of the sequence of tones 7 to 11. For example, in tone 12, the left ring finger makes an ancillary movement and then a preparatory movement. Apart from a salient ancillary movement in tone 16, the two sequences are almost exactly the same. At tone 17, the sequence is again repeated.

These detailed descriptions show the following:

- A tone in *guqin* music can be analyzed in terms of its gestural control. The action which produces a sliding tone consists of a sequence of elementary (point-to-point) movements.
- Some *guqin* tones start with low velocity, as shown in tones 1 to 5. In this type of tone, the maximum velocity tends to be obtained somewhere in the middle of the upward or downward pitch shift. The pitch shifts of tone 5 are faster than those of the other tones.
- Other *guqin* tones are characterized by starting when the velocity is maximum. This is the case for tones 8, 10, 13, 15, and 18. In order to produce this tone, the effective movement has to start in advance. Furthermore, all these tones are preceded by a countermovement, whose displacement is in the opposite direction from that of the effective movement. For example, if the effective movement goes to the left (pitch goes down), the preparatory movement goes to the right (tones 10 and 15). The duration of this countermovement is almost equal to the duration of the effective movement.[9]
- Upward and downward pitch shifts may differ with respect to their speed. Their nature is defined by the effective movements that underlie the pitch shifts. It is straightforward to assume that this aspect contributes to the particular expressiveness of the tone. Tones 8 and 13 are notable in that the pitch glissando consists of two parts that are clearly visible in the displacement and velocity curves. The change in this glissando is deliberate. It affects the expressive character of the glissando, making it lighter.

The main conclusion of this analysis is that *guqin* tones are generated by a complex interplay of very efficient corporeal articulations. They consist of preparatory (counter)movements and effective (control) movements which together define an action that leads to the production of a tone. Important and time-critical articulations are clearly anticipated and are prepared for by a movement in the opposite direction and of almost equal duration. Ineffective or ancillary movements are rarely noticed. Repetition of sequences of movements is very accurate, which is reflected in the velocity curves.

In this context, the characterization of music as moving sonic forms is useful. It captures the idea that the sound structure encodes aspects of the player's actions. The sound structure reveals the encoding of the biomechanical energy as sound energy. Behind this sound energy are movements aimed at producing tones. These tones are intended to function

in tone configurations or musical phrases. From the listener's point of view, the sound defines the proximal cues, while the player's actions define the distal cues. Clearly, not all aspects of the player's movements are encoded; preparatory, anticipated, and ancillary gestures remain hidden. Only some of the gestural controls are effective in the sense that they are reflected in the sound structure. All this suggests that the moving sonic forms are intentional and functional. The next step, then, is to investigate how well the listener is able to capture these aspects of moving forms.

6.2.4 Monitoring the Listener's Movements

In this case study, I focus on two listeners. The first listener knew the musical fragment very well but was not an expert in Chinese music. The second listener was entirely new to the music and the piece.

Their corporeal articulations were monitored with the help of a joystick (figure 6.6). The listeners were standing, and while listening to *guqin* fragments through headphones, they were requested to move a stick in accord with the music. Thus, the corporeal articulations were reduced to movement in two dimensions. The listeners listened to the *guqin* piece twice. Figure 6.7 shows the movements of two listeners (L1, L2) during two trials (T1, T2) for the second part of the piece, starting at tone 5 (P1).

Figure 6.6
The listener's recording equipment.

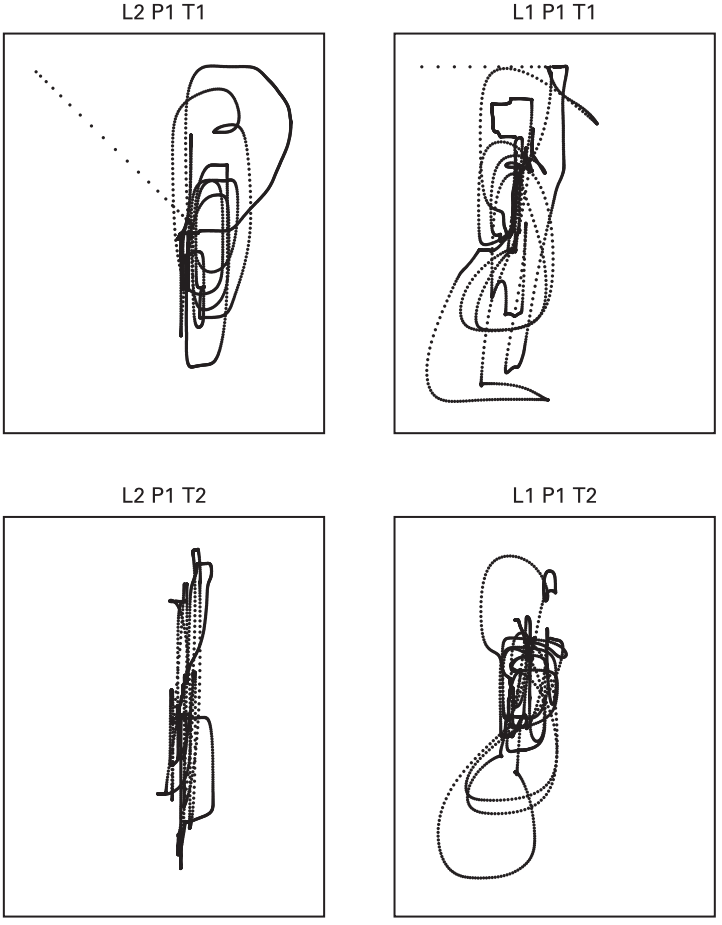


Figure 6.7
The listener's movements. L, listener; P, piece; T, trial. The X and Y axis represent horizontal and vertical movement, respectively. (See text for further explanation.) Figure 6.8 shows the corresponding velocity patterns (of the second part of this movement).

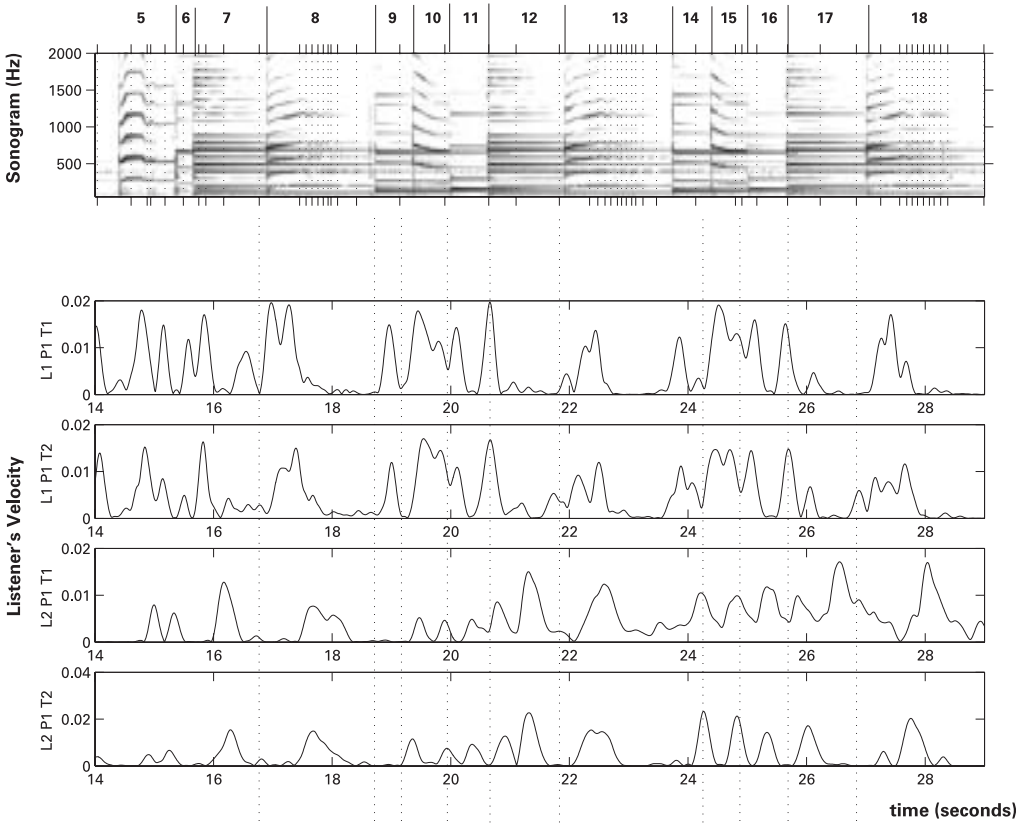


Figure 6.8
Sonogram of the second part of the *guqin* fragment and velocity curves of two listeners during two trials. (The velocity curves show the absolute values of the velocity.)

The velocity patterns were derived from these movements and are shown in figure 6.8. The top panel shows the sonogram, with tone labels, tone segmentations, and gesture segmentations marked by the vertical grid on the sonogram.

Analysis of Listener Behavior

Several observations can be made concerning consistency, synchronization, resonance, and the induction of ideomotor responses:

- In this *guqin* fragment, the beat is not prominently present, although a beat can be induced by the rhythmic structure in the second

part of the piece. Indeed, tones 9, 10, and 11 have the same duration. If we take the length of one of these notes as the unit length, then tone 12 is twice this duration and tone 13 is almost three times this duration. The pitch shift in tone 13, for example, takes about the unit length (similar to tones 8 and 18), which reinforces the rhythmic subdivision of the tones in terms of a unit length. This underlying rhythmic pattern may mark a beat to which listeners respond. The induction of a repetitive pattern in the movement is more clearly present in subject 2 than in subject 1, for example, during tones 10, 11, and 12, and during the repetition at tones 15, 16, and 17.

- Subjects seem to be rather consistent in their movements over the two trials, in the sense that they tend to replicate their movement speed when listening to the music in subsequent trials. However, over different trials, subjects tend to make different movements, as can be inferred from the movement patterns shown in figure 6.7. Thus, what is invariant here is not the displacement as such, but the speed of the movement. In addition, subjects differ in terms of the types of movements they make, as well as in terms of timing. What constitutes the start of a movement (low velocity) for subject 1 may correspond to a point of maximum velocity for subject 2. Although corporeal articulations may differ greatly from one subject to the other, the subject can replicate its own speed very accurately. This suggests that the velocity of the movement is an invariant feature of embodied perception.
- Synchronization of the movement with the characteristics of a sound is an expression of resonance behavior with musical energy. In this example, two types of synchronization can be observed: (1) synchronization related to the onset of a tone and (2) synchronization related to the characteristics inside the tone (typically, effects of sliding).

—First, concerning the onset, it can be observed that the start of a(n elementary) movement is not always in agreement with the actual start of the sound. In fact, the movement may start earlier, and it may happen that the start of a tone falls exactly on the maximum speed of the movement. This can be observed at tone 12, where the onset of the tone falls exactly on the point where the movement of subject 1 has the highest speed. For subject 2, this point corresponds to the start of an elementary movement. A similar observation can be made with respect to tones 10 and 15. Note that at tone 10, both subjects have their maximum speed on the beat, while at tone 15, only subject 1 does. Consider also tone 17, where this aspect of synchronization is salient.

The corporeal articulations of the listeners thus have an anticipatory character which is more pronounced in the experienced listener than in the nonexperienced one. The anticipation is characterized by the fact that the elementary movement may start before the tone onset and reach its maximum velocity at the tone onset.

Remarkably, such anticipated movements sometimes tend to synchronize with the movements of the player. For subject 1, this can be noticed at tone 8 (first trial), where the movement starts at the same moment as the movement of the player (indicated by the dotted vertical lines that connect with the sonogram). Also, for tones 9, 10, 11, 13, 14, and 15, the onset of the elementary movements corresponds to the onset of the player's elementary movements.

—Second, there is a clear effect of pitch-sliding on the synchronization. For example, at tone 13, the maximum speed in all trials is reached at the moment when the pitch shift becomes stable and is transformed into a vibrato. This effect can also be observed at tone 8 for subject 1, whereas subject 2 shows the opposite response, in that the movement starts during the vibrato. A similar trend can be observed in tone 18. In tones 10 and 15, the velocity of the movement tends to slow down in both subjects, while in note 13, the velocity tends to rise in synchrony with the velocity of the pitch shift. This illustrates that the corporeal articulation can be in resonance with pitch effects.

The above observations suggest that listeners are able to attune their elementary movements to characteristics of the sound energy, such as onset and pitch change. In the experienced listener, the corporeal resonance behavior has a more pronounced predictive character. The onset of the listener's elementary movement often agrees with the onset of the player's elementary movement. The speed of the movement can be influenced by the speed of the pitch change. Whether pitch goes down or up seems to affect the decrease or increase of the listener's movement velocity in this example.

The data provide evidence for the hypothesis that the listener's perception is predictive at short term. The accurate synchronization of the anticipatory character of the movements suggests that they are intentional. It is tempting to assume that aspects of this corporeal intentionality may relate to the performer's corporeal intentionality. Indeed, anticipation of the sonic moving forms seems to synchronize with the player's corporeal articulations. Whether such a close relationship be-

tween the players' articulations and the listener's articulations can be assumed, remains to be seen on the basis of more studies.

What these data suggest, however, is that the listener is capable of grasping music as an intended moving form. This can be explained by the fact that the listener's movements translate sound energy to the listener's action-based ontology. In doing this, the movements take on the character of an action (goal-directed movement) whose intentionality is in turn projected onto the sound energy and, by extension, onto the movements of the listener. Thus, player and listener are able to establish a relationship of mutual information exchange at the level of action.

Alternatively, one may consider music as a proximal cue in which the listener perceives (or forms the hypothesis of) the distal cue, that is, the action which gives rise to the proximal cue. Through corporeal attuning, the listener can express the perceived intentionality which, as shown here, can be measured and studied. This projection of human motion onto music is facilitated when the cause of that motion is itself a human actor, and thus the motion is human (Juslin et al., 2002). This mechanism is similar to the human ability to decode human action patterns from point-light displays in the visual domain (see, e.g., Pollick et al., 2001; Pollick, 2004; Troje et al., 2005).

Obviously, the movements of the listener are not (and perhaps cannot be) exactly the same as the movements of the player. What is more or less the same in both the listener and the player is the motor system that encodes and decodes moving sonic forms. The above data suggest that music perception involves a motor attuning component through which an intentional character can be attributed to music and, by extension, to the composer and performer. Motor-attuning helps the listener to read the minds of the composer and the performer, and thus to understand the music as an embodied mental phenomenon. According to Wilson and Knoblich (2005), the predictive behavior that is made possible by motor attuning (called emulation) would in turn facilitate perception. It is indeed straightforward to assume that this facilitation helps listeners to experience music as a peak experience. The predictive character of corporeal attuning is likely to be the expression of fundamental and automated patterns of communication (Bargh and Chartrand, 1999). Hawkins and Blakeslee (2004) state that predictive behavior is the essence of human intelligence.

To summarize, the case study supports the theory that the musician encodes gestures in sound, and the listener can decode particular aspects

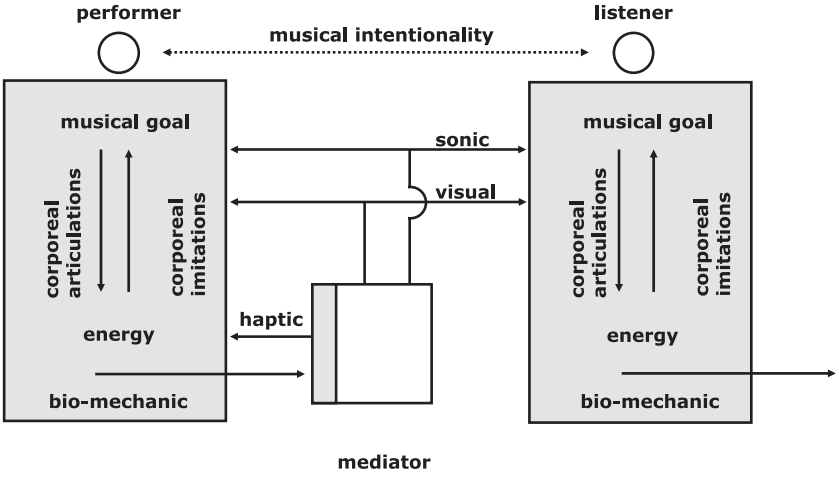


Figure 6.9
Model of musical communication between performer and listener (see text for explanation).

of them through corporeal imitation (see also Leman et al., submitted). The mechanism enables the communication of intended motion, which provides a basis for an embodied perception and mind-reading of music. This type of perception and understanding is assumed to play a key role in the listener's focus on peak experiences and the understanding of structure, emotion, and cultural significance.

6.2.5 A Model of Musical Communication

The above case study supports a model of musical communication in which the encoding and decoding of biomechanical energy allows the communication of intentions. Figure 6.9 is a schematic summary of this musical communication model in which a performer and a listener are involved, in addition to a mediator.

The starting point is the performer, who has in mind a musical goal or idea (possibly provided by a composer). This goal is realized as sound energy, using the human body and a mediation technology. More specifically, the musical goal is realized through corporeal articulations, whose biomechanical energy is transferred to the music mediation technology (the music instrument). This device in turn translates part of the biomechanical energy of the performer into sound energy, while another part of the biomechanical energy is bounced back as haptic energy (energy re-

lated to the sense of touch). The control of the musical instrument is realized in a closed loop with haptic, sonic, and perhaps visual feedback. In the mind of the performer, this physical interaction can be enhanced by corporeal imitation processes that translate the sensed energy back into the action-oriented ontology, giving meaning to the interaction. Thus, haptic energy may largely contribute to the perceptual disambiguation of the particular relationship between gestural control and sonic output.

Next, the mediator transmits the sonic and visual energy to the listener, who, through mirror processes, can make sense of it. Corporeal resonance (or imitation) thereby forms the basis of musical involvement which ultimately leads to an understanding, both corporeal and cerebral, of the music's underlying intended articulations (expressions of moving forms at local and more global levels). Obviously, the listener's understanding of the music's intentions need not necessarily be the same as the performer's. It is sufficient that the listener can relate the moving sonic forms to his or her own action-relevant ontology in order to make sense of the perceived physical energy. The listener's processing of musical information is likely to be reflected in corporeal articulations, which can be seen by other listeners.

There is a further important aspect of this model. That the mediator—for example, an acoustic instrument such as a Chinese *guqin*—can be conceived of as an extension of the performer's body. Obviously, for the performer this is an illusion, albeit a very natural one. It is generally believed that haptic feedback may largely contribute to the creation of this illusion of non-mediation. In fact, experiments show that self-attribution of body parts is based on multisensory perceptual correlations of which action-related sensing forms an important aspect (van den Bos and Jeannerod, 2002).

The model suggests a musical signification practice that is based on the encoding and decoding of patterns of corporeal articulations. Music encodes corporeal articulations in sound (moving sonic forms) which can be decoded, predicted, and understood because they rely on movements which appeal to the action-based ontology of human subjects. Although these movements may be culturally learned, they can be imitated and related to a common framework of the composer/performer and the listener. Since this framework is neurally encoded, it can be stated that musical communication is based on the sharing of neural structures that pertain to movement. This forms the power of music as universal language. However, one should always keep in mind that any

signification practice may also call upon a cerebral approach to music as a meaningful cultural phenomenon. In this practice, corporeal perception and understanding seem to be fundamental because they are based on moving forms and the resonant structures of the human body. It does not require knowledge of the cultural background of music, although such knowledge may be of great help in setting appropriate motor structures ready for action. In addition, the perceived movement may be an incentive for the activation of other processes—for example, those related to arousal and emotion. The communication of music is fundamentally based on multimodal sensing, using a motor model for encoding and decoding.

6.3 Constraints of Interactive Communication

The above model of musical communication can be extended to interactive music systems, and in particular to electronic musical instruments and digital virtual musical environments. In using interactive music systems for artistic purposes, one of the key problems is the configuration of a proper mediation technology. This configuration can be considered from four viewpoints: (1) biomechanical control and haptic feedback, (2) constraints of electronic music mediation, (3) group effects of musical communication, and (4) motivation. Taking the above model of musical communication as the base, these viewpoints define a framework for interactivity in which musical communication is embedded. The framework is complicated because it involves different aspects of action and perception, from sensorimotor interaction to intentional behavior. The primary interest in relating the theory of corporeal articulation and resonance to interactive music systems is that it offers a straightforward grounding of multimodal experience in a context of multimedia technology. The above model of musical communication can be of help in understanding the relevant issues.

6.3.1 Biomechanical Control and Haptic Feedback

As shown in the case study of *guqin* music, acoustic musical instruments draw upon the fact that the movements of the performer are tightly connected with a mechanical-energetic interface that produces sound. This connection implies that part of the performer's biomechanical energy is used to shape the microstructures of the sound energy. A smaller part of this energy is bounced back as haptic energy, which the performer can

feel by the sense of touch. Sound energy and haptic energy are basic sources of feedback on which the performer can rely for fine gestural control of the instrument and subsequently, fine control over musical expression (Winold et al., 1994).

Haptic feedback is believed to be important for the prediction, self-adaptation, and modification of sound control at the millisecond level. Recall the asymmetric shape of figure 6.2. The first part of the movement is typically unconstrained by sensory feedback, while the second part is based on sonic and haptic feedback (after touching the string). Haptic feedback to the performer is important in that it allows a disambiguation of the control unit, and thus of the perceived effects (proximal cues) of the mediator (Ernst and Bülthoff, 2004). Haptic feedback contributes to a more reliable estimate of the sonic output of the mediator. In many ways, it is a multimodal prerequisite for musical expressiveness. Indeed, research on haptic feedback is often linked with research on musical expression and performer nuances.

Of particular relevance for mediation technologies is the idea that the performance nuances subsume a layer of communication that runs in parallel with sensorimotor processing and higher-level couplings of action and perception (Sheridan, 2004). From the viewpoint of the performer, this level is more focused on the expression of a particular sensitivity or affect, while from the viewpoint of the listener, this level is more focused on the understanding of action-relevant characteristics induced by kinesthetic involvement. Much of the research on performer nuances is on the clarification of the relationship between local and global aspects of musical expressiveness.

6.3.2 Constraints of Electronic Music Mediation

Overall, haptic feedback is a natural characteristic of acoustic instruments. By their design, these instruments allow the transformation of biomechanical energy directly into sound energy. In contrast, electronic instruments have no mechanical-energetic interface that mediates between corporeal articulation and sound. In electronic music systems the energy for making sound (electricity) is independent from the haptic biomechanical energy exerted by the musician. Consequently, the interfaces and gestural control devices are decoupled from the sound production device. A consequence of a lack of haptic feedback may be that the mediator is not really experienced as part of the human body and, consequently, that corporeal articulations are badly reflected in the

microstructure of the sound energy, which in turn may be problematic for the listener.

In view of the musical communication model, this decoupling has a number of consequences for the design of effective mediators. The problem should be considered with respect to haptic feedback, mappings of control parameters, and characteristics of musical action and perception.

Simulating Haptic Feedback

The lack of haptic feedback in electronic musical instruments has been acknowledged as a central problem in their design (e.g., Rovan and Hayward, 2000; Gunther and O'Modhrain, 2003; Howard and Rimell, 2004). In recent years, much attention has been devoted to the possibility of simulating haptic and tactile feedback using electronic devices (see, e.g., Wanderley and Battier, 2000; Paradiso and O'Modhrain, 2003; Camurri and Rikakis, 2004; Johannsen, 2004). In addition to haptic feedback, visual feedback cues have been investigated (Camurri, Lagerlöf, et al., 2003; Kapur et al., 2003). This work has just begun, and needs much more elaboration in the future. It is related to a number of other problems, in particular mediator mappings and issues of music perception.

Mediator Mappings

Clearly, electronic music systems allow much more freedom for the performer, because mappings between gestural control units, on the one hand, and sound production units, on the other hand, are not constrained by any biomechanical regularities. This type of freedom has attracted the interest of many artists and researchers (Tarabella and Bertini, 2004; Karjalainen et al., 2004; Ng, 2004). It allows the playing of a complicated scale on the piano or guitar by a simple wave of the hand. The movement of the hand may be captured by a video camera and used as a trigger for playing the scale. However, as most electronic music performers know, it is exactly this freedom of mapping that may disturb the sense of contact and of non-mediation.

To better understand the problem of mapping in electronic musical instruments, figure 6.10 distinguishes among the gestural controller unit, the mapping unit, and the production unit (Wanderley and Depalle, 2004). The components in the dotted box can be seen as parts of the mediator that is depicted in the musical communication model of figure 6.9. The gestural controller is defined as the part where physical interaction with the performer takes place. It contains sensors and software for

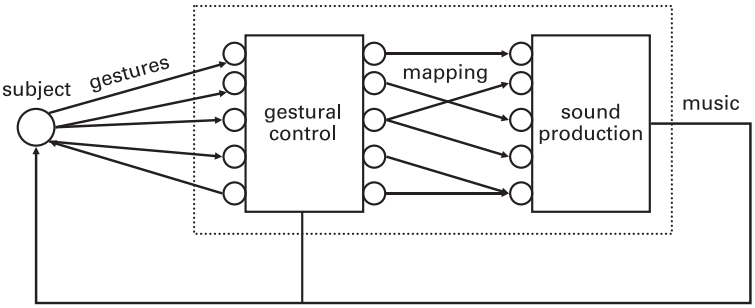


Figure 6.10
Conceptual model of an interactive system consisting of gestural controller, mapping, and production unit.

feature extraction. The gestural controller can also give feedback to the performer, for example, by means of a computer screen or simulated (electronic-based) haptic feedback. Next to the gestural controller is a mapping unit which translates gestural controls into parameters for the production of sound or other types of energy. The sound production block can be considered a sound synthesizer whose output is controlled by a number of parameters that vary in time.

In this model, the mapping unit forms an important part of the mediator in that it connects human corporeal articulations with the production of music and possibly with associated multimedia events. The mapping allows a lot of freedom, but this freedom has an obverse side: performers often feel that the information flow cannot be accurately controlled in its fine details. If this happens, it is likely that the information flow is experienced as unpredictable. As a result, it can be hard to create the illusion of non-mediation.

Mediating Musical Sound

The lack of haptic feedback may have a significant effect on the effectiveness of the mediator and, consequently, on the chain of musical communication. However, this lack is not the only problem. It is part of a set of problems that define the delicate balance between unlimited freedom in sound control, on the one hand, and constraints that limit human action and perception, on the other. After all, the purpose of a musical instrument is to allow music to be made by retaining the interest of the performer and the listener. This interest depends on the performer's ability to control the mediator in a precise way, but it also involves the listener's

ability to make sense of the situation. Mediation technology cannot be seen independently from the fact that biological organisms have a natural bias toward action-relevant cues in the environment. In music, things become even more complicated because cultural constraints have to be taken into consideration.

Therefore, an appropriate design of the mediation technology should cope with a diverse set of natural as well as culturally relevant aspects in order to keep the interest of the performer and the listener. At least four different aspects of interaction can be taken into account, related to (1) the listener's natural bias toward sonic sources, (2) structural similarities, (3) gestures, and (4) learned conventions.

Mediating sonic sources (the ecological approach) First, consider the bias toward the perception of action-relevant cues (affordances). This aspect was mentioned in the earlier discussion of the Gibsonian model of perception. The bias means that in natural environments, perception is oriented toward the action-relevant cues of the physical energy that give rise to the perception. In recent studies on ecological psychoacoustics (Rocchesso and Fontana, 2003; Neuhoff, 2004), it was shown that listeners have an impressive ability to identify very specific action-relevant characteristics of the mechanics that cause the sounds. Listeners detect the width of struck bars; the length of dropped rods; the hardness of struck mallets; the size, material, and shape of struck plates; vessels with different levels of fluid; determination of gender from footsteps; the ascending or descending of staircases; and anticipate the trajectory of approaching sound sources (Rosenblum, 2004). Listeners seem to have a natural bias for the perception of the source mechanics. This provides a basis for source-related understanding of the interactive musical instrument.

When the subject perceives the sound as being produced by a physical mechanism, it may form the impression that the sound is real; hence, that the environment in which it has been produced is real; and hence, that this environment has a high degree of presence, which in turn may facilitate the experience of being immersed. From the perspective of the performer, multisensory (sonic, visual, tactile, haptic) feedback may contribute largely to the perception of action-relevant sources and the illusion of biomechanically based control. It can be assumed, therefore, that mediators that account for action-relevant cues may yield a high degree of presence, and therefore have the potential to

be more effective in engendering a higher musical involvement for the performer and the listener.

Mediating structures and similarities (the gestalt-based or cognitive approach) There are many cases where the perceptual system of a listener may not be able to make sense of the action-relevant source. For example, in newly composed artificial sounds, often the source mechanics cannot be retraced and the perceptual system has to look for an alternative solution. In that case, the cognitive approach assumes that attention is directed toward the structural properties of musical audio, such as the relationship among intensity, pitch, timbre, and so on.

It is characteristic of music perception that the focus of attention may not be so much directed at the mechanical source of the sounds as at higher-level structural musical qualities of melodic lines, rhythm patterns, harmony, and timbre. Thus, in cases where source mechanics cannot be retraced, or where the musical context is such that the emphasis is on sonic forms rather than on the sources that produce these sounds, attention may shift toward structural characteristics. After all, these characteristics are an emerging aspect of the processing of multisensory information. Principles of perceptual organization, such as integration and segregation, will transform sound energy into musical objects as a function of disambiguation (Bregman, 1990). If action-related cues are not strong, or are put in an unusual context, the signification process may turn to the meaningful relationships between structural features. It may be assumed that the perceptual system is predisposed to organize stimuli into such structures of gestalts, independent of whether source-relevant cues are available or not. Thus mediators may be constructed so that they produce gestalt-based meaningful relationships between structures in sound energy.

Mediating gestures (the embodied cognition approach) In addition to the ecological approach and the gestalt approach, the embodied cognition approach assumes that listeners seek to give meaning to musical sounds in terms of emulated actions, that is, of corporeal articulations (grounded in the subjective action-oriented ontology) in response to sonic energy. This aspect should be distinguished from the recognition of the mechanical sound source, as well as from the cognitive involvement with sound structures (proximal cues). The sound source concerns the mechanical cause of the sound, something that is external to the

human subject. This can be called the mechanical distal cue. Instead, the gesture is about the human cause which is behind the mechanical cause of things, which is internal to the human subject. This can be called the intentional distal cue. In that sense, sounds of a symphonic orchestra, for example, even if they cannot be related to a particular external mechanical cause, can be related to intended gestures when they appeal to corporeal articulations and imitations by the listener. Therefore, mediators that take this aspect of human-related movements into account may be of interest in the context of electronic music mediation.

Mediating conventions (the cultural approach) Finally, it should be noted that listeners are sensitive to learned patterns and cultural conventions. These find their way into the articulations that are typical for a particular musical style (Hatten, 2003). Musical communication may draw on the knowledge of these patterns, which may relate to different types of conventions, including symbolic or narrative conventions such as the heralding of spring using the characteristic pitch interval of a cuckoo, or rhythmic patterns that refer to secret messages, as in Mozart's Masonic opera *Die Zauberflöte*. Therefore, conventions should be taken into account in mediation, even in electronic music performances.

To sum up, mediators for electronic interactive environments should take into account principles related to natural and cultural constraints of human action and perception. These principles range from sensorimotor processes, such as haptic feedback, to ecological and embodied aspects of perception and conventions. Human perception is rich, and it is likely that all these aspects should be taken into account in developing electronic devices that mediate between mind and matter. The above considerations clearly show that mediation is a matter not only of sensorimotor interactions but also of careful design that takes into account global constraints of musical communication.

6.3.3 Motivation

The above considerations are also related to motivation, the reasons why a subject would be interested in a human–technology interaction. Motivation brings in a number of factors, which may be external or intrinsic. For example, social group pressure is an external drive which may force a subject to attend a concert dressed in a particular way, to move and behave in particular ways, to pay attention to particular aspects of the performance, and to use particular music mediation technologies (e.g.,

iPods rather than mp3-players). External motivations may also partly drive intrinsic motivations.

Csikszentmihalyi and Csikszentmihalyi (1988) relate intrinsic motivations to intrinsically rewarding (or autotelic) experiences, which bring the subject to a state of experience that is self-motivating. They state that this state of intrinsically rewarding experience is a kind of psychic negentropy which is obtained when the contents of consciousness are in harmony with each other and with the goals that contribute to the development of the subject's self, or what I call the subject's action-oriented ontology. It is the collection of mental entities (beliefs, values, valences, motivations) by which the subject structures and anticipates its actions in the physical environment.

Music may provide an excellent context for intrinsically rewarding experiences. Through learning, challenges may be set at gradually higher levels and skills can be adapted to them. Thus, an optimal balance can be found and interest can be maintained. In that respect, self-motivation is something dynamic. If skills develop and challenges remain the same, challenges may become boring and the subject may lose interest and abandon the task. On the other hand, if the challenges are too high, the subject may become frustrated and unmotivated. According to Csikszentmihalyi and Csikszentmihalyi, it is only when challenges and skills are in balance that the subject is able to engage in an autotelic experience.

The mechanism of intrinsically rewarding experience is of particular relevance for the development of music mediation technologies. Indeed, if gestural control and sound generation are physically decoupled, skills and challenges may become decoupled and interest may be quickly lost if the subject has the impression that improvement of skills has no apparent effect on feedback from the interactive system. A major challenge for the development of a mediator is to keep track of the balance between challenges and skills.

The following example illustrates the role of intrinsic motivation in the development of a music mediation device. The device was presented at the Accenta exhibition at Ghent in 2005 (figure 6.11). Visitors to the booth heard a pure tone in a headphone and were requested to imitate the pitch of that tone by singing. This is an action that offers little apparent return, except for those interested in singing. A digital mirror showed a deformed reflection of the subject's face. It was only by repeating the sung tone correctly that the deformed image became a clearly visible mirror image. After a very brief learning period, subjects rapidly become

Figure 6.11
Interactive game called *Spiegeltje aan de wand* (Accenta exhibition, Ghent, 2005). You can see yourself in the digital mirror when you correctly imitate the tone you hear in the headphone. Otherwise, the image is deformed.

experts in imitation. Most subjects, especially children, enjoyed the game because of the intrinsic motivation to see themselves in the mirror. Low-level musical involvement was thus mediated by means of a visual feedback which stimulated self-motivation.

Recent designs of interactive music systems show that mediators can be constrained such that mirroring becomes an integrated part of the interaction. For example, Pachet (2003) developed an experimental system that interacts with a human performer by means of style imitations. The gestural controller and the sound production system are traditional keyboards with piano tones, but the system is able to learn features of the musical style of the human performer. It can reproduce responses that imitate the style. Experiments with this system support the idea that performers interact with the system at a high mental level of stylistic content exchange. This approach opens new ways to conceive the mediator technology in terms of interacting agents capable of mirror-

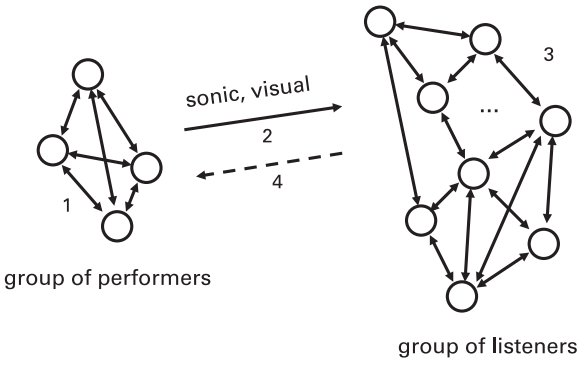


Figure 6.12
Model of social music communication, with interaction (1) among performers, (2) from performers to listeners (music-driven interaction), (3) among listeners, and (4) from listeners to performers.

ing the subject's intentionality. These initial experiments show that mirroring supports the motivation for being involved in an interaction process.

6.3.4 Group Effects in Musical Communication

In addition to the previous issues, there is another aspect of communication (figure 6.9) that is likely to be very important in music: group behavior and, more particularly, entrainment or mutual adaptive behavioral resonance. Although this aspect will not be discussed in detail in this book, it is relevant to mediation technologies and interactive multimedia systems.[10]

A model for social music communication is depicted in figure 6.12. Performers and listeners are represented by circles, and the arrows between circles suggest exchange of energy patterns through sonic, visual, olfactory, and tactile information channels. The actions of the performers (interaction type 1) generate music (sonic/visual energy). Listeners capture sonic energy and see the gestures of the performers. This sonic and visual scene may drive the movements of the listener in response to music (interaction type 2). In addition, listeners perceive each other, and in response to that (interaction type 3), they may start exerting an influence on each other. Corporeal resonance in a listener may thus result from the musical/sonic and visual energy as well as from the

multimodal energy coming from neighboring listeners. The latter may be called the social entrainment energy. In a similar way, the performers are influenced by musical/sonic energy, haptic feedback, and social entrainment energy coming from the audience (interaction type 4). This type of interaction also may entrain the action of performers. Indeed, performers report that the audience tends to be perceived as a gestalt whose action is likely to have an effect on their playing. As can be seen, the social interactions are quite complex.

The hypothesis is that in a context where groups of people are together, corporeal imitation can lead to emergent behavior, as in concert halls where masses of people start waving their hands in the air at the same time, or when, after a concert, the applause becomes synchronized (Neda et al., 2000). In both cases, the emerging behavior of the audience results from the fact that subjects are imitating their neighbors (see section 5.1.2). This phenomenon has been observed and studied in populations of insects, fish, birds, and other animals (see, e.g., Theraulaz and Spitz, 1997; Bonabeau and Theraulaz, 1997), as well as in humans (Grammer et al., 1998, 2000; Bargh and Chartrand, 1999; Lakin et al., 2003; Niedenthal et al., 2005).

It can be assumed that group imitation behavior has an enormous effect on the participants' individual experiences during this applause. Arousal, attention, and the feeling of presence may be enhanced through this type of group resonance effect. Performers, too, seem to appreciate the effects of allelo-mimetism in an audience, which they perceive through auditory, visual, and tactile channels. They perceive this effect as a global emerging effect of the audience, which can be very stimulating for their performance.

Entrainment thus forms the dynamic multimodal framework from which musical magic, a peak experience of a group of people, may emerge. Given the widespread phenomenon of this so-called magic, it must be that humans are particularly sensitive to it. Although the dynamic primitives that account for entrainment are not completely understood, it may be assumed that the exchange of corporeal articulations through different energy channels and its subsequent mirroring is a key component of this behavior. It is also likely that this magic is related to empathy, which is strongly associated with feelings of social connectedness.

Because of the social aspect involved in entrainment, I consider the development of interactive music systems that can deal with group

effects of social musical cognition to be one of the major challenges of future music research. Is it possible to design a proper mediation technology that deals with this social aspect of musical involvement? Interactive music systems could be designed to enhance this type of social resonance communication, in contexts where many subjects are confronted with many machines. In this approach, machines become social agents with which it becomes possible to exchange intentions.

To conclude this section on musical communication, it can be stated that electronic systems offer many new possibilities in human–technology interactions. Of particular interest is the idea that musical communication is situated at the level of intentions, while the primitives for that level of communication are based on the encoding, transmission, and decoding of energy. Low-level sensorimotor interactions play an important role in the encoding and decoding processes, but they stand as a function of higher-level forms of action and perception. Sounds may have a strong appeal to corporeal articulations and they may involve the subject in mirroring aspects of moving forms in sound. These corporeal articulations provide a basis for an immersing involvement with music. Consequently, beyond source-related feedback and haptic feedback, interactive music systems offer the possibility of interacting with machines at higher mental levels. The realization of those levels of interaction is based on corporeal articulations that transform physical energy into mental representations that are related to the action-oriented ontology. Mirror processes form a basic part of this kind of corporeal articulation, and mediators that incorporate mirror processes are of great potential interest for interactive music systems.

6.4 Multimedia Environments

The model of musical communication discussed so far has assumed that a musical instrument is a technology for the mediation between (1) the performer's intended corporeal articulations and (2) sound energy (sounding music). Electronic musical instruments aim at extending this paradigm to the electronic and digital domains. As suggested above, this framework should be further expanded to autonomous multimedia technology and virtual social agents. After a brief sketch of what I mean by this and how it connects with multimedia technology, I turn to a concrete example of a system that implements autonomous interactive behavior.

6.4.1 Autonomous Social Agents

In the acoustic musical instrument, the corporeal articulations of the performer are translated into sound energy using a mediator based on biomechanical energy. In the electronic musical instrument, corporeal articulations are mapped to sound production parameters using an electronic mapping device. The next step is an electronic music environment which behaves as an autonomous virtual social agent with which it is possible to communicate via the exchange of physical energies.

Such an agent is typically equipped with capabilities of synthesis and analysis, conceived in terms of artificial composer/performer modules and listening modules. These modules typically simulate the behavior of a real human musician and a real listener because this may allow interaction at the level of intentions. Clearly, this concept implies a paradigm of music production that is rather different from that of a musical instrument, even from an electronic musical instrument.

The main purpose of a musical instrument is to transmit the performer's musical actions to a listener. In contrast, the main purpose of the music environment is to establish an interaction between a human agent and a technological agent. The technology no longer forwards the musical actions of a performer. Instead, it interprets and generates actions on its own. Accordingly, the mediation is no longer based on a one-way transmission of information, but on dialogue between humans and machines. In this concept, packages of physical energy, transmitted between humans and machines, form the primitives on top of which humans can develop communication patterns at the intentional level.

This concept fits rather well with the concept of open structure in the avant-garde art and music of the twentieth century (see, e.g., Sabbe, 1987a, 1987b). Open structure in art draws upon the principle of logical and chronological indifference of the components that make up the art piece. Sabbe argues that in music, such structures were first explored by Beethoven, and later by Schoenberg, Stockhausen, and Cage. Interaction with musical environments, or artificial musicians, fits well with this tradition of open structure in art. Indeed, interactions follow no organizational hierarchy or logical preferences for structural units. Instead, the artistic result emerges from a trajectory of constrained interactions, without any need for logical foundation or for positional precedence of the multimedia objects involved. The interactions may include randomness but also imitation, and adaptation (entrainment). In my view, these interactions should allow the exchange of intentions from the actors/

performers, as well as the perception of intentional communication from the audience. Indeed, the real challenge is to build machines that cope with this high level of human communication. Such machines should be strong in anticipating human actions. In accord with the above model of musical communication, this entails that they should be equipped with a humanlike memory and prediction framework (Hawkins and Blakeslee, 2004).

6.4.2 Connection with Multimedia Technology

Multimodal interactions with musical environments or artificial musicians allow a subdivision of the problem into different steps: sensing, feature extraction, classification, and anticipation. I briefly treat these points, but I do not intend to go deeper into the involved technical aspects.

Sensing
First of all, music environments use all kinds of currently available sensor technology, such as audio and video capture, infrared, ultrasound sensing, and other techniques. Most of these techniques focus on the detection of change in energy. Given that human perception has a biological bias to human movements, it is likely to assume that sensing should focus on movements of the human body. But of course body movement is not the only form of information that may be expressive. In musical ecosystems, for example, there is no human interaction besides the potential resonances of material objects in a concert hall filled with people (e.g., Di Scipio, 2003).

Feature Extraction
Whatever the energy that is captured, it is turned into an electric and digital signal. From that signal, particular features are extracted for further processing. The extracted features may closely reflect the properties of the energy, or they may be processed to higher levels of description. The designer of a system may draw upon a whole arsenal of multimedia techniques and approaches.

Classification
Once features have been extracted, the next step is to employ them in a meaningful way. This typically involves the reduction of the dimensionality of the detected features. Given the fact that features represent

signals related to motion, most music environments will have to deal with the description of motion at more abstract levels (so-called reduced parameter spaces). Ultimately, these description levels will allow the exchange of information between multiple modalities of information-processing, allowing the transition from the visual or haptic domain to the audio domain and vice versa.

Anticipation of Action

Anticipation of human action may be based on a statistics of learned patterns (e.g., as in hidden Markov models) or a motor model that emulates the perceived action through a virtual physical simulation. Anticipation would imply that human movements can be accurately predicted at the millisecond level. However, this is something that is beyond the capacity of most existing interactive music systems that I am aware of.

6.4.3 A Platform for Musical Expressiveness

In what follows, an example is given of a platform that allows the design of music environments using modules that do sensing, feature extraction, and classification by means of multimedia technology. The platform aimed at providing a set of tools from which an artist could assemble an instrument or an environment for the processing of music-related expressiveness. The main objectives were (a) to have a better understanding of the primitives of nonverbal communication that underlie expressiveness in art; (b) to develop computational modules for sensing, feature extraction, and classification of expressiveness in real time; and (c) to exploit this understanding in artistic multimodal interactive music/dance/video applications. All this had to work in the context of a music theater, such as an opera house or cultural center, and with the potential active participation of the audience.[11]

In a context where multiple media (music, video, computer animation) were used, the focus was on the transmission of expressiveness from one domain to another, such as from music to computer animation, or from dance to music. For example, in music-to-computer animation applications, the task was to extract the expressiveness from the sonic energy and use that information to control an avatar or graphical scene that expressed sadness to a similar degree. In dance-to-music applications, the task was to extract expressiveness-related features from body movement and use them to control the expressive character of the music.

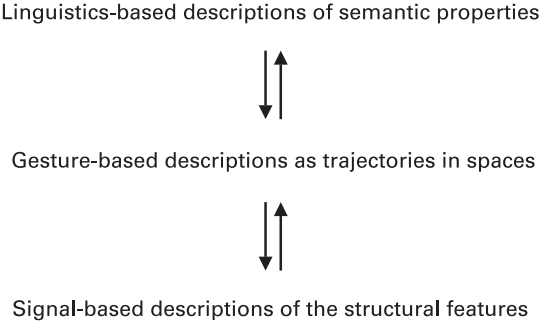


Figure 6.13
Conceptual framework for the multimodal processing of expressiveness.

Layered Conceptual Framework

Figure 6.13 sketches the conceptual framework for the platform. It aims at clarifying the possible connections among three different levels of processing: a sensory level, a gestural level, and a semantic level. The gestural level is a key component in this concept, in that it can be seen as the mediator between sensory-based descriptions and semantic descriptions. It typically contains trajectories of features which reflect aspects corporeal articulations.

Before going deeper into the nature of these layers, it is of interest to have a look at the information flow between these levels. This flow takes into account aspects related to hierarchical processing (bottom-up and top-down), as well as cross-modal processing (horizontal). In the bottom-up direction, sensory properties of physical energy are extracted and mapped into gesture trajectories, using techniques of classification. At the next higher level, these trajectories are related to semantic descriptions. In the top-down direction, a particular semantic description will be associated with a gesture trajectory. This trajectory may then be connected with parameters that control the synthesis of a particular energetic modality. In other words, the upward and downward directions of figure 6.13 represent the hierarchical analysis and synthesis of expressiveness.

The three levels comprise horizontal relationships as well, which allow the cross-modal transitions. These transitions may happen at the three levels, depending on the degree of hierarchical processing. For example, at the sensory level, it is possible to translate features from one modality directly into another. For instance, the quantity of body

movement (extracted from a video recording of a dancer) is correlated with sound intensity. Therefore, without the mediation of a gesture space, the analyzed body movement can be translated into synthesized sound intensity. The effects of this translation will be direct and will resemble the mappings used in electronic musical instruments. However, it is also possible to make this kind of cross-modal translation at a higher level of the hierarchy. For example, energy and velocity trajectories extracted from sound can be used to control the movements of an avatar. Or, at the semantic level, semantic terms extracted from dance movements can be translated to related semantic terms that control the synthesis of sound. In short, the upward and downward directions represent analysis and synthesis, whereas the horizontal directions typically represent the mappings, relationships, and correlations of patterns.

The design of this layered conceptual framework (Camurri et al., 2001; Camurri, Volpe, et al., 2005; Camurri, De Poli, et al., 2005) shows some global similarity with the memory-prediction framework for intelligent systems, as proposed by Hawkins and Blakeslee (2004). The latter framework contains a similar hierarchical as well as cross-modal processing. Yet Hawkins and Blakeslee's model is inspired by the architecture of the human brain, and it puts more emphasis on anticipation and prediction, paving the way for a fine-grained and thorough probabilistic processing. The layered framework for musical expressiveness is not that far developed in terms of a brain architecture, nor is it very apt to anticipate and predict. Yet it does offer some working modules which give a crude idea of what could be possible in a future intelligent system that allows expressive interaction with an autonomous environment. Below, I briefly summarize some of the characteristics of the different processing levels.

Sensory Level

The sensory level focused on features that were supposed to be relevant for expressiveness in music. They were extracted from various manifestations of physical energy.

In sound, for example, low-level features were related to onset, tempo (number of beats per minute), tempo variability, sound level (measured in dB), sound level variability, spectral shape (which is related to the timbre characteristics of the sound), articulation (features such as legato, staccato), articulation variability, attack velocity (which is related to the onset characteristics, which can be fast or slow), pitch, pitch density, degree of accent on structurally important notes, periodicity (related

to repetition in the energy of the signal), dynamics (intensity), roughness (sensory dissonance), noisiness, tonal tension (the correlation between local pitch patterns and global or contextual pitch patterns), and so on (Leman et al., 2005). Most of these features are based on the processing of local energy. Thus, they reflect local aspects of musical expressiveness.

When more context information is involved (typically in musical sequences that are longer than three seconds), sequences of features can be considered as trajectories representing aspects of corporeal articulations. These aspects may be related to melody, harmony, rhythm, source, and dynamics.

Whether all these features and trajectories are relevant to expressiveness is another matter, and research is needed to determine the precise relationship between features and components of expressiveness. Yet several of these features are known to work well in applications that envision the musical synthesis of basic expressions such as sad, happy, heavy, or light (Bresin and Friberg, 2000; Canazza et al., 2000). In analysis, it was possible to relate several acoustic cues to affect adjectives (Leman et al., 2005; Lesaffre, 2005) and use these in artistic applications.

In the domain of dance analysis, a similar approach has been envisaged that draws on a distinction between features calculated on different time scales (see Camurri et al., 2004a, 2004b). An example of a low-level feature is the amount of contraction/expansion that can be calculated on just one video frame or picture. The contraction index is a measure of how the dancer uses the surrounding space. This feature can be related to Laban's notion of personal space (Laban and Lawrence, 1947; see also table 7.6). Other examples of low-level features are the detected amount of movement, and the silhouette and orientation (figure 6.14). The amount of movement is based on variations in the silhouette shape. It can be considered an overall measure of the amount of detected motion, involving velocity and force. The silhouette shape and the orientation of this shape provide information about the orientation of the dancer.

Examples of descriptors of the overall direction of a movement are upward or downward movement, its directness (i.e., how much the movement followed direct paths), motion impulsiveness, and fluency. It is possible to segment a movement in terms of a sequence of elementary movements (characterizing the beginning and ending times). Then a collection of descriptors may be applied to these elementary movements, which may give statistical summaries of particular features related

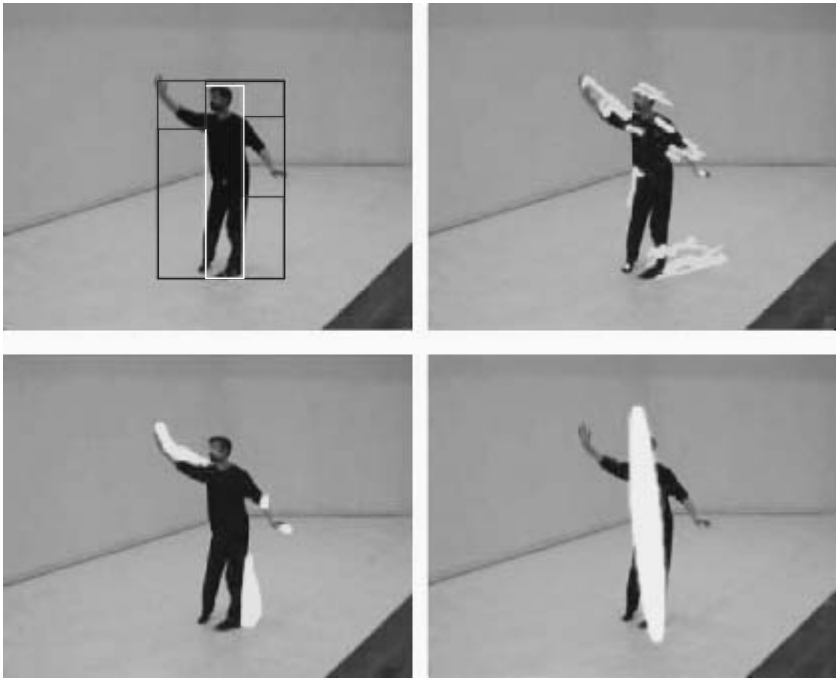


Figure 6.14
Motion descriptors extracted from expressive gestures in a dance performance. (a) Some subregions are identified as corresponding to arms and trunk. (b) Trajectories of points on the body are collected. (c) The global amount of detected movement (corresponding to the dark region around the dancer) is measured. (d) Body orientation and shape are computed starting from an ellipse approximating the body (Camurri et al., 2005). (Reprinted with permission.)

to the elementary movement, such as the average body contraction/expansion during the stroke, and so on.

Gestural Level

Often, features correlate with each other and form trajectories of values that change over time. Often these sensory-based trajectories can be reduced to trajectories in low-dimensional spaces. Different types of reduction are possible and different types of spaces can be considered (for a recent overview, see De Poli, 2004).

Kinematic tempo and energy (related to staccato/legato and intensity) were explored as representational frameworks for expressiveness (Canazza et al., 2000, 2003, 2004). In a similar context, affect spaces,

based on valence and activity, have been related to sensory features (related to legato/staccato and tempo) (Leman et al., 2005). The spaces can be derived from perceptual evaluations of different expressive music performances, using data-reduction methods (e.g., principal component analysis).

The robustness of these spaces was confirmed in the expressive synthesis of musical scores, in which the machine played the score in a sad or happy way, depending, for example, on the expression in the posture of a dancer (e.g., Bresin and Friberg, 2000). The basic idea is that expressive performance can be captured in terms of a weighted set of control parameters that influence loudness and duration, such as double duration (decrease of the interonset interval contrast for two adjacent notes having the nominal interonset interval ratio 2:1), duration contrast (long notes are played longer; short notes, shorter); faster uphill (decrease the interonset interval of notes in uphill motion of the melody), and so on. In several studies (e.g., Friberg et al., 1998; Sundberg et al., 2003), attempts were made to learn these rules by imitation of actual artistic performances. The extracted rule parameters can be mapped onto a two-dimensional space for intuitive control. The interactive platform can be used as a rapid prototyping environment for experimental setups, and the results can be used for artistic applications (Friberg, 2006).

Semantic Level

Semantic maps aim at relating kinesthetic/synesthetic and affective/emotive spaces to semantic descriptors. For example, fast tempo can be associated with semantic descriptors related to activity/excitement, happiness, potency, anger, and fear. Slow tempo can be matched with sadness, calmness, dignity, and solemnity. Loud music may determine the perception of power, anger, solemnity, and joy, whereas soft music may be associated with tenderness, sadness, and fear. High pitch may be associated with happiness, grace, excitement, and anger, fear, and activity, and low pitch may suggest sadness, dignity, and excitement, as well as boredom and pleasantness. And so on.

Semantic spaces will be discussed in more detail in chapter 7, where linguistic descriptors of music are considered in the context of music search and retrieval. It suffices here to mention that semantic spaces form a link with underlying gestural spaces and sensory spaces. Consider the semantic space that is defined by descriptors which relate to valence and activity. Valence is about positively or negatively valued affects, while activity is about the force of these affects. Recent research seems

to indicate that aspects related to the activity dimension can be more easily predicted than aspects related to the valence (pleasantness) dimension. Leman et al. (2005) addressed this question using combinations of a limited number of structural cues extracted from musical audio. It was shown that valence adjectives such as "carefree," "gay," and "hopeful" can be partly accounted for by sensory-based cues such as tempo and musical consonance, which enhance the perception of positive qualities. Activity adjectives such as "bold," "restless," and "powerful" are related to sensory-based cues such as centroid/width, pitch prominence, and loudness. The higher the loudness, the more the music is perceived as bold, restless, and powerful. The study shows that an intersubjective semantics of musical expressiveness can be partly grounded on sensory-based cues, but that activity could be better predicted than valence.[12]

To sum up, the platform for the study of musical expressiveness is grounded in a layered conceptual framework that supports both hierarchical and horizontal processing. Up and down the hierarchy corresponds with analysis and synthesis, respectively, while horizontal processing is needed for cross-modal transitions. Based on this framework, it is possible to develop a mediation technology that deals with musical expressiveness. The grounding for this is based on scientific research and correlation studies that aim at finding the relationships between measured features (third-person descriptions) and experienced, articulated, and annotated features (second-person descriptions). This correlation study takes into account the multimodal foundations of human/machine interactions. Thus far, less attention has been devoted to the anticipation and prediction of human actions. A probabilistic and brain-based architecture such as Hawkins and Blakeslee's (2004) may be helpful in transforming the present framework into a memory-predictive one that would account for the anticipation of human action. The latter may provide a key to establishing a convincing human/machine interaction.

6.5 Conclusion

Interactive music systems offer a new dimension for the classical forms of opera and music theater. They allow the microinteraction of different modalities of human expressiveness based on body movement (dance), playing music, image projections, changes of scenes, video projections, computer animation, lighting effects, visual effects, and perhaps olfactory effects. Such interactions, if well designed, open a new world of expressive possibilities in art. At the same time, they open challenging research

questions related to music mediation, multimodal perception, and multimedia technologies.

The development of interactive music systems, from musical instruments to music environments, engages music research in a number of challenging problems related to mediation technology. A basic problem concerns the adaptation of electronic equipment to the action-oriented ontology of the human subject, so that technology-based mediation can create an illusion of non-mediation. This illusion may form the basis of an interaction between minds (real and/or artificial), which is the ultimate goal of musical communication.

The relationship between physical energy and mental processing makes the study of interactive music systems challenging and places it at the edge of many new developments in science and technology. Moreover, the systems have an enormous potential for the production of art, and they can also be used for multimodal scientific research and programs in human education and therapy.