Chapter 10

Multimodal interaction

Giovanni De Poli and Federico Avanzini

In human-human communication, interpreting the mix of audio-visual signals is essential in communicating. Researchers in many fields recognize this, and thanks to advances in the development of unimodal techniques (in speech and audio processing, computer vision, etc.), and in hardware technologies (inexpensive cameras and other types of sensors), there has been a significant growth in multimodal interaction research.

As in human-human communication, however, effective communication is likely to take place when different input devices are used in combination. Multimodal interfaces have been shown to have many advantages: they prevent errors, bring robustness to the interface, help the user to correct errors or recover from them more easily, bring more bandwidth to the communication, and add alternative communication methods to different situations and environments. Disambiguation of error-prone modalities using multimodal interfaces is one important motivation for the use of multiple modalities in many systems.

A multimodal system is simply one that responds to inputs in more than one modality or communication channel (e.g., speech, gesture, writing, and others). We use a human-centred approach and by modality we mean mode of communication according to human senses and computer input devices activated by humans or measuring human qualities.The human senses are sight, touch, hearing, smell, and taste. The input modalities of many computer input devices can be considered to correspond to human senses: cameras (sight), haptic sensors (touch), microphones (hearing), olfactory(smell), and even taste. Many other computer input devices activated by humans, however, can be considered to correspond to a combination of human senses, or to none at all: keyboard, mouse, writing tablet, motion input(e.g., the device itself is moved for interaction), galvanic skin response, and other biometric sensors.

In the context of HCI, multimodal techniques can be used to construct many different types of interfaces. Of particular interest are perceptual, attentive, and enactive interfaces. Perceptual interfaces

are highly interactive, multimodal interfaces that enable rich, natural, and efficient interaction with computers. Perceptual interfaces seek to leverage sensing (input)and rendering (output) technologies in order to provide interactions not feasible with standard interfaces and common I/O devices such as the keyboard, the mouse, and the monitor, making computer vision a central component in many cases. Attentive interfaces are context-aware interfaces that rely on a person's attention as the primary input - that is, attentive interfaces use gathered information to estimate the best time and approach for communicating with the user. Since attention is epitomized by eye contact and gestures (although other measures such as mouse movement can be indicative), computer vision plays a major role in attentive interfaces. Enactive interfaces are those that help users communicate a form of knowledge based on the active use of the hands or body for apprehension tasks. Enactive knowledge is not simply multisensory mediated knowledge, but knowledge stored in the form of motor responses and acquired by the act of *doing*. Typical examples are the competence required by tasks such as typing, driving a car, dancing, playing a musical instrument, and modelling objects from clay. All of these tasks would be difficult to describe in an iconic or symbolic form.

In this chapter different perspectives on multimodal interaction, with special emphasis on sound, music and creativity, will be presented.

10.1 Research paradigms on sound and sense

In this section[1] it is shown that the modern scientific approach to sound and music computing has historical roots in research traditions that aimed at understanding the relationship between sound and sense, physics and meaning. This chapter gives a historical-philosophical overview of the different approaches that led to the current computational and empirical approaches to music cognition. It is shown that music cognition research has evolved from Cartesian dualism with a rather strict separation between sound and sense, to an approach in which sense is seen as embodied and strongly connected to sound. Along this development, music research has always been a driver for new developments that aim at bridging the gap between sound and sense. This culminates in recent studies on gesture and artistic applications of music mediation technologies.

In all human musical activities, musical sound and sense are tightly related with each other. Sound appeals to the physical environment and its moving objects, whereas sense is about private feelings and the meanings it generates. Sound can be described in an objective way, using instruments and machines. In contrast, sense is a subjective experience which may require a personal interpretation in order to explicit this experience. How are the two related with each other? And what is the aim of understanding their relationship?

10.1.1 From music philosophy to music science

Ancient Greek philosophers such as Pythagoras, Aristoxenos, Plato and Aristotle had quite different opinions about the relationship between sound and sense. Pythagoras mainly addressed the physical aspects by considering a mathematical order underlying harmonic pitch relationships. In contrast, Aristoxenos addressed perception and musical experience. Plato comes into the picture mainly because he attributed strong powers to music, thus, a strong effect from sound to sense. However, for him, it was a reason to abandon certain types of music because of the weakening effect music could have on the virtue of young people. Aristotle understood the relationship between sound and sense in terms of a mimesis theory (e.g. Politics, Part V). In this theory, he stated that rhythms and melodies

[1] adatpted from Marc Leman, Frederik Styns and Nicola Bernardini S2S2book Polotti and Rocchesso [2008]

contain similarities with the true nature of qualities in human character, such as anger, gentleness, courage, temperance and the contrary qualities. By imitating the qualities that these characters exhibit in music, our souls[2] are moved in a similar way, so that we become in tune with the affects we experience when confronted with the original. For example, when we hear imitations of men in action in music, then our feelings tend to move in sympathy with the original. When listening to music, our soul thus undergoes changes in tune with the affective character being imitated.

Understood in modern terms, Aristotle thus observed a close connection between sound and sense in that the soul would tend to move along or resonate with sound features in music that mimic dynamic processes (gestures, expressions) in humans. Anyhow, with these views on acoustics (Pythagoras' approach to music as ratios of numbers), musical experience (Aristoxenos' approach to music as perceived structure), and musical expressiveness (Aristotle's approach to music as imitation of reality), there was sufficient material for a few centuries of philosophical discussion about sound to sense relationships.

All this started again from scratch in the 17th century with the introduction of the so-called Cartesian dualism, which states that sound and sense are two entirely different things. He divided music into three basic components, namely, (i) the mathematical-physical aspect (Pythagoras), (ii) the sensory perception (Aristoxenos), and (iii) the ultimate effect of music perception on the individual listener's soul (or mind) (Aristotle). To Descartes, it is sound, and to some degree also sensory perception, that can be the subject of a scientific study. The reason is that sound, as well as the human ear, deal with physical objects. Since objects have extension and can be put into motion, we can apply our mathematical methods to them. In contrast, the effect of perception and the meaning that ensues from it resides in the soul. There it can be a subject of introspection. In that respect, sense is less suitable to scientific study because sense has no extension.

In more recent times, this concept will reappear as "body image" and "body schema". So far, Descartes' approach thus clearly distinguished sound and sense. His focus on moving objects opened the way for scientific investigations in acoustics and psychoacoustics, and it pushed matters related to sense and meaning a bit further away towards a disembodied mental phenomenon.

Like Descartes, many scientists working on music often stressed the mathematical and physical aspects, whereas the link with musical meaning was more a practical consequence. For example, the calculation of pitch tunings for clavichord instruments was often (and sometimes still is) considered to be a purely mathematical problem, yet it had consequences for the development of the harmonic and tonal system and the way it touches our sense of tone relationships and, ultimately, our mood and emotional involvement with music. Emotions and expressive gestures were not considered to be a genuine topic of scientific study. Emotions and expressive gestures were too much influenced by sound, and therefore, since they were not based on pure thinking, they were prone to error and not reliable as a basis for scientific study and knowledge. Thus, while Plato and Aristotle saw a connection between sound and sense through mimesis, Descartes claimed that sound and sense had a different ontology.

Parallel with this approach, the traditions of Aristoxenos and Aristotle also culminated in rule-based accounts of musical practices such as Zarlino's, and later Rameau's and Mattheson's. Somehow, there was the feeling that aspects of perception which closely adhere to the perception of syntax and structure had a foundation in acoustics. Yet not all aspects could be explained by it. The real experience of its existence, the associated feeling, mood and pleasure were believed to belong to the subject's private life, which was inaccessible to scientific investigation.

[2]The concept of soul is an old philosophical concept while modern philosophers tend to relate to the more modern concepts of "self", "ego", or even "mind".

Descartes' dualism had a tremendous impact on scientific thinking and in particular also on music research. The science of sound and the practice of musical experience and sense were no longer connected by a common concept. Sound was the subject of a scientific theory, while sense was still considered to be the by-product of something subjective that is done with sound. Apart from sensory perception (e.g. roughness in view of tuning systems), there was no real scientific theory of sense, and so, the gap between mind and matter, sense and sound, remained large.

10.1.2 The cognitive approach

Empirical studies of subjective involvement with music started to take place in the 19th century. Through the disciplines of psychophysics and psychology, the idea was launched that between sound and sense there is the human brain, whose principles could also be understood in terms of psychic principles and later on, as principles of information processing. With this development, the processes that underlay musical sense come into the picture.

10.1.2.1 Psychoacoustics

With the introduction of psychoacoustics by von Helmholtz (1863), the foundations were laid for an information processing approach to the sound/sense relationship. Helmholtz assumed that musical sensing, and ultimately, its experience and sense, was based on physiological mechanisms in the human ear. This idea became very influential in music research because it provided an explanation of why some very fundamental structural aspects of musical sense, such as consonance and dissonance, harmony and tonality, had an impact on our sense. This impact was no longer purely a matter of acoustics, but also of the working of our sensing system. Through scientific experiments, the causality of the mechanisms (still seen as a moving object) could be understood and mathematical functions could capture the main input/output relationships. This approach provided the foundation for experimental psychology and later for Gestalt psychology in the first half of the 20th century, and the cognitive sciences approach of the second half of the 20th century.

10.1.2.2 Gestalt psychology

The Gestalt movement gained prominence by about 1920. It had a major focus on sense as the perception and representation of musical structure, including the perception of tone distances and intervals, melodies, timbre, as well as rhythmic structures. The Gestalt approach influenced music research in that it promoted a thorough structural and cognitive account of music perception based on the idea that sense emerges as a global pattern from the information processing of patterns contained in musical sound.

10.1.2.3 Information theory

It also gradually became clear that technology would become an important methodological pillar of music research, next to experimentation. Soon after 1945, with the introduction of electronics and the collaboration between engineers and composers, electronic equipment was used for music production activities, and there was a need for tools that would connect musical thinking with sound energies. This was a major step in the development of technologies which extend the human mind to the electronic domain in which music is stored and processed. Notions such as entropy and channel capacity provided objective measures of the amount of information contained in music and the amount of information that could possibly be captured by the devices that process music. The link from

information to sense was easily made. Music, after all, was traditionally conceived of in terms of structural parameters such as pitch and duration. Information theory thus provided a measurement, and thus a higher-level description, for the formal aspects of musical sense. Owing to the fact that media technology allowed the realisation of these parameters into sonic forms, information theory could be seen as an approach to an objective and relevant description of musical sense.

10.1.2.4 Symbol-based modelling of cognition

The advent of computers marked a new area in music research. Computers could replace analogue equipment, but, apart from that, it also became possible to model the mind according to the Cartesian distinction between mind and matter. Based on information processing psychology and formal linguistics it was believed that the *res cogitans* (related to that aspect of the mind which also Descartes had separated from matter) could be captured in terms of symbolic reasoning. Computers now made it possible to mimic human ”intelligence” and develop an ”artificial intelligence”.

Cognitive science, as the new trend was called, conceived the human mind in terms of a machine that manipulates representations of content on a formal basis. The application of the symbol-based paradigm to music was very appealing. However, the major feature of this approach is that it works with a conceptualisation of the world which is cast in symbols, while in general it is difficult to predefine the algorithms that should extract the conceptualised features from the environment.

10.1.2.5 Subsymbol-based modelling of cognition

In the 1980s, based on the results of the so-called connectionist computation, a shift of paradigm from symbol-based modelling to subsymbol-based modelling was initiated. Connectionism (re)introduced statistics as the main modelling technique for making connections between sound and sense. This approach was rather appealing for music research because it could take into account the natural constraints of sound properties better than the symbol-based approach could. By including representations of sound properties (rather than focusing on symbolic descriptions which are devoid of these sound properties), the subsymbol-based approach was more in line with the naturalistic epistemology of traditional musicological thinking. It held the promise of an ecological theory of music in which sound and sense could be considered as a unity. The method promised an integrated approach to psychoacoustics, auditory physiology, Gestalt perception, self-organisation and cognition, but its major limitation, however, was that it still focused exclusively on perception.

10.1.3 Beyond cognition

The cognitive tradition was criticised for several reasons. One reason was the fact that it neglected the subjective component in the subject’s involvement with the environment. Another reason was that it neglected action components in perception and therefore remained too much focused on structure and form.

10.1.3.1 Embodied music cognition

The action-based viewpoint has generated a lot of interest and a new perspective on how to approach the sound/sense relationship. In this approach, the link between sound and sense is based on the role of action as mediator between physical energy and meaning. In the cognitive approach the sound/sense relationship was mainly conceived from the point of view of mental processing. The approach was

effective in acoustics and structural understanding of music, but it was less concerned with action, gestures and emotional involvement. In that respect, one could say that the Aristotelian component, with its focus on mimesis as binding component between sound and sense, was not part of the cognitive programme, nor was multi-modal information processing, or the issue of action-relevant perception (as reflected in the ecological psychology of Gibson).

To sum up, the embodied cognition approach states that the sound/sense relationship is mediated by the human body, and this is put as an alternative to the disembodied cognition approach where the mind is considered to be functioning on its own. The embodied cognition approach of the early 20th century is largely in agreement with recent thinking about the connections between perception and action.

10.1.3.2 Music and emotions

The study of subjective involvement with music draws upon a long tradition of experimental psychological research in which descriptions of emotion and affect are related to descriptions of musical structure. These studies take into account a subjective experience with music. Few authors, however, have been able to relate descriptions of musical affect and emotions with descriptions of the physical structure that makes up the stimulus. Most studies, indeed, interpret the description of structure as a description of perceived structure, and not as a description of physical structure. In other words, description of musical sense proceeds in terms of perceptual categories related to pitch, duration, timbre, tempo, rhythms, and so on.

10.1.3.3 Gesture modelling

During the last decade, research has been strongly motivated by a demand for new tools in view of the interactive possibilities offered by digital media technology. This stimulated the interest in gestural foundations of musical involvement. This gestural approach has been rather influential in that it puts more emphasis on sensorimotor feedback and integration, as well as on the coupling of perception and action. It is likely that more attention to the coupling of perception and action will result in more attention to the role of corporeal involvement in music, which in turn will require more attention to multi-sensory perception, perception of movement (kinaesthesia), affective involvement, and expressiveness of music.

10.1.3.4 Physical modelling

Much of the recent interest in gesture modelling has been stimulated by advances in physical modelling. A physical model of a musical instrument generates sound on the basis of the movements of physical components that make up the musical instrument. In contrast with spectral modelling, where the sound of a musical instrument is modelled using spectral characteristics of the signal that is produced by the instrument, physical modelling focuses on the parameters that describe the instrument physically, that is, in terms of moving material object components. Sound generation is then a matter of controlling the articulatory parameters of the moving components.

Physical models, so far, are good at synthesising individual sounds of the modelled instrument. And although it is still far from evident how these models may synthesise a score in a musically interesting way - including phrasing and performance nuances - it is certain that a gesture-based account of physical modelling is the way to proceed. Humans would typically add expressiveness to their interpretation, and this expressiveness would be based on the constraints of body movements

that take particular forms and shapes, sometimes perhaps learned movement sequences and gestures depending on cultural traditions. One of the goals of gesture research related to music, therefore, aims at understanding the biomechanical and psychomotor laws that characterise human movement in the context of music production and perception.

10.1.3.5 Motor theory of perception

Physical models suggest a reconsideration of the nature of perception in view of stimulus-source relationships and gestural foundations of musical engagement.

Purves and Lotto (2003), for example, argue that invariance in perception is based on statistics of proper relationships between the stimulus and the source that produces the stimulus. Their viewpoint is largely influenced by recent studies in visual perception. Instead of dealing with feature extraction and object reconstruction on the basis of properties of single stimuli, they argue that the brain is a statistical processor which constructs its perceptions by relating the stimulus to previous knowledge about stimulus-source relationships. Such a statistics, however, assumes that aspects related to human action should be taken into account because the source cannot be known unless through action. In that respect, this approach differs from previous studies in empirical modelling, which addressed perception irrespective of action related issues. Therefore, the emphasis of empirical modelling on properties of the stimulus should be extended with studies that focus on the relationship between stimulus and source, and between perception and action.

The extension of empirical modelling with a motor theory of perception is currently a hot topic of research. It has some very important consequences for the way we conceive of music research, and in particular also for the way we look at music perception and empirical modelling.

10.1.4 Embodiment and mediation technology

The embodiment hypothesis entails that meaningful activities of humans proceed in terms of goals, values, intentions and interpretations, while the physical world in which these activities are embedded can be described from the point of view of physical energy, signal processing, features and descriptors. In normal life, where people use simple tools, this difference between the subject's experiences and the physical environment is bridged by the perceptive and active capabilities of the human body. In that perspective, the human body can be seen as the natural mediator between the subject and the physical world. The subject perceives the physical world on the basis of its subjective and action-oriented ontology, and acts accordingly using the body to realise its imagined goals. Tools are used to extend the limited capacities of natural body. This idea can be extended to the notion of mediation technology.

For example, to hit a nail into a piece of wood, I will use a hammer as an extension of my body. And by doing this, I'll focus on the nail rather than on the hammer. The hammer can easily become part of my own body image, that is, become part of the mental representation of my (extended) body. My extended body then allows my mental capacities to cross the borders of my natural human body, and by doing this, I can realise things that otherwise would not be possible. Apart from hitting nails, I can ride a bike to go to the library, I can make music by playing an instrument, or I can use my computer to access digital music. For that reason, technologies that bridge the gap between our mind and the surrounding physical environment are called mediation technologies. The hammer, the bike, the musical instrument and the computer are mediation technologies. They influence the way in which connections between human experience (sense) and the physical environment (e.g. sound) can take place.

Mediation concerns the intermediary processes that bridge the semantic gap between the human approach (subject-centered) and the physical approach (object or sound-centered), but which properties should be taken into account in order to make this translation effective? The hammer is just a straightforward case, but what about music that is digitally encoded in an mp3-player? How can we access it in a natural way, so that our mind can easily manipulate the digital environment in which music is encoded? What properties of the mediation technology would facilitate access to digitally encoded energy? What mediation tools are needed to make this access feasible and natural, and what are their properties? The answer to this question is highly dependent on our understanding of the sound/sense relationship as a natural relationship. This topic is at the core of current research in music and sound computing.

10.1.4.1 An object-centered approach to sound and sense

State-of-the-art engineering solutions are far from being sufficiently robust for use in practical sense/sound applications. For example, (Paivo, 2007) demonstrates that the classical bottom-up approach (he took the melody extraction from polyphonic audio as a case study, using state-of-the-art techniques in auditory modelling, pitch detection and frame-concatenation into music notes) has reached its performance platform. Similar observations have been made in rhythm and timbre recognition. The use of powerful stochastic and probabilistic modelling techniques (Hidden Markov Chains, Bayesian modelling, Support Vector Machines, Neural Networks) (see also http://www.ismir.net/ for publications) do not really close this gap between sense and sound much further (De Mulder et al., 2006). The link between sound and sense turns out to be a hard problem. There is a growing awareness that the engineering techniques are excellent, but that the current approaches may be too narrow. The methodological problems relate to:

- *Unimodality*: the focus has been on musical audio exclusively, whereas humans process music in a multi-modal way, involving multiple senses (modalities) such as visual information and movement.
- *Structuralism*: the focus has been on the extraction of structure from musical audio files (such as pitch, melody, harmony, tonality, rhythm) whereas humans tend to access music using subjective experiences (movement, imitation, expression, mood, affect, emotion).
- *Bottom-up*: the focus has been on bottom-up (deterministic and learning) techniques whereas humans use a lot of top-down knowledge in signification practices.
- *Perception oriented*: the focus has been on the modelling of perception and cognition whereas human perception is based on action-relevant values.
- *Object/Product-centered*: research has focused on the features of the musical object (waveform), whereas the subjective factors and the social/cultural functional context in musical activities (e.g. gender, age, education, preferences, professional, amateur) have been largely ignored.

10.1.4.2 A subject-centered approach to sound and sense

Research on gesture and subjective factors such as affects and emotions show that more input should come from a better analysis of the subjective human being and its social/cultural context. That would imply: Multi-modality: the power of integrating and combining several senses that play a role in music such as auditory, visual, haptic and kinaesthetic sensing. Integration offers more than the sum of the contributing parts as it offers a reduction in variance of the final perceptual estimate. Context-based: the study of the broader social, cultural and professional context and its effect on information

processing. Indeed, the context is of great value for the disambiguation of our perception. Similarly, the context may largely determine the goals and intended musical actions. Top-down: knowledge of the music idiom to better extract higher-level descriptors from music so that users can have easier access to these descriptors. Traditionally, top-down knowledge has been conceived as a language model. However, language models may be extended with gesture models as a way to handle stimulus disambiguation. Action: the action-oriented bias of humans, rather than the perception of structural form (or Gestalt). In other words, one could say that people do not move just in response to the music they perceive, rather they move to disambiguate their perception of music, and by doing this, they signify music. User-oriented: research should involve the user in every phase of the research. It is very important to better understand the subjective factors that determine the behavior of the user.

The subject-centered approach is complementary to the object-centered approach. Its grounding in an empirical and evidence-based methodology fits rather well with the more traditional engineering approaches. The main difference relates to its social and cultural orientation and the awareness that aspects of this orientation have a large impact on the development of mediation technology. After all, the relationship between sense and sound is not just a matter of one single individual person in relation to its musical environment. Rather, this single individual person lives in contact with other people, and in a cultural environment. Both the social and cultural environment will largely determine what music means and how it can be experienced.

10.1.5 Music as innovator

The above historical and partly philosophical overview gives but a brief account of the different approaches to the sound and sense relationship. This account is certainly incomplete and open to further refinement. Yet a striking fact in this overview is that music, in spanning a broad range of domains from sound to sense and social interaction, appears to be a major driver for innovation. This innovation appears both in the theoretical domain where the relationship between body, mind, and matter is a major issue, and in the practical domain, where music mediation technology is a major issue.

The historical overview shows that major philosophical ideas, as well as technical innovations, have come from inside music thinking and engagement. Descartes' very influential dualist philosophy of mind was first developed in a compendium on music. Gestalt theory was heavily based on music research. Later on, the embodied cognition approach was first explored by people having strong roots in music playing (e.g. Truslit was a music teacher). In a similar way, the first explorations in electronic music mediation technologies were driven by composers who wanted to have better access to the electronic tools for music creation. Many of these ideas come out of the fact that music is fully embedded in sound and that the human body tends to behave in resonance with sound, whereas the "mind's I" builds up experiences on top of this. Music nowadays challenges what is possible in terms of object-centered science and technology and it tends to push these approaches more in the direction of the human subject and its interaction with other subjects. The human way in which we deal with music is a major driver for innovation in science and technology, which often approaches music from the viewpoint of sound and derived sound-features. The innovative force coming from music is related to the subject-centered issues that are strongly associated with creativity and social-cultural factors.

The idea that music drives innovation rather than vice versa should not come as completely unexpected. Music is solidly anchored to scientific foundations and as such it is an epistemological domain which may be studied with the required scientific rigour. However, music is also an art and therefore certain ways of dealing with music do not require scientific justification per se because they justify themselves directly in signification practices. The requirements of musical expression can indeed provide a formidable thrust to scientific and technological innovation in a much more efficient way than

the usual R&D cycles may ever dream of. In short, the musical research carried out in our time by a highly specialised category of professionals (the composers) may be thought as a sort of fundamental think tank from where science and technology have extracted (and indeed, may continue to extract in the future) essential, revolutionary ideas. In short, musical expression requirements depend, in general, on large scale societal changes whose essence is captured by the sensible and attuned composers. These requirements translate quickly into specific technical requirements and needs. Thus, music acts in fact as an opaque but direct knowledge transfer channel from the subliminal requirements of emerging societies to concrete developments in science and technology.

Conclusion

This section aims at tracing the historical and philosophical antecedents of sense/sound studies in view of a modern action-oriented and social-cultural oriented music epistemology. Indeed, recent developments seem to indicate that the current interest in embodied music cognition may be expanded to social aspects of music making. In order to cross the semantic gap between sense and sound, sound and music computing research tends to expand the object-centered approach engineering with a subject-centered approach from the human sciences. The subject-centered character of music, that is, its sense, has always been a major incentive for innovation in science and technology. The modern epistemology for sound and music computing is based on the idea that sound and sense are mediated by the human body, and that technology may form an extension of this natural mediator. The chapter aims at providing a perspective from which projections into the future can be made.

The section shows that the relationship between sound and sense is one of the main themes of the history and philosophy of music research. In this overview, attention has been drawn to the fact that three components of ancient Greek thinking already provided a basis for this discussion, namely, acoustics, perception, and feeling (”movement of the soul”). Scientific experiments and technological developments were first (17th - 18th century) based on an understanding of the physical principles and then (starting from the late 19th century) based on an understanding of the subjective principles, starting with principles of perception of structure, towards a better understanding of principles that underlay emotional understanding.

During the course of history, the problem of music mediation was a main motivating factor for progress in scientific thinking about the sound/sense relationship. This problem was first explored as an extension of acoustic theory to the design of music instruments, in particular, the design of scale tuning. In modern times this problem is explored as an extension of the human body as mediator between sound and sense. In the 19th century, the main contribution was the introduction of an experimental methodology and the idea that the human brain is the actual mediator between sound and sense.

In the last decades, the scientific approach to the sound/sense relationship has been strongly driven by experiments and computer modelling. Technology has played an increasingly important role, first as measuring instrument, later as modelling tool, and more recently as music mediation tools which allow access to the digital domain. The approach started from a cognitive science (which adopted Cartesian dualism) and symbolic modelling, and evolved to sub-symbolic modelling and empirical modelling in the late 1980ies. In the recent decades, more attention has been drawn to the idea that the actual mediator between sound and sense is the human body.

With regards to new trends in embodied cognition, it turns out that the idea of the human body as a natural mediator between sound and sense is not entirely a recent phenomenon, because these ideas have been explored by researchers such as Lipps, Truslit, Becking, and many others. What it offers is a possible solution to the sound/sense dichotomy by saying that the mind is fully embodied,

that is, connected to body. Scientific study of this relationship, based on novel insights of the close relationship between perception and action, is now possible thanks to modern technologies that former generations of thinkers did not have at their disposal.

A general conclusion to be drawn from this overview is that the scientific methodology has been expanding from purely physical issues (music as sound) to more subjective issues (music as sense). Scientists conceived these transition processes often in relation to philosophical issues such as the mind-body problem, the problem of intentionality and how perception relates to action. While the sound/sense relationship was first predominantly considered from a cognitive/structural point of view, this viewpoint has gradually been broadened and more attention has been devoted to the human body as the natural mediator between sound and sense. Perception is no longer conceived in terms of stimulus and extraction of structures. Instead, perception is conceived within the context of stimulus disambiguation and simulated action, with the possibility of having loops of action-driven perception. This change in approach has important consequences for the future research. Music has thereby been identified as an important driver for innovation in science and technology. The forces behind that achievement are rooted in the fact that music has a strong appeal to multi-modality, top-down knowledge, context-based influences and other subject-centered issues which strongly challenge the old disembodied Cartesian approaches to scientific thinking and technology development.

10.2 Enaction, Arts and Creativity

Enaction and Creativity are two concepts difficult to define precisely. Enaction is here understood in a large sense of considering the role of action (and further of interaction, action could not exist without interaction) at the center of the human activities whatever they are, for human biological survival as well for human cultural creation of new objects or symbols. Creativity will be considered in this section[3] according two simple meanings:

- mainly as synonymous of Artistic Creation Process
- and as creative processes in design and modelling activities.

Since the XIX century industrial revolution, creation activity in arts was mainly considered as an abstract activity starting the clearly cut separation between composer and instrumentalists in music and designers and producers in fine arts; choreographers and dancers in choreographic arts. The apogee of such period was at the middle of the XX century with the primacy of formal approaches in arts, as the serialism in music (synonymous of contemporary music) or the conceptualism in fine arts (synonymous of ”contemporary arts).

At the beginning of the use of the computer in arts (at the middle if sixties), the main stream of theories and uses focuses on the conquest of ”immateriality” allowed by computer. Keywords were ”overcome the limit of the matter”, ”reach a pure thinking of musical cues”, ”Music for mind”, ”Abstraction for visual cues”, ”breaking the real”, etc.

Recently, about ten of twenty years ago, after the relative failure of such extreme theories and points of view, and under the recent technological propositions of interactivity allowing the computers to be more and more adapted to the human senses and action, arts became more and more interactive. The ”instrumentality” is progressively re-introduced as a design locus through its sub-instance of interactivity. The role of ”gestures” has been rehabilitated, not only to produce sensorial predefined events but to properly create artistic properties. In music, performance, previously considered as the end of the musical production process, as a kind of ”sonification” of the musical pre-written score, was

[3] adapted from Luciani Enactive Interfaces NoE WP13 report (2005)

rehabilitated as a creative process in itself, not only in musical improvisations as in specific musical styles (jazz, free music, etc.) but as a creative process in itself, as in open or interactive composition. Such approaches shift the creation process from the formal organization of musical of visual events to the production process itself. Simultaneously, the role of the "instrument" as a tangible object able to feed and steer the creative process by imposing constraints and of the "instrument trade" was rehabilitated against the "free constraint approaches".

Such theoretical shift is totally in adequacy with the concept of enaction. More, Arts is probably the realm (beside biology), in which high level media of communication and of cultural data are produced by means of closed-loop sensori-motor interaction.

This historical movement is particularly clear in Music, that is an "*allographic art*", needed another way of representation and of design - the musical graphical notation -different than itself. It is less evident in others arts, that are "*autographic arts*", as visual arts or choreographic arts, meaning that they are in themselves their own tools of representation and design. However, it traverses all the Arts that we called "Dynamic Instrumental Arts". "Instrumental arts" refer to arts that need physical medium (object, body) to exist. Dynamic refer to the fact that at any stage of its production process, sensorial artistic events are evolving events. Basic Dynamic Instrumental Arts are Music, Visual Arts as animation of fine arts, Choreographic arts. Each of them addresses the question of the role of the instrument and of the interaction between artists and their instruments in specific ways according to their own historical positioning and their own particularisms.

Indeed, the word "*instrument*" is usually reserved to the musical realm. However, if we dare to use it in a more general meaning, as a physical mediator able to produce exteroceptive stimuli, visual and auditory, by an action of human body on it, we understand immediately that all the arts that need such mediator are necessarily temporally-based and interaction-based. Physical interaction, sensory-motor coupling, gesture, instrument, movement, etc. are complementary components that are always present and that are always cooperating in all sensory-based (conversely than language based) arts. Conversely, the question of the link with the interaction performance activity and the conceptual processes is raise as one of their core question. One main property of such "instrumental concept" is to reveal the implicit familiarity of artistic creation process with the "*enaction concept*".

In Musical arts, the concept of instrumentality is an ancestral concept that exists from the origin of the music and of the sound production. The pair instrument -instrumentalist is always present in music, even in computer music with the field of "Digital Musical Instruments". The main fundamental question in music is the link between the inevitable instrumental process and the musical notation and composition, and the link of the musical composition with musical perception and cognition.

In Visual Arts, two sub-domains have to be distinguished: arts that produce "static objects and events" (sculpture, paintings, etc.) and arts that produce "movements" or "moving objects and visual events" (movies, automata, animation). In the first case, as in Music, instrumentality is a native and ancient practice. The role of their matter and of the interaction with it is widely recognized and respected in such artistic style as well as in all craft practices. Differently than Music, as autographic arts, they don't need external and foreign way of notation and for their design and their composition. There is no so dramatic problem of notation and composition as it exists Music and no dramatic crack between compositional activity and other musical activities. In the second case, except in some minor cases, as shadows' theater or puppet theater, instrumentality is less difficult to define before the arrival of the computer. We cannot play with objects producing visual events as we are able to play with a violin. Movies and animation using conventional media (cinema or video) do not implement explicitly the instrumental concept. Similarly than in Music, the question of the motion notation is of a crucial and dramatic importance.

From the point of view of the novelty brought by computers, the two types of visual arts (static and moving) have been differently fed:

- computers trend to devaluate and to minorate the type of craft manual and interactional process;
- computers triggered really a revolution in the visual art of motion by allowing the designing of "objects" that can be manipulated as " violin" to produce visual evolving events.

In both cases, Enactive Interfaces and Enactive Knowledge are means to experiment and to rehabilitate the prominent role of the interaction and of the matter in the visual artistic process.

In Choreographic arts, as in theater arts, "instrumentality" is not an explicit usual concept, the human body being its own instrument. The concept of instrument has been introduced recently with the introduction of the notion of "augmented body" by external devices and equipments able at least to capture the motion of the body. Such motion, transformed in a signal, becomes an "object" that can be processed and applied to control other objects and others instruments. Computers trends to bring together musical arts, visual dynamic arts and choreographic arts, the common concept becoming the instrumental concept with all its derivatives: interaction and interactive control. As in music and in visual motion, the core difficult non-solved question is this of the notation and of the composition of such evolving events.

Summarizing in a differentiate way the major questions risen by each of the main Dynamic Instrumental Arts, we can say:

- In *Music*, the haunting question being the relation between instrumentality and composition, are new computers tools and new ways of interaction with computerized instruments, able to overcome this frontier or not? Are Enactive Interfaces able to reconcile the opposites, the enemy brothers?
- In *Visual static arts*, is the generalization of interactivity concept able to instil in the production process, as in craft process, the minimum of instrumentality required to support craft know-how?
- In *Visual Dynamic Arts*, is the notion of virtual manipulable objects able to produce visual dynamic arts with the same level of quality for the visual shapes and for the expressivity of the motion? and is the motion processing able to overcome the duality between space (autographic representations) and time (allographic representations)?
- In *choreographic arts*, is computers a step in the motion representation without the creation of a break between choreographic performance and choreographic design, that is nowadays a core and passionate question?

From such contemporary questions asked by such arts, near to the enactive concept, to computers, some relevant but non-exhaustive issues can be listed:

1. What common issue? Is the motion and the gesture as a specific motion which represents action - its processing, its rendering, its production, its notation - a common feature shared by all such instrumental dynamic arts? Could the motion and the gesture the common mean to bring them near or to merge them in a very novel and genuine way?
2. What types of computer models and computer representations and interfaces should be the best candidates to receive gestures and to produce genuine movements;
3. What type of links between the primary evolving event (gesture, movement, action) and the sensory outputs visual and auditory? Trivial links only as those proposed now in computer graphics and animation? Arbitrary links as those proposed in the mapping process in computer music? Others links? Can we speak of gestures composition independently of the 3D object

that is receiving or producing such motions? Can we apply every kind of gestures and action on every type of production process?

4. What types of design processes and link between the design process and the performance processes
5. What should be the relation between the enactive concept, well revealed by the necessity of the gestural interaction, and the artistic emotion? What is the role of the instrument and of the interaction in the shift from the production process to the aesthetic process? From the history of our artistic tools and theories, nowadays, we only know only that such mediator, such interaction cannot be totally avoided.

10.3 Some core questions about creativity: a philosophical and linguistic point of view

In this short section[4], Roberto Casati (Institut Nicod, Paris) addresses some basic issues about creativity in a question-and-answer manner.

10.3.1 Creativity: eight basic questions

What is creativity?

This question is hard to answer, as is any "what is" question - and this difficulty is crucial. But much depends on it. First of all, we talk both about creative ideas and methods, and about creative people. Obviously the sense in which we talk about creative individuals completely depends on the sense that we give to the notion of creative ideas. We say that someone is creative when it has a creative idea. We do not say, on the other hand, that an idea or a method is creative because it comes from a creative individual, as if there was a magic creativity touch. A creative individual is one with creative ideas. First comes the idea and the process, then the label. It is important to see this because otherwise one is led to think that there is a sort of magic touch, a creativity faculty, or a creativity gift, that some people have and others do not. There may be people who produce more creative ideas than others, of course, but this depends on many factors, most of which are contextual, and does not depend on a mysterious creativity system of the brain.

So, what is a creative idea, a creative thought or method?

We all seem to be able, intuitively, to distinguish between a creative idea and an idea that is not so creative. But how do we actually draw the line between creative and non-creative ideas, and is there a fact of the matter that can justify the way we draw the line?

From a cognitive point of view most of what we know about creativity comes from our understanding of how language works. Think of the sentence I just uttered: "From a cognitive point of view most of what we know about creativity comes from our understanding of how language works". It is very likely to be the first time this sentence has been uttered in the whole history of mankind. I never heard it before anyway. And I never pronounced those words before. As least as it concerns me, I created it. But so is with most sentences, for purely combinatorial reasons. Think of a simple language in which sentences are made by simple juxtaposition of words (a very simple grammar). It there were only two words in this language, and you had a rule that sentences have exactly 8 words, then you

[4] adapted from R. Casati Enactive Interfaces NoE, WP13 report (2005).

would have 256 sentences. In a real language there are so many words, and no limit to the lengths of sentences. There is, indeed, an infinity of sentences to choose from. So creativity in language is almost mandatory -you cannot help, you are almost automatically creative.

We overlook this fact because it is so familiar for us, and it is taken for granted. But think of it. Sometimes indeed we catch ourselves in the act of repeating ourselves. We immediately recognize this fact. We would be surprised. If on the other hand someone repeats himself, we are annoyed. Either way, we seem very good at detecting non-creative behavior.

So, we are linguistically creative. Each of us is creative when it comes to speaking. When speaking, no one ever repeats, with some very clear exceptions, things that one has heard. If I tried to repeat word by word the sentence I mentioned before, I would not be able to; I would probably get more or less the same thought, but not the exact wording for it.

The crucial point is that this type of creativity in turn depends on there being rules that allow us to produce certain combinations of words and not others. I stress this point because there seems to be a romantic idea around in informal discussions about creativity: that creativity is a ”breaking of the rules”. We can agree that in certain contexts creativity is partly that, but the situation is far more complex. Creativity in language only exists because there are rules that allow us to form sentences out of words. Sentences (this is an important point) that you can understand. I am not creative at all if I say, ”Grumpziso nemosenn ximadou”, or ”Subtly or John Cage”, and I am creative if I say ”the Moon is, actually, a giant dark stone”. I cannot claim creativity if no one can understand what I say because I made up an invented language or made up a forbidden (ungrammatical) sentence.

But isn’t there a difference between linguistic creativity and creativity in thought?

Indeed. I said that I may not be able to repeat the same words, but I would be able maybe to express the same thought. This means that language-creativity should be distinct from thought-creativity.

So what is it to create new thoughts, new ideas?

My suggestion is that the process is in principle not different from the process at work when we produce new sentences.

Let me state some conditions for creativity, that is, for recognition of an idea as creative.

A. When we recognize that an idea is creative, we recognize that it is new. But ”new” presupposes that we also recognize how the situation would have been without that idea.

B. There is a (limited) tolerance for simultaneous creativity and novelty: two people can come up with a new idea at the same time. It is not uniqueness that makes an idea creative.

C. Not only that. We tend to consider as creative those ideas that are solutions to problems. Someone just coming up with a novel scream is not particularly creative. ”Creative” is reserved for ideas that come in and help making a progress on a background of an existing problem space. Here I am using ”solutions” and ”problems” in a very wide sense. There may be problems in jazz, in the arts, as there are in rationalizing the work-flow and in the engineering of an apparatus. There are mathematical problems and translation problems. All these may require creative solutions -as opposed to routine solutions. This means the above-mentioned romantic notion of creativity should give way to a less ambitious, but more true to the facts, description of how people come up with new ideas. It is tricky and hard to find out how this happens and I do not think that there is enough research available to make a final statement (More about this later).

D. The existence of a set of rules is a precondition for creativity. There are for many reasons. For an idea to be recognized as creative, people must be able to see that it is new. But they can

only see its novelty if they understand the alternatives (the non-creative alternatives). Think of language again. In order for you to come up with a new sentence in English, you must know (implicitly) the rules of English.

Analogy, metaphors, and perceptual problem solving have been presented in the literature as the nuts and bolts of creative processes. For instance, problem solving has been associated to a kind of perception. It is a way to see things in a new perspective. It is a "aha!" experience, similar to the one we undergo when we recognize a face in a set of lines, or a dalmatian dog in a set of patches.

Some philosophers and cognitive scientists have tried to develop "artificial creativity". Why is this interesting?

By reflecting on how to produce machines that are recognized as creative, one may on the way find out some so far overlooked features of creativity. Among others, Douglas Hofstadter (1995) and Paul Thagard (1995) - the latter has studied how analogy enters scientific thought. In both case it is understood that some creative ideas come from the use of analogy. You look at an object, it reminds you of something else you know, then you use some of the properties of the old thing to improve on the new thing (as an example, think of the desktop analogy for running your computer.)

Margared Boden (1990, 1994), a philosopher of artificial intelligence, has distinguished a "combinational creativity" which mainly consists in a new arrangement of existing ideas; and a different creativity, in which not only there is a solution to an existing problem, but there is the creation of a new problem. This later type of creativity is "a structured, disciplined, sometimes even systematic search for the meanings promised by the new idea."

Boden thinks that creativity is basically an exploration of a conceptual space. A conceptual space can be a set of constraints. One must first accept the constraints, then explore the space. Imagine a railroad example: If you set constraints to movement (railroads force you not to make narrow turns, you have no steps), thereby allowing for lesser degrees of freedom, you will be forced to find a solution to the path between two points that must be creative. So you may end up inventing tunnels and bridges, to keep your railroad even and to minimize bending. Or, you can put constraints on colors, and decide to take pictures in black and white. Then you have to act creatively in order to enhance contrasts for colors that have the same brightness, such as red and greens. Or again, you can decide to compose in jazz. The range of what you can compose is small as compared to the range of sounds that you can concatenate, and this forces you to find interesting concatenations.

Sometimes constraints depend on social acceptance. Sometimes they depend on perception. Not everything that is possible is perceptually valid. It is no good to loosen too much the time constraints on music. You cannot sensibly listen to a piece of music whose beginning is now, whose second note is in two years, etc.

It appears then that one type of creativity consists in exploring a given, preset space, a given a set of constraints. Another type consists in inventing new constraints, thereby creating a new space. (A new constraint space is like a new style.) Of course it is of no avail to invent a new style and then completely rigidify it -the constraints should allow for a certain degree of freedom. A good new style is one with productive constraints, constraints that allow us to create within the style.

Hence I would add another principle to the conditions for creativity:

E. Creativity within a given space is fast; creating a new space is slow. In order to be creative, and idea must be recognized as creative, but if you change the language, the rules of the game, it will take time for people to see this. So you have to put great care in making sure that the transition to the new language, to the new game, is understood.)

Is creativity necessary for humans? If so, why?

Socially, creativity is clearly more than encouraged: it is taken for granted. Think, again, of language, of an ordinary conversation, such as the ones we all have and enjoy with friends. Social pressure is merciless. Apart from some very clearly defined contexts, people are required to be creative all the time. Imagine I repeat exactly, word by word, what a friend just said. Except for ironical usages, or for a request for clarification, this behavior would not be permissible, or it would be considered boring. Suppose, further, that I repeat word by word what the dominant personality of the group said. I would be stigmatized as servile. Suppose, again, that I repeated word by word something I just said, three or four times. I would be considered weird. Suppose, finally, that I say something quite articulate, and the next day some friend finds out that I reported, word by word, an opinion expressed in a newspaper. Then I would be considered a cheater. Boring, servile, weird, cheater - all these are verbal punishment for not having been creative. Again, creativity is taken for granted. It is the natural condition of us all in many social contexts.

So if there is all this pressure on creativity, it may well be that it depends on some evolutionary advantage that creative societies have had upon noncreative societies?

It may well be, but I hesitate in suggesting evolutionary explanations. Still, the explanation is suggestive. Cooperative behavior is another case for which an evolutionary explanation has been suggested. Cooperative societies may fare better than non-cooperative societies, and this is why we tend to cooperate so much, even in case in which we would be, on a local scale, better off if we defaulted on cooperation. So if creativity was necessary, because it led to improved social settings, its adaptive advantage may have made it happen that evolution "discovered" it -and reinforced it socially. But, again, these are very rough speculations.

Is there a creative " type "? Are some people predisposed to be more creative than others?

This is an important point. The problem is that it may well be the case that "creativity" does not delineate a category at all. As I suggested before, there seem not there be any "creative" system in the brain. I would like to make a comparison with other so-called "personality" traits. Common sense treats people as courageous, as cowards, as intelligent or stupid. Some scientists have tried to come up with methods for measuring these alleged "traits" of people. The results are not very encouraging: It is hard to pinpoint a stupid behavior out of very specific contexts. There appears to be no general stupidity that manifests itself in many repeated occasions. What is worse, the occasions dump systematically the alleged trait. Two very famous experiments, dating back from the '50, and oftentimes replicated, have shown this point quite dramatically. In the Millgram experiment, it turned out that most people, even those who thought of themselves as provided with a strong personality, would inflict painful experiences on fellows just because someone told them to do so. In the Good Samaritan experiment, people were shown be disposed to help someone in distress only if they had time to do it. So, unless you are prepared to say that most people are evil, you should give up personality traits as useful explanatory categories. People just tend to maximize the coherence of their actions and beliefs on a very narrow temporal scale: this is why it is easy to convince people to do even horrible things -on pain of losing face, say. If personality traits do not exist, or are trumped by contingencies of the situation, then there is no creativity trait - or if there is, we can always find a way to trump it. And then the question arises of how is it possible to stimulate creativity. The answer is that one should try and engineer situations in which people would naturally come up with new solutions.

10.4 Auditory displays and sound design

The goal of this section[5] is to provide an overview of research in Sound Design and Auditory Display, from warning design and computer auditory display to Architecture and Media.

The section organization into six main subsection reflects the topics that have been most extensively studied in the literature, i.e., warnings, earcons, auditory icons, mapping, sonification, and sound design. Indeed, there are wide overlaps between these areas (e.g., sound design and mapping can be considered as part of a task of sonification).

10.4.1 Warnings, Alerts and Audio Feedback

Auditory warnings are perhaps the only kind of auditory displays that have been thoroughly studied and for whom solid guidelines and best design practices have been formulated. We can identify five areas of applications for auditory warnings: personal devices, transport, military, control rooms and geographic-scale alerts.

The scientific approach to auditory warnings is usually divided into the two phases of hearing and understanding, the latter being influenced by training, design, and number of signals in the set. Studies in hearing triggered classic guidelines: for example alarms should be set between 15 and 25 dB above the masked threshold of environment. Moreover they faced also the issue of design for understanding, by suggesting a sound coding system that would allow mapping different levels of urgency. The problem of the legacy with traditional warnings is important: e.g. sirens are usually associated with danger, and horns with mechanical failures. The retention of auditory signals is usually limited to 4 to 7 items that can be acquired quickly, while going beyond is hard. In order to ease the recalls, it is important to design the temporal pattern accurately. Moreover, there is a substantial difference in discriminating signals in absolute or relative terms. Alarm-related behaviors can be classified as Alarm-Initiated Activities (AIA) in routine events (where ready-made responses are adequate) and critical events (where deductive reasoning is needed). Designing good warnings means balancing between attention-getting quality of sound and impact on routine performance of operators.

"Auditory warning affordances" investigates on the use of "ecological" stimuli as auditory warnings. The expectation is that sounds that are representative of the event to which they are alarming would be more easily learnt and retained. By using evocative sounds, auditory warnings should express a potential for action: for instance, sound from a syringe pump should confer the notion of replacing the drug.

Some results are that:

- Learned mappings are not easy to override;
- There is a general resistance to radical departures in alarm design practice;
- Suitability of a sound is easily outweighed by lack of identifiability of an alarm function;
- Need for participatory design practice.

However, for affordances that are learnt through long-time practice, performance may still be poor if an abstract sound is chosen.

Case study: "acqua alta" in Venice Special cases of warnings are found where it is necessary to alert many people simultaneously. Sometimes, these people are geographically spread, and new criteria for designing auditory displays come into play. Avanzini and co-workers (2005) face the problem of a system alert for the town of Venice, periodically flooded by the so-called "acqua alta",

[5] Amalia de Götzen, Pietro Polotti, Davide Rocchesso

i.e. the high tide that covers most of the town with 10-40 cm of water. Nowadays, a system of 8 electromechanical and omnidirectional sirens provide an alert system for the whole historic town.

A study of the distribution of the signal levels throughout the town was first performed. A noise map of the current alert system used in Venice was realized by means of a technique that extracts building and terrain data from digital city maps in ArcView format with reasonable confidence and limited user intervention. Then a sound pressure level map was obtained by importing the ArcView data into SoundPLAN, an integrated software package for noise pollution simulations. This software is mainly based on a ray tracing approach. The result of the analysis was a significantly non-uniform distribution of the SPL throughout the town. One of the goals of this work is, thus, the redefinition and optimization of the distribution of the loudspeakers. The authors considered a Constraint Logic Programming (CLP) approach to the problem. CLP is particularly effective for solving combinatorial minimization problems. Various criteria were considered in proposing new emission points. For instance, the aforementioned Patterson's recommendations require that the acoustic stimulus must be about 15 dB above the background noise to be clearly perceived. Also, installation and maintenance costs make it impractical to install more than 8 to 12 loudspeakers in the city area. By taking into account all of these factors, a much more effective distribution of the SPL of the alert signals was achieved. The second main issue of this work is the sound design of the alert signals. In this sense the key questions here considered are:

- How to provide information not only about the arrival of the tide but also about the magnitude of the phenomenon;
- How to design an alert sound system that would not need any listening-training, but only verbal/textual instructions.

Being Venice a tourist town, this latter point is particularly important. It would mean that any person should intuitively understand what is going on, not only local people. The choices of the authors went towards abstract signals, i.e. earcons, structured as a couple of signals, according to the concept of "attenson" (attention-getting sounds). The two sound stages specify the rising of the tide and the tide level, respectively. Also, the stimulus must be noticeable without being threatening. The criteria for designing sounds providing different urgency levels were the variation of the fundamental frequency, the sound inharmonicity and the temporal patterns.

The validation of the model concludes the paper. The subjects did not received any training but only verbal instructions. The alert signal was proved to be effective, and no difference between Venetians and not-Venetians was detected. In conclusion, a rich alert model for a very specific situation and for a particular purpose was successfully designed and validated. The model takes into account a number of factors ranging from the topography and architecture of Venice, to the need of culturally non-biased alert signal definition, as well as to the definition of articulated signals able to convey the gravity of the event in an intuitive way.

10.4.2 Earcons

Blattner, Sumikawa and Greenberg introduced the concept of *earcons*, defining them as "non-verbal audio messages that are used in computer/user interfaces to provide information to the user about some computer object, operation or interaction". These messages are called *motives*, "brief succession of pitches arranged in such a way as to produce a tonal pattern sufficiently distinct to allow it to function as an individual recognizable entity". Earcons must be learned, since there is no intuitive link between the sound and what it represents: the earcons are abstract/musical signals as opposed to auditory icons, where natural/everyday sounds are used in order to build auditory interfaces (see Section 10.4.3).

Brewster (1998) presents a new structured approach to auditory display defining composing rules and a hierarchical organization of musical parameters (timbre, rhythm, register, etc.), in order to represent hierarchical organizations of computer files and folders. Typical applications of this work are telephone-based interfaces (TBIs), where navigation is a problem due to visual display dimensions. As already mentioned, the main idea is to define a set of sound-design/composing rules for very simple "musical atoms", the earcons, with the characteristics of being easily distinguishable one from the other.

10.4.3 Auditory Icons

Another concept has been introduced in the nineties by Graver as an earcon counterpart: *auditory icons*. The basic idea is to use natural and everyday sounds to represent actions and sounds within an interface. He individuates a fundamental aspect of our way of perceiving the surrounding environment by means of our auditory system. Trying to reply to the question "what do we hear in the world?", a first and most relevant consideration emerges: a lot of research efforts were and are devoted to the study of musical perception, while our auditory system is first of all a tool for interacting with the outer world in everyday life.

When we consciously listen to or hear more or less unconsciously "something" in our daily experience, we do not really perceive and recognize sounds but rather events and sound sources. This "natural" listening behavior is denoted by Gaver as "everyday listening" as opposed to "musical listening", where the perceptual attributes are those considered in the traditional research in audition. As an example, Gaver writes: "while listening to a string quartet we might be concerned with the patterns of sensation the sounds evoke (musical listening), or we might listen to the characteristics and identities of the instruments themselves (everyday listening). Conversely, while walking down a city street we are likely to listen to the sources of sounds - the size of an approaching car, how close it is and how quickly it is approaching." Despite the importance of non-musical and non-speech sounds, the research in this field is scarce. What Gaver writes is true: we do not really know how our senses manage to gather so much information from a situation like the one of the approaching car described above.

10.4.4 Mapping

Auditory Display in general, and Sonification in particular, are about giving an audible representation to information, events, and processes. These entities may take a variety of forms and can be reduced to space- or time-varying data. In any case, the main task of the sound designer is to find an effective mapping between the data and the auditory objects that are supposed to represent them in a way that is perceptually and cognitively meaningful.

Kramer (1994) gave a first important contribution describing the role of mediating structures between the data and the listener or, in other words, of mapping. It is important to investigate the role of mediating structures between the data and the listener or, in other words, of mapping. The term audification was proposed to indicate a "direct translation of a data waveform to the audible domain for purposes of monitoring and comprehension". Examples are found in electroencephalography, seismology and in sonar signal analysis. In sonification, instead, data are used to control a sound generation, and the generation technique is not necessarily in direct relationship to the data. For instance, we may associate pitch, loudness, and rhythm of a percussive sound source with the physical variables being read from sensors in an engine.

A major problem is how to recall the mappings. This can be done via metaphors (e.g., high pitch = up) or feelings (e.g., harsh = bad situation), and the interactions between the two. These aspects are still very hot and open for further research nowadays.

Direct mapping (Audification) The most straightforward kind of mapping is the one that takes the data to feed the digital-to-analog converters directly, thus playing back the data at an audio sampling rate. This can be of some effectiveness only if the data are temporal series, as it is the case in seismology.

The idea of listening to the data produced by seismograms to seek relevant phenomena and improve understanding is quite old. If the seismic signals are properly conditioned and transposed in frequency, they sound pretty natural to our ears, and we can use our abilities in interpreting noises in everyday conditions.

One of the main motivations for using auditory display is that there are important events that are difficult to detect in visual time-series displays of noisy data, unless using complex spectral analyzes. Conversely, these events are easily detected by ear. There are several problems that have to be faced when trying to sonify seismic data, especially related with the huge dynamic range (> 100 dB) and with the frequency bandwidth which, albeit restricted below 40 Hz, spans more than 17 octaves. Many of the mentioned problems cause headaches to visual analysts as well. In order to let relevant events audible, the recorded signals have to be subject to a certain amount of processing, like gain control, time compression, frequency shift or transposition, annotation, looping, stereo placement.

Naturalistic mapping In some cases, it is possible to use natural or mechanical sounds to convey information of various kinds. This is especially effective when the information is physically related to the reference sound sources, so that our everyday physical experience can be exploited in interpreting the sounds.

Abstract mapping Sonification often implies an abstract mapping of nonacoustic events onto acoustic events. An example of abstract mapping is Geiger counters which are used to detect and sonify ionizing radiation. An inert gas-filled tube (usually helium, neon or argon with halogens added) briefly conducts electricity when a particle or photon of radiation makes the gas conductive. The tube amplifies this conduction by a cascade effect and outputs a current pulse, which is then often displayed as audible clicks.

A good mapping can be the key to demonstrate the superiority of auditory over other forms of display for certain applications. Indeed, researchers in Sonification and Auditory Display have long been looking for the killer application for their findings and intuitions. This is especially difficult if the data are not immediately associable with sound objects, and abstract mappings have to be devised, as for example stock market data.

Musical mapping Music has its own laws and organizing principles, but sometimes these can be bent to follow flows of data.

10.5 Sonification

Sonification can be considered as the auditory equivalent of graphic representation in the visual domain. The main goal of sonification is to define a way for representing reality by means of sound.

Scaletti (1994) proposed a working definition of sonification as "a mapping of numerically represented relations in some domain under study to relations in an acoustic domain for the purpose of interpreting, understanding, or communicating relations in the domain under study."

Another less restrictive interpretation of sonification is found by transposition of what people currently intend with the word visualization.

10.5.1 Information Sound Spaces (ISS)

In his thesis work, Stephen Barras aims at defining a methodology for representing information by means of sonification processes. The initial motivation of Barrass' work could be summarized in the following quotation: "The computer-based workplace is unnaturally quiet...and disquietingly unnatural...". In other words, the starting point of his work was the problem of the development of auditory displays for the computer. The first goal becomes, thus, to solve the contrast between the informative soundscape of the everyday world and the silence of the computer-based workplace. On the other side the danger is that a "noisy/musical" computer could easily become an annoying element. This concern, according to Barrass, highlights the need to design useful but not intruding/obsessive sounds.

More into detail, his thesis addresses the problems pointed out by previous researchers in the field of auditory display, as:

- The definition of a method for evaluating the usefulness of the sounds for a specific activity;
- The definition of methods for an effective representation of data relations by means of sounds;
- The achievement of a psychoacoustic control of auditory displays;
- The development of computer aided tools for auditory information design.

Barrass illustrates a set of already existing approaches to auditory display design. A possible classification of these approaches is:

- Syntactic and grammar-based (eg. Morse code, Earcons);
- Pragmatic: materials, lexicon and/or palette;
- Semantic: the sound is semantically related to what is meant to represent. In particular, the semantic relationships can be subdivided in:
  - Symbolic: the signifier does not resemble the signified;
  - Indexical: the signified is causally related to the signifier (e.g. the sound of a tennis ball);
  - Iconical: the signifier resembles the signified that is the case of a picture/a photograph.

Barrass proposes different approaches for auditory display design. Among the others, a Pragmatic approach and a Task-oriented approach are discussed. The Pragmatic approach concerns design principles of warnings and alarms (see also Section 10.4.1). A set of rules are asserted as:

1. Use two stages signals a) attention demanding b) designation signal;
2. Use interrupted or variable signals;
3. Use modulated signals;
4. Do not provoke startling;
5. Do not overload the auditory channel.

A Task-oriented approach takes a particular role in the following developments of the Thesis, in terms of Sound Design for Information display. Task analysis is a method developed in Human-Computer Interaction (HCI) design to analyze and characterize the information required in order to manipulate events, modes, objects and other aspects of user interfaces. The methodology is based on Task analysis

and Data characterization (TaDa). According to this analysis, the information requirements necessary for an information representation on a certain display addressing a specific kind of user are defined. One of the possible strategies to take into consideration the user from the very first step of the design process is to use a story to describe a problem. The tools become storyboards, scenarios, interviews, and case studies.

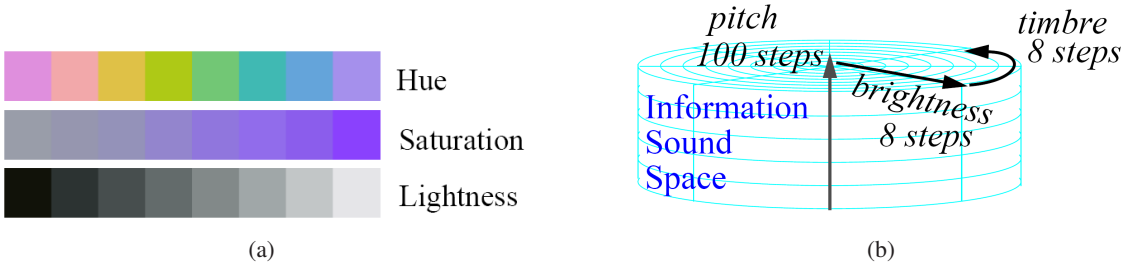

Figure 10.1: *left: 8 steps in hue, saturation and lightness; right: The TBP prototype of an Information-Sound Space (ISS).*

Another important part of Barrass' thesis is the definition of an Information-Sound Space, what he calls a cognitive artefact for auditory design. Barrass starts from the example of the Hue, Saturation, Brightness (HSB) model for the representation of the "color space" and from its representation of a color choosing tool by means of a circle with the hues corresponding to different sectors, the saturation levels mapped along the rays and the brightness controlled by means of a separated slider.

The ISS (Information Sound –Space) representation is a cylinder with the following dimensions:

- categorical (not ordered) organization of the information (the sectors of the circle)
- perceptual metric (ordered) along the radial spokes
- perceptual metric along vertical axle

The three dimensions of the ISS are related to timbre, brightness and pitch, the brightness corresponding to the radial dimension and the pitch to the height dimension.

10.5.2 Interactive Sonification

Interactive Sonification acknowlegdes the importance of human interaction for understanding and using auditory feedback. Interaction allows users to continuously query different "auditory views" of objects in the case of sonification the data under analysis) which help to give a more complete overview.

For example Rath and Rocchesso (2005) explain how continuous sonic feedback made by physical models can be used in HCI. The control metaphor which is used to demonstrate this statement is balancing a ball along a tiltable track. The idea of using a continuous feedback comes out by simply analyzing the natural behavior: trigger sounds are less usual, while we often refer to continuous sounds to get information about what is happening around us. Sonic feedback has the advantage that it can help us without changing our focus of attention: the audio channel can improve the effectiveness and naturalness of the interaction.

The physical model of a rolling ball is used: this sound is particularly informative, conveying information about direction, velocity, shape and surface textures of the contacting objects. The main characteristic of this model is its reactivity and dynamic behavior: "the impact model used produces

complex transients that depend on the parameters of the interaction and the instantaneous states of the contacting objects". The physics based algorithm developed in this work involves a degree of simplification and abstraction that implies efficient implementation and control. The approach is the cartoonification of sounds: a simplification of the models aimed at retaining perceptual invariants, useful for an efficient interaction, instead of producing the exact replica of the original sounds.

This investigation suggests that continuous feedback of a carefully designed sound can be used for sensory substitution of haptic or visual feedback in embodied interfaces. Many multimodal contexts can benefit from cartoon sound models to improve the effectiveness of the interaction: video games and virtual environments are the more obvious ones.

10.6 Interactive sounds

Most of Virtual Reality (VR) applications built to date make use of visual displays, haptic devices, and spatialized sound displays. Multisensory information is essential for designing immersive virtual worlds, as an individual's perceptual experience is influenced by interactions among sensory modalities. As an example, in real environments visual information can alter the haptic perception of object size, orientation, and shape. Similarly, being able to hear sounds of objects in an environment, while touching and manipulating them, provides a sense of immersion in the environment not obtainable otherwise. Properly designed and synchronized haptic and auditory displays are likely to provide much greater immersion in a virtual environment than a high-fidelity visual display alone. Moreover, by skewing the relationship between the haptic and visual and/or auditory displays, the range of object properties that can be effectively conveyed to the user can be significantly enhanced.

The importance of multimodal feedback in computer graphics and interaction has been recognized for a long time and is motivated by our daily interaction with the world. Streams of information coming from different channels complement and integrate each other, with some modality possibly dominating over the remaining ones, depending on the task. Research in ecological acoustics demonstrates that auditory feedback in particular can effectively convey information about a number of attributes of vibrating objects, such as material, shape, size, and so on (see also Chapter 10.4).

10.6.1 Ecological acoustics

Perception refers to how animals, including humans, can be aware of their surroundings. Perception involves motions of receptor systems (often including the whole body), and action involves motion of effectors (often including the whole body). Thus, the perception and control of behavior is largely equivalent to the perception and control of motion. Movements are controlled and stabilized relative to some referents. To watch tennis, the eyes must be stabilized relative to the moving ball. To stand, the body must be stabilized relative to the gravito-inertial force environment. Action and perception can be controlled relative to a myriad of referents. We must select referents for the control of action. The selection of referents should have a functional basis, that is, it should depend on the goals of action (e.g., a pilot who controls orientation relative to the ground may lose aerodynamic control, and a pilot who controls navigation relative to gravito-inertial force will get lost). One aspect of learning to perform new tasks will be the determination of which referents are relevant.

The ecological approach to perception, originated in the work of Gibson, refers to a particular idea of how perception works and how it should be studied. The label "ecological" reflects two main themes that distinguish this approach from more established views. First, perception is an achievement of animal-environment systems, not simply animals (or their brains). What makes up the environment

of a particular animal is part of this theory of perception. Second, the main purpose of perception is to guide action, so a theory of perception cannot ignore what animals do. The kinds of activities that a particular animal does, e.g. how it eats and moves, are part of this theory of perception.

10.6.1.1 The ecological approach to perception

Direct versus indirect perception

The ecological approach is considered controversial because of one central claim: perception is direct. To understand the claim we can contrast it with the more traditional view.

Roughly speaking, the classical theory of perception states that perception and motor control depend upon internal referents, such as the retina for vision and cochlea for audition. These internal, psychological referents for the description and control of motion are known as sensory reference frames. Sensory reference frames are necessary if sensory stimulation is ambiguous (i.e., impoverished) with respect to external reality; in this case, our position and motion relative to the physical world cannot be perceived *directly*, but can only be derived *indirectly* from motion relative to sensory reference frames. Motion relative to sensory reference frames often differs from motion relative to physical reference frames (e.g., if the eye is moving relative to the external environment). For this reason, sensory reference frames provide only an indirect relation to physical reference frames. For example, when objects in the world reflect light, the pattern of light that reaches the back of the eye (the retina) has lost and distorted a lot of detail. The role of perception is then fixing the input and adding meaningful interpretations to it so that the brain can make an inference about what caused that input in the first place. This means that accuracy depends on the perceiver's ability to "fill in the gaps" between motion defined relative to sensory reference frames and motion defined relative to physical reference frames, and this process requires inferential cognitive processing.

A theory of *direct* perception, in contrast, argues that sensory stimulation is determined in such a way that there exists a 1:1 correspondence between patterns of sensory stimulation and the underlying aspects of physical reality. This is a very strong assumption, since it basically says that reality is fully specified in the available sensory stimulation. Gibson provides the following example in the domain of visual perception, which supports, in his opinion, the direct perception theory. If one assumes that objects are isolated points in otherwise empty space, then their distances on a line projecting to the eye cannot be discriminated, as they stimulate the same retinal location. Under this assumption it is correct to state that distance is not perceivable by eye alone. However Gibson argues that this formulation is inappropriate for describing how we see. Instead he emphasizes that the presence of a continuous background surface provides rich visual structure.

Including the environment and activity into the theory of perception allows a better description of the input, a description that shows the input to be richly structured by the environment and the animal's own activities. According to Gibson, this realization opens up the new possibility that perception might be *veridical*. A relevant consequence of the direct perception approach is that sensory reference frames are unnecessary: if perception is direct, then anything that can be perceived can also be measured in the physical world.

Energy flows and invariants

Consider the following problem in visual perception: how can a perceiver distinguish object motion from his or her own motion? Gibson provides an ecological solution to this problem, from which some general concepts can be introduced. The solution goes as follows: since the retinal input is ambiguous, it must be compared with other input. A first example of additional input is the information on whether

any muscle commands had been issued to move the eyes or the head or the legs. If no counteracting motor commands is detected, then object motion can be concluded; on the contrary, if such motor commands are present then this will allow the alternative conclusion of self-motion. When the observer is moved passively (e.g. in a train), other input must be taken into account: an overall (global) change in the pattern of light indicates self-motion, while a local change against a stationary background indicates object motion.

This argument opened a new field of research devoted to the study of the structure in changing patterns of light at a given point of observation: the *optic flow*. The goal of this research is to discover particular patterns, called *invariants*, which are relevant to perception and hence to action of an animal immersed in an environment. Perceivers exploits invariants in the optic flow, in order to effectively guide their activities. For example: a waiter, who rushes towards the swinging door of the restaurant kitchen, adjusts his motion in order to control the collision with the door: he maintains enough speed to push through the door, and the same time he is slow enough not to hurt himself. In order for his motion to be effective he must know when a collision will happen and how hard the collision will be. One can identify structures in the optic flow that are relevant to these facts: these are examples of quantitative invariants.

The above considerations apply not only to visual perception but also to other senses, including audition (see Section 10.6.2 next). Moreover, recent research has introduced the concept of *global array*. According to this concept, individual forms of energy (such as optic or acoustic flows) are subordinate components of a higher-order entity, the global array, which consists of spatio-temporal structure that extends across many dimensions of energy. The general claim underlying this concept is that observers are not separately sensitive to structures in the optic and acoustic flows but, rather, observers are directly sensitive to patterns that extend across these flows, that is, to patterns in the global array.

Affordances

The most radical contribution of Gibson's theory is probably the notion of *affordance*. Gibson uses the term affordance as the noun form of the verb "to afford". The environment of a given animal affords things for that animal. What kinds of things are afforded? The answer is that behaviors are afforded. A stair with a certain proportion of a person's leg length affords climbing (is climbable); a surface which is rigid relative to the weight of an animal affords stance and traversal (is traversable); a ball which is falling with a certain velocity, relative to the speed that a person can generate in running toward it, affords catching (is catchable), and so on. Therefore, affordances are the possibilities for action of a particular animal-environment setting; they are usually described as "-ables", as in the examples above. What is important is that affordances are not determined by absolute properties of objects and environment, but depend on how these relate to the characteristics of a particular animal, e.g. size, agility, style of locomotion, and so on.

The variety of affordances constitute ecological reformulations of the traditional problems of size, distance, and shape perception. Note that affordances and events are not identical and, moreover, that they differ from one another in a qualitative manner. Events are defined without respect to the animal, and they do not refer to behavior. Insted, affordances are defined relative to the animal and refer to behavior (i.e., they are animal-environment relations that afford some behavior). The concept of affordance thus emphasizes the relevance of activity to defining the environment to be perceived.

10.6.2 Everyday sounds and the acoustic array

Ecological psychology has traditionally concentrated on visual perception. There is now interest in auditory perception and in the study of the *acoustic array*, the auditory equivalent of the optic array.

The majority of the studies in this field deal with the perception of properties of environment, objects, surfaces, and their changing relations, which is a major thread in the development of ecological psychology in general. In all of this research, there is an assumption that properties of objects, surfaces, and events are perceived as such. Therefore studies in audition investigate the identification of sound source properties, such as material, size, shape, and so on.

Musical listening versus everyday listening

Gaver introduces the concept of *everyday listening*, as opposed to *musical listening*. When a listener hears a sound, he might concentrate on attributes like pitch, loudness, and timbre, and their variations over time. Or he might notice that its masking effect on other sounds. Gaver refers to these as examples of *musical listening*, meaning that the considered perceptual dimensions and attributes have to do with the sound itself, and are those used in the creation of music.

On the other hand, the listener might concentrate on the characteristics of the sound source. As an example, if the sound is emitted by a car engine the listener might notice that the engine is powerful, that the car is approaching quickly from behind, or even that the road is a narrow alley with echoing walls on each side. Gaver refer to this as an example of *everyday listening*, the experience of listening to events rather than sounds. In this case the perceptual dimensions and attributes have to do with the sound-producing event and its environment, rather than the sound itself.

Everyday listening is not well understood by traditional approaches to audition, although it forms most of our experience of hearing the day-to-day world. Descriptions of sound in traditional psychoacoustics are typically based on Fourier analysis and include frequency, amplitude, phase, and duration. Traditional psychoacoustics takes these "primitive" parameters as the main dimensions of sound and tries to map them into of corresponding "elemental" sensations (e.g., the correspondence between sound amplitude and perceived loudness, or between frequency and perceived pitch). This kind of approach does not consider higher-level structures that are informative about events.

Everyday listening needs a different theoretical framework, in order to understand listening and manipulate sounds along source-related dimensions instead of sound-related dimensions. Such a framework must answer two fundamental questions. First, it has to develop an account of ecologically relevant perceptual attributes, i.e. the features of events that are conveyed through listening. Thus the first question asked by Gaver is: "What do we hear?". Second, it has to develop an ecological acoustics, that describes which acoustic properties of sounds are related to information about the sound sources. Thus the second question asked by Gaver is: "How do we hear it?"

Acoustic flow and acoustic invariants

Any source of sound involves an interaction of materials. Let us go back to the above example of hearing an approaching car: part of the energy produced in the engine produces vibrations in the car, instead of contributing to its motion. Mechanical vibrations, in turn, produce waves of acoustic pressure in the air surrounding the car, where the waveforms follows the movement of the car's surfaces (within limits determined by the frequency-dependent coupling of the surface's vibrations to the medium). These pressure waves then contain information about the vibrations that caused them, and result in a sound signal from which a listener might obtain such information. More in general, the patterns of vibration produced by contacting materials depend both on contact forces, duration of

contact, and time-variations of the interaction, as well as sizes, shapes, materials, and textures of the objects.

Sound also conveys information about the environment in which the event have occurred. In everyday conditions, a listener's ear is reached not only by the direct sound but also by the reflections of sound over various other objects in the environment, resulting in a coloration of the spectrum. In addition, the transmitting medium also has an influence on sound signals: dissipation of energy, especially at high-frequency, increases with the path travelled by the sound waves and thus carries information about the distance of the source. Another example is Doppler effect, which is produced when sound sources and listeners in relative motion, and results in a shift of the frequencies. Changes in loudness caused by changes in distance from a moving sound source may provide information about time-to-contact in a fashion analogous to changes in visual texture. The result is an *acoustic array*, analogous to the optical array described previously.

Several *acoustic invariants* can be associated to sound events: for instance, several attributes of a vibrating solid, including its size, shape, and density, determines the frequencies of sound it produces. It is quite obvious that the a single physical parameters can influence simultaneously many different sound parameters. As an example, changing the size of an object will scale the sound spectrum, i.e. whill change the frequencies of the sound but not their pattern. On the other hand, changing the object shape results in a change of both the frequencies and their relationships. Gaver argues that these complex patterns of change may serve as information distinguishing the physical parameters responsible: ecological acoustics focuses of discovering this kind of acoustic invariants.

Maps of everyday sounds

As already mentioned, Gaver has proposed an ecological categorization of everyday sounds.

A first category includes sounds generated by solid objects. The pattern of vibrations of a given solid is structured by a number of its physical attributes. Properties can be grouped in terms of attributes of the *interaction* that has produced the vibration, those of the *material* of the vibrating objects, and those of the *geometry* and configuration of the objects.

Aerodynamic sounds are caused by the direct introduction and modification of atmospheric pressure differences from some source. The simplest aerodynamic sound is exemplified by an exploding balloon. Other aerodynamic sounds, e.g. the noise of a fan, are caused by more continuous events. Another sort of aerodynamic event involves situations in which changes in pressure themselves transmit energy to objects and set them into vibration (for example, when wind passes through a wire).

Sound-producing events involving liquids (e.g., dripping and splashing) are similar to those of vibrating solids: they depend on an initial deformation that is counter-acted by restoring forces in the material. The difference is that no audible sound is produced by the vibrations of the liquid. Instead, the resulting sounds are created by the resonant cavities (bubbles) that form and oscillate and in the surface of the liquid. As an example, a solid object that hits a liquid pushes it aside and form a cavity that resonates to a characteristic frequency, amplifying and modifying the pressure wave formed by the impact itself.

Although all sound-producing events involve any of the above categories (vibrating solids, aerodynamic, or liquid interactions), many also depend on complex patterns of simpler events. As an example, footsteps are temporal patterns of impact sounds. The perception of these *patterned* sounds is also related to the timing of successive events, (e.g. successive footstep sounds must occur within a range of rates and regularities in order to be perceived as walking). A slightly more complex example is a door slam, which involves the squeak of scraping hinges and the impact of the door on its frame. This kind of *compound* sounds involve mutual constraints on the objects that participate in related

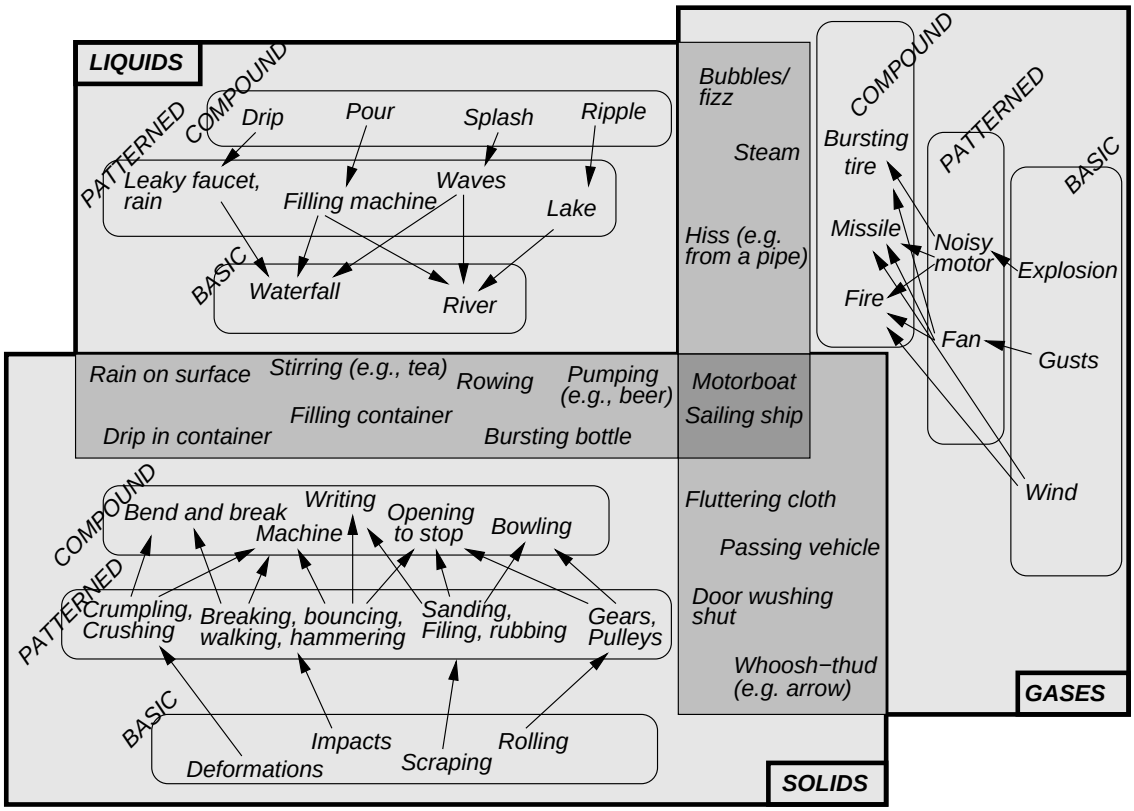

Figure 10.2: *A map of everyday sounds. Complexity increases towards the center. Figure based on (Gaver, 1993).*

events: concatenating the creak of a heavy door closing slowly with the slap of a light door slammed shut would probably not sound natural.

Starting from these considerations, Gaver derived a tentative map of everyday sounds, which is shown in figure 10.2 and discussed in the following.

- Basic Level Sources: consider, for example, the region describing sounds made by vibrating solids. Four different sources of vibration in solids are indicated as basic level events: deformation, impacts, scraping and rolling.
- Patterned Sources involve temporal patterning of basic events. For instance walking, as described above, but also breaking, spilling, and so on, are all complex events involving patterns of simpler impacts. Similarly, crumpling or crushing are examples of patterned deformation sounds. In addition, other sorts of information are made available by their temporal complexity. For example, the regularity of a bouncing sound provides information about the symmetry of the bouncing object.
- Compound events involve more than one type of basic level event. An example is the slamming door discussed above. Other examples are the sounds made by writing, which involve a complex series of impacts and scrapes over time, while those made by bowling involve rolling followed by impact sounds.
- Hybrid events involve yet another level of complexity in which more than one basic types of

material is involved. As an example, the sounds resulting from water dripping on a reverberant surface are caused both by the surface's vibrations and the quickly-changing reverberant cavities, and thus involve attributes both of liquid and vibrating solid sounds.

10.7 Multimodal perception and interaction

10.7.1 Combining and integrating auditory information

Humans achieve robust perception through the combination and integration of information from multiple sensory modalities. According to some authors, multisensory perception emerges gradually during the first months of life, and experience significantly shapes multisensory functions. By contrast, a different line of thinking assumes that sensory systems are fused at birth, and the single senses differentiate later. Empirical findings in newborns and young children have provided evidence for both views. In general experience seems to be necessary to fully develop multisensory functions.

Sensory combination and integration

Looking at how multisensory information is combined, two general strategies can be identified: the first is to maximize information delivered from the different sensory modalities (sensory *combination*). The second strategy is to reduce the variance in the sensory estimate to increase its reliability (sensory *integration*).

Sensory combination describes interactions between sensory signals that are not redundant: they may be in different units, coordinate systems, or about complementary aspects of the same environmental property. Disambiguation and cooperation are examples for this kind of interactions: if a single modality is not enough to provide a robust estimate, information from several modalities can be combined. As an example, object recognition is achieved through different modalities that complement each other and increase the information content.

By contrast, *sensory integration* describes interactions between redundant signals. For example, when knocking on wood at least three sensory estimates about the location of the knocking event can be derived: visual, auditory and proprioceptive. In order for these three location signals to be integrated, they first have to be transformed into the same coordinates and units. For this, the visual and auditory signals have to be combined with the proprioceptive neck-muscle signals to be transformed into body coordinates. The process of sensory combination might be non-linear. At a later stage the three signals are then integrated to form a coherent percept of the location of the knocking event.

There are a number of studies that show that vision dominates the integrated percept in many tasks, while other modalities (in particular audition and touch) have a less marked influence. This phenomenon of visual dominance is often termed *visual capture*. As an example, it is known that in the spatial domain vision can bias the perceived location of sounds whereas sounds rarely influence visual localization. One key reason for this asymmetry seems to be that vision provides more accurate location information. In general, however, the amount of cross-modal integration depends on the features to be evaluated or the tasks to be accomplished.

Auditory capture and illusions

Psychology has a long history of studying intermodal conflict and illusions in order to understand mechanisms of multisensory integration. Much of the literature on multisensory perception has focused on spatial interactions: an example is the ventriloquist effect, in which the perceived location

of a sound shifts towards a visual stimulus presented at a different position. Identity interactions are also studied: an example is the already mentioned McGurk effect, in which what is being heard is influenced by what is being seen (for example, when hearing /ba/ but seeing the speaker say /ga/ the final perception may be /da/).

As already noted, the visual modality does not always win in such crossmodal tasks. In particular, the senses can interact in time, i.e they interact in determining not what is being perceived or where it is being perceived, but *when* it is being perceived. The temporal relationships between inputs from the different senses play an important role in multisensory integration. Indeed, a window of synchrony between auditory and visual events is crucial even in the spatial ventriloquist effect, which disappears when the audio-visual asynchrony exceeds approximately 300 ms. This is also the case in the McGurk effect, which fails to occur when the audio-visual asynchrony exceeds $200 - 300$ ms.

There is a variety of crossmodal effects that demonstrate that, outside the spatial domain, audition can bias vision.

10.7.2 Perception is action

Embodiment and enaction

According to traditional mainstream views of perception and action, perception is a process in the brain where the perceptual system constructs an internal representation of the world, and eventually action follows as a subordinate function. This view of the relation between perception and action makes then two assumptions. First, the causal flow between perception and action is primarily one-way: perception is input from world to mind, action is output from mind to world, and thought (cognition) is the mediating process. Second, perception and action are merely instrumentally related to each other, so that each is a means to the other. If this kind of "input-output" picture is right, then it must be possible, at least in principle, to disassociate capacities for perception, action, and thought.

Although everyone agrees that perception depends on processes taking place in the brain, and that internal representations are very likely produced in the brain, more recent theories have questioned such a modular decomposition in which cognition interfaces between perception and action. The ecological approach discussed in section 10.6.1 reject the one-way assumption, but not the instrumental aspect of the traditional view, so that perception and action are seen as instrumentally interdependent. Others argue that a better alternative is to reject both assumptions: the main claim of these theories is that it is not possible to disassociate perception and action schematically, and that every kind of perception is intrinsically active and thoughtful: perception is not a process in the brain, but a kind of skilful activity on the part of the animal as a whole. Only a creature with certain kinds of bodily skills (e.g. a basic familiarity with the sensory effects of eye or hand movements, etc.) could be a perceiver.

One of the most influential contributions in this direction is due to Varela et al. (1991). They presented an "enactive conception" of experience, which is not regarded as something that occurs inside the animal, but rather as something that the animal *enacts* as it explores the environment in which it is situated. In this view, the subject of mental states is the *embodied*, environmentally situated animal. The animal and the environment form a pair in which the two parts are coupled and reciprocally determining. Perception is thought of in terms of activity on the part of the animal. The term "embodied" is used by the authors as a mean to highlight two points: first, cognition depends upon the kinds of experience that are generated from specific sensorimotor capacities. Second, these individual sensorimotor capacities are themselves embedded in a biological, psychological, and cultural context. Sensory and motor processes, perception and action, are fundamentally inseparable in cognition.

10.8 Multimodal and Cross-Modal Approaches to Control of Interactive Systems

10.8.1 Introduction

This section[6] briefly surveys some relevant aspects of current research into control of interactive (music) systems, putting into evidence research issues, achieved results, and problems that are still open for the future. A particular focus is on multimodal and cross-modal techniques for expressive control of sound and music processing and synthesis. The section will discuss a conceptual framework, the methodological aspects, the research perspectives.

The problem of effectively controlling sound generation and processing has always been relevant for music research in general and for Sound and Music Computing in particular. Research into control concerns perceptual, cognitive, affective aspects. It ranges from the study of the mechanisms involved in playing traditional acoustic instruments to the novel opportunities offered by modern digital music instruments. More recently, the problem of defining effective strategies for real-time control of multimodal interactive systems, with particular reference to music but not limited to it, is attracting growing interest from the scientific community because of its relevance also for future research and applications in broader fields of human-computer interaction.

In this framework, research into control extends its scope to include for example analysis of human movement and gesture (not only gestures of musicians playing an instrument but also gestures of subjects interacting with computer systems), analysis of the perceptual and cognitive mechanisms of gesture interpretation, analysis of the communication of non-verbal expressive and emotional content through gesture, multimodality and cross-modality, identification of strategies for mapping the information obtained from gesture analysis onto real-time control of sound and music output including high-level information (e.g. real-time control of expressive sound and music output).

A key issue in this research is its cross-disciplinary nature. Research can highly benefit from cross-fertilisation between scientific and technical knowledge on the one side, and art and humanities on the other side. Such need of cross-fertilisation opens new perspectives to research in both fields: if from the one side scientific and technological research can benefit from models and theories borrowed from psychology, social science, art and humanities, on the other side these disciplines can take advantage of the tools that technology can provide for their own research, i.e. for investigating the hidden subtleties of human beings at a depth that was hard to reach before. The convergence of different research communities such as musicology, computer science, computer engineering, mathematics, psychology, neuroscience, arts and humanities as well as of theoretical and empirical approaches bears witness to the need and the importance of such cross-fertilisation.

10.8.2 Multisensory Integrated Expressive Environments

Multisensory Integrated Expressive Environments (MIEEs) are a framework for mixed reality applications in the performing arts, culture-oriented applications, and future applications such as home entertainment, therapy, and rehabilitation. Paradigmatic contexts for applications of MIEEs are multimedia concerts, interactive dance/music/video installations, interactive museum exhibitions, and distributed cooperative environments for theatre and artistic expression.

Imagine a home high-fidelity (hi-fi) music system that not only has the standard controls for volume, treble, bass, balance, and so forth, but also features *expressive knobs* possibly controlled by

[6] adapted from Camurri, De Poli, Leman and Volpe / IEEE Multimedia 2005 and from Antonio Camurri, Carlo Drioli, Barbara Mazzarino, Gualtiero Volpe, Polotti and Rocchesso [2008], chapt. 6

your movement, such as dancing, in your living room. The system lets you actively listen to, say, a Chopin piece, by changing the agogics, that is, the music interpretation. For example, light and smooth movements might influence a more intimate legato phrasing in the music performance, while jumpy and joyful movements might change the phrasing to a faster tempo and staccato.

MIEEs address the expressive aspects of non-verbal human communication. In the above example, we mentioned a few terms from music performance theory (such as staccato and legato, as well as adjectives such as light, heavy, joyful, and jumpy) which are often used in humanistic theories of the performing arts. In MIEEs, real and virtual subjects interact with each other through the exchange of information that represents the communicative expressiveness in different sensory modalities (auditory, visual, tactile, and so on). The main goal of MIEEs is to establish a framework that accounts for the relationship between the current state-of-the-art in audio-visual technology on the one hand and humanistic theories on expressive actions and aesthetic experience on the other hand.

For the artistic aspect, MIEEs provide a deeper and enhanced experience of the artistic content. In the music research community, for example, MIEEs are sometimes conceived of as a new generation of musical instruments based on real-time and intelligent human-machine interaction. These new musical instruments are a holistic human-machine concept based on an assembly of modular input/output devices and musical software components arranged according to essential human musical content processing capabilities. The focus on technology and multisensory expressive content processing adds a new dimension to the mediation of art. MIEEs introduce a level of cross-modality that interconnects microscopic scales of information processing with human capabilities in synaesthesia (multimodal perception of features of objects) and kinesthesia (perception of movement).

The main motivation in introducing MIEEs is to adapt the new mediating technology to basic human forms of communication. Many typical communication modes are nonlinguistic and based on movement, action, gestures, and mimetic activities. Unfortunately, the current state of the art in MIEEs suffers from a serious lack of advanced content processing capabilities in the cognitive, affective/emotive, and motoric domains. For example, although advances have been made in the processing of musical pitch, timbre, texture, and rhythm, the results are mainly restricted to low-level features. We can say the same about movement gestures. It is difficult to characterize a musical object, or to specify a movement gesture characteristics. If MIEEs have to interact intelligently and spontaneously with users, then their communication capabilities should rely on a set of advanced musical and gestural content processing tools. In many situations, users might want to interact in a spontaneous and expressive way with these systems, using descriptions of perceived qualities or making expressive movements. Moreover, making machines useful in artistic contexts that rely on different sensory modalities implies that the often subtle nuances of artistic expression should be dealt with in these different modalities. A technology focusing on affect, emotion, expressiveness, and cross-modality interactions is thus required.

10.8.3 Cross-modal expressiveness

This section briefly surveys some relevant aspects of current research on control, putting into evidence research issues, achieved results, and problems that are still open for the future. A particular focus is on multimodal and cross-modal techniques for expressive control. Multimodal analysis enables the integrated analysis of information coming from different multimedia streams (audio, video) and affecting different sensorial modalities (auditory, visual). Cross-modal analysis enables exploiting potential similarities in the approach for analyzing different multimedia streams: so, for example techniques developed for analysis in a given modality (e.g. audio) can also be used for analysis in another modality (e.g. video); further, commonalities at mid-and high-level in representations of

different sensory channels are an important perspective for developing models for control and mapping based on a-modal, converging representations.

The physical stimuli that make up an artistic environment contain information about expressiveness. That information can, to some extent, be extracted and then communicated among a MIEE's virtual and real subjects. With multiple sensory modalities (auditory, visual, motoric, gestural), this allows the transmission of expressiveness parameters from one domain to another - for example, from music (auditory) to computer animation (visual), or from dance (motoric) to music (auditory). That is, expressive parameters are an example of parameters emerging from modalities and independent from them. In other words, expressive parameters define a cross-modal control space that is at a higher level with respect to the single modalities.

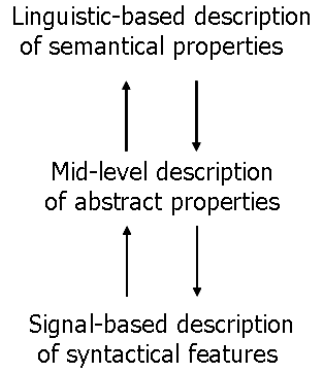

Figure 10.3: *The layered conceptual framework distinguishes between syntax and semantics, and in between, a connection layer that consists of affect/emotion expressiveness spaces and mappings.*

Figure 10.3 shows a way of conceiving the transmission of cross-modal expressiveness. This layered framework will be analysed more in detail in Section 10.8.4. The signal-based level represents the analysis and synthesis of physical properties (bottom). The symbolic level represents the descriptions of meanings, affects, emotions, and expressiveness in terms of linguistic or visual symbolic entities (top). The connection layer (gesture-based level) represents spaces in which trajectories allow the connection from signal-based descriptions to symbolic-based descriptions (middle). In most cases, the signal-based descriptions pertain to the signal syntactical properties, while the symbol-based descriptions pertain to its semantic properties. The latter may include cognitive, emotive, affective, and expressive evaluations.

The flow of expressiveness may go in two directions (upward and downward). Taking the expressive hi-fi music system as an example, physical properties of human movement (of the system user) may be extracted and gestures mapped as a trajectory on a space. That trajectory describes the expressive content in terms of linguistic/semantic descriptors such as how much the movement is fluent, smooth, heavy, rigid, and so on. Starting from this linguistic-semantic description (in the downward direction), a particular gesture-based trajectory may be used to synthesize physical properties of that expressive content in another modality, in our case music (in terms of legato/staccato, amplitude, shapes of notes, and so on). The gesture-based mappings describe properties of expressiveness that are independent from any particular sensory modality. This level of mapping introduces flexibility to the multimodal representational model.

10.8.4 A conceptual framework

A relevant foundational aspect for research in sound and music control concerns the definition of a conceptual framework envisaging control at different levels under a multimodal and cross-modal perspective: from low-level analysis of audio signals, toward high-level semantic descriptions including affective, emotional content. This Section presents a conceptual framework worked out in the EU-IST Project MEGA (2000-2003)[7] that can be considered as a starting point for research on this direction.

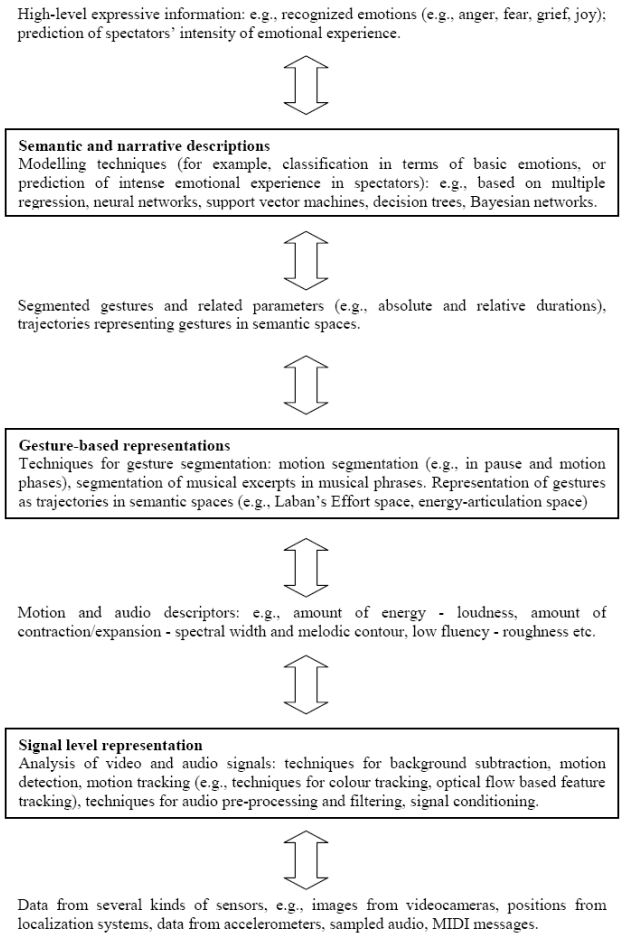

Figure 10.4: *The layered conceptual framework makes a distinction between syntax and semantics, and in between, a connection layer that consists of affect / emotion / expressiveness (AEE) spaces and mappings.*

A main question thus relates to the nature of the physical cues that carry expressiveness, and a second question is how to set up cross-modal interchanges (as well as person/machine interchanges) of expressiveness. These questions necessitated the development of a layered conceptual framework for affect processing that splits up the problem into different sub-problems. The conceptual framework aims at clarifying the possible links between physical properties of a particular modality, and the affective/emotive/expressive (AEE) meaning that is typically associated with these properties. Figure 10.4 sketches the conceptual framework in terms of

[7] `www.megaproject.org`

- a syntactical layer that stands for the analysis and synthesis of physical properties (bottom),
- a semantic layer that contains descriptions of affects, emotions, and expressiveness (top),
- a layer of AEE mappings and spaces that link the syntactical layer with the semantic layer (middle).

The syntactical layer contains different modalities, in particular audio, movement, and animation and arrows point to flows of information. Communication of expressiveness in the cross-modal sense could work in the following way. First, (in the upward direction) physical properties of the musical audio are extracted and the mapping onto an AEE-space allows the description of the affective content in the semantic layer. Starting from this description (in the downward direction), a particular AEE-mapping may be selected that is then used to synthesise physical properties of that affect in another modality, such as animation. This path is followed, for example, when sadness is expressed in a piece of music, and correspondingly an avatar is displaying this sadness in his posture.

<table>
<tr><td colspan="2">Structure</td><td colspan="2">Concept level</td><td colspan="5">Musical content features</td></tr>
<tr><td rowspan="5">Contextual</td><td rowspan="2">Global beyond 3 seconds</td><td rowspan="2">High II</td><td rowspan="2">Expressive</td><td colspan="5"></td></tr>
<tr><td colspan="5">Cognition | Emotion | Affect = Syntactic + semantic concepts</td></tr>
<tr><td rowspan="2">High I</td><td rowspan="2">Formal</td><td>Melody</td><td>Harmony</td><td>Rhythm</td><td>Source</td><td>Dynamics</td></tr>
<tr><td>Key
Profile</td><td>Tonality
Cadence</td><td>Rhythmic patterns
Tempo</td><td>Instrument
Voice</td><td>Trajectory
Articulation</td></tr>
<tr><td>Global < 3 seconds</td><td>Mid</td><td>Perceptual</td><td>Successive intervallic pattern</td><td>Simultaneous intervallic pattern</td><td>Beat
Inter-onset interval</td><td>Spectral envelope</td><td>Dynamic range
Sound level</td></tr>
<tr><td rowspan="3">Non contextual</td><td rowspan="2">Local + spatial</td><td rowspan="2">Low II</td><td rowspan="2">Sensorial</td><td colspan="2">Pattern</td><td>Time</td><td>Timbre</td><td>Loudness</td></tr>
<tr><td colspan="2" rowspan="2">Periodicity pitch
Pitch deviations
Fundamental frequency</td><td rowspan="2">Note duration
Onset
Offset</td><td rowspan="2">Roughness
Spectral flux
Spectral centroid</td><td rowspan="2">Neural energy
Peak</td></tr>
<tr><td>Local + temporal</td><td>Low I</td><td>Physical</td></tr>
<tr><td colspan="2"></td><td colspan="2">Frequency</td><td>Duration</td><td>Spectrum</td><td>Intensity</td></tr>
</table>

Figure 10.5: *Taxonomy of musical syntactical cues.*

10.8.4.1 Syntactic layer

The syntactic layer is about the extraction of the physical features that are relevant for affect, emotion and expressiveness processing. In the domain of musical audio processing, Lesaffre and colleagues worked out a useful taxonomy of concepts that gives a structured understanding of this layer in terms of a number of justified distinctions (Figure 10.5). A distinction is made between low-level, mid-level, and high-level descriptors of musical signals. In this viewpoint, the low-level features are related to very local temporal and spatial characteristics of sound. They deal with the physical categories of frequency, duration, spectrum, intensity, and with the perceptual categories of pitch, time, timbre,

and perceived loudness. Low-level features are extracted and processed (in the statistical sense) in order to carry out a subsequent analysis related to expression. For example, in the audio domain, these low-level features are related to tempo (i.e. number of beats per minute), tempo variability, sound level, 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 structural important notes, periodicity, dynamics (intensity), roughness (or sensory dissonance), tonal tension (or the correlation between local pitch patterns and global or contextual pitch patterns), and so on.

When more context information is involved (typically in musical sequences that are longer than 3 seconds), then other categories emerge, in particular, categories related to melody, harmony, rhythm, source, and dynamics. Each of these categories has several distinct specifications, related to an increasing complexity, increasing use of contextual information, and increasing use of top-down knowledge. The highest category is called the expressive category. This layer can in fact be developed into a separate layer because it involves affective, emotive and expressive meanings that cannot be directly extracted from audio structures. Figure 10.4 introduced this layer as a separate layer that is connected with the syntactical cues using a middle layer of mappings and spaces. Examples of mappings and spaces will be given below.

In the domain of movement (dance) analysis, a similar approach can be envisaged that leans on a distinction between features calculated on different time scales. In this context also, it makes sense to distinguish between (i) low-level features, calculated on a time interval of a few milliseconds (e.g. one or a few frames coming from a video camera), (ii) mid-level features, calculated on a movement stroke (in the following also referred as "motion phase"), i.e. on time durations of a few seconds, and (iii) high-level features that are related to the conveyed expressive content (but also to cognitive aspects) and referring to sequences of movement strokes or motion (and pause) phases. An example of a low-level feature is the amount of contraction/expansion that can be calculated on just one frame, i.e. on 40 ms with the common sample rate of 25 fps. Other examples of low-level features are the detected amount of movement, kinematic measures (e.g. velocity and acceleration of body parts), measures related to the occupation of the space surrounding the body. Examples of mid-level descriptors are the overall direction of the movement in the stroke (e.g. upward or downward) or its directness (i.e. how much the movement followed direct paths), motion impulsiveness, and fluency. At this level it is possible to obtain a first segmentation of movement in strokes that can be employed for developing an event-based representation of movement. In fact, strokes or motion phases can be characterised by a beginning, an end, and a collection of descriptors including both mid-level features calculated on the stroke and statistical summaries (e.g. average, standard deviation), performed on the stroke, of low-level features (e.g. average body contraction/expansion during the stroke).

10.8.4.2 Semantic layer

The semantic layer is about the experienced meaning of affective, emotive, expressive processing. Apart from aesthetic theories of affect processing in music and in dance, experimental studies were set up that aim at depicting the underlying structure of affect attribution in performing arts (see next sections). Affect semantics in music has been studied by allowing a large number of listeners to use adjectives (either on a completely free basis, or taken from an elaborate list) to specify the affective content of musical excerpts. Afterwards, the data are analysed and clustered into categories. There seems to be a considerable agreement about two fundamental dimensions of musical affect processing, namely Valence and Activity. Valence is about positively or negatively valued affects, while Activity

is about the force of these affects. A third dimension is often noticed, but its meaning is less clearly specified. These results provided the basis for the experiments performed along the project.

10.8.4.3 Connecting syntax and semantics: Maps and spaces

Different types of maps and spaces can be considered for connecting syntax and semantics. One type is called the semantic map because it relates the meaning of affective/emotive/expressive concepts with physical cues of a certain modality. In the domain of music, for example, several cues have been identified and related to affect processing. For example, tempo is considered to be the most important factor affecting emotional expression in music. Fast tempo is associated with various expressions of activity/excitement, happiness, potency, anger and fear while slow tempo with various expressions of sadness, calmness, solemnity, dignity. Loud music may be determinant for the perception of expressions of intensity, power, anger and joy whereas soft music may be associated with tenderness, sadness, solemnity, and fear. High pitch may be associated with expressions such as happy, graceful, exciting, angry, fearful and active, and low pitch may suggest sadness, dignity, excitement as well as boredom and pleasantness, and so on. Kinematics spaces or energy-velocity spaces are another important type of space. They have been successfully used for the analysis and synthesis of the musical performance. This space is derived from factor analysis of perceptual evaluation of different expressive music performances. Listeners tend to use these coordinates as mid level evaluation criteria. The most evident correlation of energy-velocity dimensions with syntactical features is legato-staccato versus tempo. The robustness of this space is confirmed in the synthesis of different and varying expressive intentions in a musical performance. The MIDI parameters typically control tempo and key velocity. The audio-parameters control legato, loudness, brightness, attack time, vibrato, and envelope shape.

10.8.5 Methodologies of gesture analysis

The definition of suitable scientific methodologies for investigating - within the conceptual framework and under a multimodal perspective - the subtleties involved in sound and music control is a key issue. An important topic for control research is gesture analysis of both performers and interacting subjects. Gestures are an easy and natural way for controlling sound generation and processing. For these reasons, this section discusses methodologies and approaches focusing on full-body movement and gesture. Nevertheless, the concepts here discussed can be easily generalised to include other modalities. Discovering the key factors that characterise gesture, and in particular expressive gesture, in a general framework is a challenging task. When considering such an unstructured scenario one often has to face the problem of the poor or noisy characterisation of most movements in terms of expressive content. Thus, a common approach consists in starting research from a constrained framework where expressiveness in movement can be exploited to its maximum extent.

10.8.5.1 Bottom-up approach

Let us consider the dance scenario (consider, however, that what we are going to say also applies to music performance). A possible methodology for designing repeatable experiments is to have a dancer performing a series of dance movements (choreographies) that are distinguished by their expressive content. We use the term "micro-dance" for a short fragment of choreography having a typical duration in the range of 15-90 s. A microdance is conceived as a potential carrier of expressive information, and it is not strongly related to a given emotion (i.e. the choreography has no explicit gestures

denoting emotional states). Therefore, different performances of the same micro-dance can convey different expressive or emotional content to spectators: e.g. light/heavy, fluent/rigid, happy/sad, emotional engagement, or evoked emotional strength. Human testers/spectators judge each micro-dance performance. Spectators' ratings are used for evaluation and compared with the output of developed computational models (e.g. for the analysis of expressiveness). Moreover, micro-dances can also be used for testing feature extraction algorithms by comparing the outputs of the algorithms with spectators' ratings of the same micro-dance performance (see for example the work by Camurri et al. (2004b) on spectators' expectation with respect to the motion of the body center of gravity). In case of music performances, we have musical phrases (corresponding to micro-dances above) and the same approach can be applied.

10.8.5.2 Subtractive approach

Micro-dances can be useful to isolate factors related to expressiveness and to help in providing experimental evidence with respect to the cues that choreographers and psychologists identified. This is obtained by the analysis of differences and invariants in the same micro-dance performed with different expressive intentions. With the same goal, another approach is based on the live observation of genuinely artistic performances, and their corresponding audiovisual recordings. A reference archive of artistic performances has to be carefully defined for this method, chosen after a strict intensive interaction with composers and performers. Image (audio) processing techniques are utilised to gradually subtract information from the recordings. For example, parts of the dancer's body could be progressively hidden until only a set of moving points remain, deforming filters could be applied (e.g. blur), the frame rate could be slowed down, etc. Each time information is reduced, spectators are asked to rate the intensity of their emotional engagement in a scale ranging from negative to positive values (a negative value meaning that the video fragment would rise some negative feeling in the spectator). The transitions between positive and negatives ratings and a zero-rating (i.e. no expressiveness was found by the spectator in the analysed video sequence) would help to identify what are the movement features carrying expressive information. An intensive interaction is needed between the image processing phase (i.e. the decisions on which information has to be subtracted) and the rating phase.

10.8.6 Examples of multimodal and cross-modal analysis

Here we provide some concrete examples of multimodal and cross-modal analysis with reference to the above mentioned conceptual framework. Multi-modal and cross-modal analysis can be applied both in a bottom-up approach and in a subtractive approach. In the latter, they are used for extracting and comparing features among subsequent subtraction steps.

10.8.6.1 Analysis of human full-body movement

A major step in multimodal analysis of human full-body movement is the extraction of a collection of motion descriptors. With respect to the approaches discussed above, such descriptors can be used in the bottom-up approach for characterizing motion (e.g. micro-dances). The top-down approach can be used for validating the descriptors with respect to their role and contribute in conveying expressive content.

With respect to the conceptual framework, at Layer 1 consolidated computer vision techniques (e.g. background subtraction, motion detection, motion tracking) are applied to the incoming video frames. Two kinds of outputs are usually generated: trajectories of points on the dancers' bodies

(motion trajectories) and processed images. As an example a Silhouette Motion Image (SMI) is an image carrying information about variations of the shape and position of the dancer's silhouette in the last few frames. We also use an extension of SMIs taking into account the internal motion in silhouettes.

From such outputs a collection of motion descriptors are extracted including:

- Cues related to the amount of movement (energy) and in particular what we call Quantity of Motion (QoM). QoM is computed as the area (i.e. number of pixel) of SMI. It can be considered as an overall measure of the amount of detected motion, involving velocity and force.
- Cues related to body contraction/expansion and in particular the Contraction Index (CI), conceived as a measure, ranging from 0 to 1, of how the dancer's body uses the space surrounding it. The algorithm to compute the CI combines two different techniques: the individuation of an ellipse approximating the body silhouette and computations based on the bounding region.
- Cues derived from psychological studies such as amount of upward movement, dynamics of the Contraction Index (i.e. how much CI was over a given threshold along a time unit);
- Cues related to the use of space, such as length and overall direction of motion trajectories.
- Kinematical cues, such as velocity and acceleration of motion trajectories.

A relevant task for Layer 2 is motion segmentation. A possible technique for motion segmentation is based on the measured QoM. The evolution in time of the QoM resembles the evolution of velocity of biological motion, which can be roughly described as a sequence of bell-shaped curves (motion bells, see Figure 6.3). In order to segment motion by identifying the component gestures, a list of these motion bells and their features (e.g. peak value and duration) is extracted. An empirical threshold is defined to perform segmentation: the dancer is considered to be moving if the QoM is greater than 2.5% of the total area of the silhouette. It is interesting to notice that the motion bell approach can also be applied to sound signal analysis.

Segmentation allows extracting further higher-level cues at Level 2. A concrete example is the Directness Index (DI), calculated as the ratio between the length of the straight trajectory connecting the first and the last point of a motion trajectory and the sum of the lengths of each segment constituting the trajectory. Furthermore, motion fluency and impulsiveness can be evaluated. Fluency can be estimated from an analysis of the temporal sequence of motion bells. A dance fragment performed with frequent stops and restarts will result less fluent than the same movement performed in a continuous, "harmonic" way. The hesitating, bounded performance will be characterised by a higher percentage of acceleration and deceleration in the time unit (due to the frequent stops and restarts). A first measure of impulsiveness can be obtained from the shape of a motion bell. In fact, since QoM is directly related to the amount of detected movement, a short motion bell having a high pick value will be the result of an impulsive movement (i.e. a movement in which speed rapidly moves from a value near or equal to zero, to a peak and back to zero). On the other hand, a sustained, continuous movement will show a motion bell characterised by a relatively long time period in which the QoM values have little fluctuations around the average value (i.e. the speed is more or less constant during the movement).

One of the tasks of Layer 4 is to classify dances with respect to their emotional/expressive content. For example, in a study carried in the framework of the EU-IST Project MEGA results were obtained on the classification of expressive gestures with respect to their four basic emotions (anger, fear, grief, joy) by an automatic system based on decision trees.

10.8.7 Examples of Multisensory Integrated Expressive Environments

Here we present a few examples of MIEEs developed with the EyesWeb open software platform in the European Union Information Society Technologies (IST) Multisensory Expressive Gesture Applications (MEGA) project (`http://www.megaproject.org`).

Interactive concert We developed a MIEE to model and implement the interaction between an actress and her vocal clone. This piece, named *Allegoria dell'opinione verbale* (Allegory of the spoken opinion) by the composer Roberto Doati based on poetry by Gianni Revello, was conceived in the Department of Communication, Computer, and System Sciences, or DIST, InfoMus Lab. It was performed on stage during the 2001-2002 season of Gran Teatro La Fenice (Opera House of Venice) at the Teatro Malibran in Venice, within a musical theatre production that performed works from Aperghis, Cage, Casale, Kagel, Pachini, and Schnebel. The piece was also performed at the Opera House of Genova Teatro Carlo Felice in Genova, Italy.

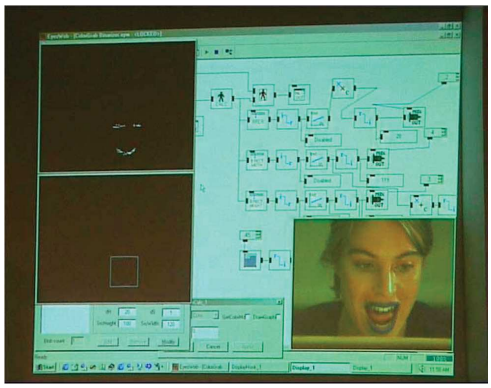

Figure 10.6: *The EyesWeb application for* Allegoria dell'opinione verbale *by R. Doati.*

During the concert the actress (Francesca Faiella) sits on a stool placed in the front of the stage near the left side. She is turned toward the left backstage so that the audience sees her profile. A large screen projects a frontal view of her face. A video camera is placed (hidden) in the left part of the backstage. The camera captures images of the actress' face to be projected on the large screen and acquires her lip and facial movements. The actress enacts the poetry in front of the camera. EyesWeb extracts and processes the movements of her lips and face. The system uses expressive cues to process her voice in real time and diffuse spatialized electroacoustic music on eight loudspeakers placed within the auditorium. The signals reproduced by the loudspeakers are derived only from the actress' voice. Previous recordings of her voice reciting the Revello poetry are resynthetized and processed in real time with parameters controlled by lips movements. The audience can observe the movements of the actress' face on the large screen while listening to the piece and thus perceiving the interaction of her movements with sound changes coming from the loudspeakers. Figure 10.6 shows the EyesWeb patch employed in the concert;

Medea: Exploring sound-movement expressiveness. Since 1959, when electronic music was established as a new way of music composition, the rules of traditional music performance and enjoyment have changed to include space, motion, and gesture as musical parameters. For example, musicians are often located somewhere other than the stage, sometimes even in the audience, and where the music will be performed often influences compositional thinking. Loudspeakers move sound through

the space at varying speeds (based on other musical parameters). In addition, the development of live electronics- that is, computers applied to real-time processing of instrumental sounds-has allowed space as a musical instrumental practice to flourish. Electro-acoustic technologies let composers explore new listening dimensions and consider the sounds coming from loudspeakers as possessing different logical meanings from the sounds produced by traditional instruments.

Medea, Adriano Guarnieri's "video opera," is an innovative work stemming from research in multimedia that demonstrates the importance and amount of research dedicated to sound movement in space[8]. Among Medea's intentions, derived from artistic and musical suggestions and needs, is a desire to establish an explicit connection between sound movement and expressiveness and to show how engagement can be enhanced acoustically in multimodality environments, for example, through the motion of sound through virtual spaces. Whereas sound positioning and movement have seldom been used in concert settings, the ear has great detection capabilities connected to its primary role (a signalling device for invisible or unseen cues); music is now trying to put these capabilities to creative use.

Sound motion through space is an established tradition in much of contemporary music, much of which exploits multimodality to enhance performance. Music-specifically sound motion in space-conveys expressive content related to performance gestures. Although composers have investigated the connection between music and emotion for traditional parameters, such as intensity, timbre, and pitch, spatialization is still a new research path. The use of space as a musical parameter in an expressive dimension requires new paradigms for interaction, mapping strategies, and multimedia interfaces based on real-time analysis and synthesis of expressive content in music and gesture. Researchers have developed and applied models and algorithms for extracting high-level, qualitative information about expressive content to real-time music and multimedia applications. Analysis of expressive gestures has sought to extract expressive information from human movements and gestures and to control the generation of audio content depending on the analysis. Figure 10.7 diagrams the link between physical and spatial movement. Medea offers real-world examples of such multimodal extensions.

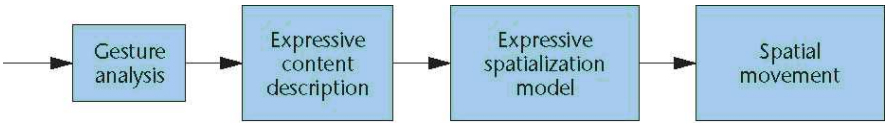

Figure 10.7: *Connection between physical and spatial movements.*

Medea's score cites sound spatialization as a fundamental feature of the opera. The musicians in the hall should be considered a sonic body living among the audience to create a sort of gravitational centre for the trumpets located on either side of the audience. The presence of trombones with their gestural posture becomes a central expressive feature. The live-electronics performers executed all movements and transformations following the conductor and the score (each instrument has its own spatialization modes, and the score marks each transformation and movement precisely), with all sound movements except for those coming from the trombones having been predetermined. Guarnieri defined 11 modalities for the trombone movements. An EyesWeb patch controlled the trombones' random space movements, a webcam captured movements derived from trombone players' gestures, and the EyesWeb program digitally processed them to provide each movement's speed parameter using a gesture-speed mapping. This method's functionality derives from a translation of the image bitmap in terms of speed: intense instrumental gestural activity (rocking off) leads to a large bitmap varia-

[8] adapted from De Gotzen, IEEE Multimedia 2004.

tion and therefore to high speed, while reduced gestural activity corresponds to a moderate movement speed.

Figure 10.8 shows one of the four trombone players during Medea's premiere performance. In the context of Medea as a video opera, the expressive matching between physical movement (by the instrumentalist) and sound movement through space clearly plays the metaphorical role of a "camera car," where the public enters the physical movement through the movement of sound itself. It should be noted that all of these considerations are subtle and subliminal as is most of the experience of listening to contemporary music.

Figure 10.8: *Trombone during the premiere performance of Adriano Guarnieri's Medea.*

10.8.8 Perspectives

Multimodal and cross-modal approaches for integrated analysis of multimedia streams offers an interesting challenge and opens novel perspectives for control of interactive music systems. Moreover, they can be exploited in the broader fields of multimedia content analysis, multimodal interactive systems, innovative natural and expressive interfaces.

This section presented a conceptual framework, research methodologies and concrete examples of cross-modal and multimodal techniques for control of interactive music systems. Preliminary results indicate the potential of such approach: cross-modal techniques enable to adapt to the analysis in a given modality approaches originally conceived for another modality, allowing in this way the development of novel and original techniques. Multi-modality allows integration of features and use of complementary information, e.g. use of information in a given modality for supplementing lack of information in another modality or for reinforcing the results obtained by analysis in another modality.

While these preliminary results are encouraging, further research is needed for fully exploiting cross-modality and multimodality. For example, an open problem which is currently under investigation at DIST - InfoMus Lab concerns the development of high-level models allowing the definition of cross-modal features. That is, while the examples in this chapter concern cross-modal algorithms, a research challenge consists of identifying a collection of features that, being at a higher-level of abstraction with respect to modal features, are in fact independent of modalities and can be considered cross-modal since they can be extracted from and applied to data coming from different modalities. Such cross-modal features are abstracted from the currently available modal features and define higher-level feature spaces allowing for multimodal mapping of data from one modality to another.

Another, more general, open research issue is how to exploit the information obtained from multimodal and cross-modal techniques for effective control of future interactive music systems. That

is, how to define suitable strategies for mapping the information obtained from the analysis of users' behavior (e.g. performer's expressive gestures) onto real-time generation of expressive outputs (e.g. expressive sound and music output). This issue includes the development of mapping strategies integrating both fast adaptive and reactive behavior and more high-level decision-making processes. Current state-of-the-art control strategies often consist of direct associations, without any dynamics, of features of analyzed (expressive) gestures with parameters of synthesised (expressive) gestures (e.g. the actual position of a dancer on the stage may be mapped onto the reproduction of a given sound). Such direct associations are usually employed for implementing statically reactive behavior. The objective is to develop high-level indirect strategies, including reasoning and decision-making processes, and related to rational and cognitive processes. Indirect strategies implement adaptive and dynamic behavior and are usually characterised by a state evolving over time and decisional processes. Production systems and decision-making algorithms may be employed to implement this kind of strategies. Multimodal interactive systems based on a dialogical paradigm may employ indirect strategies only or a suitable mix of direct and indirect strategies.

As a final remark, it should be noticed that control issues in the Sound and Music Computing field are often related to aesthetic, artistic choices. To which extent can a multimodal interactive (music) system make autonomous decisions? That is, does the system have to follow the instructions given by the director, the choreographer, the composer, (in general the creator of a performance or of an installation) or is it allowed to have some degree of freedom in its behavior? The expressive autonomy of a multimodal interactive system is defined as the amount of degrees of freedom that a director, a choreographer, a composer (or in general the designer of an application involving communication of expressive content) leaves to the system in order to make decisions about the most suitable expressive content to convey in a given moment and about the way to convey it. In general, a multimodal interactive system can have different degrees of expressive autonomy and the required degree of expressive autonomy is crucial for the development of its multimodal and cross-modal control strategies.

References

P. Polotti and D. Rocchesso. *Sound to Sense - Sense to Sound: A state of the art in Sound and Music Computing*. Logos Verlag, Berlin, Germany, 2008.

Contents