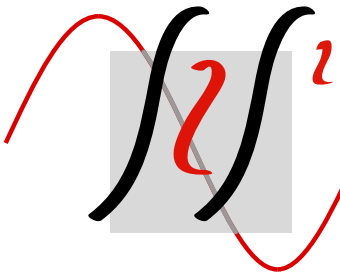

Sound to Sense, Sense to Sound

Sound to Sense, Sense to Sound: A State-of-the-Art

*** Draft Version 0.09 ***

The first S2S^2 Summer School

Genova — July 2005

Ed. Marc Leman, Damien Cirotteau

Contents

Introduction

S2S^2 has the following overall objective: to bring together the state-of-the-art research in the sound domain and in the proper combination of human sciences, technological research and neuropsychological sciences that does relate to sound and sense. Reaching this objective can foster a new generation of research topics such as higher-level sound analysis, the so-called "engaging" synthesis, an integrated sound-music research field, etc.

Nowadays, there is a wide variety of techniques that can be used to generate and analyze sounds. However, urgent requirements (coming from the world of ubiquitous, mobile, pervasive technologies and mixed reality in general) trigger some fundamental yet unanswered questions:

- how to synthesize sounds that are perceptually adequate in a given situation (or context)?
- how to synthesize sound for direct manipulation or other forms of control?
- how to analyze sound to extract information that is genuinely meaningful?
- how to model and communicate sound embedded in multimodal content in multisensory experiences?
- how to model sound in context-aware environments?

As a specific core research emerging and motivated by the above depicted scenario, essentially sound and sense are two separate domains and there is a lack of methods to bridge them with two-way paths: From Sound to Sense, from Sense to Sound (S2S^2). The coordination action S2S^2 has been conceived to prepare the scientific grounds on which to build the next generation of scientific research on sound and its perceptual/cognitive reflexes. So far, a number of fast-moving

sciences ranging from signal processing to experimental psychology, from acoustics to cognitive musicology, have tapped the S2S^2 arena here or there. What we are still missing is an integrated multidisciplinary and multidirectional approach. Only by coordinating the actions of the most active contributors in different subfields of the S2S^2 arena we can hope to elicit fresh ideas and new paradigms. The potential impact on society is terrific, as there is already a number of mass application technologies that are stagnating because of the existing gap between sound and sense. Just to name a few: sound/music information retrieval and data mining (whose importance exceeds P2P exchange technologies), virtual and augmented environments, expressive multimodal communication, intelligent navigation, etc. S2S^2 overall objective can be further specified in the following short-term objectives:

- to establish a reference framework for all the thematic areas (See Thematic areas) catered by the best experts available in Europe and abroad by setting up the appropriate communication and information sharing tools on all fields (website, mailing lists, publications, good practice statements and references, etc.);
- to develop a research roadmap for Sound-related Sciences and their research applications, to assess the current state of the art and likely future research directions (propose industrial projects, joint projects, ..);
- to promote advanced scientific studies and extensive reporting on all thematic areas through the organization of dedicated and specialized international thematic workshops;
- to assist in training and development of new researchers in this area through the constitution of a program of training and mobility dedicated to professionals and post-graduate students;
- to disseminate the research activity coordinated within S2S^2 to potential beneficiaries and external collaborators such as industry and international groups (participating to conferences, contributing to international standards, etc.);
- to promote extensive scientific development through the active participation to international conferences calling for papers on all aspects relating to the thematic areas developed by S2S^2;
- to create a distributed publishing activity capable of answering the needs for publication in several domains belonging to sound research activities through both traditional publishing

and new Free Publishing items addressed to the specialized and the layman public alike on all thematic areas;

- to extend the awareness of scientific research in sound and its related social implications (e.g. ecological acoustics, increase of life quality, etc.) to socially relevant areas (such as non-governmental organizations, recently-industrialized countries, equal-opportunity situations, etc.) through specific dissemination in these areas.

Partners

- Media Innovation Unit - Firenze Tecnologia, Firenze - Italy (Coordinator)
- Kungl Tekniska Högskolan, Stockholm - Sweden
- CSC DEI, Università di Padova, Padova - Italy
- DI-VIPS, Università di Verona, Verona - Italy
- DIST, Università di Genova, Genova - Italy
- Helsinki University of Technology, Helsinki - Finland
- PECA DEC, Ecole Normale Supérieure, Paris - France
- IPEM, Ghent University - Belgium
- LEAD, Université de Dijon - France
- Fundació Universitat Pompeu Fabra, Barcelona - Spain
- OFAI Austrian Research Institute for Artificial Intelligence, Wien - Austria

Purpose of this book

This book aims at giving a state-of-the-art in research related to sound and sense.

Chapter 1

Sound, sense, and music mediation: A historical/philosophical perspective

Marc Leman and Frederik Styns
IPEM, Dept. of Musicology, Ghent University, Ghent

1.1 Introduction

This chapter gives a historical/philosophical overview of the sense/sound relationship from the perspective of music mediation. Sense is thereby associated with musical signification practice, while sound is associated with physical energy or matter. Music mediation is about intermediary processes that account for the transition of musical sound into sense, or sense into sound. This overview shows that in the past, the sound/sense relationship was first considered from a cognitive/structural point of view. Only recently, this viewpoint has been broadened and more attention has been devoted to the human body as mediator between sound and sense. This change in approach has important consequences for future research. The overview aims at providing a perspective for the current state-of-the-art in music research, from which projections into the future can be made.

1.2 Stating the problem

Musical sound can have a large impact on a human being, and this impact may be beneficial but in some cases also harming. For example, it is generally known that music can be beneficial for the personal development such as the forming of a personal self or identity, or for social bonding such as the forming of a group identity Hargreaves and North [1999], McNeill [1995]. It is believed that music may enhance sports activities, consumption Wilson [2003], and it can have healing effects on human subjects Walker [2003], Stradling [2002]. On the other hand, there is also evidence that certain types of music can have a harming effect on people, even driving people to self-destruction and suicide (e.g. Maguire and Snipes [1994], Wintersgill [1994], Gowensmith and Bloom [1997], Scheel and Westefeld [1999], Stack [2000], Lacourse et al. [2001], Rustad et al. [2003]).

In this paper, we take for granted that music can have a powerful effect on humans. Yet, a better understanding of this effect is necessary for two reasons, first, for the development of technologies for music mediation , and second, for the enhancement of possible beneficial effects.

Technologies for music mediation aim at bridging the gap between a human approach and a physical approach. Humans think and act in terms of goals, values, interpretation, while the physical approach considers music from the point of view of physical energy and signal processing. Mediation is about the intermediary processes that link the human approach with the physical approach.

The question is what properties should a music mediation technology have? For example, electronic music instruments may translate a meaningful musical idea (sense) into sound but which properties should be taken into account in order to make this translation effective? Or when music is contained on a mobile wearable device, how can we access it in a natural way? What properties of the mediation technology would facilitate access to digitally encoded energy?

Indeed, modern digital technology raises many questions concerning to access to music. Reference can be made to digital music libraries, the use of interactive multimedia systems, or digital audio effects (sonifications), control of sound synthesis and many other applications. Traditional mediators, based on bio-mechanical devices are no longer sufficient and need a counterpart in the digital domain. But what mediation tools are needed to make this access feasible and natural, and what are their properties? The answer to this question is highly depending on our understanding of the sound/sense relationship as a natural relationship. It is therefore of interest to have a look at how this relationship has been perceived in the past.

In that past, sense has often been related to mental representations and the activity of the human mind. But how is this mind connected to matter? It is known that this relationship motivated philosophical thinking in the ancient times (Plato, Aristotle, Aristoxenos), up to the development of scientific thinking in modern times (Descartes, Spinoza), and actual philosophical thinking Dennett [1991], Damasio [1999]. Should we adopt the view that sense and sound form part of two parallel worlds (mind and matter), as Descartes thought? Or as two sides of the same coin, as Spinoza thought? Or is mind just an epiphenomenon of matter, as modern philosophy suggest? Or is there just mind that counts, as postmodern philosophy claim? And what about new trends in embodied cognition, ecological theories and trends to multi-modal perception?

This chapter aims at tracing historical and philosophical antecedents of sense/sound studies in view of an action-oriented music epistemology. This epistemology is grounded 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 is not intended to contribute to philosophy, nor to history but to focus on the major important issues that should be taken into account when thinking about future activities in music research. The chapter aims at providing a perspective from which projections into the future can be made.

1.3 From music philosophy to music science

The roots of the modern views on sense/sound relationships can be traced back ancient Greek philosophers, such as Pythagoras, Aristoxenos, Plato and Aristotle. Pythagoras focused attention on the mathematical order underlying harmonic musical relations, while Aristoxenos concerned himself with perception and musical experience Barker [1984]. This distinction between acoustics and practice, is still relevant today, as it reflects the basic distinction between sound and sense. Plato comes into the picture mainly because he attributed strong powers to music, which for him was a reason to abandon certain types of music because of the weakening effect it has on the virtue of young people.

Of particular relevance to the modern viewpoint is Aristotle's famous mimesis theory (Politics, Part V). In this theory, he states that rhythms and melodies contain similarities with the true nature of qualities in human character, such as anger, gentleness, courage, temperance, and the contrary qualities. When we hear imitations of that in music – and according to Aristotle, the objects of imitation in art are men in action, emotions and characters – our feelings move in

sympathy with the original. When listening to music, our soul thus undergoes changes in tune with the affective character being imitated. Aristotle assumes that by imitating the qualities that these characters exhibit in music, our souls are moved in a similar way, so that we become in tune with the affects we experience when confronted with the original.

With these views on acoustics (music as ratios of numbers), musical experience (music as perceived structure), and musical expressiveness (music as imitation of reality), there was sufficient material for a few centuries of philosophical discussion.

This would last until renaissance, when science and art were inspired by a new freedom of thinking. In the early 17th Century, scientific thinking became more prominent, and this had an effect on music research. In his Musicae compendium (1618), the young Descartes gives a good summary of the state-of-the-art. He divided music into three basic components, each of which can be isolated for study, firstly, the mathematical-physical aspect of sound, secondly, the nature of sensory perception, and thirdly, the ultimate effect of such perception on the individual listener's soul. The first is clearly about sound as physical energy, while the second and third are about sense, namely, sense as perceived structure and sense as affect. To Descartes, the impact of sound on a listener's emotions or 'soul' was a purely subjective, irrational element and therefore incapable of being scientifically measured.

Indeed, the scientific revolution, of which Descartes was a component next to other towering figures such as J. Kepler, S. Steven, G. Galilei, M. Mersenne, I. Beeckman, C. Huygens, and others, had a major focus on the mathematical and physical aspect of music, whereas the link with musical sense was more a practical consequence of this aspect, namely the calculation of pitch tunings for clavichord instruments Cohen [1984]. In line with this is Euler's "gradus suavitatis", which is an algorithm that assigns to a frequency ratio a number that corresponds with the "degree of pleasure" of this ratio. Structural aspects of perception, such as pitch scales and consonance, were clearly at the borderline of mathematical and physical enquiries. Emotions or expressive gestures were not yet considered to be a genuine subject of scientific study.

Parallel with this scientific approach to sound, the traditions of Aristoxenes and Aristotle culminated in rule-based accounts of musical practices such as Zarlino's, and later J. Rameau's and Mattheson's. In Der Volkommene Kapelmeister (1739), for example, Mattheson offers a manual of how to compose in a convincing way music that is expressive of certain affects. This work of Mattheson focuses on the way people deal with music and on the way they experience the musical sounds as something that tangles their most intimate feelings. These composition recipes can be seen as handbooks for creating music that makes sense. Obviously, this approach

was based on musical intuition.

In the 18th Century, the science of sound and the practice of musical sense were not clearly 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 that is done with sound. There was no real scientific theory of sense, and so, the gap between sound and sense remained huge.

1.4 The cognitive approach

The idea that a scientific study of subjective involvement with music was possible dates from the late 19th Century. Psychophysics and psychology launched the idea that between sound and sense there is the human brain, whose principles can be understood as processing of information.

1.4.1 Pioneers

The first stage is characterized by the pioneering work of scientists such H. v. Helmholtz , W. Wundt, F. Brentano, who provided the foundations of psychoacoustics, psychology, and phenomenology respectively. With the introduction of psychoacoustics by Helmholtz von Helmholtz [1863/1968], the foundations were led for an information processing approach to the sound/sense relationship. This approach assumed that musical sense could be seen as the result from neurophysiological mechanisms that function as resonators in response to sound input. This approach became very influential in music research because it provided an explanation of some very fundamental structural aspects of musical sense, such as consonance and dissonance, harmony and tonality. Euler's numerical principle could now be exchange by physiological mechanisms whose principles can be known by doing scientific experiments. Mathematical functions can capture the main input/output relationships of these physiological mechanisms. This approach provided the physiological grounding for Gestalt psychology in the first half of the 20th Century, and the cognitive sciences approach of the second half of the 20th Century.

1.4.2 Gestalt psychology and systematic musicology

The second important step was the Gestalt movement, which dates back to the work of Stumpf and Brentano in the late 19th century, and which gained prominence by about 1920, thanks to

the work of scholars such as Wertheimer, Kohler and Koffka. At about 1930, music psychology had already reached a solid base of knowledge. This was based, among others, on the elaborate books on Tonpsychologie by C. Stumpf Stumpf [1883, 1890], G. Révész Révész [1946], or E. Kurth's Kurth [1913/1973]theory of energetic musical experience, and a lot of empirical work by many other scholars that had been directed to the perception of tone distances and intervals, melodies, timbre, as well as rhythmic structures. After 1945, the Gestalt theory lost much of its attractiveness and internationally acclaimed innovative position (see Leman and Schneider [1997]). Instead, it met severe criticisms especially from behavioristic and operationalistic quarters. There had been too many Gestalt laws, and perhaps not enough hardcore explanations to account for these, notwithstanding the great account of experimental work that had been done over decades.

Though its sparkle by 1950 was gone, Gestalt psychology never really disappeared, and instead continued to produce works of prime importance to both general psychology and music psychology in particular. Gestalt thinking gradually gained a new impetus, and was found to be of particular importance in combination with then up-to-date trends in cybernetics and information science.

Gestalt theory was also one of the pillars of systematic musicology. One may just point to Stumpf's many experiments on "Verschmelzung" and consonance Schneider [1997], to Köhlers extensive experiments on timbre, that led to the identification of formants, to Koffka's experiments on rhythm perception, or to various experiments set out by v. Hornbostel and Abraham on tonal distances and tonal brightness. What emerged in this approach is a thorough cognitive account of music perception based on the idea that sense emerges as a global pattern from the information processing of pattern contained in musical sound. Much of the older literature on systematic musicology is summarized in Wellek Wellek [1963] and in Elschek Elschek [1992]. The later contains a comprehensive catalogue of systematic musicology.

1.4.3 Information theoretical accounts of sense

The pioneers and Gestalt theory introduced a methodology based on experiment. Gradually also, it became clear that technology would become the next important methodological pillar. 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 theories that related sound to sense from the viewpoint of music production,

that is, from sense to sound. So far, the main contributions had come from the viewpoint of music perception, that is, from sound to sense.

The approach which took music technology seriously into account was conceived of in terms of *information theory* (e.g. Moles [1952, 1958], Winckel [1960]). 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.

1.4.4 Phenomenology and new media technology

Schaeffer Schaeffer [1966], however, noticed that an objective description of music does not always correspond with our subjective perception. In line with phenomenology and gestalt theory, he felt that the description of musical structure, based on information theory, does not always tell us how music is actually perceived by subjects. Measurements of structures are certainly useful and necessary, but these measurements don't always reveal relationships with subjective understanding. Schaeffer therefore related perception of sounds to the manipulation of the analogue electronic sound-generating equipment of that time. He conceived of musical sense in view of the new media technology of his time. Schaeffer, therefore, drew attention to the role of new media as mediators between sound and sense.

1.4.5 Computational modelling of music cognition

The symbol-oriented approach to music description was launched by the appeal of the *information processing psychology* and *formal linguistics* of the late 1950s (see e.g. Lindsay and Norman [1977]), but now in combination with computer modelling. It offered an approach to musical sense that drew upon the notion of *simulation* of mental information processing mechanisms. Cognitive science, as the new trend was soon called, introduced the point of view that the human mind and thus sense could be conceived of in terms of a machine that manipulates representations of

content on a formal basis Fodor [1981].

The application of the symbol-based paradigm to music (see e.g. Longuet Higgins [1987], Laske [1975], Baroni and Callegari [1984] and other researchers, see also Balaban et al. [1992]) 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 pre-define the algorithms that should extract the conceptualised features from the environment. The predefinition of knowledge atoms and the subsequent manipulation of those knowledge atoms in order to generate further knowledge is a main characteristic of a *Cartesian* or *rationalist* conception of the world. Symbol systems, when used in the context of rationalist modelling, should therefore be used with caution.

In the 1980ies, a shift of paradigm from symbol-based modelling towards subsymbol-based modelling was initiated by the results of the so-called *connectionist* computation Rumelhart et al. [1987], Kohonen [1995]. Connectionism, in fact, (re-)introduced statistics as the main modelling technique for making connections between sound and sense. Given the limitations of rationalist modelling, this approach was rather appealing for music research Todd et al. [1999]. The subsymbolic approach is now regarded as an appropriate tool to evaluate cognitive science above the status of mechanized folk psychology. It offers a computational methodology that is in line with the naturalistic epistemology of traditional systematic musicology. It promises an integrated approach to psychoacoustics, auditory physiology, Gestalt perception, self-organization and cognition. In the search for universal psychological laws, methodology and empirical foundation are required to be as hard as in the physical sciences.

1.5 Beyond cognition

The cognitive tradition became criticized for the fact that it neglected the subjective component in the subject's involvement with the environment. Criticism came from many different corners, first of all from inside cognitive science, in particular from scholars that stressed the phenomenological and embodied aspects of cognition Maturana and Varela [1987], Varela et al. [1992] and later also from the so-called postmodern musicology.

1.5.1 Subjectivism and postmodern musicology

David Huron Huron [1999] defines New Musicology as "a methodological movement in music scholarship of the past two decades, that is loosely guided by a recognition of the limits of human understanding, an awareness of the social milieu in which scholarship is pursued, and a realization of the political arena in which the fruits of scholarship are used and abused". DeNora DeNora [2003] argues that, in response to developments in other disciplines such as literary theory, philosophy, history, anthropology, and sociology, *new* musicologists have called into question the separation of historical issues and musical form and that they have focused on the role of music as a social medium. *New Musicology*, like *Postmodern* thinking, assumes that there is no absolute truth to be known. More precisely truth ought to be understood as a social construction that relates to a local or partial perspective on the world. So the focus of new musicology is on the socio-cultural contexts in which music is produced, perceived and studied and how such contexts guide the way people approach, experience and study music. Aspects of this school of thinking are certainly relevant to the sound/sense relationship (e.g. Hatten [1994], Lidov [2005], Cumming [2000], but the hermeneutic methodology draws on subjective projections and interpretations. The methodology is less easily concealed with the scientific approach of modern music research. In addition, there is less attention to the problem of music mediation technologies.

1.5.2 Embodied music cognition

The embodied view by Maturana and Varela and others Varela et al. [1992], Maturana and Varela [1987] has generated a lot of interest and a new perspective for how to approach the sound/sense relationship. In this approach, the link between sound and sense is based on the role of the human body as mediator between physical energy and meaning. In the cognitive approach the sound/sense relationship was mainly conceived from the point if view of mental processing. The approach was effective in acoustics and structural understanding of music but it was less concerned with gestures and emotional involvement. The Aristotelian component, related to imitation an aesthetic experiences, was not part of the main cognitive program, nor was multi-modal information processing.

Yet the idea that musical involvement is based on the imitation of moving sonic forms (and thus multi-modal) has a certain tradition. In fact, this tradition has been gradually re-discovered

in the last decennium. In systematic musicology, a school of researchers in the late 19th and early 20th centuries had already a conception of musical involvement based on corporeal articulations Lipps [1903], Meirsmann [1922/23], Heinitz [1931], Becking [1928], Truslit [1938].

This approach differs from the well-known gestalt theoretical idea (e.g. Wertheimer, Köhler, Koffka) in that it puts more emphasis on action. Like gestalt theory, this approach may be traced back to open problems in Kant's [1790] aesthetic theory, in particular the idea that beauty is in the formal structure. Unlike gestalt theory, the emphasis was less on brain processes and the construction of good forms, but rather more on the phenomenology of the empathic relationship with these forms through movement and action.

For example, Lipps Lipps [1903] argues that the understanding of an expressive movement (*Ausdrucksbewegung*) in music is based on empathy (*inneren Mitmachen, Einfülung*). While being involved with moving sonic forms, we imitate the movements as expressions. By doing this, we practice the motor muscles which are involved when genuine emotions are felt. As such, we have access to the intended emotional meaning of the music. According to Lipps, the act of (free or unbounded) imitation gives pleasure because it is an expression of the self (Lipps, 1903, p. 111).[1] As such, sad music may be a source of pleasure (*Lust*) because the moving sonic forms allow the subject to express an imitative movement (sadness). This imitation allows the subject to participate in the expressive movement without being emotionally involved, that is, without experiencing an emotional state of sadness.

Also Truslit Truslit [1938], Repp [1993] sees corporeal articulations as manifestations of the inner motion heard in music. He says that "provided the sound has the dynamo-agogic development corresponding to a natural movement, it will evoke the impression of this movement in us" Repp [1993]. Particularly striking is the example he gives of Beethoven who, while composing, would hum or growl up and down in pitch without singing specific notes. This is also a phenomenon often heard when jazz musicians are playing. Truslit used the technology of his time to extract information from acoustic patterns, as well as information from body movements with the idea of studying their correlations.

In Gestaltung und Bewegung in der Musik, Alexander Truslit argues that in order to fully experience music, it is essential to understand its most crucial characteristic. According to Truslit this characteristic, the driving force of the music, is the expression of inner movement. The

[1] Similar ideas are found in the theory of optimal experience of Csikszentmihalyi Csikszentmihalyi [1990]. Any expression of the self, or anything that contributes to its ordering, gives pleasure.

composer makes music that is full of inner movement. The musician gives shape to these inner movements by translating them into proper body gestures and the 'good' music listener is able to trace and imitate these movements in order to experience and understand the music properly.

According to Truslit, not all music listeners are able to perceive the inner movements of the music. However, some music listeners have a special capacity to couple the auditive information to visual representations. Such visual representations are referred to as synoptic pictures. Listeners possessing this capability have a great advantage for understanding the musical inner movement.[2]

In accordance with Truslit, Becking Becking [1928], Nettheim [1996] makes also a connection between music and movement, based on the idea of a dynamic rhythmic flow beyond the musical surface. This flow, a continuous up-down movement, connects points of metrical gravitude that vary in relative weight. Beckings most original idea was that these metrical weights vary from composer to composer. The analytical method Becking worked out in order to determine these weights was his method of accompanying movements, conducted with a light baton. Like Truslit, Becking determined some basic movements. These basic movements form the basic vocabulary that allowed him to classify the personal constants of different composers in different eras.

The embodied cognition approach states that the sound/sense relationship is mediated by the human body. This is largely in agreement with recent thinking about the connections between perception and action Prinz and Hommel [2002], Dautenhahn and Nehaniv [2002].

[2]Central in Truslits approach of musical movement are the notions of dynamics (intensity) and agogics (duration). If the music has the dynamo-agogic development corresponding to a natural movement, it will evoke the impression of this movement. Four basic movements are being distinguished in order to identify and understand musical movement. These basic movements are: straight, open, closed and winding. Furthermore, it is stated that, based on this basic vocabulary of movements, it is possible to determine the shape of the inner movements of the music in an objective way. Once the shapes of the movements are determined, it is found useful to make graphical representations of them. Such graphical representations can be used by musicians and music listeners as guidelines for understanding and examining music's inner movement. Truslit sees the inner movement of music first of all as something that is presented in the musical melody. The addition of rhythmic, metric or harmonic elements can only refine this inner movement. A distinction is made between rhythmic movement and the inner movement of the music that Truslit focuses on. In contrast to rhythmic movement, which is related to individual parts of the body, the inner movement forms the melody and is, via the labyrinth (is situated in the vestibular system), related to the human body as a whole.

1.5.3 Music and emotions

The study of subjective involvement with music draws upon a long tradition of experimental psychological research, initiated by Wundt in the late 19th Century. Reference can be made to research in experimental psychology in which descriptions of emotion and affect are related to descriptions of musical structure (see e.g. Hevner [1936], Watson [1942], Reinecke [1964], Imberty [1976], Wedin [1972], Juslin and Sloboda [2001], Gabrielsson and Juslin [2003] for an overview). 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.

In that respect, Berlyne's work Berlyne [1971] on experimental aesthetics is important for having specified a relationship between subjective experience (e.g. arousal) and objective descriptions of *complexity*, *uncertainty* or *redundancy*. In Berlyne's concept, the latter provide an information-theoretic account of symbolic structures (e.g. melodies). They are not just based on perceived structures but are extracted directly from the stimulus (as symbolically represented). However, up to the present, most research has been based on a comparison between perceived musical structure and experienced musical affect. What is needed are comparisons of structure as perceived and structure which is directly extracted from the physical energy Leman et al. [2003].

1.5.4 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.[3] With the advent of powerful computing tools, and in particular real-time interactive music systems Pressing [1992], Rowe [1992], gradually more attention has been devoted to the role of gesture in music Wanderley and Battier [2000], Camurri et al. [2001], Sundberg [2000], Camurri et al. [2005]. This gestural approach has been rather influential in that it puts more emphasis on sensorimotor feedback and integration, as

[3] In 2004, the ConGAS COST-287 action, supported by the EU, established a European network of laboratories that focus on issues related to gesture and music.

well as on the coupling of perception and action. With new sensor technology, gesture-based research has meanwhile become a vast domain of music research Paradiso and O'Modhrain [2003], Johannsen [2004], Camurri and Rikakis [2004], Camurri and Volpe [2004], with consequences for the methodological and epistemological foundations of music cognition research. There is now convincing evidence that much of what happens in perception can be understood in terms of action (see e.g. Jeannerod [1994], Berthoz [1997], Prinz and Hommel [2002]. Pioneering studies in music Clynes [1977], Todd et al. [1999], Friberg and Sundberg [1999] had addressed this coupling of perception and action in musical activity, yet the epistemological and methodological consequences of this approach have not been fully worked out in terms of a musicological paradigm Leman [1999]. It is likely that more attention to the coupling of perception and action will result in more attention to the role of corporeal involvement with in music, which in turn will require more attention to multi-sensory perception, perception of movement (kinaesthesia), affective involvement, and expressiveness of music Leman and Cammuri [In Press].

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 (for an overview, see Karjalainen et al. [2001]). 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 D'haes [2004]. 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 Camurri and Volpe [2004].

Motor theory of perception:

Physical models further suggest a reconsideration of the nature of perception in view of stimulus-source relationships and gestural foundations of musical engagement. Purves and Lotto Purves and Lotto [2003], for example, argue that invariance in perception is based on a 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. It 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. Liberman and Mattingly Liberman and Mattingly [1989] had already assumed that the speech production-perception system is, in effect, an *articulatory synthesiser*. In the production mode, the synthesiser is activated by an abstract gestural pattern from which the synthesiser computes a series of articulatory movements that are needed to realise the gestures into muscle movements of the vocal tract. In the perception mode, then, the synthesiser computes the series of articulatory movements that *could have* produced the signal, and from this articulatory representation, the intended gestural pattern, contained in the stimulus, is obtained. Liberman and Mattingly assumed a specialised module responsible for both perception and production of phonetic structures. The perceptual side of this module converts automatically from acoustic signal to gesture. Perception of sound comes down to finding the proper parameters of the gesture that would allow the re-synthesis of what is heard. So, features related to sound are in fact picked up as parameters for the control of the articulatory system. Perception of a sound, in that view, is an inhibited re-synthesis of that sound, inhibited in the sense that the re-synthesis is not actually carried out but simulated. The things that need to be stored in memory, then, are not auditory images, but gestures, sequences of parameters that control the human articulatory (physical) system. The view also assumes that perception and action share a common representational system. Such models thus receive input from the sensors and produce appropriate actions as output and, by doing this, stimuli thus become meaningful in relation to their sources which are objects of action Varela et al. [1992]. Action, in other words, guarantees that the stimuli

are connected to the object, the source of the physical energy that makes up the stimulus. 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.

1.6 Conclusion

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 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) gradually on an understanding of the subjective principles, starting with principles of perception of structure, towards a better understanding of principles that underly 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 driven by experiments and computer modelling. Technology has played an increasing important role, first as measuring instrument, later as modelling tool and music mediation tools. The approach started from cognitive science and symbolic modelling, and turned to sub-symbolic modelling and empirical modelling in the late 1980ies. In the recent decades, more attention has been draws 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 connected to matter by means of the 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 disposition.

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).

Bibliography

Mira Balaban, Kemal Ebciöglu, and Otto E. Laske, editors. *Understanding music with AI: perspectives on music cognition*. AAAI Press, Cambridge (Mass.), 1992.

Andrew Barker. *Greek musical writings*. Cambridge readings in the literature of music. Cambridge University Press, Cambridge, 1984.

M. Baroni and L. Callegari. *Musical grammars and computer analysis*. Quaderni della Rivista italiana di musicologia 8. Olschki, Firenze, 1984.

Gustav Becking. *Der musikalische Rhythmus als Erkenntnisquelle*. B. Filser, Augsburg,, 1928.

D. E. Berlyne. *Aesthetics and psychobiology*. Appleton-Century-Crofts, New York, 1971.

A. Berthoz. *Le sens du mouvement*. Editions O. Jacob,, Paris, 1997.

A. Camurri, G. De Poli, M. Leman, and G. Volpe. A multi-layered conceptual framework for expressive gesture applications. In X. Serra, editor, *Intl EU-TMR MOSART Workshop*, Barcelona, 2001. Univ Pompeu Fabra.

A. Camurri and T. Rikakis. Multisensory communication and experience through multimedia. *Ieee Multimedia*, 11(3):17–19, 2004.

A. Camurri and G. Volpe, editors. *Gesture-based communication in human-computer interaction. Selected revised papers of the 5th Intl Gesture Workshop (GW2003)*. Lecture Notes in Artificial Intelligence, LNAI. Springer-Verlag, Berlin, 2004.

A. Camurri, G. Volpe, G. De Poli, and M. Leman. Communicating expressiveness and affect in multimodal interactive systems. *IEEE Multimedia*, 12(1):43–53, 2005.

Manfred Clynes. *Sentics: the touch of emotions*. Anchor Press, New York, 1977.

H.F. Cohen. *Quantifying music: the science of music at the first stage of the scientific revolution, 1580-1650*. Reidel, Dordrecht, 1984.

Mihaly Csikszentmihalyi. *Flow: the psychology of optimal experience*. Harper & Row, New York, 1990.

Naomi Cumming. *The sonic self: musical subjectivity and signification*. Advances in semiotics. Indiana University Press, Bloomington, 2000.

Antonio R. Damasio. *The feeling of what happens: body and emotion in the making of consciousness*. Harcourt Brace, New York, 1999.

Kerstin Dautenhahn and Chrystopher L. Nehaniv. *Imitation in animals and artifacts*. Complex adaptive systems. MIT press, Cambridge (Mass.), 2002.

Daniel Clement Dennett. *Consciousness explained*. Little, Brown and Co., Boston, 1991.

Tia DeNora. *After Adorno: rethinking music sociology*. Cambridge University Press, Cambridge, 2003.

W. D'haes. *Automatic estimation of control parameters for musical synthesis algorithms*. Phd thesis, Universiteit Antwerpen, 2004.

O. Elschek. *Die Musikforschung der Gegenwart, ihre Systematik, Theorie und Entwicklung*. Dr. E. Stiglmayr, Wien-Föhrenau, 1992.

Jerry A. Fodor. *Representations: philosophical essays on the foundations of cognitive science*. MIT Press, Cambridge, Mass., 1st mit press edition, 1981.

A. Friberg and J. Sundberg. Does music performance allude to locomotion? a model of final ritardandi derived from measurements of stopping runners. *Journal of the Acoustical Society of America*, 105(3):1469–1484, 1999.

A. Gabrielsson and P. N. Juslin. Emotional expression in music. In K. R. Scherer Goldsmith H. H., R. J. Davidson, editor, *Handbook of affective sciences*, pages 503–534. Oxford University Press, New York, 2003.

W. N. Gowensmith and L. J. Bloom. The effects of heavy metal music on arousal and anger. *Journal of Music Therapy*, 34(1):33–45, 1997.

D. J. Hargreaves and A. C. North. The functions of music in everyday life: redefining the social in music psychology. *Psychology of Music*, 27:71–83, 1999.

Robert S. Hatten. *Musical meaning in Beethoven markerdess, correlation, and interpretation*. Advances in semiotics. Indiana university press, Bloomington (Ind.), 1994.

W. Heinitz. *Strukturprobleme in Primitiver Musik*. Friederichsen, De Gruyter & Co. M. B. H., Hamburg, 1931.

K. Hevner. Experimental studies of the elements of expression in music. *American Journal of Psychology*, 48:246–248, 1936.

D. Huron. The new empiricism: systematic musicology in a postmodern age, 1999.

M. Imberty. Signification and meaning in music, 1976.

M. Jeannerod. The representing brain - neural correlates of motor intention and imagery. *Behavioral and Brain Sciences*, 17(2):187–202, 1994.

G. Johannsen, editor. *Engineering and music-supervisory, control and auditory communication (special issue of Proceedings of the IEEE 92(4))*. IEEE, 2004.

Patrik N. Juslin and John A. Sloboda. *Music and emotion: theory and research*. Series in affective science. Oxford University Press, Oxford, 2001.

M. Karjalainen, T. Tolonen, V. Valimaki, C. Erkut, M. Laurson, and J. Hiipakka. An overview of new techniques and effects in model-based sound synthesis. *Journal of New Music Research*, 30 (3):203–212, 2001.

Teuvo Kohonen. *Self organizing maps*. Springer series in information sciences 30. Springer, Berlin, 1995.

E. Kurth. *Die Voraussetzungen der Theoretischen Harmonik und der tonalen Darstellungssysteme*. Musikverlag Emil Katzbichler, München, 1913/1973.

E. Lacourse, M. Claes, and M. Villeneuve. Heavy metal music and adolescent suicidal risk. *Journal of Youth and Adolescence*, 30(3):321–332, 2001.

Otto E. Laske. *Introduction to a generative theory of music*. Sonological reports 1B. Utrecht state university, Institute of sonology, Utrecht, 1975.

M. Leman. Naturalistic approaches to musical semiotics and the study of causal musical signification. In I. Zannos, editor, *Music and Signs, Semiotic and Cognitive Studies in Music*, pages 11–38. ASKO Art & Science, Bratislava, 1999.

M. Leman and A. Cammuri. Understanding musical expressiveness using interactive multimedia platforms. *Musicae scientiae*, In Press.

M. Leman and A. Schneider. Origin and nature of cognitive and systematic musicology: An introduction. In M. Leman, editor, *Music, Gestalt, and Computing - Studies in Cognitive and Systematic Musicology*, pages 13–29. Springer-Verlag, Berlin, Heidelberg, 1997.

M. Leman, V. Vermeulen, L. De Voogdt, J. Taelman, D. Moelants, and M. Lesaffre. Correlation of gestural musical audio cues and perceived expressive qualities. *Gesture-Based Communication in Human-Computer Interaction*, 2915:40–54, 2003.

A. M. Liberman and I. G. Mattingly. A specialization for speech-perception. *Science*, 243:489–494, 1989.

D. Lidov. *Is language a Music? Writings on Musical Form and Signification*. Indiana University Press, Bloomington, 2005.

Peter H. Lindsay and Donald A. Norman. *Human information processing: An introduction to psychology*. Academic Press, New York, 2nd edition, 1977.

Theodor Lipps. *Ästhetik Psychologie des Schönen und der Kunst*. L. Voss, Hamburg und Leipzig,, 1903.

H. Christopher Longuet Higgins. *Mental processes studies in cognitive science*. Explorations in cognitive science 1. MIT press, Cambridge (Mass.), 1987.

E. R. Maguire and J. B. Snipes. Reassessing the link between country-music and suicide. *Social Forces*, 72(4):1239–1243, 1994.

Humberto R. Maturana and Francisco J. Varela. *The tree of knowledge: the biological roots of human understanding*. New Science Library, Boston, 1987.

William Hardy McNeill. *Keeping together in time: dance and drill in human history*. Harvard University Press, Cambridge (Mass.), 1995.

H. Meirsmann. Versuch einer phänomologie der musik. *Zeitschrift für Musikwissenschaft*, 25: 226–269, 1922/23.

A. Moles. *Physique et technique du bruit*. Dunod, Paris, 1952.

Abraham Moles. *Théorie de l'information et perception esthétique*. Etudes de radio télévision. Flammarion, Paris, 1958.

N. Nettheim. How musical rhythm reveals human attitudes: Gustav becking's theory. *International Review of the Aesthetics and Sociology of Music*, 27(2):101–122, 1996.

J. A. Paradiso and S. O'Modhrain, editors. *New interfaces for musical performance and interaction (special issue of the Journal of New Music Research 32(4))*. Journal of New Music Research. Swets Zeitlinger, Lisse, 2003.

Jeff Pressing. *Synthesizer performance and real-time techniques*. The Computer music and digital audio series; v. 8. A-R Editions, Madison, Wis, 1992.

Wolfgang Prinz and Bernhard Hommel, editors. *Common mechanisms in perception and action*. Attention and performance 19. Oxford university press, Oxford, 2002.

Dale Purves and R. Beau Lotto. *Why we see what we do: an empirical theory of vision*. Sinauer Associates, Sunderland (Mass.), 2003.

H. P. Reinecke. *Experimentelle Beiträge zur Psychologie des musikalischen Hörens*. Universität Hamburg, Hamburg, 1964.

B. H. Repp. Music as motion: a synopsis of alexander truslit's (1938) gestaltung und bewegung in der music. *Psychology of Music*, 12(1):48–72, 1993.

Géza Révész. *Einführung in die Musikpsychologie*. A. Francke, Bern, 1946.

Robert Rowe. Interactive music systems: machine listening and composing, 1992.

David E. Rumelhart, James L. McLelland, Chisato Asanuma, and PDP research group. *Parallel distributed processing explorations in the microstructure of cognition*. Computational models of cognition and perception. MIT press, Cambridge (Mass.), 6th print. edition, 1987.

R. A. Rustad, J. E. Small, D. A. Jobes, M. A. Safer, and R. J. Peterson. The impact of rock videos and music with suicidal content on thoughts and attitudes about suicide. *Suicide and Life-Threatening Behavior*, 33(2):120–131, 2003.

Pierre Schaeffer. *Traité des objets musicaux essai interdisciplines*. Pierres vives. Seuil, Paris, 1966.

K. R. Scheel and J. S. Westefeld. Heavy metal music and adolescent suicidality: An empirical investigation. *Adolescence*, 34(134):253–273, 1999.

A. Schneider. Verschmelzung, tonal fusion, and consonance: Carl stumpf revisited. In M. Leman, editor, *Music, Gestalt, and Computing. Studies in Cognitive and Systematic Musicology*. Springer-Verlag, Berlin, 1997.

S. Stack. Blues fans and suicide acceptability. *Death Studies*, 24(3):223–231, 2000.

R. Stradling. Music as medicine: The history of music healing since antiquity. *Social History of Medicine*, 15(2):341–342, 2002.

Carl Stumpf. *Tonpsychologie*. S. Hirzel, Leipzig, 1883.

Carl Stumpf. *Tonpsychologie II*. S. Hirzel, Leipzig, 1890.

J. Sundberg, editor. *Music and Motion (Special issue of the Journal of New Music Research 29(3))*. Swets and Zeitlinger, Lisse, 2000.

N. P. M. Todd, D. J. O'Boyle, and C. S. Lee. A sensory-motor theory of rhythm, time perception and beat induction. *Journal of New Music Research*, 28(1):5–28, 1999.

Alexander Truslit. *Gestaltung und bewegung in der musik; ein tönendes buch vom musikalischen vortrag und seinem bewegungserlebten gestalten und hören*. C.F. Vieweg, Berlin-Lichterfelde,, 1938.

Francisco J. Varela, Eleanor Rosch, and Evan Thompson. *The embodied mind, cognitive science and human experience*. MIT press, Cambridge (Mass.), 2nd print. edition, 1992.

H. von Helmholtz. *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik*. Georg Olms Verlagsbuchhandlungbib, Hildesheim, 1863/1968.

M. Walker. Music as knowledge in shamanism and other healing traditions of siberia. *Arctic Anthropology*, 40(2):40–48, 2003.

M. Wanderley and M. Battier, editors. *Trends in Gestural Control of Music*. IRCAM, Paris, 2000.

K. B. Watson. The nature and measurement of musical meanings. *Psychological Monographs*, 54: 1–43, 1942.

L. Wedin. Multidimensional study of perceptual-emotional expression in music. *Scandinavian Journal of Psychology*, 13:241–257, 1972.

Albert Wellek. *Musikpsychologie und Musicästhetik; Grundriss der systematischen Musikwissenschaft*. Akademische Verlagsgesellschaft, Frankfurt am Main,, 1963.

S. Wilson. The effect of music on perceived atmosphere and purchase intentions in a restaurant. *Psychology of Music*, 31(1):93–112, 2003.

F. Winckel. *Vues Nouvelles sur le Monde des Sons*. Dunod, Paris, 1960.

P. Wintersgill. Music and melancholia. *Journal of the Royal Society of Medicine*, 87(12):764–766, 1994.

Chapter 2

Learning music: prospects about implicit knowledge in music, new technologies and music education

Abstract

Research in auditory cognition domain has shown that even nonmusician listeners have knowledge about the Western tonal musical system. Acquired by mere exposure to musical pieces, this implicit knowledge guides and shapes music perception. The first part of our article presents some research studying implicit learning processes, which are at the origin of musical knowledge of nonmusician listeners, and the perception of musical structures. The second part makes the link between findings in cognitive psychology and the use of multimedia. It presents some examples of applications for the instruction of learning and perceiving Western tonal music and contemporary music.

Introduction

The present article proposes an overview on the musical learning by underlining the force of the cognitive system, able to learning and treating complex information at an implicit level. The first part summarizes research in cognitive sciences, which study the processes of implicit learning and

the musical perception in listener nonmusician. These studies show that the nonmusicians obtain generally good performances, very often comparable to those of the musicians. The second part illustrates by means of some examples the use of multimedia tools for music (learning, perception, understanding). These illustrations are based on the recent advances in cognitive psychology concerning the acquisition of knowledge, their representation, influence of the attention as well as the interaction between visual and auditive modalities.

2.1 Implicit processing of musical structures

2.1.1 How do non-musician listeners acquire implicit knowledge of music?

Implicit learning processes enable the acquisition of highly complex information and without complete verbalizable knowledge of what has been learned (Seger, 1994). Two examples of highly structured systems in our environment are language and music. Listeners become sensitive to the underlying regularities just by mere exposure to linguistic and musical material in everyday life. The implicitly acquired knowledge influences perception and interaction with the environment. This capacity of the cognitive system is studied in the laboratory with artificial material containing statistical structures, such as finite state grammars or artificial languages (i.e., Altmann, Dienes & Goode, 1995; Reber, 1967, 1989; Saffran, Newport & Aslin, 1996). Tonal acculturation is one example of the cognitive capacity to become sensitive to regularities in the environment. Francès (1958) was one of the first underlining the importance of statistical regularities in music for tonal acculturation, suggesting that mere exposure to musical pieces is sufficient to acquire tonal knowledge, even if it remains at an implicit level. In music cognition domain, numerous research has provided evidence for nonmusicians' knowledge about the tonal system.

2.1.2 Implicit learning of Western pitch regularities

Western tonal music constitutes a constrained system of regularities (i.e., regularities of co-occurrence, frequency of occurrence and psychoacoustic regularities) based on a limited number of elements. This section presents the tonal system from the perspective of cognitive psychology: it underlines the basic regularities between musical events, which appear in most musical styles of occidental everyday life (e.g., classical music, pop music, jazz music, Latin music etc.) and

which can be acquired by implicit learning processes. The Western tonal system is based on 12 pitches repeated cyclically over octaves. Strong regularities of co-occurrence and frequencies of occurrence exist among these 12 pitch classes (referred to as the tones C, C#/Db, D, D#/Eb, E, F, F#/Gb, G, G#/Ab, A, A#/Bb, B): tones are combined into chords and into keys, forming a three-level organizational system (Figure 1). Based on tones and chords, keys (tonalities) define a third level of musical units. Keys have more or less close harmonic relations to each other. Keys sharing numerous tones and chords are said to be harmonically related. The strength of harmonic relations depends on the number of shared events. In music theory, major keys are conceived spatially as a circle (i.e., the circle of fifths), with harmonic distance represented by the number of steps on the circle. Inter-key distances are also defined between major and minor keys. The three levels of musical units (i.e., tones, chords, keys) occur with strong regularities of co-occurrence. Tones and chords belonging to the same key are more likely to co-occur in a musical piece than tones and chords belonging to different keys. Changes between keys are more likely to occur between closely related keys (e.g., C and G major) than between less-related ones (e.g., C and E major). Within each key, tones and chords have different tonal functions creating tonal and harmonic hierarchies. These within-key hierarchies are strongly correlated with the frequency of occurrence of tones and chords in Western musical pieces. Tones and chords used with higher frequency (and longer duration) correspond to events that are defined by music theory as having more important functions in a given key (Budge, 1943; Francès, 1958; Krumhansl, 1990a).

This short description reveals a fundamental characteristic of the Western tonal music: functions of tones and chords depend on the established key. The same event can define an in-key or an out-of-key event and can take different levels of functional importance. For listeners, understanding context dependency of musical events' functions is crucial for the understanding of musical structures, notably the multitude of musical structures that can be created on the basis of twelve pitch classes. Music cognition research suggests that mere exposure to Western musical pieces suffices to develop implicit, but nevertheless sophisticated, knowledge of the tonal system. Just by listening to music in everyday life, listeners become sensitive to the regularities of the tonal system without being necessarily able to verbalize them (Dowling & Harwood, 1986; Francès, 1958; Krumhansl, 1990a). The seminal work by Krumhansl, Bharucha and colleagues has investigated the perception of relations between tones and between chords as well as the influence of a changing tonal context on the perceived relations (see Krumhansl 1990a for a review). The data showed the cognitive reality of tonal and harmonic hierarchies for listeners and the context dependency of musical tones and chords in perception and memorization.

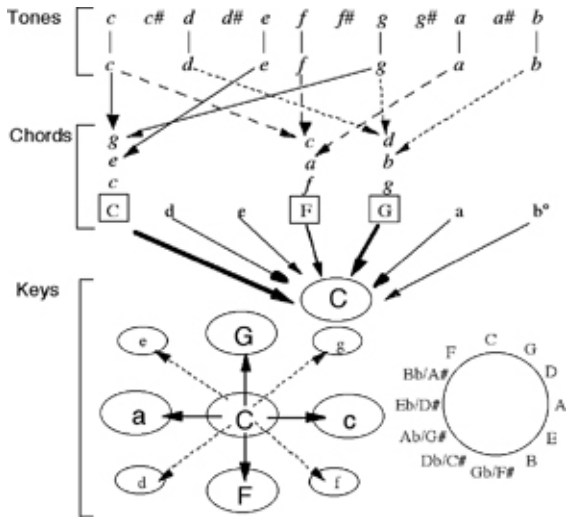

Figure 2.1: Schematic representations of the three organizational levels of the tonal system. Top) 12 pitch classes, followed by the diatonic scale in C Major. Middle) construction of three major chords, followed by the chord set in the key of C Major key. Bottom) relations of the C Major key with close major and minor keys (left) and with all major keys forming the circle of fifths (right). (Tones are represented in italics, minor and major chords/keys in lower and upper case respectively). (from Tillmann et al. 2001, Implicit Learning of Regularities in Western Tonal Music by Self-Organization (pp. 175-184), Figure 1, in: Proceedings of the Sixth Neural Computation and Psychology Workshop: Evolution, Learning, and Development, Springer)

2.1.3 Connectionist model of musical knowledge representation and its acquisition

Bharucha (1987) proposed a connectionist account of tonal knowledge representation. In the MUSACT model (i.e., MUSical ACTivation), tonal knowledge is conceived as a network of interconnected units (Figure 2). The units are organized in three layers corresponding to tones, chords, and keys. Each tone unit is connected to the chords of which that tone is a component. Analogously, each chord unit is connected to the keys of which it is a member. Musical relations emerge from the activation that reverberates via connected links between tone, chord and key units. When a chord is played to MUSACT, the units representing the sounded component tones are activated and activation reverberates between the layers until equilibrium is reached (see Bharucha, 1987; Bigand et al., 1999 for more details). The emerging activation patterns

reflect tonal and harmonic hierarchies of the established key: for example, units representing harmonically related chords are activated more strongly than units representing unrelated chords. The context dependency of musical events in the tonal system is thus not stored explicitly for each of the different keys, but emerges from activation spreading through the network.

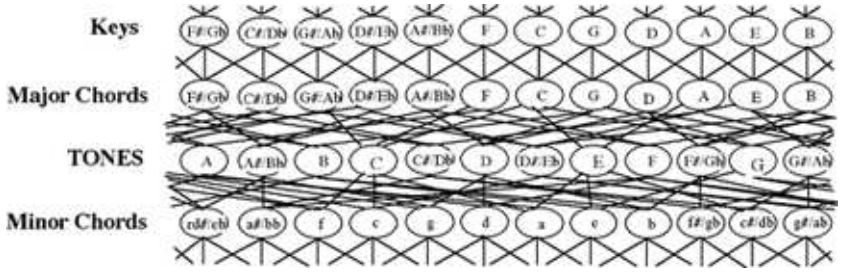

Figure 2.2: MUSACT model of tonal knowledge activation. The tone layer is the input layer, which is connected to the chord layer (consisting of major and minor chords). The chord layer is connected to the key layer (third layer). Adapted from Bharucha (1987).

In Tillmann et al. (2000), we take advantage of the learning possibilities of artificial neural networks (e.g., connectionist models) to simulate tonal knowledge acquisition in nonmusician listeners. For this purpose, unsupervised learning algorithms seem to be well suited: they extract statistical regularities via passive exposure and encode events that often occur together (Grossberg, 1970, 1976; Kohonen, 1995; Rumelhart & Zipser, 1985; von der Malsberg, 1973). Self-organizing maps (Kohonen, 1995) are one version of unsupervised learning algorithms that leads to a topological organization of the learned information.

To simulate tonal acculturation, a hierarchical network composed of two self-organizing maps was exposed to short musical sequences (i.e., chord sequences). After learning, the connections in the network have changed and the units have specialized for the detection of chords and keys (the input layer coded the tones of the input material). The learned architecture is associated with a spreading activation process (as in MUSACT) to simulate top-down influences on the activation patterns. Interestingly, the learned connections and the activation patterns after reverberation mirror the outcome of the hardwired network MUSACT, which has been conceived as an idealized end-state of implicit learning processes (see Tillmann et al., 2000). In collaboration with Michel Paindavoine (LE2I-CNRS, Dijon) and Charles Delbé (LEAD-CNRS, Dijon), we are currently working on several extensions of this connectionist approach. One of the projects concerns the construction of a set of "artificial musical ears" for this modeling approach.

A step of auditory pre-processing allows decoding sound files with the acoustic complexity of the musical stimuli. On the basis of this rich input, a network will be trained with a corpus of real recordings containing a variety of musical pieces.

2.1.4 Studying implicit learning processes with artificial materials

Implicit learning processes are supposed to be at the origin of listeners' tonal knowledge, acquired in everyday life. Implicit learning processes are studied more closely in the laboratory with artificial materials containing statistical regularities. In the seminal work by Reber (1967), participants are asked to memorize grammatical letter strings in a first phase of the experiment, but are unaware that any rules exist. During the second phase of the experiment, they are informed that the previously seen sequences are produced by a rule system (which is not described) and are asked to judge the grammaticality of new letter strings. Participants differentiate grammatical letter strings from new ungrammatical ones at better than chance level. Most of them are unable to explain the rules underlying the grammar in free verbal reports (e.g., Altmann et al., 1995; Dienes, Broadbent & Berry, 1991; Reber, 1967, 1989).

Various findings are convergent in demonstrating the cognitive capacity to learn complex structures and regularities. The acquisition of regularities in the experimental material is not restricted to visual events (e.g., letters, lights, shapes), but has been extended to auditory events, such as sine waves (Altmann et al., 1995), musical timbres (e.g., gong, trumpet, piano, violin, voice in Bigand, Perruchet & Boyer, 1998) or environmental sounds (e.g., drill, clap, steam in Howard & Ballas, 1980, 1982). Recent studies started to consider the acoustical characteristics of the sound, such as prosodic cues (Johnson & Jusczyk, 2001; Saffran et al., 1996, Experiment 2; Thiessen & Saffran, 2003) or acoustical similarities (Tillmann & McAdams, 2004). The hypothesis is to test whether the relation between the statistical regularities and regularities inherent to the acoustical material could influence learning: conflicting information might hinder statistical learning, while converging information might facilitate learning. Tonal acculturation might represent a beneficial configuration: musical events appearing frequently together are also linked acoustically since they share (real and virtual) harmonics. To investigate whether convergence with acoustical features represent a facilitatory or even necessary condition for statistical learning, Tillmann and McAdams (2004) systematically manipulated acoustical similarities between timbres so that they either underline the statistical regularities of the timbre units, contradict these regularities or a neutral to them. The outcome shows that listeners learned the statistical regularities of

the complex auditory material and the manipulated surface characteristics did not affect this statistical learning. The surface characteristics only affected grouping and overall preference bias for the different materials. This outcome suggests that tonal acculturation does not necessarily need the convergence between statistical and acoustical regularities. Supporting evidence can be found in acculturation to Arabic music, which is lacking the convergence between statistical and acoustic features (Ayari &McAdams, 2004). Together with the implicit learning study on twelve-tone music (Bigand, D'Adamo et al., 2003), the data emits the rather encouraging hypothesis about the possibility to learn regularities of new musical styles.

2.1.5 Implicit learning of new musical systems

Music is an interesting medium to investigate implicit learning processes for several reasons. It is a highly complex structure of our environment that is too complex to be apprehended through explicit thoughts and deductive reasoning. Musical events per se are of no importance, yet musical pieces are more than a pleasing succession of coloured sounds. The psychological effect of musical sounds comes from the complex multilevel relationships the musical events pertain in a given piece (Meyer, 1956; Lerdahl & Jackendoff, 1983). The abstract associative and architectonic relations that pertain between events that are not close in time define relevant structures in music. These relations are difficult to articulate in an explicit way. Despite a considerable tradition in music history, as well as in contemporary music theory, to formalize the relevant structure of Western music (see Lerdahl & Jackendoff, 1983; Lerdahl 2001; Narmour, 2000), none of these frameworks provides a complete and satisfactory account of the Western tonal musical grammar. A further interesting feature of music for research on implicit learning is that musical structures are not conceived for explicit processing. It is even of crucial importance for composers that listeners are sensitive to the structures that underlie a musical piece while still being unaware of them. And in fact, the most common impression among a general audience is of being unable to verbally describe what they perceive. In some instances, people are even convinced that they no not perceive any underlying structure. The fact that musical events do not refer to any specific object in the external world probably contributes to the difficulty of apprehending musical structures in an explicit way.

A final interesting feature is that musical systems constantly evolve towards new musical grammars. Being faced with masterpieces that derive from an entirely new musical system is not an artificial situation for contemporary listeners and this raises a challenging issue for

implicit learning theories. The considerable and persistent confusion reported by listeners to contemporary music suggests that some musical grammars may be too artificial to be internalized through passive exposure (McAdams, 1989). As a consequence, several cognitive constraints have been delineated, which musical grammars should obey in order to be learnable (Lerdahl, 1988, 2001). Contemporary music challenges the ability of the human brain to internalize every type of regularity. This raises a question with implications for cognitive science, music cognition, and contemporary music research.

To the best of our knowledge, very little research has directly addressed implicit learning with musical material (Bigand, Perruchet & Boyer, 1998; Dienes et al. 1991). Much research in music cognition, however, indirectly deals with implicit learning processes by showing that explicit learning is not necessary for the development of a sensitivity to the underlying rules of Western music[1] (see section above). Only a few studies have addressed the implicit learning of new musical systems. Most of them have focused on the learning of serial music, a system that has evolved in the West at the beginning of the 20th century. During this period, the tonal musical system gradually waned and it was overtaken by serial systems of composition developed, in particular, by Schoenberg (Griffiths, 1978). Serial works of music obey compositional rules that differ from those that govern tonal music.

A serial musical piece is based on a specific temporal ordering of the twelve tones of the chromatic scale, (i.e., the tones C, C#, D, D#, E, F, F#, G, G#, A, A#, B), irrespective of their octave placement . The specific ordering of these tones defines the tone row of the piece. Each tone of the row should be played before a given tone occurs for the second time. For example, if the piece is made of the following row B-D#-D-F-C#-E-G-F#-Bb-A-C-G#, the C note should not be repeated before all the other notes have been sounded, irrespective of octave placement[2]. In theory, each tone of the row should have roughly the same frequency of occurrence on the overall of the piece. This principle defines the most basic feature of the new musical system and was applied to the musical stimuli of the present study.

The serial musical system defines several types of transformation that can be applied to the tone row. First, the tone row can be transposed to each of the twelve tones it contains. The

[1]The tonal system designates the most usual style of music in the West, including, Baroque (Bach), Classic (Mozart) and Romantic (Chopin) music , as well as folk music such as pop-music, jazz and latin-music.

[2]The 12 tones of the chromatic scale are repeated at different pitch heights, spaced by the interval called an octave. A lower C and a higher C both belong to the same pitch class C. Listeners perceive tones of the same pitch class as perceptually equivalent.

other transformations in serial music consist in playing the row in retrograde order, in inversion (intervals change in direction), and played in retrograde-inversion of the original row.

Each serial composition results from a complex combination of all of these transformations that are applied on a specific tone row. Schoenberg argued that these manipulations would produce an interesting balance between perceptual variety and unity. A critical point on which he insisted was that the initial row must remain unchanged throughout the entire piece (1925). In other words, Schoenberg's cognitive intuition was that the perceptual coherence deriving from the serial grammar was unlikely to be immediately perceived but would result from a familiarization with the row.

Several experimental studies have addressed the psychological reality of the organization resulting from serial musical grammar. The oldest, Francès (1958, exp. 6), consisted in presenting participants with 28 musical pieces based on a specific tone row and requiring participants to detect four pieces that violate the row. These odd pieces were actually derived from another row (the foil row). The analysis of accurate response revealed that participants had considerable difficulty in detecting the four musical pieces that violate the initial row. Moreover, the fact that music theorists specialized in serial music did not respond differently from musically untrained participants suggests that extensive exposure to serial works is not sufficient for the internalization of this new musical system. Although Francès' research is remarkable as pioneer work in this domain, the study contained several weaknesses relative to the experimental design as well as to the analysis of the data and this detracts from the impact of his conclusion. The most noticeable problem concerns the foil row as it was strongly related to the tested row.

Empirical evidence supporting the perceptual reality of the rules of serial music was reported by Dowling (1972) with short melodies of 5 tones. In Dowling's experiment, participants were trained to identify reversed, retrograde and retrograde-inversion of standard melodies of 5 tones of equal duration. The melodies were deliberately made with small pitch intervals in order to improve performance. Dowling found that musically untrained participants managed to identify above chance the rules of the serial music, with highest accuracy for the reversed transform and the lowest for the retrograde inversion. Given that Dowling's musical stimuli were extremely short and simple, it is difficult to conclude that the rules of serial music may be internalized from a passive hearing of serial music. Moreover, in a very similar experiment using 12 tones instead of 5, DeLannoy (1972) reported that participants did not success above chance in distinguishing legal transformations of a standard musical sequence from those that violate the serial rules.

More recently, Dienes and Longuet-Higgins (2001) attempted to train participants in the grammar of serial music by presenting them with 50 musical sequences that illustrated one of the transformation rules of serial music. The last 6 notes of the row were a transform of the first 6 (i.e., a reverse, a retrograde or a retrograde inversion transformation). After this familiarization phase, participants were presented with a new set of 50 sequences, some of them violating the rules of serial music (i.e., the last 6 notes were not a legal transformation of the first 6). Participants were required to differentiate grammatical pieces (according to serial rules) from nongrammatical ones. Accuracy rates generally did not differ from chance level, which is consistent with Francès (1956) and Delannoy (1972)'s finding.

A critical feature of the experiment of Dienes et al (1991) is that participants were never exposed to a single tone row. Participants were trained with the transformational rules of serial music, but these rules were always instantiated with a new set of tones. The temporal order of the first 6 notes was chosen at random. As a consequence, the referential row is constantly moving from one trial to the other. Such a procedure is very demanding since it consists in requiring participants to learn abstract rules that are illustrated by a constantly changing alphabet. To the best of our knowledge, there is no evidence in the domain of implicit learning domain to show that learning can occur in such a situation. If participants do not have the opportunity to be exposed to an invariant tone row in the training phase, it is not surprising that they fail to exhibit sensitivity to the serial grammar in the test phase. It should be noticed that this situation violates the basic principle of serial music, which postulates that only one row should be used for one piece (Schoenberg, 1925). Krumhansl, Sandler and Sergeant' study (1987) has provided the strongest support for the psychological relevance of serial rules. Experiments 1 and 2 were run with simple forms of two tone rows. That is to say, the tone rows were played with isochronous tones that never exceeded the pitch range of one octave. Experiments 3 and 4 were run with excerpts of Wind Quintets op. 26 and String Quartet op. 37 by Schoenberg. The results of the classification tasks used in Experiments 2 and 3 demonstrated that participants discriminated, at a level above chance, inversion, retrograde, and retrograde inversion of the two tone rows with correct responses varying from 73% to 85% in Experiment 2, and from 60% to 80% in Experiment 3. At first glance, this high accuracy is surprising. However, it should be noticed that participants were exposed to very simple forms of the tone rows a great number of times during Experiment 1. It seems likely that this previous exposure helps to explain the good performance. In other words, the peculiar importance of this study lies in the suggestion that previous exposure to a tone row can be a critical feature for the perception of the rules of serial music. The question remains however, as to the type of learning that actually occurred during

this previous exposure. Given that all participants had received formal instruction in music, we cannot rule out the possibility that they used their explicit knowledge of musical notation to mentally represent the structures of the two rows. In order to define the nature (implicit/explicit) of the knowledge in learning serial music rules, Bigand, D'Adamo and Poulin (2003) tested the ability of musically untrained and trained listeners to internalize serial music rules with 80 canons, especially designed by a professional composer. A set of 40 pieces were varied instantiations (transpositions) of an first tone row (grammatical pieces). The other set of 40 pieces derived from another row (nongrammatical pieces) but they were matched to the previous ones according to all superficial features (rhythm, pitch ranges, overall form of melodic contour, duration, dynamics). In a learning phase, half of these canons were presented two times to participants, who had simply to indicate whether a given piece was heard for the first or for the second time. In a test phase, 20 pairs of canons were played to the participants. In Experiment 1, each pair contained a new canon composed from the same dodecaphonic series and a matched foil (Figure 3). Matched foils had the same musical surface (i.e., same pitch range, melodic contour and rhythm), but derived from another dodecaphonic series. As a consequence, foils sounded very much like the canon to which they were matched. The participants' task was to indicate which canon of the pair was composed in the same fashion as those listened during the learning phase of the study. All participants reported extreme difficulties in performing the grammatical task. Numerous participants complained that it was difficult to differentiate the two pieces of the pairs. Both experimental groups nevertheless performed above chance, with 61% correct response for nonmusicians and 62% of correct response for musicians, with no significant difference between the two groups. In a second experiment (run with musically untrained listeners only), the stimuli of the familiarization phase were identical to those of Experiments 1, whereas the stimuli of the test phase were pairs in which one of the pieces was derived from a retrograde-inversion of the tested row. The striking finding was that musically untrained listeners continued to discriminate canons from foils above chance level (60% of correct responses), suggesting that even musically untrained listeners are able to internalize through passive exposure complex regularities deriving from the serial compositional rules. This conclusion is consistent with other finding showing that the structures of Western contemporary music are processed in similar way by both groups of listeners. Although a short exposition phase, listeners were sensitive to the structure of contemporary twelve-tone music, which is based on frequency distributions of tone intervals. These results shed some light on the implicit versus explicit nature of the acquired knowledge, and the content of the information internalized through hearing the pieces.

Probably, the knowledge internalized during the listening of serial musical pieces was

inaccessible to explicit thought of participants. If knowledge internalized through exposure was represented at an explicit level, then experts should be more able than non experts participants to explicitly use this knowledge. This should result in a clear-cut advantage of musical experts over musically untrained listeners. If in turn the knowledge is represented at an implicit level, no strong difference should be observed between musically expert and novice participants. The present study converges with conclusions drawn from several other studies run with Western tonal music and argues in favor of the implicit nature of the knowledge learnt.

2.2 Perspectives in musical learning: using multimedia technologies

2.2.1 How to optimize learning of Western tonal music with the help of multimedia technologies?

Explaining the theoretical core of the Western musical system is one of the most difficult tasks for music teachers, and it is generally assumed, at least in France, that this explanation should only occur at the end of the curriculum in both music conservatoire and university departments. Lerdahl's Tonal Pitch Space theory (TPST) is likely however to contribute to the development of music tools that would help music lovers as well as those at an early stage of musical study to improve their understanding of Western music. The TPST can be considered as an idealized knowledge representation of tonal hierarchy. The psychological representation of knowledge poses a certain number of problems for which different solutions have been proposed (Krumhansl, Bharucha, & Castellano, 1982; Krumhansl, Bharucha, & Kessler, 1982; Krumhansl & Kessler, 1982; Longuet-Higgins, 1978; Shepard, 1982). For all these approaches, tonal hierarchies are represented in the form of a multidimensional space in which the distances of chords from the instantiated tonic correspond to their relative hierarchical importance. The more important the chord is, the smaller the distance. Lerdahl successfully explains the way in which the TPST synthesizes various existing musicological and psychological models and suggests new solutions. In my opinion, the crucial contribution of the model is the description of a formal means of quantifying the tonal distance maintained between any two events belonging to any key, a quantification which no other approach has accomplished.

The proposed model outlines many developments to the one initially described in an earlier

series of articles (Lerdahl, 1988; Lerdahl, 1991; Lerdahl, 1996). For readers from a psychological background who may not be familiar with this type of approach, I will summarize the basic ideas[3] . According to the theory, tonal hierarchy is represented in three embedded levels. The first two (the pitch class level and chordal level) represent within-key hierarchies between tones and chords. The third level represents the distances between keys (Region level). The pitch class level (basic space) represents the relation between the 12 pitch classes. It contains five sublevels (from level a to e), corresponding to the chromatic level (level e), diatonic level (level d), triadic level (level c), fifth level (level b) and the tonic level (level a). In a given context, a tonic tone, part of a tonic chord, will be represented at all five levels. The dominant and the third tones of a tonic chord will be represented at four levels (from b to e) and three levels (from c to e) respectively. A diatonic but non-chordal tone will be represented at two levels (from d to e). A non-diatonic chord will be represented at only one level (level e). The level at which a given pitch class is represented thus reflects its importance in the tonal context. For example, in the context of a C major chord in the C major key, the tone C would be represented at all levels (from a to e), the tone G, at four levels (from b to e), the tone E, at three levels (from c to e) and the diatonic of the C major scale will be represented at two levels only (from d to e).

This representation has two implications. First it allows an understanding as to why notes which are distant in interval (C-E-G-C) can nevertheless be perceived to be as close as adjacent notes (C-D-E-F-G): though forming distant intervals, these notes are adjacent at the triadic level in the representational space (level c). Moreover, the model of musical tension bound to these forces of attraction constitutes a very promising development for psychology. The second implication concerns the computation of distances between chords. Let's return to the following example. If the C major chord was played in the context of G major, the tone F# will be represented at two levels (from d to e), while the tone F would remain at only one level (level e). This would produce one change in pitch class. The central idea of the TPST is to consider the number of changes that occurs in this basic space when the musical context is changed (as in the present example) as a way to define the pitch space distance between two musical events.

The second level of the model involves the chordal level, that is the distance between chords of a given key. The model computes the distances separating the seven diatonic chords taking into account the number of steps that separate the roots of the chords along the circle of fifths (C-G-D-A-E-B-F) and the number of changes in pitch class level created by the second chord. Let us consider the distance between the C and G major chords in the key of C major.

[3]See Pineau and Tillmann (2001) and Bigand (1993b) for an introduction in French.

The G major chord induces 4 changes in the pitch class level. The dominant tone D is now represented at 2 more levels (from b to e), the third tone B, at one supplementary level (from c to e) and the tonic tone at one supplementary level (from a to e). The number of steps that separate the two chords on the circle of fifths equals 1. As a consequence the tonal pitch space distance between these two chords in this key context equals 5. Following the same rationale, the distance in pitch space between the tonic and the subdominant chords equals 5. The distance between the tonic and the vi equals 7, as does the distance between the tonic and the mediant chord (iii). The distance between the tonic chord and the supertonic (ii) equals 8 as does the distance between the tonic and the diminished seventh chord. This model quantifies the strength of relations in harmonic progression. Accordingly the succession I-vi corresponds to a harmonic progression that is tenser than the succession I-IV.

The third level of the TPS model involves the regional level. It evaluates distances between chords of different regions by taking into account the distances between regions as well as the existence of a pivot region. The regional space of the TPST is created by combining the cycle of fifth and the parallel/relative major-minor cycle. That is to say, the shortest distance in regional space (i.e., 7) is found between a given major key (say C major) and its dominant (G) its subdominant (F), its parallel minor (C minor) and its relative minor keys (A minor). The greatest distance (30) is found between a major key and the augmented fourth key (C and F#). The tonal distance between two chords of different keys depends upon how the second chord is musically interpreted. For example the distance between a C major chord in the context of C major and a C# minor chord would equal 23 if the C# is interpreted as a vi of the E major key. The distance equals 30 if the C# is understood as the tonic chord of Db minor key. As a consequence, the distance in pitch space between two events that belong to distant keys depends upon the selected route between the two events. In most cases, the selected route is defined by the overall musical context. By default, the model computes this distance according to the principle of shortest path: "the pitch-space distance between two events is preferably calculated for the smallest value" (p. 74). The shortest path principle is psychologically plausible. It has the heuristic merit of being able to influence the analysis of time span and prolongational reduction by preferring analyses that reduce the value of these distances. The implementation of this principle in an artificial system should fairly easily lead to "intelligent" systems capable of automatic harmonic analysis.

One of the main features of the TPST for an efficient learning tool is to bridge the intuitive mental representations of the untrained with the mental representations of experts. Current developments in multimedia offer considerable opportunity to evolve the naive representation

of the inexpert in a given domain. The basic strategy consists in combining different modes of knowledge representation (e.g.,sounds, image, language, animation, action) to progressively transform the initial mental representation into a representation of the domain that fits as closely as possible with that of experts. In the present case, the use of a space to describe the inner structure of the Western tonal system considerably facilitates this transformation. The mental representation of a complex system in a two or three-dimensional space is a metaphor that is intuitively accessible even for a child, and which is common in a large variety of domains. A musical learning tool may thus consist in videotape (or animation) that illustrates how music progresses through pitch space. When listening to a musical piece, the video displays in real time every distance traveled through pitch space. After having listened several times to the piece, the journey through pitch space of the piece would be stored in memory in both visual and auditory format. After the hearing of several pieces of the same stylistic period, the journeys through pitch space specific to this style would be stored in memory. After hearing several pieces of the Western music repertoire, the listener would have created a mental representation of the overall structure of the tonal pitch space that fits with that of the expert. From a teaching perspective, the interesting point is that this mental representation will emerge from mere exposure to musical pieces presented with this music tool. In other words, the tool allows a passive exploration of the tonal pitch space by visualizing in a comprehensible format the deep harmonic structure of the heard pieces. Only a few musical terms will be required to understand one of the most critical features of Western music.

The structure of the space can be adapted at will and should notably be adjusted to suit the age of the user. At this early stage of the developmental process, we chose a structure that mimics real space with planets and satellites. Given the circularity of the Western musical space, only a portion of the space can be seen at a given time point, but this portion will progressively change when the music is moving from one region to another. A planet metaphorically represents a key, while the satellites represent the seven diatonic chords. Satellites corresponding to played chords are lit up in yellow, thus representing the route of the harmonic progressions within each key. The colour of the planet representing the key intensifies when several chords from the key are played, thus imitating the fact that a feel for tonality increases with its duration. When the music modulates to another key, chords from both the initial key and the new key light up, and the animation turns towards the new key, and then another portion of the tonal pitch space is discovered. When the piece of music progresses rapidly towards distant keys, as in the case of Chopin's Prelude in E major, the pivot keys are briefly highlighted and passed quickly. The journey depends upon the modulations that have occurred. With the present tool, the user

can associate the visual journey through tonal pitch space with the auditory sensation created by the music. The animation contains sufficient music theoretic information to allow the user to describe this musical journey in terms that are close to those employed by musicologists. Of course, this animation may also bring to the fore other important elements for the comprehension of harmonic processes, such as the arrangement of chords and voice leading. Connected to a MIDI instrument, it may equally be transformed into a tool for tonal music composition. By chaining chords together, the user can follow his or her journey through tonal space, and explore the structure of the tonal space.

2.2.2 Creating learning multimedia tools for music with the contribution of cognitive sciences and ergonomics

The first multimedia works comprising music began to emerge at the beginning of the 90th. Since then, the number and diversity of musical multimedia products (CD-Rom, DVD-Rom, Web site) have been increasing considerably. However, multimedia products helping the user to integrate musical structures are rare. If we want to propose successful learning multimedia tools the main question is how and why to use multimedia resources. We are going to present some principles of cognitive ergonomics that seem fundamental for the multimedia dedicated to music education. These principles will be illustrated by two multimedia learning tools invented by the LEAD.

The first principle is that the vantage point of music learning tools should be nonexperts' immediate representation formats. They have to combine in an advantageous way the multimodal possibilities of representation to make these initial representations evolve towards those of experts (principle of availability). Multimodality should be used as a powerful means to clarify the structure of complex systems, and to allow the user to easily develop a mental representation of the system compatible with the one of experts. The aim of the project carried out by the LEAD, was to give listeners the access to a musical system often considered as complex (contemporary music) but potentially of high educational value. It was an ideal opportunity to try out a multimedia approach of a complex system.

Reduction of information and optimisation of presentation forms

One of the principal problems regarding multimedia is an overload of presentation forms. Multiplication of presentation forms (text, picture, animation, video, sound, etc.) often entails a

cognitive cost that is high compared to the benefits in terms of training. This profusion of presentation forms often led to an explosion of the quantity of information presented to the user, without an objective analysis of the adaptation of presentation forms used or of the combination of the modes of learning (visual & auditory, verbal & visual, etc.) available in the tools proposed. Such information overload is often accompanied by an organization of knowledge based on models that are not adapted to the initial knowledge of a user. The second fundamental principle in producing multimedia learning tools is thus the reduction of the quantity of information and the optimization of the form in which it is presented. Properly used multimodality, particularly concerning the interaction between vision and audition, improves attending processes, memorization of musical material, and develops the capacity to represent the musical structures. In music, there are several forms of presentation of the sound phenomenon. The score is the best known of them. However, it requires user's specific knowledge and a regular musical practice. There are other forms of music presentation: tablature of string instruments, sonagram (spectrum), wave form (amplitude), tracks of a music software or piano-roll of a sequencer, etc. All these presentation modes can certainly be employed in multimedia, but they often require user's expert knowledge.

For the project carried out by the LEAD, we sought graphic representations that could advantageously replace the experts presentation forms. These graphics consist of simple forms symbolizing one or more elements of musical structure (melody contour, texture, harmonic density, rhythmic pattern, etc). The principal constraint is that these forms should not require additional coding, otherwise they would go against the aims in view, but induce the musical structure in an intuitive and direct way. Other presentation forms, which require expert knowledge, never intervene in the initial presentation of a musical excerpt. Figures 5 and 6 show two forms of presentation of a sequence of chords in the piece Couleur de la Cité Céleste by Oliver Messiaen. The constitution in terms of tones is identical for the 13 chords. It is the register, duration and the change in instrumentation between the various instruments which give listener an impression of a succession of sound colours (klangfarbenmelodie). The excerpt is represented by blocks of colours whose width corresponds to the duration of chords. Their height symbolizes the extent of chords (from the lowest to the highest) and the position on the scale (on the left) represents the register. Blocks appear in synchrony with sound. With this type of representation, it is easy, for any non expert listener, to perceive a degree of similarity between certain chords. Based on this type of representation, a user may intuitively become aware of the external structure of a sequence, but it is not sufficient yet to form a precise representation of the musical structure. Figure 6 represents the same sequence of chords in a score. In order to better focus listener's

attention on the harmonic structure, the real duration of chords was replaced by the duration equalized for all chords. However, in contrast with a graphic representation of sound that was privileged here, this mode of representation is to give the users an opportunity to deconstruct musical structure. They can choose what they will listen to: the whole sequence, each chord separately, groups of instruments within a chord or each note of a chord.

Synthesis of knowledge and implementation of continuity

Learning multimedia tools should synthesize the knowledge of music in order to make it available to nonexperts. Thus it is necessary to implement this knowledge in a way adapted to the initial knowledge of the user. In the case of music (complex music in particular), it is important to raise the question of perceptibility of musical structures. It is a question of knowing exactly what is to be heard. In our project, the pieces were selected according to the cognitive problems they present (relating to their aesthetic differences). For example, Couleur de la Cité Céleste by Messiaen, is representative of the aesthetics where colour and timbre are of major concern. It is composed of a large variety of musical elements that follow one another to form a sound mosaic. Globally, a multimedia learning tool must favour categorization and memorization of musical material, in order to allow emergence of the mental representation of temporal organization of a piece. Figure 7 shows the main page of the multimedia learning tool of Messiaen's piece. One can see the representation of the formal structure of the excerpt in the centre of the screen. Eight icons on the right and on the left of the screen give access to eight links (history of the piece, composer's biography, the orchestra, a large-scale structure of the excerpt, and four of the main materials of the piece: texts and metaphors of the Apocalypse, Gregorian chant, colours, and bird songs).

One of the crucial problems posed to the pedagogy of listening is that of attention. Perception of musical structures is strongly dependent on the attending processes. In the simplest cases, these processes are guided by the music itself (when, for example, a composer emphasizes the principal melody by a discrete accompaniment) or by the performance of a player (when this one chooses to emphasize such element of structure). However, most of the time, music has a complex and deliberately ambiguous structure. Contrary to the traditional methods in music education, that exert these capacities with sorrow, multimedia tools make it possible to focus listener's attention on internal or external elements of musical structure.

A part of our project consisted in seeking in the resources of multimedia the means of

guiding attending processes and, beyond, of favouring the memorization and the comprehension of the musical structures. The schema of the formal structure of the beginning of Couleur de la Cité Céleste (Figure 7) was conceived to facilitate mental representation of a complex formal structure in which short musical sequences follow one another in form of a mosaic that would emerge while we perceive it. This page does not contain any oral or textual explanation. Awareness of the structure should emerge solely from the visual and auditive interaction. The choice of a circular representation corresponds to the shape of a work which has no beginning or end and whose elements return in a recurrent way. Each piece of the mosaic corresponds to a short musical sequence. Each colour represents a type of musical material (e.g. blue for bird songs). Nuances of colours differentiate the variations inside each category of sequence. Animation takes into account the cognitive processes of attention and memorization. At the beginning of animation, the stained-glass scheme is empty. Progressively, as music unfolds, empty spaces are filled until the stained glass is completed (all the sequences were played). When a sequence is finished, stained glass that represents it is obscured gradually (approximately 6 to 7 seconds, according to the maximum duration of the perceptual present; Fraisse, 1957). The clearness of the piece of the stained glass solidifies at a very low rate as if there remained a remote trace in memory about it. When an identical sequence returns, the part previously activated is briefly reactivated then turns over in stand-by. This multimedia artifice supports the categorization and the memorization of materials. It also makes it possible to establish bonds of similarity and consequently gives direction to the formal structure which proceeds under the eyes of the user. This example showed how it is possible to focus the attention on sequential events. It is also useful to focus the attention of the user on simultaneous events. Figure 8 shows the animated representation of the formal structure of the beginning of Eight Lines by S. Reich. Contrary to the Messiaen's piece represented by a stained glass, the one by Reich, whose unfolding follows a linear trajectory, is represented by 8 rectangular boxes. Coloured rectangular paving stones indicate the moment of appearance and the duration of intervention of each instrumental part. Synchronization between the sound and the image is visualized by a vertical marker. When an instrument is active, its coloured paving stone is cleared up. In figure 8, the active instrumental parts are the parts of viola and violoncello (in bottom), low clarinet and flute (in top). Inside the paving stone, graphic animations, always synchronized with music, emerge to focus the attention on melody and the rhythmic structure of the instrumental part. The points indicate the number, the duration and the height of the notes, the features profile melody contour.

New technologies of sound processing may provide other new possibilities for multimedia music learning tools . The sound files obtained with these techniques or especially developed

software can be integrated into the multimedia in order to improve the learning tools. The possibility of an interactive deconstruction and reconstruction of the musical structures combined to specific visual interface is certainly the most promising perspective for the closest future.

Conclusion

The power of the implicit learning is without doubt a major contribution to the research on the musical cognition. It reduces the distance between listener nonmusicians and experts and leads to question the common practices in music education. Implicit learning supplies a solid scientific base, with the contributions of cognitive psychology (memory and attending process), ergonomics and new technologies in order to create creating multimedia learning tools for music.

References

Altmann, G. T., Dienes, Z., & Goode, A. (1995). Modality independence of implicitly learned grammatical knowledge. Journal of Experimental Psychology: Learning, Memory and Cognition, 21, 899-912.

Ayari, M., & McAdams, S. (2003). Aural analysis of Arabic improvised instrumental music (tagsim). Music Perception, 21, 159-216.

Bharucha, J. J. (1987). Music cognition and perceptual facilitation: A connectionist framework. Music Perception, 5(1), 1-30.

Bigand, E., Madurell, F., Tillmann, B., & Pineau, M. (1999). Effect of global structure and temporal organization on chord processing. Journal of Experimental Psychology: Human Perception and Performance, 25(1), 184-197.

Bigand, D'Adamo et al., (2003) = paper of serial music implicit learning

Bigand, E., Perruchet, P., & Boyer, M. (1998). Implicit learning of an artificial grammar of musical timbres. Cahiers de Psychologie Cognitive/Current Psychology of Cognition, 17(3), 577-600.

Budge, H. (1943). A study of chord frequencies. Teacher College.

Dienes, Z., Broadbent, D., & Berry, D. C. (1991). Implicit and explicit knowledge bases

in artificial grammar learning. Journal of Experimental Psychology: Learning, Memory and Cognition, 17, 875-887.

Dowling, W. J., & Harwood, D. L. (1986). Music Cognition. Orlando, Florida: Academic Press.

Francès, R. (1958). La perception de la musique (2¡ ed.). Paris: Vrin.

Grossberg, S. (1970). Some networks that can learn, remember and reproduce any number of complicated space-time patterns. Studies in Applied Mathematics, 49, 135-166.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. Biological Cybernetics, 23, 121-134.

Howard, J. H. J., & Ballas, J. A. (1980). Syntactic and semantic factors in the classification of nonspeech transient patterns. Perception & Psychophysics, 28(5), 431-439.

Howard, J. H. J., & Ballas, J. A. (1982). Acquisition of acoustic pattern categories by exemplar observation. Organization, Behavior and Human Performance, 30, 157-173.

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. Journal of Memory and Language, 44(4), 548-567.

Kohonen, T. (1995). Self-Organizing Maps. Berlin: Springer.

Krumhansl, C. L. (1990a). Cognitive foundations of musical pitch. New York: Oxford University Press.

Krumhansl, C. L., Bharucha, J., & Castellano, M. A. (1982). Key distance effects on perceived harmonic structure in music. Percept Psychophys, 32(2), 96-108.

Krumhansl, C. L., Bharucha, J. J., & Kessler, E. J. (1982). Perceived harmonic structures of chords in three related keys. Journal of Experimental Psychology: Human Perception and Performance, 8, 24-36.

Krumhansl, C. L., & Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. Psychol Rev, 89(4), 334-368.

Krumhansl, C. L. (1990). Cognitive foundations of musical pitch. Oxford, England: Oxford University Press. Lerdahl, F. (1988). Tonal pitch space. Music Perception, 5(3), 315-349.

Lerdahl, F. (1989). Structure de prolongation dans l'atonalite. In S. McAdams & I. Deliege (Eds.), La musique et les sciences cognitives (pp. 171-179). Liège: Mardaga.

Lerdahl, F. (1991). Pitch-space journeys in two Chopin Preludes. In M. R. Jones & S.

Longuet-Higgins, H. (1978). The perception of music. Interdisciplinary Science Review, 3, 148-156.

Reber, A. S. (1967). Implicit learning of artificial grammars. Journal of Verbal Learning and Verbal Behavior, 6, 855-863.

Reber, A. S. (1989). Implicit learning and tacit knowledge. Journal of Experimental Psychology: General, 118, 219-235.

Rumelhart, D. E., & Zipser, D. (1985). Feature discovery by competitive learning. Cognitive Science, 9, 75-112.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation : The role of distributional cues. Journal of Memory and Language, 35(4), 606-621.

Seger, C. A. (1994). Implicit learning. Psychological Bulletin, 115, 163-169.

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. Developmental Psychology, 39(4), 706-716.

Tillmann, B., & McAdams, S. (2004). Implicit Learning of musical timbre sequences : statistical regularities confronted with acoustical (dis)similarities. Journal of Experimental Psychology: Learning, Memory & Cognition, 30, 1131-1142.

Tillmann, B., Bharucha, J. J., & Bigand, E. (2000). Implicit learning of tonality: a self-organizing approach. Psychological Review, 107(4), 885-913.

von der Malsberg, C. (1973). Self-organizing of orientation sensitive cells in the striate cortex. Kybernetic, 14, 85-100.

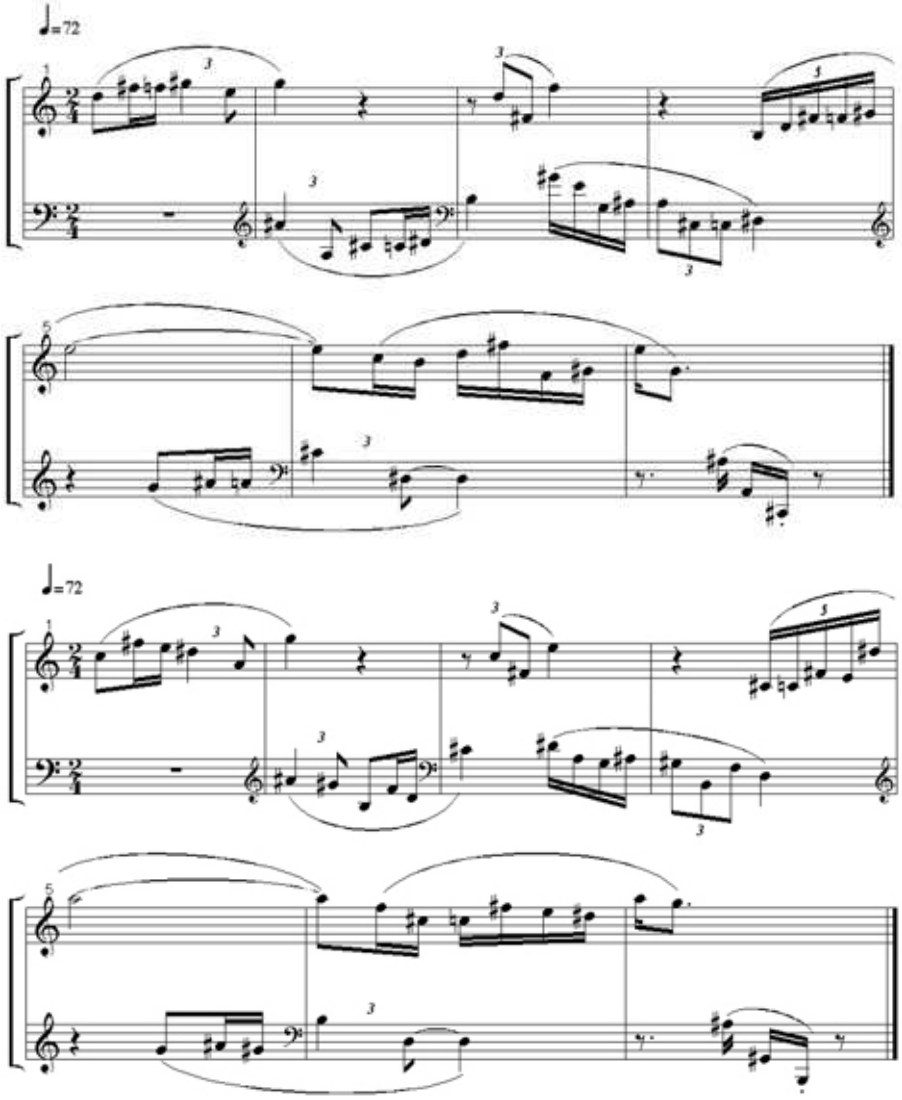

Figure 2.3: Example of pairs of matched canons composed with two different rows (grammatical piece in the higher panel, nongrammatical piece in the lower panel), but both with the same superficial features (rhythm, pitch ranges, overall form of melodic contour, duration, dynamics).

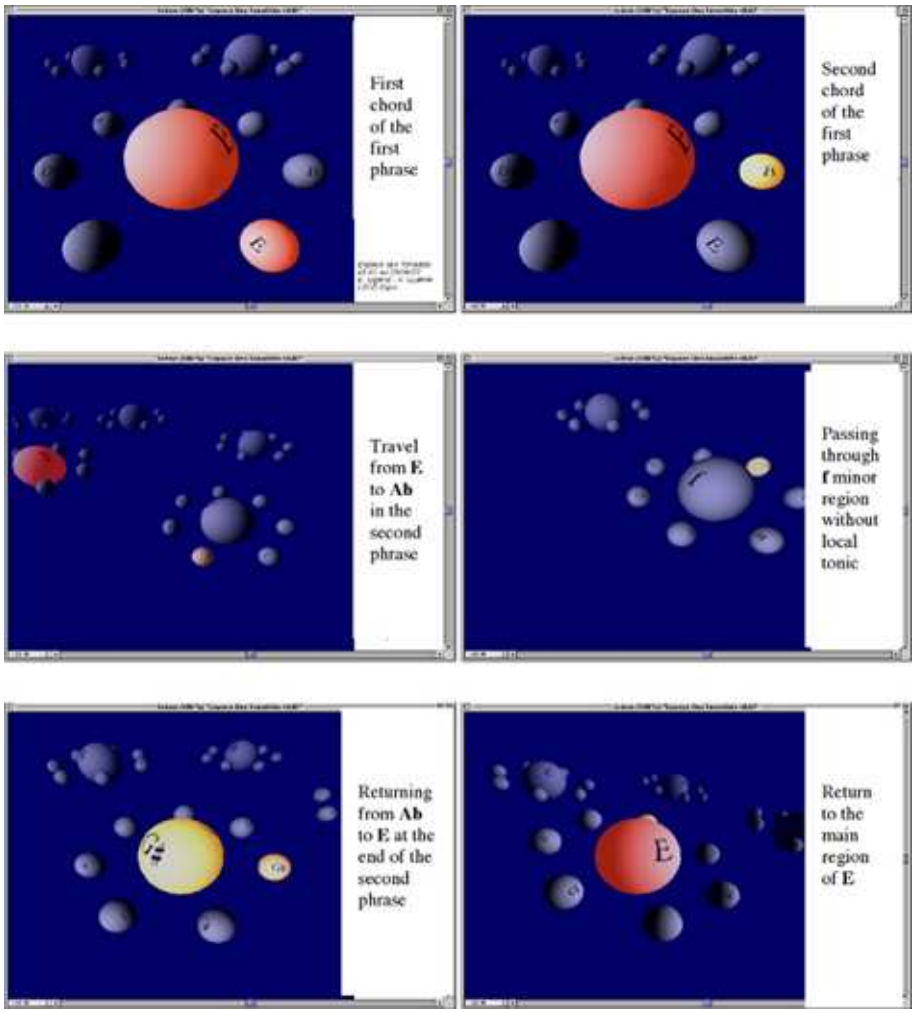

Figure 2.4: Figure 4 illustrates a musical tool directly derived from TPST and which is currently being developed in our team.

[4].

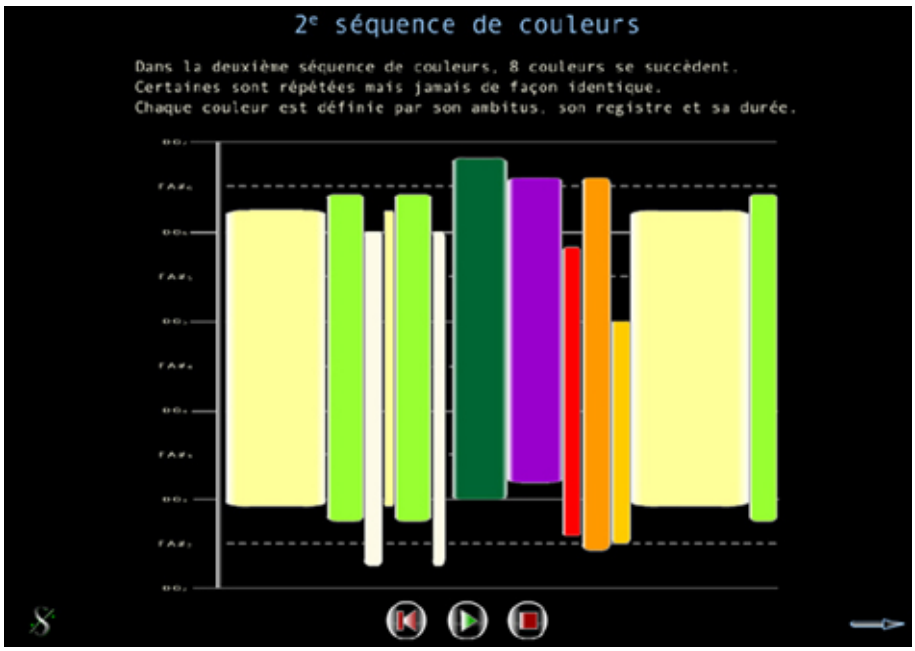

Figure 2.5: Graphic representation of a chord sequence of Couleurs de la Cité céleste by Olivier Messiaen

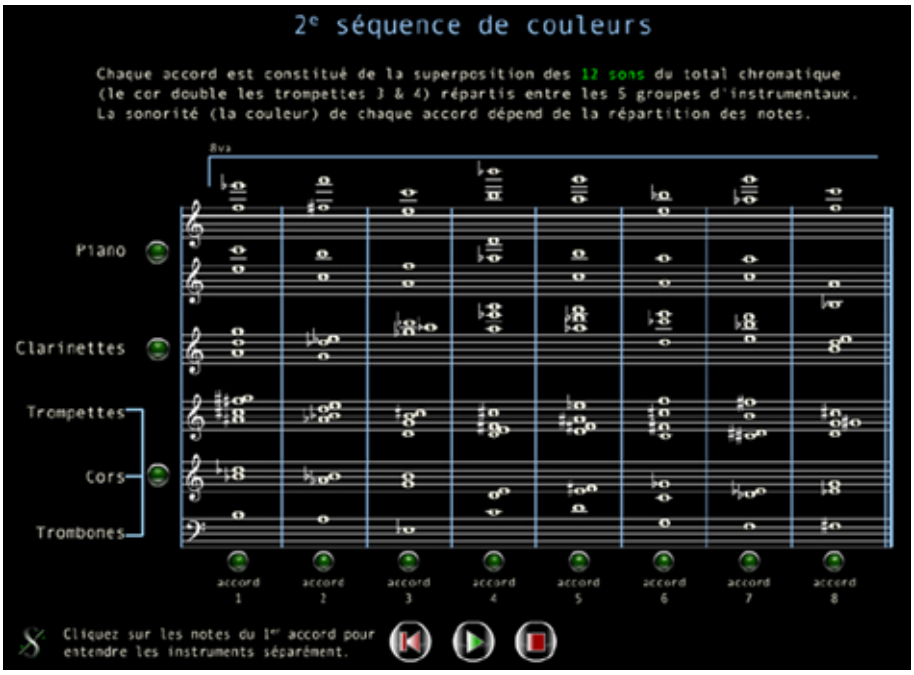

Figure 2.6: Score representation of the same chord sequence as 2.5 of Couleurs de la Cité céleste by Olivier Messiaen

Figure 2.7: Main page of the multimedia learning tool of Couleurs de la Cité céleste by Olivier Messiaen

Figure 2.8: Formal structure represented in the multimedia learning tool of Eight Lines by S. Reich

Chapter 3

From Sound to "Sense" via Feature Extraction and Machine Learning: Deriving High-Level Descriptors for Characterising Music

Gerhard Widmer[1,2], Simon Dixon[1], Peter Knees[2], Elias Pampalk[1], Tim Pohle[1]

[1] Austrian Research Institute for Artificial Intelligence, Vienna, Austria
[2] Department of Computational Perception, Johannes Kepler University Linz, Austria

3.1 Introduction

Research in intelligent music processing is experiencing an enormous boost these days due to the emergence of the new application and research field of *Music Information Retrieval (MIR)*. The rapid growth of digital music collections and the concomitant shift of the music market towards digital music distribution urgently call for intelligent computational support in the automated handling of large amounts of digital music. Ideas for a large variety of content-based music services are currently being developed in music industry and in the research community. They range from content-based music search engines to automatic music recommendation services,

from intuitive interfaces on portable music players to methods for the automatic structuring and visualisation of large digital music collections, and from personalised radio stations to tools that permit the listener to actively modify and 'play with' the music as it is being played.

What all of these content-based services have in common is that they require the computer to be able to 'make sense of' and 'understand' the actual content of the music, in the sense of being able to recognise and extract musically, perceptually and contextually meaningful ('semantic') patterns from recordings, and to associate descriptors with the music that make sense to human listeners.

There is a large variety of musical descriptors that are potentially of interest. They range from low-level features of the sound, such as its bass content or its harmonic richness, to high-level concepts such as "hip hop" or "sad music". Also, semantic descriptors may come in the form of atomic, discrete labels like "rhythmic" or "waltz", or they may be complex, structured entities such as harmony and rhythmic structure. As it is impossible to cover all of these in one coherent chapter, we will have to limit ourselves to a particular class of semantic desciptors.

This chapter, then, focuses on methods for automatically extracting high-level atomic descriptors for the characterisation of music. It will be shown how high-level terms can be inferred via a combination of bottom-up audio descriptor extraction and the application of *machine learning* algorithms. Also, it will be shown that meaningful descriptors can be extracted not just from an analysis of the music (audio) itself, but also from extra-musical sources, such as the internet (via 'web mining').

Systems that learn to assign labels must be evaluated in systematic, controlled experiments. The most obvious and direct way is via *classification experiments*, where the labels to be assigned are interpreted as distinct classes. In particular, *genre classification*, i.e., the automatic assignment of an appropriate style label to a piece of music, has become a popular benchmark task in the MIR community (for many reasons, not the least of them being the fact that genre labels are generally much easier to obtain than other, more intuitive or personal descriptors). Accordingly, the current chapter will very much focus on genre classification as the kind of benchmark problem that measures the efficacy of machine learning (and the underlying descriptors) in assigning meaningful terms to music. However, in principle, one can try to predict any other high-level labels from low-level features, as long as there is a sufficient number of training examples with given labels. Some experiments regarding non-genre concepts will be briefly described in section 3.3.4, and in section 3.4.2 we will show how textual characterisations of music artists can be automatically derived from the Web.

The chapter is structured as follows. Section 3.2 deals with the extraction of music descriptors (both very basic ones like timbre and more abstract ones like melody or rhythm) from recordings via audio analysis. It focuses in particular on features that have been used in recent genre classification research. Section 3.3 shows how the gap between what can be extracted bottom-up and more abstract, human-centered concepts can be partly closed with the help of inductive machine learning. New approaches to infer additional high-level knowledge about music from extra-musical sources (the Internet) are presented in section 3.4. Section 3.5, finally, discusses current research and application perspectives and identifies important questions that will have to be addressed in the future.

3.2 Bottom-up Extraction of Descriptors from Audio

Extracting descriptors from audio recordings to characterise aspects of the audio content is not a new area of research. Much effort has been spent on feature extraction in areas like speech processing or audio signal analysis. It is impossible to give a comprehensive overview of all the audio descriptors developed over the past decades. Instead, this chapter will focus solely on descriptors that are useful for, or have been evaluated in, music classification tasks, in the context of newer work in Music Information Retrieval. The real focus of this chapter is on extracting or predicting higher-level descriptors via machine learning. Besides, a more in-depth presentation of audio and music descriptors is offered in another chapter of this book [REF. TO UPF CHAPTER], so the following sections only briefly recapitulate those audio features that have played a major role in recent music classification work.

Connected to the concept of classification is the notion of music or generally sound *similarity*. Obviously, operational similarity metrics can be used directly for audio and music classification (e.g., via nearest-neighbour algorithms), but also for a wide variety of other tasks. In fact, some of the music description schemes presented in the following do not produce features or descriptors at all, but directly compute similarities; they will also be mentioned, where appropriate.

3.2.1 Simple Audio Descriptors for Music Classification

This section describes some common simple approaches to describe properties of audio (music) signals. For all algorithms discussed here, the continuous stream of audio information is cut into small, possibly overlapping fragments of equal length, called *frames*. The typical length of a frame is about 20 ms. Usually, for each frame one scalar value per descriptor is calculated, which can be done either on the time-domain or the frequency-domain representation of the signal. To obtain a (scalar) descriptor that pertains to an entire audio track, the values of all frames can be combined by, for example, applying simple statistics such as mean and standard deviation of all individual values.

Time-Domain Descriptors

On the time-domain representation of the audio signal, several descriptors can be calculated. An algorithm that mainly describes the power envelope of the audio signal is *Root Mean Square (RMS)*: The individual values appearing in each frame are squared, and the root of the mean of these values is calculated. These values might be combined as described above, or by calculating which fraction of all RMS values is below (e.g.) the average RMS value of a piece (*Low Energy Rate*). Comparable to the RMS values are the *Amplitude Envelope* values, which are the maximum absolute values of each frame. The amplitude envelope and RMS descriptors are commonly used as a first step in algorithms that detect rhythmic structure.

The time-domain representation might also be used to construct measures that model the concept of *Loudness* (i.e. the perceived "volume"). For example, a simple and effective way is to take the 0.23th power of the RMS values.

Another possibility is to approximately measure the perceived *brightness* with the *Zero Crossing Rate*. This descriptor simply counts how often the signal passes zero-level.

Also, the time-domain representation can be used to extract periodicity information from it. Common methods are autocorrelation and comb filterbanks. Autocorrelation gives for each given time lag the amount of self-similarity of the time domain samples by multiplying the signal with a time-lagged version of itself. In the comb filterbank approach, for each periodicity of interest, there is a comb filter with the appropriate resonance frequency.

Frequency-Domain Descriptors

A number of simple measures are commonly applied to describe properties of the frequency distribution of a frame:

- The *Band Energy Ratio* is the relation between the energy in the low frequency bands and the energy of the high frequency bands. This descriptor is vulnerable to producing unexpectedly high values when the energy in the low energy bands is close to zero.
- The *Spectral Centroid* is the center of gravity of the frequency distribution. Like the zero crossing rate, it can be regarded as a measure of perceived brightness or sharpness.
- The *Spectral Rolloff* frequency is the frequency below which a certain amount (e.g. 95%) of the frequency power distribution is concentrated.

These descriptors are calculated individually for each frame. The *Spectral Flux* is modeled to describe the temporal change of the spectrum. It is the Euclidean distance between the (normalised) frequency distributions of two consecutive frames, and can be regarded as a measure of the rate at which the spectrum changes locally.

The descriptors mentioned so far represent rather simple concepts. A more sophisticated approach are the *Mel Frequency Cepstral Coefficients (MFCCs)*, which model the shape of the spectrum in a compressed form. They are calculated by representing the spectrum on the perceptually motivated Mel-Scale, and taking the logarithms of the amplitudes to simulate loudness perception. Afterwards, the discrete cosine transformation is applied, which results in a number of coefficients (MFCCs). Lower coefficients describe the coarse envelope of the frame's spectrum, and higher coefficients describe more detailed properties of the spectrum envelope. Usually, the higher-order MFCCs are discarded, and only the lower MFCCs are used to describe the music.

A popular way to compare two recorded pieces of music using MFCCs is to discard the temporal order of the frames, and to summarise them by clustering (e.g., Logan and Salomon [2001], Aucouturier and Pachet [2002a]). In the case of Aucouturier and Pachet [2002a], for instance, the clustered MFCC representations of the frames are described by Gaussian Mixture Models (GMMs), which are the features for the piece of music. A way to compare GMMs is sampling: one GMM is used to produce random points with the distribution of this GMM, and the likelihood that the other GMM produces these points is checked.

It might seem that discarding the temporal order information altogether ignores highly important information. But recent research Flexer et al. [2005] has shown that MFCC-based description models using Hidden Markov Models (which explicitly model the temporal structure of the data) do not improve classification accuracy (as already noted in Aucouturier and Pachet [2004]), though they do seem to better capture details of the sound of musical recordings (at least in terms of statistical likelihoods). Whether this really makes a difference in actual applications remains still to be shown.

The interested reader is referred to Chapter XXX/UPF of this book for a much more comprehensive review of audio descriptors and music description schemes.

3.2.2 Extracting Higher-level Musical Patterns

The basic intuition behind research on classification by higher-level descriptors is that many musical categories can be defined in terms of high-level *musical* concepts. To some extent it is possible to define musical genre, for example, in terms of the melody, rhythm, harmony and instrumentation which are typical of each genre. Thus genre classification can be reduced to a set of subproblems: recognising particular types of melodies, rhythms, harmonies and instruments. Each of these subproblems is interesting in itself, and has attracted considerable research interest, which we review here.

Early work on music signal analysis is reviewed by Roads Roads [1996]. The problems that received the most attention were pitch detection, rhythm recognition and spectral analysis, corresponding respectively to the most important features of music: melody, rhythm and timbre (harmony and instrumentation).

Pitch detection is the estimation of the fundamental frequency of a signal, usually assuming it to be monophonic. Common methods include: time domain algorithms such as counting of zero-crossings and autocorrelation; frequency domain methods such as Fourier analysis and the phase vocoder; and auditory models which combine time and frequency domain information based on an understanding of human auditory processing. Recent work extends these methods to find the predominant pitch (usually the melody note) in polyphonic mixtures Goto and Hayamizu [1999], Gómez et al. [2003].

The problem of extracting *rhythmic content* from a musical performance, and in particular finding the rate and temporal location of musical beats, has attracted considerable interest. A

review of this work is found in Gouyon and Dixon [2005]. Initial attempts focussed on rhythmic parsing of musical scores, that is without the tempo and timing variations that characterise performed music, but recent tempo and beat tracking systems work quite successfully on a wide range of performed music. The use of rhythm for classification of dance music was explored in Dixon et al. [2003, 2004].

Spectral analysis examines the time-frequency content of a signal, which is essential for extracting information about *instruments* and *harmony*. Short time Fourier analysis is the most widely used technique, but many others are available for analysing specific types of signals, most of which are built upon the Fourier transform. MFCCs, already mentioned in section 3.2.1 above, model the spectral contour rather than examining spectral content in detail, and thus can be seen as implicitly capturing the instruments playing (rather than the notes that were played). Specific work on instrument identification can be found in Herrera et al. [2003].

Regarding *harmony*, extensive research has been performed on the extraction of multiple simulateous notes in the context of automatic transcription systems, which are reviewed by Klapuri Klapuri [2004]. Transcription typically involves the follow steps: producing a time-frequency representation of the signal, finding peaks in the frequency dimension, tracking these peaks over the time dimension to produce a set of partials, and combining the partials to produce a set of notes. The differences between systems are usually related to the assumptions made about the input signal (for example the number of simultaneous notes, types of instruments, fastest notes, or musical style), and the means of decision making (for example using heuristics, neural nets or probabilistic reasoning).

Despite considerable successes, the research described above makes it increasingly clear that precise, correct, and general solutions to problems like automatic rhythm identification or harmonic structure analysis are not to be expected in the near future — the problems are simply too hard and would require the computer to possess the kind of broad musical experience and 'knowledge' that human listeners seem to apply so effortlessly when listening to music. Recent work in the field of Music Information Retrieval has thus started to focus more on *approximate* solutions to problems like melody extraction Eggink and Brown [2004] or chord transcription Yoshioka et al. [2004], or on more *specialised* problems, like the estimation of global tempo Alonso et al. [2004] or tonality Gómez and Herrera [2004], or the identification of drum patterns Yoshii et al. [2004].

Each of these areas provides a limited high level musical description of an audio signal. Systems have yet to be defined which combine all of these aspects, but this is likely to be seen in

the near future.

3.3 Closing the Gap: Prediction of High-level Descriptors via Machine Learning

While the bottom-up extraction of features and patterns from audio continues to be a very active research area, it is also clear that there are strict limits as to the kinds of music descriptions that can be directly extracted from the audio signal. When it comes to intuitive, human-centered characterisations such as 'peaceful' or 'aggressive music' or highly personal categorisations such as 'music I like to listen to while working', there is little hope of analytically defining audio features that unequivocally and universally define these concepts. Yet such concepts play a central role in the way people organise and interact with and 'use' their music.

That is where *automatic learning* comes in. The only way one can hope to build a machine that can associate such high-level concepts with music items is by having the machine learn the correct associations between low-level audio features and high-level concepts, from examples of music items that have been labeled with the appropriate concepts. In this section, we give a very brief introduction to the basic concepts of *machine learning* and *pattern classification*, and review some typical results with machine learning algorithms in musical classification tasks. In particular, the automatic labeling of music pieces with *genres* has received a lot of interest lately, and section 3.3.3 focuses specifically on genre classification. Section 3.3.4 then reports on recent experiments with more subjective concepts, which clearly show that a lot of improvement is still needed. One possible avenue towards achieving this improvement will then be discussed in section 3.4.

3.3.1 Classification via Machine Learning

Inductive learning as the automatic construction of classifiers from pre-classified training examples has a long tradition in several sub-fields of computer science. The field of *statistical pattern classification* Duda et al. [2001], Hastie et al. [2001] has developed a multitude of methods for deriving classifiers from examples, where a 'classifier', for the purposes of this chapter, can be regarded as a black box that takes as input a new object to be classified (described via a set of features) and outputs a prediction regarding the most likely class the object belongs to. Classifiers

are automatically constructed via *learning algorithms* that take as input a set of example objects labeled with the correct class, and construct a classifier from these that is (more or less) consistent with the given training examples, but also makes predictions on new, unseen objects — that is, the classifier is a *generalisation* of the training examples.

In the context of this chapter, training examples would be music items (e.g., songs) characterised by a list of audio features and labeled with the appropriate high-level concept (e.g., "this is a piece I like to listen to while working"), and the task of the learning algorithm is to produce a classifier that can predict the appropriate high-level concept for new songs (again represented by their audio features).

Common training and classification algorithms in statistical pattern classification Duda et al. [2001] include nearest neighbour classifiers (k-NN), Gaussian Mixture Models, neural networks (mostly multi-layer feed-forward perceptrons), and support vector machines Cristianini and Shawe-Taylo [2000].

The field of Machine Learning Mitchell [1997] is particularly concerned with algorithms that induce classifiers that are *interpretable*, i.e., that explicitly describe the criteria that are associated with or define a given class. Typical examples of machine learning algorithms that are also used in music classification are decision trees Quinlan [1986] and rule learning algorithms Fürnkranz [1999].

Learned classifiers must be evaluated empirically, in order to assess the kind of prediction accuracy that may be expected on new, unseen cases. This is essentially done by testing the classifier on new (labeled) examples which have not been used in any way in learning, and recording the rate of prediction errors made by the classifier. There is a multitude of procedures for doing this, and a lot of scientific literature on advantages and shortcomings of the various methods. The basic idea is to set aside a part of the available examples for testing (the 'test set'), then inducing the classifier from the remaining data (the 'training set'), and then testing the classifier on the test set. A systematic method most commonly used is known as *n-fold cross-validation*, where the available data set is randomly split into n subsets ('folds'), and the above procedure is carried out n times, each time using one of the n folds for testing, and the remaining $n-1$ folds for training. The error (or conversely, accuracy) rates reported in most learning papers are based on experiments of this type.

A central issue that deserves some discussion is the *training data* required for learning. Attractive as the machine learning approach may be, it does require (large) collections of rep-

resentative labeled training examples, e.g., music recordings with the correct categorisation attached. Manually labeling music examples is a very laborious and time-consuming process, especially when it involves listening to the pieces before deciding on the category. Additionally, there is the copyright issue. Ideally, the research community would like to be able to share common training corpora. If a researcher wants to test her own features in classification experiment, she needs access to the actual audio files.

There are some efforts currently being undertaken in the Music Information Retrieval community to compile large repositories of labeled music that can be made available to all interested researchers without copyright problems. Noteworthy examples of this are Masataka Goto's RWC Music Database (http://staff.aist.go.jp/m.goto/RWC-MDB), the IMIRSEL (International Music Information Retrieval System Evaluation Laboratory) project at the University of Illinois at Urbana-Champaign (http://www.music-ir.org/evaluation — see also Downie et al. [2004]), and the new FreeSound Initiative (http://freesound.iua.upf.edu).

3.3.2 Learning Algorithms Commonly Used in Music Classification

In this section, we briefly review some of the most common learning algorithms that are used in music classification and learning tasks.

Decision trees Quinlan [1986] are probably the most popular class of classification models in machine learning, and they are widely used also in Music Information Retrieval. In West and Cox [2004], for instance, decision tree learning algorithms have been used to build a model of the distribution of frame values.

Because of its known merits, k-NN classification is widely used. Sometimes, the feature values – possibly after feature selection – of each piece are regarded as a vector, and the distance used for k-NN classifier is the euclidean distance between individual pieces (e.g. Costa et al. [2004], Gouyon et al. [2004]) or to representative reference vectors (e.g. Hellmuth et al. [2004], Kastner et al. [2004]).

Support Vector Machines (SVMs) are also applied to music classification: e.g. Xu et al. [2003] use them for genre classification, and Li and Ogihara [2003] train several SVMs to recognise mood labels, where each SVM decides if one specific label is present in the music.

Gaussian Mixture Models (GMMs) are useful for estimating the distribution of feature values. They can be used as a classifier by modeling each class as a GMM; an instance is then

classified by calculating, for each class (GMM), the likelihood that the instance was produced by the respective GMM, and predicting the class with the maximum likelihood. In Liu et al. [2003], mood detection in classical music is done based on this approach. GMM classifiers have also been used in Burred and Lerch [2003], Tzanetakis and Cook [2002] for genre classification.

Neural Networks have also been applied to music classification: Costa et al. [2004] use a multilayer perceptron to determine the class of a piece given its feature vector. Hellmuth et al. [2004] use a more elaborate approach by training a separate neural network for each class, and an additional one that combines the outputs of these networks.

3.3.3 Genre Classification: Typical Experimental Results

The experimental results found in the literature on genre classification are not easy to compare, as researchers use many different music collections to evaluate their methods. Also, the ways of annotating the collections vary: some researchers label the pieces according to their own judgment, while others use online databases for the assignment of genre labels. Additionally, different authors often tackle slightly different problems (such as categorical vs. probabilistic classification), which makes a comparison of the results even more difficult. These facts should be kept in mind when assessing the examples given in this section.

Generally, when trying to separate the classes Pop and Classical, very high accuracies are reached, suggesting that this task is not too difficult. E.g., Costa et al. [2004] achieve up to 90.3% classification accuracy, and Mierswa and Morik [2005] report even 100% on 200 pieces. In both cases, the baseline is one half. Although Xu et al. [2003] report a classification accuracy of 93% for four genres, in general the classification accuracy decreases when the number of genres grows.

For classification into dance music genres, Gouyon et al. [2004] obtain up to 78,9% accuracy (15.9% baseline) when classifying 698 pieces of music into eight classes. This classification is based on a number of rhythmic descriptors and a rule-based classifier whose rules were designed manually. For a wider range of musical contents, divided into eleven genres, Uhle and Dittmar [2004] report a classification accuracy of 67.6%, also based on rhythm features.

At the ISMIR 2004 conference, a comparison of different audio description algorithms was conducted in the form of a contest[1]. For the section of genre classification, the winning algorithm

[1] http://ismir2004.ismir.net/ISMIR_Contest.html

achieved a classification accuracy of 84.07% correct answers. The test collection consisted of 729 pieces, divided into six classes, with a baseline of 43.9%.

3.3.4 Trying to Predict Labels Other Than Genre

Genre or style is a descriptor that is useful for many applications, especially in commercial settings. Even though the concept of 'genre' is not well defined (see, e.g., Aucouturier and Pachet [2003]), it is still much more 'objective' than the kinds of personal characterisations human listeners attach to their music. But it is precisely these personal, subjective categorisations ("happy music", "aggressive music", "music I like whan I am sad", "music that one can dance to") that, if learnable by computers, would open new possibilities for intelligent and rewarding musical interactions between humans and machines.

A small preliminary experiment on the learnability of subjective, non-genre categorisations is reported in this section. As will be seen, the results are rather poor, and a lot of improvement is still needed. Web-based learning about music is a promising alternative that might help overcome the current limitations; that is the topic of the next section (Section 3.4).

The experiment presented here aimed to investigate the learnability of the categorisations *mood* (happy / neutral / sad), *perceived tempo* (very slow / slow / medium / fast / very fast / varying), *complexity* (low / medium / high), *emotion* (soft / neutral / aggressive), *focus* (vocal / both / instruments), and *genre* (blues / classical / electronica / folk / jazz / new age / noise / rock / world). To this end, each piece in a music collection of 729 pieces was labeled with the according value.

This data basis was used to examine the discriminative power of several descriptor sets in combination with a number of machine learning algorithms. The descriptor sets consisted mainly of descriptors that are widely used for music classification tasks (see section 3.2.1 above). Three different descriptor sets were tested: The set that was also used in Tzanetakis and Cook [2002], a set made from some Mpeg7 Low Level Descriptors, and a set that contained all descriptors of the above sets, together with some additional ones for rhythm and melody description.

To train the machine learning algorithms, mean and variance of the descriptors' values for a 30-second excerpt of the piece of music were taken as attributes. Table 3.1 shows the highest classification accuracies that were achieved with different learning algorithms; accuracy was estimated via stratified tenfold cross validation. The evaluated learning algorithms were J48 (a decision tree learner, available — like all the other learning algorithms mentioned here

— in the machine learning toolkit WEKA[2]), SMO (a support vector machine), Naive Bayes, Naive Bayes with Kernel estimation, Boosting, Boosting with J48, Regression with MP5, Linear Regression, and k-NN with k = 1, 3, 5, 10. The table also lists the results obtained when applying the algorithm from Aucouturier and Pachet [2004] with to the same categorisations. For this algorithm, the best values obtained for k-NN classification with k = 1, 3, 5, 10 are shown. The other learning algorithms were not applicable to its feature data. Also, the baseline is given (i.e. the classification accuracy achieved when always guessing the most frequent class).

	mood	perceived tempo	complexity	emotion	focus
Baseline	50.00 %	42.53 %	75.66 %	44.46 %	68.92 %
Set from Tzanetakis and Cook [2002]	50.00 %	42.53 %	76.63 %	45.06 %	71.08 %
Some Mpeg7 LLDs	50.00 %	43.13 %	76.14 %	46.75 %	70.00 %
"Large" Set	51.08 %	44.70 %	76.87 %	47.47 %	71.20 %
Best from Aucouturier and Pachet [2004]	50.24 %	48.67 %	78.55 %	57.95 %	75.18 %

Table 3.1: Best classification accuracies for the different categorisations in the small preliminary experiment.

These results show that with the examined techniques, in some cases it is even not possible to get classification accuracies higher than the baseline. For all categorisations except *mood*, the algorithm from Aucouturier and Pachet [2004] performed better than the other approaches. There is a number of ways in which this experiment could be improved, e.g., by the application of feature selection algorithms or the development of dedicated descriptors for each different task. Still, the results point to some fundamental limitations of the feature-based learning approach; concepts like the emotional quality of a piece of music seem to elude a purely audio-based approach.

3.4 A New Direction: Inferring High-level Descriptors from Extra-Musical Information

Listening to and 'making sense of' music is much more than decoding and parsing an incoming stream of sound waves into higher-level objects such as onsets, notes, melodies, harmonies, etc.

[2]Software freely available from http://www.cs.waikato.ac.nz/ml/

Music is embedded in a rich web of cultural, historical, cultural, and social (and marketing) contexts that influence how music is heard, interpreted, and categorised. That is, many qualities or categorisations attributed to a piece of music by listeners cannot solely be explained by the content of the audio signal itself.

Also, recent research on genre classification is showing clearly that purely audio-based approaches to music classification may be hitting a kind of 'glass ceiling' Aucouturier and Pachet [2004]: there seem to be strict limits to the level of classification accuracy that can be obtained with purely audio-based features, no matter how sophisticated the audio descriptors. From a pragmatic point of view, then, it is clear that, if at all, high-quality automatic music annotation and classification can only be achieved by also taking into account and exploiting information sources that are external to the music itself.

The Internet is a rich, albeit unstructured, source of potential information, where millions of music lovers and experts discuss, describe, and exchange music. Possible information sources include personal web pages, music and concert reviews published on the Web, newspaper articles, discussion forums, chat rooms, playlists exchanged through peer-to-peer networks, and many more. A common term for denoting all the musically relevant information that is potentially "out there" is 'community metadata' Whitman and Lawrence [2002]. Recent approaches to high-level music characterisation try to automatically extract relevant descriptors from the Internet — mostly from general, unstructured web pages —, via the use of information retrieval, text mining, and information extraction techniques (e.g., Baumann and Hummel [2003], Whitman and Ellis [2004], Whitman and Lawrence [2002], Whitman and Smaragdis [2002]). In a sense, this is like learning about music without ever listening to it, by analysing the way people talk about and describe music, rather than what the music actually sounds like.

In the following, two research projects are briefly presented that show in a prototypical way how the Internet can be exploited as a source of information about – in this case – music artists. Section 3.4.1 shows how artists can be probabilistically related to genres via web mining, and section 3.4.2 presents an approach to the hierarchical clustering of music artists, and the automatic labeling of the individual clusters with descriptive terms gleaned from the Web.

3.4.1 Assigning Artists to Genres via Web Mining

In this section we will explain how to extract features (words) related to artists from web pages and how to use these features to construct a probabilistic genre classifier. This permits the

computer to classify new artists present on the web using the Internet community's 'collective knowledge'. To learn the concept of a genre the method requires a set of typical artists for each genre in advance. Based on these artists and a set of web pages that talk about these artists, a characteristic profile is created for each genre. Using this profile (i.e. a weighted list of typical keywords) any artist can be classified by simple evaluation of word occurrences on related web pages. The following is a simplified account of the basic method; the details can be found in Knees et al. [2004].

To obtain useful data for genre profile generation, Internet search engines like *Google* are queried with artist names, along with some constraints (e.g., *+music +review*) that should filter out non-musical pages, and the top ranked pages are retrieved. (Without these constraints, a search for groups such as *Kiss* would result in many unrelated pages). The retrieved pages tend to be common web pages such as fan pages, reviews from online music magazines, or music retailers. The first N available top-ranked webpages for each query are retrieved, all HTML markup tags are removed, so that only the plain text content is left, and common English stop word lists are used to remove frequent terms (e.g. a, and, or, the).

The *features* by which artists are characterised are the individual words that occur in any of the pages. In order to identify those words that may indicate what genre an artist belongs to, the next important step is feature weighting. A common method for this comes from the field of *Information Retrieval* and is known as term frequency × inverse document frequency ($tf \times idf$) Salton and Buckley [1988]. For each artist a and each term t appearing in the retrieved pages, we count the number of occurrences tf_{ta} (term frequency) of term t in documents related to a, and df_t, the number of pages the term occurred in (document frequency). These are combined by multiplying the term frequency with the inverse document frequency. Basically, the intention of the $tf \times idf$ function is to assign a high score to terms that occur frequently, but also to reduce the score if these terms occur on many different pages and thus do not contain useful information.

In the approach described in Knees et al. [2004], an additional step is performed to find those terms that are most discriminative for each genre: a χ^2 test is used to select those terms that are least independent of (i.e., are likely to be predictive of) the classes. Selecting the top N terms for each category and scaling all χ^2 values per category such that the score for the top ranked term equals 1.0, gives a list of terms that seem to be typical of a given genre. An example of such a list for the genre heavy metal/hard rock is shown in Table 3.2. Note that neither of the constraint words (music and review) are included (they occur in all the pages, but they do not help in discriminating the genres).

The top 4 words are all (part of) artist names which were queried. However, many artists which are not part of the queries are also in the list, such as Phil Anselmo (Pantera), Hetfield, Hammett, Trujillo (Metallica), and Ozzy Osbourne. Furthermore, related groups such as Slayer, Megadeth, Iron Maiden, and Judas Priest are found as well as album names (Hysteria, Pyromania, ...) and song names (Paranoid, Unforgiven, Snowblind, St. Anger, ...) and other descriptive words such as evil, loud, hard, aggression, and heavy metal.

To classify previously unseen artists we simply query *Google* with the artist name, count the occurrences of the characteristic genre terms on the retrieved web pages, and multiply these numbers with their respective scores for each genre. The scores in each genre are summed up, and the probability of membership of an artist to a genre is then computed as the fraction of the achieved score of each genre over the sum of scores over all genres.

In Knees et al. [2004], this procedure was tested using a genre taxonomy of 14 genres, and it was shown that correct genre recognition rates of 80% and better are achievable with this purely web-based approach, which compares very favourably with audio-based classification (see section 3.3.3 above).

On top of this classification system, an interactive demo applet (the "GenreCrawler") was implemented that permits the user to experiment with the system by typing in arbitrary new artists. In fact, the words to be typed in need not be artist names at all — they could be anything. The learned classifier can relate arbitrary words to genres, if that makes sense at all. For example, a query for "Pathétique" results in an unambiguous answer: *Classical Music*. A screenshot of the *GenreCrawler* at work can be seen in Figure 3.1.

3.4.2 Learning Textual Characterisations

It is easy to convert the linguistic features (words) identified with the above method into a *similarity measure*, again using standard methods from information retrieval. Similarity measures have a wide range of applications, and one is presented in this section: learning to group music artists into meaningful categories, and describing these categories with characteristic words. Again, this is exploiting the Internet as an information source and could not be achieved on an audio basis alone.

More precisely, the goal is to find words to describe what a group of artists has in common, or what distinguishes it from other groups. Such information can be used for hierarchical user

1.00 *sabbath	0.26 heavy	0.17 riff	0.12 butler
0.97 *pantera	0.26 ulrich	0.17 leaf	0.12 blackened
0.89 *metallica	0.26 vulgar	0.17 superjoint	0.12 bringin
0.72 *leppard	0.25 megadeth	0.17 maiden	0.12 purple
0.58 metal	0.25 pigs	0.17 armageddon	0.12 foolin
0.56 hetfield	0.24 halford	0.17 gillan	0.12 headless
0.55 hysteria	0.24 dio	0.17 ozzfest	0.12 intensity
0.53 ozzy	0.23 reinventing	0.17 leps	0.12 mob
0.52 iommi	0.23 lange	0.16 slayer	0.12 excitable
0.42 puppets	0.23 newsted	0.15 purify	0.12 ward
0.40 dimebag	0.21 leppards	0.15 judas	0.11 zeppelin
0.40 anselmo	0.21 adrenalize	0.15 hell	0.11 sandman
0.40 pyromania	0.21 mutt	0.15 fairies	0.11 demolition
0.40 paranoid	0.20 kirk	0.15 bands	0.11 sanitarium
0.39 osbourne	0.20 riffs	0.15 iron	0.11 *black
0.37 *def	0.20 s&m	0.14 band	0.11 appice
0.34 euphoria	0.20 trendkill	0.14 reload	0.11 jovi
0.32 geezer	0.20 snowblind	0.14 bassist	0.11 anger
0.29 vinnie	0.19 cowboys	0.14 slang	0.11 rocked
0.28 collen	0.18 darrell	0.13 wizard	0.10 drummer
0.28 hammett	0.18 screams	0.13 vivian	0.10 bass
0.27 bloody	0.18 bites	0.13 elektra	0.09 rocket
0.27 thrash	0.18 unforgiven	0.13 shreds	0.09 evil
0.27 phil	0.18 lars	0.13 aggression	0.09 loud
0.26 lep	0.17 trujillo	0.13 scar	0.09 hard

Table 3.2: The top 100 terms with highest χ^2_{tc} values for genre "heavy metal/hard rock" defined by 4 artists (Black Sabbath, Pantera, Metallica, Def Leppard). Words marked with * are part of the search queries. The values are normalised so that the highest score equals 1.0.

interfaces to explore music collections an the artist level Pampalk et al. [2005]. A simple text-based interface is shown in Figure 3.2 below.

As a first step, artists must be clustered hierarchically, and then appropriate terms (words)

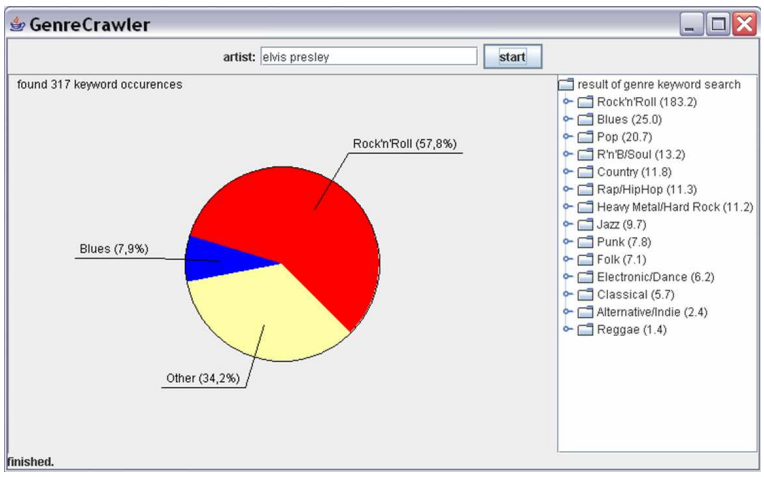

Figure 3.1: The GenreCrawler (cf. Knees et al. [2004]) trying to classify Elvis Presley.

must be selected to describe these clusters. The basis of clustering is a similarity measure, which in our case is based on the linguistic features (characteristic words) extracted from Web pages by the GenreCrawler. There is a multitude of methods for hierarchical clustering. In the system described here Pampalk et al. [2005], basically, a one-dimensional self organising map (SOM) Kohonen [2001] is used, with extensions for hierarhical structuring Miikkulainen [1990], Koikkalainen and E.Oja [1990]. Overlaps between the clusters are permitted, such that an artist may belong to more than one cluster. To obtain a multi-level hierarchical clustering, for each cluster found another one-dimensional SOM is trained (on all artists assigned to the cluster) until the cluster size falls below a certain limit.

The second step is the selection of characteristic terms to describe the individual clusters. The goal is to select those words that best summarise a group of artists. The assumption underlying this application is that the artists are mostly unknown to the user (otherwise we could just label the clusters with the artists' names).

There are a number of approaches to select characteristic words Pampalk et al. [2005]. One of these was developed by Lagus and Kaski (LK) Lagus and Kaski [1999] for labeling large document collections organised by SOMs. LK only use the term frequency tf_{ta} for each term t and artist a. The heuristically motivated ranking formula (higher values are better) is,

$$f_{tc} = (tf_{tc} / \sum_{t'} tf_{t'c}) \cdot \frac{(tf_{tc} / \sum_{t'} tf_{t'c})}{\sum_{c'} (tf_{tc'} / \sum_{t'} t f_{t'c'})}, \tag{3.1}$$

where tf_{tc} is the average term frequency in cluster c. The left side of the product is the importance

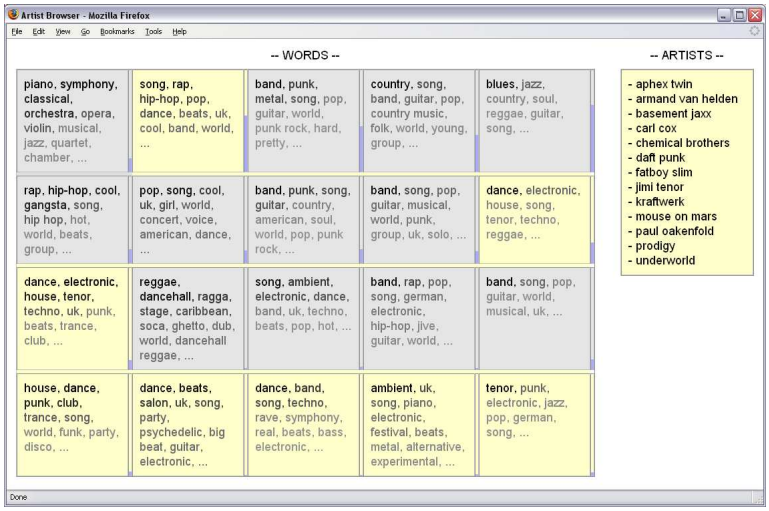

Figure 3.2: Screen shot of the HTML user interface to a system that automatically infers textual characterisations of artist clusters (cf. Pampalk et al. [2005]).

of t in c defined through the frequency of t relative to the frequency of other terms in c. The right side is the importance of t in c relative to the importance of t in all other clusters.

To illustrate, Figure 3.2 shows a simple HTML interface that permits a user to explore the cluster structure learned by the system. There are two main parts to it: the hierarchy of clusters visualised as a grid of boxed texts and, just to the right of it, a display of a list of artists mapped to the currently selected cluster. The clusters of the first level in the hierarchy are visualised using the five boxes in the first (top) row. After the user selects a cluster, a second row appears which displays the children of the selected cluster. The selected clusters are highlighted in a different color. The hierarchy is displayed in such a way that the user can always see every previously made decision on a higher level. The number of artists mapped to a cluster is visualised by a bar next to the cluster. Inside a text box, at most the top 10 terms are displayed. The value of the ranking function for each term is coded through the color in which the term is displayed. The best term is always black and as the values decrease the color fades out. In the screenshot, at the first level the second node was selected, on the second level the fifth node, and on the third level, the first node. More details about method and experimental results can be found in Pampalk et al. [2005].

To summarise, the last two sections were meant to illustrate how the Internet can be used as a rich source of information about music. These are just simple first steps, and a lot of research

on extracting richer music-related information from the Web can be expected.

A general problem with web-based approaches is that many new and not so well known artists or music pieces do not appear on web pages. That limits the approach to yesterday's mainstream western culture. Another issue is the dynamics of web contents (e.g. Lawrence and Giles [1999]). This has been studied in Knees et al. [2004] and the study was continued in Knees [2004]. The experiments reported there indicate that, while the web may indeed be unstable, simple approaches like the ones described here may be highly robust to such fluctuations in web contents. Thus, the web mining approach may turn out to be an important pillar in research on music categorisation, if not music 'understanding'.

3.5 Research and Application Perspectives

Building computers that can 'make sense' of music has long been a goal topic that inspired scientists, especially in the field of Artificial Intelligence (AI). For the past 20 or so years, research in AI and Music has been aiming at creating systems that could in some way mimic human music perception, or to put it in more technical terms, that could recognise musical structures like melodies, harmonic structure, rhythm, etc. at the same level of competence as human experts. While there has been some success in specialised problems such as beat tracking, most of the truly complex musical capabilities are still outside of the range of computers. For example, no machine is currently capable of correctly transcribing an audio recording of even modest complexity, or of understanding the high-level *form* of music (e.g., recognising whether a classical piece is in sonata form, identifying a motif and its variations in a Mozart sonata, or unambiguously segmenting a popular piece into verse and chorus and bridge).

The new application field of Music Information Retrieval has led to, or at least contributed to, a shift of expectations: from a practical point of view, the real goal is not so much for a computer to 'understand' music in a human-like way, but simply to have enough 'intelligence' to support intelligent musical services and applications. Perfect musical understanding may not be required here. For instance, genre classification need not reach 100% accuracy to be useful in music recommendation systems. Likewise, a system for quick music browsing (e.g., Goto [2003]) need not perform a perfect segmentation of the music — if it finds roughly those parts in a recording where some of the interesting things are going on, that may be perfectly sufficient. Also, relatively simple capabilities like classifying music recordings into broad categories (genres)

or assigning other high-level 'semantic' labels to pieces can be immensely useful.

As has been indicated in this chapter, some of these capabilities are within reach, and indeed, some highly interesting real-world applications of this technology are currently emerging in the music market. ¿From the research point of view, it is quite clear that there is still ample room for improvement, even within the relatively narrow domain of learning to assign high-level descriptors and labels to music recordings, which was the topic of this chapter. For instance, recent work on musical web mining has shown the promise of using extra-musical information for music classification, but little research has so far been performed on *integrating* different information sources — low-level audio features, higher-level structures automatically extracted from audio, web-based features, and possibly lyrics (which can also be recovered automatically from the Internet Knees et al. [2005]) — in non-trivial ways.

A concept of central importance to MIR is *music similarity measures*. These are useful not only for classification, but for a wide variety of practical application scenarios, e.g., the automatic structuring and visualisation of large digital music collections Pampalk et al. [2002, 2004], automatic playlist generation (e.g., Aucouturier and Pachet [2002b]), automatic music recommendation, and many more. Current music similarity measures are usually based on lower-level descriptors which are somehow averaged over a whole piece, so that a Euclidean distance metric can be applied to them. More complex approaches like clustering and distribution modelling via mixtures give a slightly more detailed account of the contents of a piece, but still ignore the temporal aspect of music. While preliminary experiments with Hidden Markov Models Aucouturier and Pachet [2004], Flexer et al. [2005], which do model temporal dependencies, do not seem to lead to improvements when based on low-level timbral features (like MFCCs), there is no reason to assume that the integration of higher-level descriptors (like melody, harmony, etc.) and temporal modelling will not permit substantial improvement. A lot of research on these issues is to be expected in the near future, driven by the sheer practical potential of music similarity measures. To put it simply: computers equipped with good music similarity measures may not be able to *make sense* of music in any human-like way, but they will be able to do more and more *sensible things* with music.

Acknowledgments

This work is supported by the European Union in the context of the projects S2S^2 ("Sound to Sense, Sense to Sound", IST-2004-03773) and SIMAC ("Semantic Interaction with Music Audio Contents", FP6 507142). Further support for ÖFAI's research in the area of intelligent music processing is currently provided by the following institutions: the European Union (project COST 282 KnowlEST "Knowledge Exploration in Science and Technology"); the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung* (FWF; projects Y99-START "Artificial Intelligence Models of Musical Expression" and L112-N04 "Operational Models of Music Similarity for MIR"); and the Viennese *Wissenschafts-, Forschungs- und Technologiefonds* (WWTF; project CI010 "Interfaces to Music"). The Austrian Research Institute for Artificial Intelligence also acknowledges financial support by the Austrian Federal Ministries of Education, Science and Culture and of Transport, Innovation and Technology.

Bibliography

M. Alonso, B. David, and G. Richard. Tempo and beat estimation of musical signals. In *5th International Conference on Music Information Retrieval (ISMIR 2004)*, pages 158–163, 2004.

J.J. Aucouturier and F. Pachet. Music similarity measures: What's the use? In *Proceedings of the Third International Symposium on Music Information Retrieval (ISMIR 2002)*, pages 157–163, Paris, France, 2002a.

J.J. Aucouturier and F. Pachet. Scaling up music playlist generation. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME 2002)*, Lausanne, Switzerland, 2002b.

J.J Aucouturier and F. Pachet. Musical genre: A survey. *Journal of New Music Research*, 32(1): 83–93, 2003.

J.J. Aucouturier and F. Pachet. Improving timbre similarity: How high is the sky? *Journal of Negative Results in Speech and Audio Sciences*, 1(1), 2004.

S. Baumann and O. Hummel. Using cultural metadata for artist recommendation. In *Proceedings of the International Conference on Web Delivery of Music (WedelMusic)*, Leeds, UK, 2003.

J.-J. Burred and A. Lerch. A hierarchical approach to automatic musical genre classification. In *Proceedings of th 6th International Conference on Digital Audio Effects (DAFx-03), London, UK, September 8-11, 2003*, London, UK, September 8-11 2003.

C. H. L. Costa, J. D. Valle Jr., and A. L. Koerich. Automatic classification of audio data. In *Proceedings of the 2004 IEEE International Conference on Systems, Man and Cybernetics - SMC*, Hague, Netherlands, October, 10-13 2004.

N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*. Cambridge University Press, 2000.

S. Dixon, F. Gouyon, and G. Widmer. Towards characterisation of music via rhythmic patterns. In *5th International Conference on Music Information Retrieval*, pages 509–516, 2004.

S. Dixon, E. Pampalk, and G. Widmer. Classification of dance music by periodicity patterns. In *4th International Conference on Music Information Retrieval*, pages 159–165, 2003.

J.S. Downie, J. Futrelle, and D. Tcheng. The international music information retrieval systems evaluation laboratory: Governance, access and security. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR'04)*, Barcelona, Spain, 2004.

R. Duda, P Hart, and D. Stork. *Pattern Classification (2nd Edition)*. John Wiley & Sons, New York, 2001.

J. Eggink and G. Brown. Extracting melody lines from complex audio. In *5th International Conference on Music Information Retrieval (ISMIR 2004)*, pages 84–91, 2004.

A. Flexer, E. Pampalk, and G. Widmer. Hidden markov models for spectral similarity of songs. In *Submitted*, 2005.

J. Fürnkranz. Separate-and-conquer rule learning. *Artificial Intelligence Review*, 13(1):3–54, 1999.

E. Gómez and P. Herrera. Extimating the tonality of polyphonic audio files: Cognitive versus machine learning modelling strategies. In *5th International Conference on Music Information Retrieval (ISMIR 2004)*, pages 92–95, 2004.

E. Gómez, A. Klapuri, and B. Meudic. Melody description and extraction in the context of music content processing. *Journal of New Music Research*, 32(1):23–41, 2003.

M. Goto. Smartmusickiosk: Music listening station with chorus-search function. In *Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology (UIST 2003)*, pages 31–40, 2003.

M. Goto and S. Hayamizu. A real-time music scene description system: Detecting melody and bass lines in audio signals. In *Working Notes of the IJCAI-99 Workshop on Computational Auditory Scene Analysis*, pages 31–40. International Joint Conference on Artificial Intelligence, 1999.

F. Gouyon and S. Dixon. A review of automatic rhythm description systems. *Computer Music Journal*, 29(1):34–54, 2005.

F. Gouyon, S. Dixon, E. Pampalk, and G. Widmer. Evaluating rhythmic descriptors for musical genre classification. In *Proceedings of the AES 25th International Conference*, London, UK, June 17-19 2004.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Verlag, New York, 2001.

O. Hellmuth, E. Allamanche, J. Herre, T. Kastner, N. Lefebvre, and R. Wistorf. Music genre estimation from low level audio features. In *Proceedings of the AES 25th International Conference*, London, UK, June 17-19 2004.

P. Herrera, G. Peeters, and S. Dubnov. Automatic classification of musical instrument sounds. *Journal of New Music Research*, 32(1):3–22, 2003.

T. Kastner, J. Herre, E. Allamanche, O. Hellmuth, C. Ertel, and M. Schalek. Automatic optimization of a music similarity metric using similarity pairs. In *Proceedings of the AES 25th International Conference*, London, UK, June 17-19 2004.

A. Klapuri. Automatic music transcription as we know it today. *Journal of New Music Research*, 33(3):269–282, 2004.

P. Knees. Automatische Klassifikation von Musikkünstlern basierend auf Web-Daten (automatic classification of music artists based on web-data). Master´s thesis, Vienna University of Technology, Vienna, 2004.

P. Knees, E. Pampalk, and G. Widmer. Artist classification with web-based data. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR'04)*, Barcelona, Spain, 2004.

P. Knees, M. Schedl, and G. Widmer. Multiple lyrics alignment: Automatic retrieval of song lyrics. In *Submitted*, 2005.

T. Kohonen. *Self-Organizing Maps*. Springer Verlag, 2001.

P. Koikkalainen and E.Oja. Self-organizing hierarchical feature maps. In *Proceedings of the International Joint Conference on Neural Networks*, San Diego, CA, 1990.

K. Lagus and S. Kaski. Keyword selection method for characterizing text document maps, volume 1. In *Proceedings of ICANN99, Ninth International Conference on Artificial Neural Networks*, pages 371–376, London, 1999. IEEE.

S. Lawrence and C. Lee Giles. Accessibility of information on the web. *Nature*, 400(6740):107–109, 1999.

T. Li and M. Ogihara. Detecting emotion in music. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR'03)*, Baltimore, MD, USA, October 26-30 2003.

D. Liu, L. Lu, and H.-J. Zhang. Automatic mood detection from acoustic music data. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR'03)*, Baltimore, MD, USA, October 26-30 2003.

B. Logan and A. Salomon. A music similarity function based on signal analysis. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2001.

I. Mierswa and K. Morik. Automatic feature extraction for classifying audio data. *Machine Learning*, 58(2-3):127–149, 2005.

R. Miikkulainen. *Script recognition with hierarchical feature maps*. Connection Science, 1990.

T.M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

E. Pampalk, S. Dixon, and G. Widmer. Exploring music collections by browsing different views. *Computer Music Journal*, 28(2):49–62, 2004.

E. Pampalk, A. Flexer, and G. Widmer. Hierarchical organization and description of music collections at the artist level. In *Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005)*, 2005.

E. Pampalk, A. Rauber, and D. Merkl. Content-based organization and visualization of music archives. In *Proceedings of the 10th ACM International Conference on Multimedia*, pages 570–579, Juan les Pins, France, 2002.

J.R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

C. Roads. *The Computer Music Tutorial*. MIT Press, Cambridge MA, 1996.

G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

C. Uhle and C. Dittmar. Drum pattern based genre classification of popular music. In *Proceedings of the AES 25th International Conference*, London, UK, June 17-19 2004.

K. West and S. Cox. Features and classifiers for the automatic classification of musical audio signals. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR'04)*, Barcelona, Spain, October, 10-14 2004.

B. Whitman and D. Ellis. Automatic record reviews. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004)*, Barcelona, Spain, 2004.

B. Whitman and S. Lawrence. Inferring descriptions and similarity for music from community metadata. In *Proceedings of the 2002 International Computer Music Conference (ICMC 2002)*, Göteborg, Sweden, 2002.

B. Whitman and B. Smaragdis. Combining musical and cultural features for intelligent style detection. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR 2002)*, Paris, France, 2002.

C. Xu, N. C. Maddage, X. Shao, and F. C. Tian. Musical genre classification using support vector machines. In *Proceedings of IEEE ICASSP03*, Hong Kong, China, April 6-10 2003.

K. Yoshii, M. Goto, and H. Okuno. Automatic drum sound description for real-world music using template adaptation and matching methods. In *5th International Conference on Music Information Retrieval (ISMIR 2004)*, pages 184–191, 2004.

T. Yoshioka, T. Kitahara, K. Komatani, T. Ogata, and H. Okuno. Automatic chord transcription with concurrent recognition of chord symbols and boundaries. In *5th International Conference on Music Information Retrieval (ISMIR 2004)*, pages 100–105, 2004.

Chapter 4

"Sense" in Expressive Music Performance: Data Acquisition, Computational Studies, and Models

Werner Goebl[1], Simon Dixon[1], Giovanni De Poli[2], Anders Friberg[3], Roberto Bresin[3], and Gerhard Widmer[1,4]

[1]Austrian Research Institute for Artificial Intelligence (OFAI), Vienna

[2]Department of Information Engineering, University of Padova;

[3]Department of Speech, Music, and Hearing, Royal Institute of Technology, Stockholm;

[4]Department of Computational Perception, Johannes Kepler University, Linz

This chapter gives an introduction into basic strands of current research in expressive music performance. A special focus is given on the various methods to acquire performance data either during a performance (e.g., through computer-monitored instruments) or from audio recordings. We then overview the different computational approaches to formalise and model the various aspects in expressive music performance. Future challenges and open problems are tackled briefly at the end of this chapter.

4.1 Introduction

Millions of people are regularly attending live music events or listening to recordings of music performances. What drives them to do so is hard to pin down with certainty, and the reasons for it might be manifold. But while enjoying the music, they are all listening to (mostly) human-made music that contains a specific human expression, whatever kind it might be — what they hear makes sense to them. Without this expressivity the music would not attract people; it is an integral part of the music.

Given the central importance of expressivity (not only in music, but in all communication modes and interaction contexts), it is not surprising that human expression and expressive behaviour have become a domain of intense scientific study. In the domain of music, much research has focused on the act of *expressive music performance*, as it is commonly and most typically found in classical music: the deliberate shaping of the music by the performer, the imposing of expressive qualities onto an otherwise 'dead' musical score via controlled variation of parameters such as intensity, tempo, timing, articulation, etc. Early attempts at quantifying this phenomenon date back to the beginning of the 20th century, and even earlier than that.

If we wish to precisely measure and analyse every detail of an expressive music performance (onset timing, timbre and intensity, duration, etc), we end up with huge amounts of data that quickly become unmanageable. Since the first large-scale, systematic research into expression in music performance (usually of classical music) in the 1930s, this has always been a main problem in this field that was controlled either by reducing the amount of music investigated to some seconds of music, or by limiting the number of performances studied to one or two. Recent approaches try to overcome this drawback by using modern computational methods in order to study, model, and understand this complex interaction of performed events and other information of the performance (e.g., the score and the music structure in the case of "classical music").

In the past ten years, quite some very comprehensive overview papers have been published on the various aspects of music performance research. The probably most cited is Alf Gabrielsson's chapter in Diana Deutsch's book "Psychology of Music" [Gabrielsson, 1999] in which he reviewed over 600 papers in this field published until approximately 1995. In a follow-up paper, he added and discussed another 200 peer-reviewed contributions that appeared until 2002 [Gabrielsson, 2003]. A cognitive-psychological review has been contributed by Palmer [1997] summarising empirical research that concentrate on cognitive aspects of music perfor-

mance such as memory retrieval, anticipatory planning, or motor control. The musicologist's perspective is represented by two major edited books devoted exclusively to music performance research [Rink, 1995, 2002]. Lately, more introductory chapters highlight the various methodological issues of systematic musicological performance research [Rink, 2003, Clarke, 2004, Cook, 2004, Windsor, 2004]. Two recent contributions surveyed the diversity of computational approaches in modeling expressive music performance [De Poli, 2004, Widmer and Goebl, 2004]. Parncutt and McPherson [2002] attempted to bridge the gap between the research on music performance and the music practice by bringing together two authors from each of the two sides for each chapter of this book.

Having this variety of overview papers in mind, we aim in this chapter to give a systematic overview on the more technological side of accessing, measuring, analysing, studying, and modeling expressive music performances. As an outset, we screened the current literature of the past century on the various ways of obtaining expressive data of music performances. Then, we review current computational models for expressive music performance. In a final section we briefly sketch possible future strands and open problems that might be tackled by future research in this field.

4.2 Data Acquisition and Preparation

This section is devoted to very practical issues of obtaining data of various kinds on expressive performance and the basic processing therof. We can distinguish basically two different kinds of obtaining information on music performance. The first is to monitor performances during the production process with various measurement devices (MIDI pianos, accelerometers, movement sensors, video systems, etc.). Specific performance parameters can be accessed directly (hammer velocity of each played tone, bow speed, fingering, etc.). The other way is to extract all these relevant data from the recorded audio signal. This method has the disadvantage that some information easily to extract during performance is almost impossible to gain from the audio domain (think for instance of the right pedal at the piano). The advantage, however, is that we have now over a century of recorded music at our disposal that could serve as valuable sources for various kinds of scientific investigation. In the following sub-sections, we discuss the various approaches for monitoring and measuring music performance.

4.2.1 Using Specially Equipped Instruments

Before computers and digital measurement devices were invented and easily available for everyone, researchers employed a vast variety of mechanical and electric measurement apparati to capture all sorts of human or mechanical movements at musical instruments.

Historical Measurement Devices

Mechanical and Electro-Mechanical Setups Of the first to record the movement of piano keys were Binet and Courtier [1895] who used a 6-mm caoutchouc rubber tube placed under the keys that was connected to a cylindric graphical recorder that captured continuous air pressure resulting from striking different keys on the piano. They investigated some basic pianistic tasks such as playing trills, connecting tones, or passing-under of the thumb in scales with exemplary material. In the first of the two contributions of this study, Ebhardt [1898] mounted metal springs on a bar above the strings that closed a electrical shutter when the hammer was about to touch the strings. The electric signal was recorded with a kymograph and timed with a 100-Hz oscillator. He studied the timing precision of simple finger tapping and playing scales. Further tasks with binary and ternary metrum revealed lengthening the IOI of an accentuated onset. Onset and offset timing of church hymn performances were investigated by Sears [1902]. He equipped a reed organ with mercury contacts that registered key depression of 10 selected keys. This information was recorded on four tracks on the surface of a smoked kymograph drum. He studied several temporal aspects of performances by four organ players, such as duration of the excerpts, bars, and individual note values, accent behavior, or note overlap (articulation).

A multitude of mechanical measurement devices introduced Ortmann [1925, 1929] in studies on physiological determinants of piano playing. To investigate the different behaviors of the key, he mounted a tuning fork aside one piano key that scribe wave traces into smoked paper that vary with the speed of the key. With this setup, he was one of the first to study the response of the key on different pianistic playing techniques. For assessing finger movements, Ortmann [1929, p. 230] used a purpose-built mechanical apparatus that comprises non-flexible aluminum strips that are on one side connected to either the finger (proximal phalanx) or the key surface and on the other side they write on a revolving drum. With this apparatus continuous displacement of finger and key could be recorded and analysed. Another mechanical system was the “Pantograph” [Ortmann, 1929, p. 164], a parallelogram lever construction to record lateral arm movement. For other types of movement, he used active optical systems. The motion of a

tiny light bulb attached to the wrist or the finger leaves a clear trace on a photo plate (the room in very subdued light), when the shutter of the photo camera remains open for entire duration of the movement.

Similar active markers mounted on head, shoulder, elbow, and wrist were used by Bernstein and Popova in their important 1930 study [reported by Kay et al., 2003] to study the complex interaction and coupling of the limbs in piano playing. They used their "kymocyclographic camera" to record the movements of the active markers. A rotating shutter allows the light of the markers to impinge on the constantly moving photographic film. With this device they could record up to 600 instances of the movement per second.

Piano Rolls as Data Source A source of expression data are piano rolls for reproducing pianos that exist from different manufacturers (e.g., Welte-Mignon, Hupfeld, Aeolian Duo-Art, Ampico) and of performances of a manifold of renowned pianists [Bowers, 1972, Hagmann]. They were the first means to record and store artistic music performances before the gramophone has been invented. Starting in the late 1920s, scientists took advantage of this source of data and investigated various aspects of performance. Heinlein [1929a,b, 1930] used Duo-Art rolls of the Aeolian company to study pedal use of four pianists playing Schumann's *Träumerei*. Rolls of the same company were the basis of Vernon's 1936 study. He investigated vertical synchronisation of the tones in a chord [see Goebl, 2001]. Hartmann [1932] used Hupfeld "Animatic Rolls" and provided a very detailed study on tone and bar durations as well as note onset asynchronies in two recordings of the first movement of Beethoven's Op. 27 No. 2 (Josef Pembaur, Harold Bauer). Since the precise recording procedures by these companies are still unknown because they deliberately were hold back for commercial reasons, the authenticity of these rolls is sometimes questionable [Hagmann, Gottschewski, 1996]. For example, the Welte-Mignon system were able to simultaneously control dynamics only for keyboard halves. Hence, to emphasise the melody note and to play the rest of the chord tones softer was only possible on such a system when the melody tone was played at a different point in time than the others [Gottschewski, 1996, pp. 26–42]. Although we know today that pianists anticipate melody notes [Palmer, 1996b, Repp, 1996a, Goebl, 2001], the Welte-Mignon rolls cannot be taken literally as a source for studying note asynchronies [as done by Vernon, 1936]. The interpretation of piano rolls need to carefully performed having in mind the conditions of their production. There are currently some private attempts to systematically scan in piano rolls and transform them into standard symbolic format (e.g., MIDI). However, we are not aware of any scientific project concerned with this.

The Iowa Piano Camera During the 1930's, Carl E. Seashore guided a research group that focused on different aspects of music performance, namely the singing voice, violin playing, and piano performance [Seashore, 1932, 1936a,b]. They developed various measurement setups for scientific investigation, among those most prominently the "Iowa Piano Camera" [Henderson et al., 1936] that captured optically onset and offset times and hammer velocity of each key and additionally the movement of the two pedals. It was therefore a complete and comparably very precise device that was not topped until the present days computer-controlled pianos [such as the Disklavier or the SE, see Goebl and Bresin, 2003]. Each hammer is equipped with a shutter that controls light exposure onto a moving film. The hammer shutter interrupts (as in later computer-control reproducing pianos) twice the light exposure on the film: a first time from 24 to 12 mm before the hammer touches the strings and a second time at hammer–string contact. The average hammer speed of the last 12 mm of the hammer's travel can be inferred from the distance on the film between these two interrupts (today's computer-controlled pianos take the average speed of the final 5 mm). According to Skinner and Seashore [1936], the temporal resolution goes down to 10 ms. The hammer velocity gets quantised into 17 dynamics categories [Henderson, 1936]. With this system, the IOWA group performed several studies with professional pianists. Henderson [1936] had two professionals playing the middle section of Chopin's Nocturne Op. 15 No. 3. In this very comprehensive study, they examine temporal behavior, phrasing, accentuation, pedalling, and chord asynchronies. Skinner and Seashore [1936] analysed repeated performances of pieces by Beethoven and Chopin and found high timing consistency within the pianists.

Contemporary Measurement Devices

Henry Shaffer's Photocell Bechstein After the efforts of Seashore's research group at Iowa, it took over 40 years before a new initiative included modern technology to capture piano performance. It was L. Henry Shaffer at Exeter who equipped each of the 88 tones of a Bechstein grand piano with pairs of photocells and the two pedals to capture the essential expressive parameters of piano performance [Shaffer, 1980, 1981, 1984, Shaffer et al., 1985, Shaffer and Todd, 1987, Shaffer, 1992]. The optical registration of the action's movements had the advantage not to affect the playability of the piano. The photocells were mounted into the piano action in pairs, each capturing the moment of the hammer's transit. One was placed to register the instant of hammer–string contact, the other one the resting position of the hammer. The position of the two pedals were monitored by micro switches and stored as 12-bit words on the computer. Each such event was assigned a time stamp rounded to the nearest microsecond and stored on a computer.

The sensor at the strings yielded the note onset time, the one at the hammer's resting position (when the hammer returns) the note offset time. The time difference between the two sensors is an inverse estimate of the force at which the key was depressed. Already then, the introduced technology was in principle identical to the commercially available computer-monitored pianos until now (e.g., the Yamaha Disklavier series or the Bösendorfer SE). This device was used also by other members of that laboratory [e.g., Clarke, 1982, 1985, Todd, 1985, 1989, 1992]

Studies with Synthesiser Keyboards or Digital Pianos Before computer-monitored acoustic pianos became widely distributed and easily available, simple synthesiser keyboards or digital pianos were used to capture expressive data from music performances. These devices provide timing and loudness data for each performed event through the standardised digital communications protocol MIDI (Musical Instrument Digital Interface) that can be stored in files on computer hard-disks [Huber, 1999] and used as an ideal data source for expression. However, such keyboards do not provide a realistic performance setting for advanced pianists, because the response of the keys is very different from an acoustic piano and the synthesised sound (especially with extensive use of the right pedal) does not satisfy trained ears of highly-skilled (classical) pianists.

Such electronic devices were used for various general expression studies [e.g., Palmer, 1989, 1992, Repp, 1994a,b, 1995a, Desain and Honing, 1994]. Bruno Repp repeated two of his studies that were first performed with data from a digital piano (one concerned with legato articulation, Repp, 1995a, the other with the use of the right pedal, Repp, 1996b) later on a computer-controlled grand piano [Repp, 1997a,d, respectively]. Interestingly, the results of both pairs of studies were similar to each other, even though the acoustic properties of the digital piano were considerably different from the grand piano.

The Yamaha Disklavier System Present performance studies dealing with piano performances make generally use of commercially available computer-controlled acoustic pianos. Apart from systems that can be built into a piano [e.g., Autoklav, Pianocorder, see Coenen and Schäfer, 1992], the most common is the Disklavier system by Yamaha. The first computer-controlled grand pianos was available from 1989 onwards (e.g., MX100A/B, DGP); a revised version was issued in 1992 (e.g., MX100II, DGPII, all informations derived from personal communication with Yamaha Rellingen, Germany). The Mark II series was retailed since 1997, the Mark III series followed approximately in 2001. Currently, the Mark IV series can be purchased that includes also a computer with screen and several high-level functions such as an automatic accompaniment

system. From 1998, Yamaha introduced their high-end PRO series of Disklaviers that involves an extended MIDI format to store more than 7-bit velocity information (values from 0 to 127) and information on key release.

There were few attempts to assess the Disklavier's accuracy of recording and reproducing performances. Coenen and Schäfer [1992] compared various reproducing systems (among them a Disklavier DG2RE and a SE225) on their applicability for compositorical purposes (reproducing compositions for mechanical instruments). Maria [1999] had a Disklavier DS6 Pro at his disposal and tested its precision in various ways. More systematic tests on recording and reproduction accuracy were performed by Goebl and Bresin [2001, 2003] using accelerometer registration to inspect key and hammer movements during recording and reproduction.

Yamaha delivers both upright and grand piano versions of its Disklavier system. One of the first to investigate an early upright Disklavier (MX100A) was Bolzinger [1995] who found a logarithmic relationship between MIDI velocity values and sound pressure level (dB). This upright model was used for several performance studies [Palmer and van de Sande, 1993, Palmer and Holleran, 1994, Repp, 1995b,c, 1996a,c,d, 1997b,c].

The Yamaha Disklavier grand piano was even more widely used in performance research. Moore [1992] combined data from a Disklavier grand piano with electromyographic recordings of the muscular activity of four performers playing trills. Behne and Wetekam [1994] recorded student performances of the theme of Mozart's K.331 on a Disklavier grand piano and studied systematic timing variations of the Siciliano rhythm. As mentioned above, Repp repeated his work on legato and pedalling on a Disklavier grand piano Repp [1997a,d, respectively]. Juslin and Madison [1999] used a Disklavier grand piano to record and play back different (manipulated) performances of two melodies to assess listeners' ability to recognise simple emotional categories. Bresin and Battel [2000] analysed multiple performances recorded on a Disklavier grand piano of Mozart's K.545 in terms of articulation strategies. Clarke and Windsor [2000] used recordings made on a Disklavier grand piano for perceptual evaluation of real and artificially created performances. A short piece by Beethoven was recorded on a Disklavier grand piano played by either one professional pianist [Windsor et al., 2001] or by 16 professional pianists [Timmers et al., 2002, Timmers, 2002] in different tempi. Timing characteristics of the different types of grace notes were investigated. Riley-Butler [2002] used a Disklavier grand piano in educational settings. She showed piano roll representations of student's performances to them and observed considerable increase of learning effectivity with this method.

Bösendorfer's SE System The SE ("Stahnke Electronics") System dates back to the early 1980s when the engineer Wayne Stahnke developed a reproducing system in cooperation with the MIT Artificial Intelligence Laboratory built into a Bösendorfer Imperial grand piano [Roads, 1986, Moog and Rhea, 1990]. A first prototype was ready in 1985; the system had been officially sold by Kimball (at that time owner of Bösendorfer) starting from summer 1986. This system was very expensive and only few academic institution could afford it. Until the end of its production, only about three dozen of these systems have been built and sold. The SE works in principle like the Disklavier system (optical sensors register hammershank speed and key release and linear motors reproduce final hammer velocity, see for details Goebl and Bresin, 2003). However, its recording and reproducing capabilities are superior even compared with other much younger systems [Goebl and Bresin, 2003]. Despite its rare occurrence in academic institutions, it was used for performance research in some cases.

Palmer and Brown [1991] performed basic tests on the relation of hammer velocity and peak amplitude of the outcoming sound. Repp [1993] tried to estimate peak sound level of piano tones from the two lowest partials as measured in the spectrogram and compared a digital piano, a Disklavier MX100A upright piano with the Bösendorfer SE. Studies in music performance were accomplished at Ohio State University [Palmer and van de Sande, 1995, Palmer, 1996b,a], at Musichochschule Karlsruhe [e.g., Mazzola and Beran, 1998, Mazzola, 2002, p. 833], or at the grand piano located at the Bösendorfer company in Vienna [Goebl, 2001, Widmer, 2001, 2002b, 2003, Goebl and Bresin, 2003, Widmer, 2005].

Currently (June 2005), the Bösendorfer company in Vienna is developing a new computer-controlled reproducing piano called "CEUS" (personal communication with Bösendorfer Vienna) that introduces among other features sensors that register the continuous motion of each key. This data might be extremely valuable for performance studies into the pianists' touch and tone control.

4.2.2 Measuring Audio By Hand

In contrast to measuring music expression during performance through any kind of sensors placed in or around the performer or the instrument (see previous section), the other approach is to analyse the recorded sound of music performances. It has the essential advantage that any type of recording may serve as a basis for investigation, e.g., commercially available CDs, historic recordings, or recordings from ethnomusicological research. One has simply to go into a record

store and buy all the famous performances by the great pianists of the past century.[1]

However, to extract discrete performance information from audio is difficult and sometimes impossible. The straight-forward method is to inspect the wave form of the audio signal with computer software and mark manually with a cursor the onset times of selected musical events. Though this method is time consuming, it delivers timing information with a reasonable precision. To extract data on dynamics is a bit more complicated (e.g., by reading peak energy values from the root-mean-square of the signal averaged over a certain time window), but only possible for overall dynamics. We are not aware of a successful procedure to extract individual dynamics of simultaneous tones [for an attempt, see Repp, 1993]. Many other signal processing problems have not been solved as well (e.g., extracting pedal information, tone length, etc., see also McAdams et al., 2004).

First studies that extracted timing information directly from sound used oscillogram filming (e.g., Bengtsson and Gabrielsson, 1977, for more references see Gabrielsson, 1999, p. 533). Povel [1977] analysed gramophone records of three performances of Johann Sebastian Bach's first prelude of WTC I. He determined the note onsets "by eye" from two differently obtained oscillograms of the recordings (that were transferred on analog tape). He reported a temporal precision of 1–2 ms (!). Recordings of the same piece were investigated by Cook [1987] who obtained timing (and intensity) data already through a computational routine. The onset detection was automated by a threshold procedure applied to the digitised sound signal (8 bit, 4 kHz) and post corrected by hand. He reported a timing resolution of 10 ms. He also stored intensity values, but did not specify in more detail what exactly was measured here.

Gabrielsson et al. [1983] analysed timing patterns of performances from 28 different monophonic melodies played by 5 performers. The timing data were measured from the audio recordings with a precision of ±5 ms (p. 196). In a later study, Gabrielsson [1987] extracted both timing and (overall) intensity data from the theme of Mozart's K.331. In this study, a digital sampling system was used that allowed a temporal precision of 1–10 ms (p. 87). The dynamics were estimated by reading peak amplitudes of each score event (in voltages). Nakamura [1987] used a Brüel & Kjær level recorder to register dynamics of solo performances played on a violin, oboe, and recorder. He analysed the produced dynamics in relation to the perceived intensity of the music.

[1] In analysing recordings the researcher has to be aware that almost all records are glued together from several takes so the analysed performance might never have taken place in this particular rendition [see also Clarke, 2004, p. 88].

The first larger corpus of recordings was measured by Repp [1990] who fed 19 recordings of the third movement of Beethoven's Op. 31 No. 3 into a VAX 11/780 computer and read off the note onsets from waveform displays. In cases of doubt, he played the sound until the onset and moved the cursor stepwise back in time, until the following note was no longer audible [Repp, 1990, p. 625]. He measured the performances on quarter-note level[2] and reported an absolute mean error of 6.5 ms for repeated measurements (equivalent to 1% of the inter-onset intervals, p. 626). In a further study, Repp [1992] had 28 recordings of Schumann's "Träumerei" by 24 renowned pianists at his disposal. This time, he used a standard waveform editing program to hand-measure the 10-kHz sampled audio files. The rest of the procedure was identical (aural control of ambiguous onsets). He reported an average absolute measurement error of 4.3 ms (or less than 1%). In his later troika on the "microcosm of musical expression" [Repp, 1998, 1999a,b], he applied the same measurement procedure on 115 performances of the first five bars of Chopin's Op. 10 No. 3 Etude collected from libraries and record stores. He used "SoundEdit16" software to measure the onset on sixteenth note level. In addition to previous work, he extracted overall intensity information as well [Repp, 1999a] by taking the peak sound levels (pSPL in dB) extracted from the root-mean-square (RMS) integrated sound signal (over a rectangular window of 30 ms).

Nettheim [2001] measured parts of recordings of four historical performances of Chopin e-minor Nocturne Op. 72 No. 1 (Pachmann, Godowsky, Rubinstein, Horowitz). He used a time-stretching software ("Musician's CD Player," par. 8) to reduce the playback speed by factor 7 (without changing the pitch of the music). He then simply took the onset times from a time display during playback. Tone onsets of all individual tones were measured with this method.[3] In repeated measurements, he reported accuracy of the order of 14 ms. In addition to note onset timing, he assigned arbitrary intensity values to each tone ranging from 1 to 100 by ear (par. 11). He reports about the difficulties arising from that approach.

In recent contributions on timing and synchronisation in Jazz performances, the timing of the various instruments of Jazz ensembles were investigated. Friberg and Sundström [2002] measured cymbal onsets from spectrogram displays with a reported precision of ±3 ms. Ashley [2002] studied the synchronisation of the melody instruments with the double bass line. He repeatedly measured onsets of both lines from wave form plots of the digitised signal with usual differences between the measurements of 3-5 ms. About the same consistency (typically

[2] In the second part of this paper, he measured and analysed eight-note and sixteenth-note values as well.

[3] Obviously, the chosen excerpts were slow pieces with a comparatively low note density.

2 ms) was achieved by Collier and Collier [2002] through a likewise measurement procedure ("CoolEdit 96," manual annotation of physical onsets in trumpet solos). They exemplified an equivocal situation where the trumpet tone "emerges from the band" (p. 468). In those cases, they aurally determined the onset. Lisboa et al. [2005] used "Pro Tools" wave editor to extract onset timing oc Cello solo performances; Moelants [2004] made use of a speech transcription software ("Praat") to assess trill and ornament timing in solo string performances.

In a recent commercial enterprise, John Q. Walker and colleagues have been trying to extract the complete performance information out of historical (audio) recordings in order to play them back on a modern Disklavier.[4] Their commercial aim is to re-sell old recordings with modern sound quality or live performance feel. They computationally extract as much performance information as possible and add the missing information (e.g., tone length, pedalling) to an artificially created MIDI file. They use it to control a modern Disklavier grand piano and compare this performance to the original recording. Then they modify the added information in the MIDI files and play it back again and repeat this process iteratively until the Disklavier's reproduction sounds identical to the original recording [see also Midgette, 2005].

Another way of assessing temporal content of recordings is by repeatedly tapping along with the music recording e.g., on a MIDI drum pad or the like and recording this information [Cook, 1995, Bowen, 1996, Bachmann, 1999]. This is a comparably fast method to gain rough timing data on a tappable beat level. However, perceptual studies on tapping along with expressive music showed that tappers — even after repeatedly tapping along with the same short piece of music — still underestimate abrupt tempo changes or systematic variations [Dixon et al., 2005].

4.2.3 Computational Extraction of Expression from Audio

Several approaches exist for the extraction of expression from audio data, or equivalently, annotating audio data with content-based metadata. The most general approach is to attempt to extract as much musical information as possible, using an automatic transcription system, but such systems are not robust enough to provide the level of precision and accuracy required for analysis of expression [Klapuri, 2004]. Nevertheless, some systems were developed with the specific goal of expression extraction, in an attempt to relieve some of the painstaking effort of manual annotation [e.g., Dixon, 2000]. Since the score is often available for the musical performances being analysed, Scheirer [1997] recognised that much better performance could be

[4] `http://www.zenph.com`

obtained by incorporating score information into the audio analysis algorithms, but the system was never developed to be sufficiently general or robust to be used in practice. One thing that was lacking from music analysis software was an interface for interactive editing of partially correct automatic annotations, without which the use of the software was not significantly more efficient than manual annotation.

The first system with such an interface was BeatRoot [Dixon, 2001a,b], an automatic beat tracking system with a graphical user interface which visualised (and auralised) the audio and derived beat times, allowing the user to edit the output and retrack the audio data based on the corrections. BeatRoot produces a list of beat times, from which tempo curves and other representations can be computed. Although it has its drawbacks, this system has been used extensively in studies of musical expression [Goebl and Dixon, 2001, Dixon et al., 2002, Widmer, 2002a, Widmer et al., 2003, Goebl et al., 2004]. Recently, Gouyon et al. [2004] implemented a subset of BeatRoot as a plugin for the audio editor WaveSurfer [Sjölander and Beskow, 2000].

A similar methodology was applied in the development of JTranscriber [Dixon, 2004], which was written as a front end for an existing transcription system [Dixon, 2000]. The graphical interface shows a spectrogram scaled to a semitone frequency scale, with the transcribed notes superimposed over the spectrogram in piano roll notation. The automatically generated output can be edited with simple mouse-based operations, with audio playback of the original and the transcription, together or separately, possible at any time.

These tools provide a better approach than manual annotation, but since they have no access to score information, they still require a significant amount of interactive correction, so that they are not suitable for very large scale studies. An alternative approach is to use existing knowledge, such as from previous annotations of other performances of the same piece of music and transfer the metadata after aligning the audio files. The audio alignment system MATCH [Dixon and Widmer, 2005] finds optimal alignments between pairs of recordings, and is then able to transfer annotations from one recording to the corresponding times in the second. This proves to be a much more efficient method of annotating multiple performances of the same piece, since manual annotation needs to be performed only once. Further, audio alignment algorithms are much more accurate than techniques for direct extraction of expressive information from audio data, so the amount of subsequent correction for each matched file is much less.

Taking this idea one step further, the initial step of annotation can be avoided entirely if the musical score is available in a symbolic format, by synthesising a mechanical performance from the score and matching the audio recordings to the synthetic performance. For analysis of ex-

pression in audio, e.g. absolute measurements of tempo, the performance data must be matched to the score, so that the relationship between actual and nominal durations can be computed. Several score-performance alignment systems have been developed for various classes of music [Cano et al., 1999, Soulez et al., 2003, Turetsky and Ellis, 2003, Shalev-Shwartz et al., 2004].

Other relevant work is the on-line version of the MATCH algorithm, which can be used for tracking live performances with high accuracy [Dixon, 2005b,a]. This system is being developed for real time visualisation of performance expression. The technical issues are similar to those faced by score following systems, such as those used for automatic accompaniment [Dannenberg, 1984, Orio and Déchelle, 2001, Raphael, 2004], although the goals are somewhat different. Matching involving purely symbolic data has also been explored. Cambouropoulos developed a system for extracting score files from expressive performances in MIDI format [Cambouropoulos, 2000]. After manual correction, the matched MIDI and score files were used in detailed studies of musical expression. Various other approaches to symbolic score-performance matching are reviewed by Heijink et al. [2000b,a].

4.2.4 Extracting Expression from Performers Movements

While the previous sections dealt with the extraction of expression contained in music performances, this section is devoted to expression as represented in all kinds of movements that occur when performers interact with their instruments during performance [for an overview, see Davidson and Correia, 2002, Clarke, 2004]. Performers' movements are a powerful communication channel of expression to the audience, sometimes even overriding the acoustic information [Behne, 1990, Davidson, 1994].

There are several ways to monitor performers' movements. One possibility is to connect mechanical devices to the playing apparatus of the performer [e.g., Ortmann, 1929] that has the disadvantage to inhibit the free execution of the movements. More common are optical tracking systems that either simply video-tape performers movements or record special passive or active markers placed on particular joints of the performers' body. We already mentioned an early study by Berstein and Poppova (1930) who introduced an active photographical tracking system [Kay et al., 2003]. These systems use light-emitting makers placed on the various limbs and body parts of the performer. They are recorded by video cameras that are connected to software that extracts the position of the markers [e.g., the Selspot System, as used by Dahl, 2004, 2005]. The disadvantage of those systems is that the participants need to be cabled which is a

time-consuming process and the cables might inhibit the participants to move as they would move normally. Passive systems use reflective markers that are illuminated by external lamps. In order to create a three-dimensional picture of movement, the data from several cameras are coupled by software [e.g., Palmer and Dalla Bella, 2004].

Even less intrusive are video systems that simply record performance movements without any particular marking of the performer's limbs. Elaborated software systems are able to track defined body joints directly from the plain video signal (e.g., EyesWeb[5], see Camurri et al., 2004, 2005 or Camurri and Volpe, 2004 for an overview in gesture-related research). Perception studies on communication of expression through performers gestures use simpler point-light video recordings (reflective markers on body joints recorded in a darkened room) to present them to participants for ratings [Davidson, 1993].

4.2.5 Extraction of Emotional Content from MIDI and Audio

For listeners and musicians, an important aspect of music is its ability to express emotions [Juslin and Laukka, 2004]. An important research question has been to investigate the coupling between emotional expression and the underlying musical parameters. Two important distinctions have to be made. The first distinction is between perceived emotional expression ("what is communicated") and induced emotion ("what you feel"). Here, we will concentrate on the perceived emotion which has been the focus of most of the research in the past. The second distinction is between compositional parameters (pitch, melody, harmony, rhythm) and performance parameters (tempo, phrasing, articulation, accents). The influence of compositional parameters has been investigated during a long time starting with the important work of Hevner [1937]. A comprehensive summary is given in Gabrielsson and Lindström [2001]. The influence of performance parameters has recently been investigated in a number of studies [for overviews see Juslin and Sloboda, 2001, Juslin, 2003]. These studies indicate that for basic emotions such as happy, sad or angry, there is a simple and consistent relationship between the emotional description and the parameter values. For example, a sad expression is characterised by slow tempo, low sound level, legato articulation and a happy expression is characterised by fast tempo, moderate sound level and staccato articulation.

Predicting the emotional expression is usually done using a two-step process [see also Lindström et al., 2005]: (1) Parameter extraction The first step extracts the basic parameters from

[5] `http://www.megaproject.org`

the incoming signal. The selection of parameters is a trade-off between what is needed in terms of emotion mapping and what is possible. MIDI performances are the simplest case in which the basic information in terms of notes, dynamics and articulation is already available. From this data it is possible to deduce for example the tempo using beat-tracking methods as described above. Audio from monophonic music performances can also be analyzed on the note-level giving similar parameters as for the MIDI case (with some errors). In addition, using audio a few extra parameters are available such as the spectral content and the attack velocity. The CUEX algorithm by Friberg et al. [2005], including a real-time version [Friberg et al., 2002], was specifically designed for prediction of emotional expression yielding eight different parameters for each recognised note. Polyphonic audio is the most difficult case which has only recently been considered. Due to the analysis difficulty several approaches can be envisioned. One possibility is to first make a note extraction using the recent advances in polyphonic transcription mentioned above [e.g., Klapuri, 2004] and then extract the parameters. Due to the lack of precision of polyphonic transcription there will be many errors. However, this may not be too important for the prediction of the emotion in the second step below since preferably the mapping is redundant and insensitive to small errors in the parameters. A more straight-forward approach is to extract overall parameters directly from audio, such as using auditory-based measures for pitch, rhythm and timbre [Leman et al., 2004, Liu et al., 2003]. (2) Mapping The second step is the mapping from the extracted parameters to the emotion character. The selection of method is dependent on the use (research or real time control) and the desired behaviour of the output data. A typical data-driven method is to use listener ratings (the "right" answer) for a set of performances to train a model. Common statistical/mathematical models are used such as regression [Leman et al., 2004, Juslin, 2000], bayesian networks [Canazza et al., 2003], or hidden markov models [Dillon, 2003]. An alternative approach more suitable for real time control is to directly implement qualitative data from previous studies using a fuzzy logic model [Seif El-Nasr et al., 2000, Friberg, 2005], see also Section **??**.

4.3 Computational Models of Music Performance

Models describe relations among different kinds of observable (and often measurable) information about a phenomenon, discarding details that are felt to be irrelevant. They serve to generalise the findings and have both a descriptive and predictive value. Often the information is quantitative and we can distinguish input data, supposedly known, and output data, which are inferred

by the model. In this case, inputs can be considered as the causes and output the effect of the phenomenon. When a model can be implemented on a computer, it is called computational model and it allows deducing the values of output data corresponding to the provided values of inputs. This process is called simulation and it is widely used to predict the behaviour of the phenomenon in different circumstances and can be used to validate the model, by comparing the predicted results with actual observations.

In music performance modelling, the information that can be considered is not only quantitative, as *physical information*, e.g. timing or performer's movements. We have also *symbolic information* that refers more to a cognitive organization of the music than to an exact physical value and *expressive information* more related to the affective and emotional content of the music. Recently computer science and engineering started paying attention to expressive information and developing suitable theories and processing tools giving rise to the field of affective computing and Kansei information processing. Music and music performance in particular, attracted the interest of researchers for developing and testing such tools. Music indeed is the more abstract of the arts and has a long tradition of formalization. Moreover it combines in an interesting way all these aspects.

4.3.1 Modeling Strategies

We may distinguish some strategies in developing the structure of the model and in finding its parameters. The most prevalent ones are analysis-by-measurement and analysis-by-synthesis. Recently some methods from artificial intelligence started being developed: machine learning and case based reasoning. We may distinguish local models, that acts at note level and try to explain the observed facts in a local context, and global models that take into account the higher level of the musical structure or more abstract expression pattern. The two approaches often require different modelling strategies and structures. In certain cases, it is possible to devise a combination of both approaches with the purpose being to obtain better results. The composed models are built by several components, each one aiming to represent the different sources of expression. However, a good combination of the different parts is still quite challenging.

Analysis By Measurements

The first strategy, analysis-by-measurements, is based on the analysis of deviations from the musical notation measured in recorded human performances. The analysis aims to recognise regularities in the deviation patterns and to describe them by means of a mathematical model, relating score to expressive values (see Gabrielsson 1999 and Gabrielsson 2003 for an overview of the main results). The method starts by selecting the performances to be analyzed. Often rather small set of carefully selected performances are used. Then the physical properties of every note are measured using the methods seen in section 4.2 and the data so obtained are checked for reliability and consistency. The most relevant variables are selected and analyzed by statistical methods. The analysis assumes an interpretation model that can be confirmed or modified by the results of the measurements. Often the hypothesis that deviations deriving from different patterns or hierarchical levels can be separated and then added is implicitly assumed. This hypothesis helps the modelling phase, but may be oversimplified. Several methodologies of approximation of human performances were proposed using neural network techniques or fuzzy logic approach or using a multiple regression analysis algorithm or linear vector space theory. In these cases, the researcher devises a parametric model and then estimates its parameters that best approximate a set of given performances.

Many models address very specific aspects of expressive performance, for example, the final ritard and its relation to human motion [Kronman and Sundberg, 1987, Todd, 1995, Friberg and Sundberg, 1999, Sundberg, 2000, Friberg et al., 2000b, **?**]; the timing of grace notes [Timmers et al., 2002]; vibrato [Desain and Honing, 1996, Schoonderwaldt and Friberg, 2001]; melody lead [Goebl, 2001, 2003]; legato [Bresin and Battel, 2000]; or staccato and its relation to local musical context [Bresin and Widmer, 2000, Bresin, 2001].

A global approach was pursued by Todd in his phrasing model [Todd, 1992, 1995]. This model assumes that the structure of a musical piece can be decomposed in a hierarchical sequence of segments, where each segment is on its turn decomposed in a sequence of segments. The performer emphasises the hierarchical structure by an accelerando-ritardando pattern and by a crescendo-decrescendo pattern for each segment. These patterns are superimposed (summed) onto each other and describe from the global variation over the whole to local fluctuations at the note level.

Analysis By Synthesis

While analysis by measurement develop models that best fit quantitative data, the analysis-by-synthesis paradigm takes into account the human perception and subjective factors. First, the analysis of real performances and the intuition of expert musicians suggest hypotheses that are formalised as rules. The rules are tested by producing synthetic performances of many pieces and then evaluated by listeners. As a result the hypotheses are refined, accepted or rejected. This method avoids the difficult problem of objective comparison of performances, including subjective and perceptual elements in the development loop. On the other hand, this method depends too much on the personal competences and taste of few experts.

The most important one is the KTH rule system [Friberg, 1991, 1995, Friberg et al., 1998, 2000a, Sundberg et al., 1983, 1989, 1991]. In the KTH system, the rules describe quantitatively the deviations to be applied to a musical score, in order to produce a more attractive and human-like performance than the mechanical one that results from a literal playing of the score. Every rule tries to predict (and to explain with musical or psychoacoustic principles) some deviations that a human performer is likely to insert. Many rules are based on low-level structural analysis of the text. The KTH rules can be grouped according to the purposes that they apparently have in music communication. Differentiation rules appear to facilitate categorization of pitch and duration, whereas grouping rules appear to facilitate grouping of notes, both at micro and macro level.

Machine Learning

In the traditional way of developing models, the researcher normally makes some hypothesis on the performance aspects s/he want to model and then s/he tries to establish the empirical validity of the model by testing it on real data or on synthetic performances. A different approach, pursued by Widmer and coworkers [Widmer, 1995a,b, 1996, 2000, 2002b, Widmer and Tobudic, 2003, Widmer, 2003, Widmer et al., 2003, Widmer, 2005, Tobudic and Widmer, 2005], instead tries to extract new and potentially interesting regularities and performance principles from many performance examples, by using machine learning and data mining algorithms. The aim of these methods is to search for and discover complex dependencies on very large data sets, without any preliminary hypothesis. The advantage is the possibility of discover new (and possibly interesting) knowledge, avoiding any musical expectation or assumption. Moreover, these algorithms normally allow describing discoveries in intelligible terms. The main criteria

for acceptance of the results are generality, accuracy, and simplicity.

Models were developed to predict local, note-level expressive deviations and higher-level phrasing patterns. Moreover, these two types of models can be combined to yield an integrated, multi-level model of expressive timing and dynamics.

Case-Based Reasoning

An alternative approach, much closer to the observation-imitation-experimentation process observed in humans, is that of directly using the knowledge implicit in human performances samples. Case-based reasoning (CBR) is based on the idea of solving new problems by using (often with some kind of adaptation) similar previously solved problems. An example in this direction is the SaxEx system for expressive performance of Jazz ballads [Arcos et al., 1998, López de Mántaras and Arcos, 2002] which predicts expressive transformations to saxophone phrases recordings by looking at how other, similar phrases were played by a human musician. The success of this approach greatly depends on the availability of a large amount of well-distributed previously solved problems, that are not easy to collect.

Mathematical Theory Approach

A rather different model based mainly on mathematical considerations is the Mazzola model [Mazzola, 1990, Mazzola and Zahorka, 1994, Mazzola et al., 1995, Mazzola, 2002, Mazzola and Göller, 2002]. This model basically consists of an analysis part and a performance part. The analysis part involves computer-aided analysis tools, for various aspects of the music structure, that assign particular weights to each note in a symbolic score. The performance part, that transforms structural features into an artificial performance, is theoretically anchored in the so-called Stemma Theory and Operator Theory (a sort of additive rule-based structure-to-performance mapping). It iteratively modifies the performance vector fields, each of which controls a single expressive parameter of a synthesised performance.

4.3.2 Perspectives

Comparing Performances

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

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

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

Modeling Different Expressive Intentions

The models discussed in the previous sections aim at explaining and simulating performances which is played accordingly to appropriate rules imposed by a specific musical praxis. The focus is on aspects that most performances have in common. Recently research started paying attention to aspects that differentiate performances and performers styles [Repp, 1992, Widmer, 2003]. The same piece of music can be performed trying to convey different expressive intentions [Gabrielsson and Lindström, 2001], changing the style of the performance. The CARO model [Canazza et al., 2004] is able to modify a neutral performance (i.e. played without any specific expressive intention) in order to convey different expressive intentions. Bresin and Friberg [2000]

developed some macro rules for selecting appropriate values for the parameters of the KTH rule system in order to convey different emotions.

Expression Recognition Models

The methods seen in the previous sections aim at explaining how expression is conveyed by the performer and how it is related to the musical structure. Recently these accumulated research results started giving rise to models that aim to extract and recognise expression from a performance [Dannenberg et al., 1997, Friberg et al., 2002, Mion and De Poli, 2004].

4.4 Open Problems and Future Paths

Although computational modelling of expressive human performance has been developing quickly during the past decade, there is ample room for further research, and the field of computational performance modelling continues to be active. However, the idea of a creative activity being predictable and, more specifically, the notion of a direct "quasi-causal" relation between the musical score and the performance is quite problematic. The person and personality of the artist as a mediator between music and listener is totally neglected in basically all models discussed above. There are some severe general limits to what any predictive model can describe. For instance, very often performers intentionally play the repetition of the same phrase or section totally differently the second time around. Being able to predict this would presuppose models of aspects that are outside the music itself, such as performance context, artistic intentions, personal experiences, listeners' expectations, etc.

Although it might sound quaint, there are concrete attempts to elaborate computational models of expressive performance to a complexity so that they are able to compete with human performers. Since 2002, a scientific initiative brings together scientists from all over the world for a competition of artificially created performances (RENCON, contest for performance rendering systems, the next one to be held at the ICMC'05 in Barcelona[6]). Their aim is to construct computational systems that are able to pass ax expressive performance Turing Test [that is an artificial performance sounds indistinguishable to a human performance, Hiraga et al., 2004]. One ambitious goal is a computer system to win the Chopin competition in 2050 [Hiraga et al.,

[6] http://www.icmc2005.org

2004].

It is very hard to imagine that this will ever be possible, not only because the organisers of such a competition wont accept a computer to participate, but also because a computational model would have to take into account the complex social and cognitive contexts in which, like any human intellectual and artistic activity, a music performance is situated. But even if complete predictive models of such phenomena are strictly impossible, they advance our understanding and appreciation of the complexity of artistic behaviour, and it remains an intellectual and scientific challenge to probe the limits of formal modelling and rational characterisation.

Acknowledgements

This research is supported by the European Union (project FP6 IST-2004-03773 S2S2 "Sound to Sense, Sense to Sound"); the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung (FWF; START project Y99-INF "Computer-Based Music Research: Artificial Intelligence Models of Musical Expression"), and the Viennese Science and Technology Fund (WWTF; project CI010 "Interfaces to Music"). The Austrian Research Institute for Artificial Intelligence (OFAI) acknowledges basic financial support by the Austrian Federal Ministries for Education, Science, and Culture, and for Transport, Innovation and Technology.

Bibliography

Josep Lluís Arcos, Ramon López de Mántaras, and Xavier Serra. SaxEx: A case-based reasoning system for generating expressive performances. *Journal of New Music Research*, 27(3):194–210, 1998.

Richard Ashley. Do[n't] change a hair for me: The art of Jazz Rubato. *Music Perception*, 19(3): 311–332, 2002.

Kai Bachmann. *Das Verhältnis der Tempi in mehrsätzigen Musikwerken: ein Beitrag zur musikalischen Aufführungsanalyse am Beispiel der Symphonien Ludwig van Beethovens*. Unpublished doctoral thesis, Institut für Musikwissenschaft, Universität Salzburg, Salzburg, 1999.

Klaus-Ernst Behne. "Blicken Sie auf die Pianisten?!" Zur bildbeeinflußten Beurteilung von Klaviermusik im Fernsehen. *Medienpsychologie*, 2(2):115–131, 1990.

Klaus-Ernst Behne and Burkhard Wetekam. Musikpsychologische Interpretationsforschung: Individualität und Intention. In Klaus-Ernst Behne, Günter Kleinen, and Helga de la Motte-Haber, editors, *Musikpsychologie. Empirische Forschungen, ästhetische Experimente*, volume 10, pages 24–32. Noetzel, Wilhelmshaven, 1994.

Ingmar Bengtsson and Alf Gabrielsson. Rhythm research in Uppsala. In *Music, Room, Acoustics*, volume 17, pages 19–56. Publications issued by the Royal Swedish Academy of Music, Stockholm, 1977.

Alfred Binet and Jules Courtier. Recherches graphiques sur la musique. *L'Année Psychologique*, 2: 201–222, 1895. URL `http://www.musica-mechana.de`. Available also in a German translation by Schmitz, H.-W. (1994), Das Mechanische Musikinstrument 61, 16–24.

Simon Bolzinger. *Contribution a l'étude de la rétroaction dans la pratique musicale par l'analyse de l'influence des variations d'acoustique de la salle sur le jeu du pianiste*. Unpublished doctoral thesis, Institut de Mécanique de Marseille, Université Aix-Marseille II, Marseille, 1995.

José A. Bowen. Tempo, duration, and flexibility: Techniques in the analysis of performance. *Journal of Musicological Research*, 16(2):111–156, 1996.

Q. David Bowers. *Encyclopedia of Automatic Musical Instruments*. Vestal Press Ltd., New York, 13th edition, 1972.

Roberto Bresin. Articulation rules for automatic music performance. In Andrew Schloss and Roger Dannenberg, editors, *Proceedings of the 2001 International Computer Music Conference, Havana, Cuba*, pages 294–297. International Computer Music Association, San Francisco, 2001.

Roberto Bresin and Giovanni Umberto Battel. Articulation strategies in expressive piano performance. *Journal of New Music Research*, 29(3):211–224, 2000.

Roberto Bresin and Anders Friberg. Emotional coloring of computer-controlled music performances. *Computer Music Journal*, 24(4):44–63, 2000.

Roberto Bresin and Gerhard Widmer. Production of staccato articulation in Mozart sonatas played on a grand piano. Preliminary results. *Speech, Music, and Hearing. Quarterly Progress and Status Report*, 2000(4):1–6, 2000.

Emilios Cambouropoulos. Score Extraction from MIDI Files. In *In Proceedings of the 13th Colloquium on Musical Informatics (CIM'2000)*. L'Aquila, Italy, 2000.

Antonio Camurri, Carol L. Krumhansl, Barbara Mazzarino, and Gualterio Volpe. An exploratory study of aniticipation human movement in dance. In *Proceedings of the 2nd International Symposium on Measurement, Analysis, and Modeling of Human Functions*. Genova, Italy, 2004.

Antonio Camurri and Gualterio Volpe, editors. *Gesture-Based Communication in Human-Computer Interaction*. Springer, Berlin, 2004. LNAI 2915.

Antonio Camurri, Gualterio Volpe, Giovanni De Poli, and Marc Leman. Communicating expressiveness and affect in multimodal interactive systems. *IEEE Multimedia*, 12(1):43–53, 2005.

Sergio Canazza, Giovanni De Poli, Carlo Drioli, Antonio Rodà, and Alvise Vidolin. Modeling and control of expressiveness in music performance. *Proceedings of the IEEE*, 92(4):686–701, 2004.

Sergio Canazza, Giovanni De Poli, G. Mion, Antonio Rodà, Alvise Vidolin, and Patrick Zanon. Expressive classifiers at CSC: An overview of the main research streams. In *Proceedings of the XIV Colloquium on Musical Informatics (XIV CIM 2003) May 8–10*. Firenze, 2003.

P. Cano, A. Loscos, and J. Bonada. Score-performance matching using HMMs. In *Proceedings of the International Computer Music Conference*, pages 441–444. International Computer Music Association, 1999.

Eric F. Clarke. Timing in the performance of Erik Satie's 'Vexations'. *Acta Psychologica*, 50(1): 1–19, 1982.

Eric F. Clarke. Some aspects of rhythm and expression in performances of Erik Satie's "Gnossienne No. 5". *Music Perception*, 2:299–328, 1985.

Eric F. Clarke. Empirical methods in the study of performance. In Eric F. Clarke and Nicholas Cook, editors, *Empirical Musicology. Aims, Methods, and Prospects*, pages 77–102. University Press, Oxford, 2004.

Eric F. Clarke and W. Luke Windsor. Real and simulated expression: A listening study. *Music Perception*, 17(3):277–313, 2000.

Alcedo Coenen and Sabine Schäfer. Computer-controlled player pianos. *Computer Music Journal*, 16(4):104–111, 1992.

Geoffrey L. Collier and James Lincoln Collier. A study of timing in two Louis Armstrong solos. *Music Perception*, 19(3):463–483, 2002.

Nicholas Cook. Structure and performance timing in Bach's C major prelude (WTC I): An empirical study. *Music Analysis*, 6(3):100–114, 1987.

Nicholas Cook. The conductor and the theorist: Furtwängler, Schenker and the first movement of Beethoven's Ninth Symphony. In John Rink, editor, *The Practice of Performance*, pages 105–125. Cambridge University Press, Cambridge, UK, 1995.

Nicholas Cook. Computational and comparative Musicology. In Eric F. Clarke and Nicholas Cook, editors, *Empirical Musicology. Aims, Methods, and Prospects*, pages 103–126. University Press, Oxford, 2004.

Sophia Dahl. Playing the accent – comparing striking velocity and timing in an ostinato rhythm performed by four drummers. *Acta Acustica*, 90(4):762–776, 2004.

Sophia Dahl. Movements and analysis of drumming. In Eckart Altenmüller, J. Kesselring, and M. Wiesendanger, editors, *Music, Motor Control and the Brain*, page in press. University Press, Oxford, 2005.

R.B. Dannenberg. An on-line algorithm for real-time accompaniment. In *Proceedings of the International Computer Music Conference*, pages 193–198, 1984.

R.B. Dannenberg, B. Thom, and D. Watson. A machine learning approach to musical style recognition. In *Proceedings of the 1997 International Computer Music Conference*, pages 334–347, San Francisco, CA, Oct. 1997. International Computer Music Association.

Jane W. Davidson. Visual perception of performance manner in the movements of solo musicians. *Psychology of Music*, 21(2):103–113, 1993.

Jane W. Davidson. What type of information is conveyed in the body movements of solo musician performers? *Journal of Human Movement Studies*, 26(6):279–301, 1994.

Jane W. Davidson and Jorge Salgado Correia. Body movement. In Richard Parncutt and Gary McPherson, editors, *The Science and Psychology of Music Performance. Creating Strategies for Teaching and Learning*, pages 237–250. University Press, Oxford, 2002.

Giovanni De Poli. Methodologies for expressiveness modelling of and for music performance. *Journal of New Music Research*, 33(3):189–202, 2004.

Peter Desain and Henkjan Honing. Does expressive timing in music performance scale proportionally with tempo? *Psychological Research*, 56:285–292, 1994.

Peter Desain and Henkjan Honing. Modeling continuous aspects of music performance: Vibrato and Portamento. In Bruce Pennycook and Eugenia Costa-Giomi, editors, *Proceedings of the 4th International Conference on Music Perception and Cognition (ICMPC'96)*. Faculty of Music, McGill University, Montreal, Canada, 1996.

Roberto Dillon. A statistical approach to expressive intention recognition in violin performances. In Roberto Bresin, editor, *Proceedings of the Stockholm Music Acoustics Conference (SMAC'03), August 6–9, 2003*, pages 529–532. Department of Speech, Music, and Hearing, Royal Institute of Technology, Stockholm, Sweden, 2003.

S. Dixon. Extraction of musical performance parameters from audio data. In *Proceedings of the First IEEE Pacific-Rim Conference on Multimedia*, pages 42–45, Sydney, 2000. University of Sydney.

S. Dixon. Analysis of musical content in digital audio. In J. DiMarco, editor, *Computer Graphics and Multimedia: Applications, Problems, and Solutions*, pages 214–235. Idea Group, Hershey PA, 2004.

S. Dixon. Live tracking of musical performances using on-line time warping. 2005a. Submitted.

S. Dixon. An on-line time warping algorithm for tracking musical performances. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2005b. to appear.

S. Dixon and G. Widmer. MATCH: A music alignment tool chest. In *6th International Conference on Music Information Retrieval*, page Submitted, 2005.

Simon Dixon. Automatic extraction of tempo and beat from expressive performances. *Journal of New Music Research*, 30(1):39–58, 2001a.

Simon Dixon. Learning to detect onsets of acoustic piano tones. In Claudia Lomeli Buyoli and Ramon Loureiro, editors, *MOSART Workshop on current research directions in computer music, November 15–17, 2001*, pages 147–151. Audiovisual Institute, Pompeu Fabra University, Barcelona, Spain, 2001b.

Simon Dixon, Werner Goebl, and Emilios Cambouropoulos. Smoothed tempo perception of expressively performed music. *Music Perception*, 23:in press, 2005.

Simon Dixon, Werner Goebl, and Gerhard Widmer. The Performance Worm: Real time visualisation based on Langner's representation. In Mats Nordahl, editor, *Proceedings of the 2002 International Computer Music Conference, Göteborg, Sweden*, pages 361–364. International Computer Music Association, San Francisco, 2002.

Kurt Ebhardt. Zwei Beiträge zur Psychologie des Rhythmus und des Tempo. *Zeitschrift für Psychologie und Physiologie der Sinnesorgane*, 18:99–154, 1898.

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

Anders Friberg. *A Quantitative Rule System for Musical Performance*. Doctoral dissertation, Department of Speech, Music and Hearing, Royal Institute of Technology, Stockholm, 1995.

Anders Friberg. A fuzzy analyzer of emotional expression in music performance and body motion. In Johan Sundberg and B. Brunson, editors, *Proceedings of Music and Music Science, October 28–30, 2004*. Royal College of Music in Stockholm, Stockholm, 2005. CD-ROM.

Anders Friberg, Roberto Bresin, Lars Frydén, and Johan Sundberg. Musical punctuation on the mircolevel: Automatic identification and performance of small melodic units. *Journal of New Music Research*, 27(3):271–292, 1998.

Anders Friberg, Vittorio Colombo, Lars Frydén, and Johan Sundberg. Generating musical performances with Director Musices. *Computer Music Journal*, 24(3):23–29, 2000a.

Anders Friberg, Erwin Schoonderwaldt, and Patrik N. Juslin. CUEX: An algorithm for extracting expressive tone variables from audio recordings. *Acoustica united with Acta Acoustica*, in press, 2005.

Anders Friberg, Erwin Schoonderwaldt, Patrik N. Juslin, and Roberto Bresin. Automatic Real-Time Extraction of Musical Expression. In *Proceedings of the 2002 International Computer Music Conference, Göteborg, Sweden*, pages 365–367. International Computer Music Association, San Francisco, 2002.

Anders Friberg and Johan Sundberg. Does music performance allude to locomotion? A model of final ritardandi derived from measurements of stopping runners. *Journal of the Acoustical Society of America*, 105(3):1469–1484, 1999.

Anders Friberg, Johan Sundberg, and Lars Frydén. Music from motion: Sound level envelopes of tones expressing human locomotion. *Journal of New Music Research*, 29(3):199–210, 2000b.

Anders Friberg and A. Sundström. Swing ratios and ensemble timing in jazz performance: Evidence for a common rhythmic pattern. *Music Perception*, 19(3):333–349, 2002.

Alf Gabrielsson. Once again: The Theme from Mozart's Piano Sonata in A Major (K.331). In Alf Gabrielsson, editor, *Action and Perception in Rhythm and Music*, volume 55, pages 81–103. Publications issued by the Royal Swedish Academy of Music, Stockholm, Sweden, 1987.

Alf Gabrielsson. Music Performance. In Diana Deutsch, editor, *Psychology of Music*, pages 501–602. Academic Press, San Diego, 2nd edition, 1999.

Alf Gabrielsson. Music performance research at the millenium. *Psychology of Music*, 31(3):221–272, 2003.

Alf Gabrielsson, Ingmar Bengtsson, and Barbro Gabrielsson. Performance of musical rhythm in 3/4 and 6/8 meter. *Scandinavian Journal of Psychology*, 24:193–213, 1983.

Alf Gabrielsson and Eric Lindström. The influence of musical structure on emotional expression. In Patrik N. Juslin and John A. Sloboda, editors, *Music and Emotion: Theory and Research*, pages 223–248. Oxford University Press, New York, 2001.

Werner Goebl. Melody lead in piano performance: Expressive device or artifact? *Journal of the Acoustical Society of America*, 110(1):563–572, 2001.

Werner Goebl. *The Role of Timing and Intensity in the Production and Perception of Melody in Expressive Piano Performance*. Doctoral thesis, Institut für Musikwissenschaft, Karl-Franzens-Universität Graz, Graz, Austria, 2003. available online at http://www.ofai.at/music.

Werner Goebl and Roberto Bresin. Are computer-controlled pianos a reliable tool in music performance research? Recording and reproduction precision of a Yamaha Disklavier grand piano. In Claudia Lomeli Buyoli and Ramon Loureiro, editors, *MOSART Workshop on Current Research Directions in Computer Music, November 15–17, 2001*, pages 45–50. Audiovisual Institute, Pompeu Fabra University, Barcelona, Spain, 2001.

Werner Goebl and Roberto Bresin. Measurement and reproduction accuracy of computer-controlled grand pianos. *Journal of the Acoustical Society of America*, 114(4):2273–2283, 2003.

Werner Goebl and Simon Dixon. Analyses of tempo classes in performances of Mozart piano sonatas. In Henna Lappalainen, editor, *Proceedings of the Seventh International Symposium on Systematic and Comparative Musicology, Third International Conference on Cognitive Musicology, August 16–19, 2001*, pages 65–76. University of Jyväskylä, Jyväskylä, Finland, 2001.

Werner Goebl, Elias Pampalk, and Gerhard Widmer. Exploring expressive performance trajectories: Six famous pianists play six Chopin pieces. In Scott D. Lipscomp, Richard Ashley, Robert O. Gjerdingen, and Peter Webster, editors, *Proceedings of the 8th International Conference on Music Perception and Cognition, Evanston, IL, 2004 (ICMPC8)*, pages 505–509. Causal Productions, Adelaide, Australia, 2004. CD-ROM.

Hermann Gottschewski. *Die Interpretation als Kunstwerk. Musikalische Zeitgestaltung und ihre Analyse am Beispiel von Welte-Mignon-Klavieraufnahmen aus dem Jahre 1905*. Freiburger Beiträge zur Musikwissenschaft, Bd. 5. Laaber-Verlag, Laaber, 1996.

F. Gouyon, N. Wack, and S. Dixon. An open source tool for semi-automatic rhythmic annotation. In *Proceedings of the 7th International Conference on Digital Audio Effects*, pages 193–196, 2004.

Peter Hagmann. *Das Welte-Mignon-Klavier, die Welte-Philharmonie-Orgel und die Anfänge der Reproduktion von Musik*. Peter Lang, Bern, Frankfurt am Main, New York. URL `http://www.freidok.uni-freiburg.de/volltexte/608/`.

A. Hartmann. Untersuchungen über das metrische Verhalten in musikalischen Interpretationsvarianten. *Archiv für die gesamte Psychologie*, 84:103–192, 1932.

Hank Heijink, Peter Desain, Henkjan Honing, and Luke Windsor. Make me a match: An evaluation of different approaches to score-performance matching. *Computer Music Journal*, 24 (1):43–56, 2000a.

Hank Heijink, Luke Windsor, and Peter Desain. Data processing in music performance research: Using structural information to improve score-performance matching. *Behavior Research Methods, Instruments and Computers*, 32(4):546–554, 2000b.

C. P. Heinlein. A discussion of the nature of pianoforte damper-pedalling together with an experimental study of some individual differences in pedal performance. *Journal of General Psychology*, 2:489–508, 1929a.

C. P. Heinlein. The functional role of finger touch and damper-pedalling in the appreciation of pianoforte music. *Journal of General Psychology*, 2:462–469, 1929b.

C. P. Heinlein. Pianoforte damper-pedalling under ten different experimental conditions. *Journal of General Psychology*, 3:511–528, 1930.

M. T. Henderson. Rhythmic organization in artistic piano performance. In Carl Emil Seashore, editor, *Objective Analysis of Musical Performance*, volume IV of *University of Iowa Studies in the Psychology of Music*, pages 281–305. University Press, Iowa City, 1936.

M. T. Henderson, J. Tiffin, and Carl Emil Seashore. The Iowa piano camera and its use. In Carl Emil Seashore, editor, *Objective Analysis of Musical Performance*, volume IV, pages 252–262. University Press, Iowa City, 1936.

Kate Hevner. The affective value of pitch and tempo in music. *American Journal of Psychology*, 49: 621–630, 1937.

Rumi Hiraga, Roberto Bresin, Keiji Hirata, and Haruhiro Katayose. Rencon 2004: Turing Test for musical expression. In *Proceedings of the 2004 Conference on New Interfaces for Musical Expression (NIME04)*, pages 120–123. Hamamatsu, Japan, 2004.

David Miles Huber. *The MIDI Manual*. Butterworth-Heinemann, Boston, MA, 1999.

Patrik N. Juslin. Cue utilization in communication of emotion in music performance: Relating performance to perception. *Journal of Experimental Psychology: Human Perception and Performance*, 26(1797–1813), 2000.

Patrik N. Juslin. Studies of music performance: A theoretical analysis of empirical findings. In Roberto Bresin, editor, *Proceedings of the Stockholm Music Acoustics Conference (SMAC'03), August 6–9, 2003*, volume II, pages 513–516. Department of Speech, Music, and Hearing, Royal Institute of Technology, Stockholm, Sweden, 2003.

Patrik N. Juslin and Petri Laukka. Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research*, 33(3): 217–238, 2004.

Patrik N. Juslin and Guy Madison. The role of timing patterns in recognition of emotional expression from musical performance. *Music Perception*, 17(2):197–221, 1999.

Patrik N. Juslin and John A. Sloboda. *Music and Emotion: Theory and Research*. Oxford University Press, New York, 2001.

Bruce A. Kay, Michael T. Turvey, and Onno G. Meijer. An early oscillator model: Studies on the biodynamics of the piano strike (Bernstein & Popova, 1930). *Motor Control*, 7(1):1–45, 2003.

A. Klapuri. Automatic music transcription as we know it today. *Journal of New Music Research*, 33(3):269–282, 2004.

U. Kronman and Johan Sundberg. Is the musical retard an allusion to physical motion? In Alf Gabrielsson, editor, *Action and Perception in Rhythm and Music*, volume 55, pages 57–68. Publications issued by the Royal Swedish Academy of Music, Stockholm, Sweden, 1987.

Marc Leman, V. Vermeulen, De Voogdt L., J. Taelman, Dirk Moelants, and M. Lesaffre. Correlation of gestural musical audio cues and perceived expressive qualities. In Antonio Camurri and Gualterio Volpe, editors, *Gesture-based Communication in Human-Computer Interaction*, pages xx–yy. Springer, Berlin, 2004. LNAI 2915.

Eric Lindström, Antonio Camurri, Anders Friberg, Gualterio Volpe, and Marie-Louise Rinman. Affect, attitude and evaluation of multi-sensory performances. *Journal of New Music Research*, in press, 2005.

Tânia Lisboa, Aaron Williamon, Massimo Zicari, and Hubert Eiholzer. Mastery through imitation: A preliminary study. *Musicae Scientiae*, 9(1), 2005.

Dan Liu, Lu Lie, and Hong-Jiang Zhang. Automatic mood detection from acoustic music data. In *Proceedings of the International Symposium on Music Information Retrieval*. 2003.

Ramon López de Mántaras and Josep Lluís Arcos. AI and music: From composition to expressive performances. *AI Magazine*, 23(3):43–57, 2002.

Marco Maria. Unschärfetests mit hybriden Tasteninstrumenten. In Bernd Enders and Joachim Stange-Elbe, editors, *Global Village – Global Brain – Global Music. KlangArt Kongreß 1999*. Osnabrück, Germany, 1999.

Guerino Mazzola. *Geometrie der Töne. Elemente der Mathematischen Musiktheorie*. Birkhäuser Verlag, Basel, 1990.

Guerino Mazzola, editor. *The Topos of Music — Geometric Logic of Concepts, Theory, and Performance*. Birkhäuser Verlag, Basel, 2002.

Guerino Mazzola and Jan Beran. Rational composition of performance. In Reinhard Kopiez and Wolfgang Auhagen, editors, *Controlling Creative Processes in Music*, volume 12 of *Schriften zur Musikpsychologie and Musikästhetik, Bd. 12*, pages 37–68. Lang, Frankfurt/M., 1998.

Guerino Mazzola and Stefan Göller. Performance and interpretation. *Journal of New Music Research*, 31(3):221–232, 2002.

Guerino Mazzola and Oliver Zahorka. Tempo curves revisited: Hierarchies of performance fields. *Computer Music Journal*, 18(1):40–52, 1994.

Guerino Mazzola, Oliver Zahorka, and Joachim Stange-Elbe. Analysis and performance of a dream. In Anders Friberg and Johan Sundberg, editors, *Proceedings of the KTH Symposion on Grammars for Music Performance*, pages 59–68. Department of Speech Communication and Music Acoustics, Stockholm, Sweden, 1995.

Stephen McAdams, Philippe Depalle, and Eric F. Clarke. Analyzing musical sound. In Eric F. Clarke and Nicholas Cook, editors, *Empirical Musicology. Aims, Methods, and Prospects*, pages 157–196. University Press, Oxford, 2004.

Anne Midgette. Play it again, Vladimir (via computer). *The New York Times*, June 5 2005.

L. Mion and G. De Poli. Expressiveness detection of music performances in the kinematics energy space. In *Proc. Sound and Music Computing Conf. (JIM/CIM 04)*, pages 257–261, Paris, Oct. 2004.

Dirk Moelants. Temporal aspects of instrumentalists' performance of tremolo, trills, and vibrato. In *Proceedings of the International Symposium on Musical Acoustics (ISMA'04)*, pages 281–284. The Acoustical Society of Japan, Nara, Japan, 2004.

Robert A. Moog and Thomas L. Rhea. Evolution of the keyboard interface: The Bösendorfer 290 SE recording piano and the Moog multiply-touch-sensitive keyboards. *Computer Music Journal*, 14(2):52–60, 1990.

George P. Moore. Piano trills. *Music Perception*, 9(3):351–359, 1992.

T. Nakamura. The communication of dynamics between musicians and listeners through musical performance. *Perception and Psychophysics*, 41(6):525–533, 1987.

Nigel Nettheim. A musical microscope applied to the piano playing of Vladimir de Pachmann, 2001. http://users.bigpond.net.au/nettheim/pachmic/microsc.htm.

N. Orio and F. Déchelle. Score following using spectral analysis and hidden Markov models. In *Proceedings of the International Computer Music Conference*, pages 151–154, 2001.

Otto Ortmann. *The Physical Basis of Piano Touch and Tone*. Kegan Paul, Trench, Trubner; J. Curwen; E. P. Dutton, London, New York, 1925.

Otto Ortmann. *The Physiological Mechanics of Piano Technique*. Kegan Paul, Trench, Trubner, E. P. Dutton, London, New York, 1929. Paperback reprint: New York: E. P. Dutton 1962.

Caroline Palmer. Mapping musical thought to musical performance. *Journal of Experimental Psychology: Human Perception and Performance*, 15(12):331–346, 1989.

Caroline Palmer. The role of interpretive preferences in music performance. In Mari Riess Jones and Susan Holleran, editors, *Cognitive Bases of Musical Communication*, pages 249–262. American Psychological Association, Washington DC, 1992.

Caroline Palmer. Anatomy of a performance: Sources of musical expression. *Music Perception*, 13(3):433–453, 1996a.

Caroline Palmer. On the assignment of structure in music performance. *Music Perception*, 14(1): 23–56, 1996b.

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

Caroline Palmer and Judith C. Brown. Investigations in the amplitude of sounded piano tones. *Journal of the Acoustical Society of America*, 90(1):60–66, 1991.

Caroline Palmer and Simone Dalla Bella. Movement amplitude and tempo change in piano performance. *Journal of the Acoustical Society of America*, 115(5):2590, 2004.

Caroline Palmer and Susan Holleran. Harmonic, melodic, and frequency height influences in the perception of multivoiced music. *Perception and Psychophysics*, 56(3):301–312, 1994.

Caroline Palmer and Carla van de Sande. Units of knowledge in music performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(2):457–470, 1993.

Caroline Palmer and Carla van de Sande. Range of planning in music performance. *Journal of Experimental Psychology: Human Perception and Performance*, 21(5):947–962, 1995.

Richard Parncutt and Gary McPherson, editors. *The Science and Psychology of Music Performance. Creating Strategies for Teaching and Learning*. University Press, Oxford, New York, 2002.

Dirk-Jan Povel. Temporal structure of performed music: Some preliminary observations. *Acta Psychologica*, 41(4):309–320, 1977.

C. Raphael. A hybrid graphical model for aligning polyphonic audio with musical scores. In *Proceedings of the 5th International Conference on Musical Information Retrieval*, pages 387–394, 2004.

Bruno Heinrich Repp. Patterns of expressive timing in performances of a Beethoven minuet by nineteen famous pianists. *Journal of the Acoustical Society of America*, 88(2):622–641, 1990.

Bruno Heinrich Repp. Diversity and commonality in music performance: An analysis of timing microstructure in Schumann's "Träumerei". *Journal of the Acoustical Society of America*, 92(5): 2546–2568, 1992.

Bruno Heinrich Repp. Some empirical observations on sound level properties of recorded piano tones. *Journal of the Acoustical Society of America*, 93(2):1136–1144, 1993.

Bruno Heinrich Repp. On determining the basic tempo of an expressive music performance. *Psychology of Music*, 22:157–167, 1994a.

Bruno Heinrich Repp. Relational invariance of expressive microstructure across global tempo changes in music performance: an exploratory study. *Psychological Research*, 56(4):269–284, 1994b.

Bruno Heinrich Repp. Acoustics, perception, and production of legato articulation on a digital piano. *Journal of the Acoustical Society of America*, 97(6):3862–3874, 1995a. URL `http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m{\&}form=6{\&}dopt=r{\&}uid=000779066`

Bruno Heinrich Repp. Detectability of duration and intensity increments in melody tones: a partial connection between music perception and performance. *Perception and Psychophysics*, 57(8):1217–1232, 1995b.

Bruno Heinrich Repp. Expressive timing in Schumann's "Träumerei:" An analysis of performances by graduate student pianists. *Journal of the Acoustical Society of America*, 98(5):2413–2427, 1995c.

Bruno Heinrich Repp. Patterns of note onset asynchronies in expressive piano performance. *Journal of the Acoustical Society of America*, 100(6):3917–3932, 1996a.

Bruno Heinrich Repp. Pedal timing and tempo in expressive piano performance: A preliminary investigation. *Psychology of Music*, 24(2):199–221, 1996b.

Bruno Heinrich Repp. The art of inaccuracy: Why pianists' errors are difficult to hear. *Music Perception*, 14(2):161–184, 1996c.

Bruno Heinrich Repp. The dynamics of expressive piano performance: Schumann's "Träumerei" revisited. *Journal of the Acoustical Society of America*, 100(1):641–650, 1996d. URL `http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m{\&}form=6{\&}dopt=r{\&}uid=000867585`

Bruno Heinrich Repp. Acoustics, perception, and production of legato articulation on a computer-controlled grand piano. *Journal of the Acoustical Society of America*, 102(3):1878–1890, 1997a.

Bruno Heinrich Repp. Expressive timing in a Debussy Prelude: A comparison of student and expert pianists. *Musicae Scientiae*, 1(2):257–268, 1997b.

Bruno Heinrich Repp. The Aesthetic Quality of a Quantitatively Average Music Performance: Two Preliminary Experiments. *Music Perception*, 14(4):419–444, 1997c.

Bruno Heinrich Repp. The effect of tempo on pedal timing in piano performance. *Psychological Research*, 60(3):164–172, 1997d. URL http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m{\&}form=6{\&}dopt=r{\&}uid=000934296

Bruno Heinrich Repp. A microcosm of musical expression. I. Quantitative analysis of pianists' timing in the initial measures of Chopin's Etude in E major. *Journal of the Acoustical Society of America*, 104(2):1085–1100, 1998. URL http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m{\&}form=6{\&}dopt=r{\&}uid=000971492

Bruno Heinrich Repp. A microcosm of musical expression: II. Quantitative analysis of pianists' dynamics in the initial measures of Chopin's Etude in E major. *Journal of the Acoustical Society of America*, 105(3):1972–1988, 1999a. URL http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m{\&}form=6{\&}dopt=r{\&}uid=001008961

Bruno Heinrich Repp. A microcosm of musical expression: III. Contributions of timing and dynamics to the aesthetic impression of pianists' performances of the initial measures of Chopin's Etude in E major. *Journal of the Acoustical Society of America*, 106(1):469–478, 1999b. URL http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m{\&}form=6{\&}dopt=r{\&}uid=001042063

Kathleen Riley-Butler. Teaching expressivity: An aural–visual feedback–replication model. In *ESCOM 10th Anniversary Conference on Musical Creativity, April 5–8, 2002*. Université de Liège, Liège, Belgium, 2002. CD-ROM.

John Rink, editor. *The Practice of Performance: Studies in Musical Interpretation*. University Press, Cambridge UK, 1995.

John Rink, editor. *Musical Performance. A Guide to Understanding*. Cambridge University Press, Cambridge, UK, 2002.

John Rink. In respect of performance: The view from Musicology. *Psychology of Music*, 31(3): 303–323, 2003.

Curtis Roads. Bösendorfer 290 SE computer-based piano. *Computer Music Journal*, 10(3):102–103, 1986.

E.D. Scheirer. Using musical knowledge to extract expressive performance information from audio recordings. In H. Okuno and D. Rosenthal, editors, *Readings in Computational Auditory Scene Analysis*. Lawrence Erlbaum, 1997.

Erwin Schoonderwaldt and Anders Friberg. Towards a rule-based model for violin vibrato. In Claudia Lomeli Buyoli and Ramon Loureiro, editors, *MOSART Workshop on Current Research Directions in Computer Music, November 15–17, 2001*, pages 61–64. Audiovisual Institute, Pompeu Fabra University, Barcelona, Spain, 2001.

Charles H. Sears. A contribution to the psychology of rhythm. *American Journal of Psychology*, 13 (1):28–61, 1902.

Carl Emil Seashore, editor. *The Vibrato*, volume I of *University of Iowa Studies in the Psychology of Music*. University Press, Iowa City, 1932.

Carl Emil Seashore, editor. *Objective Analysis of Musical Performance*, volume IV of *University of Iowa Studies in the Psychology of Music*. University Press, Iowa City, 1936a.

Carl Emil Seashore, editor. *Psychology of the Vibrato in Voice and Instrument*, volume III of *University of Iowa Studies. Studies in the Psychology of Music Volume III*. University Press, Iowa City, 1936b.

M. Seif El-Nasr, J. Yen, and T. R. Iorger. FLAME – Fuzzy logic adaptive mode of emotions. *Autonomous Agents and Multi-Agent Systems*, 3(3):219–257, 2000.

L. Henry Shaffer. Analysing piano performance. In George E. Stelmach and Jean Requin, editors, *Tutorials in Motor Behavior*. North-Holland, Amsterdam, 1980.

L. Henry Shaffer. Performances of Chopin, Bach and Bartòk: Studies in motor programming. *Cognitive Psychology*, 13(3):326–376, 1981.

L. Henry Shaffer. Timing in solo and duet piano performances. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, 36A(4):577–595, 1984.

L. Henry Shaffer. How to interpret music. In Mari Riess Jones and Susan Holleran, editors, *Cognitive Bases of Musical Communication*, pages 263–278. American Psychological Association, Washington DC, 1992.

L. Henry Shaffer, Eric F. Clarke, and Neil P. McAngus Todd. Metre and Rhythm in Pianoplaying. *Cognition*, 20(1):61–77, 1985.

L. Henry Shaffer and Neil P. McAngus Todd. The interpretative component in musical performance. In Alf Gabrielsson, editor, *Action and Perception in Rhythm and Music*, volume 55, pages 139–152. Publications issued by the Royal Swedish Academy of Music, Stockholm, Sweden, 1987.

S. Shalev-Shwartz, J. Keshet, and Y. Singer. Learning to align polyphonic music. In *5th International Conference on Music Information Retrieval*, pages 381–386, 2004.

K. Sjölander and J. Beskow. WaveSurfer – an open source speech tool. In *Proceedings of the International Conference on Spoken Language Processing*, 2000.

Laila Skinner and Carl Emil Seashore. A musical pattern score of the first movement of the Beethoven Sonata, Opus 27, No. 2. In Carl Emil Seashore, editor, *Objective Analysis of Musical Performance*, volume IV of *Studies in the Psychology of Music*, pages 263–279. University Press, Iowa City, 1936.

F. Soulez, X. Rodet, and D. Schwarz. Improving polyphonic and poly-instrumental music to score alignment. In *4th International Conference on Music Information Retrieval*, pages 143–148, 2003.

Johan Sundberg. Four years of research on music and motion. *Journal of New Music Research*, 29 (3):183–185, 2000.

Johan Sundberg, Anders Askenfelt, and Lars Frydén. Musical performance. A synthesis-by-rule approach. *Computer Music Journal*, 7:37–43, 1983.

Johan Sundberg, Anders Friberg, and Lars Frydén. Rules for automated performance of ensemble music. *Contemporary Music Review*, 3:89–109, 1989.

Johan Sundberg, Anders Friberg, and Lars Frydén. Threshold and preference quantities of rules for music performance. *Music Perception*, 9(1):71–92, 1991.

Renee Timmers. *Freedom and Constraints in Timing and Ornamentation*. Shaker Publishing, Maastricht, 2002.

Renee Timmers, Richard Ashley, Peter Desain, Henkjan Honing, and Luke W. Windsor. Timing of ornaments in the theme of Beethoven's Paisello Variations: Empirical data and a model. *Music Perception*, 20(1):3–33, 2002.

Asmir Tobudic and Gerhard Widmer. Relational IBL in classical music. *Machine Learning*, to appear, 2005.

Neil P. McAngus Todd. A model of expressive timing in tonal music. *Music Perception*, 3(1):33–58, 1985.

Neil P. McAngus Todd. A computational model of Rubato. *Contemporary Music Review*, 3:69–88, 1989.

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

Neil P. McAngus Todd. The kinematics of musical expression. *Journal of the Acoustical Society of America*, 97(3):1940–1949, 1995.

R. Turetsky and D. Ellis. Ground-truth transcriptions of real music from force-aligned MIDI syntheses. In *4th International Conference on Music Information Retrieval*, pages 135–141, 2003.

Leroy Ninde Vernon. Synchronization of chords in artistic piano music. In Carl Emil Seashore, editor, *Objective Analysis of Musical Performance*, volume IV of *Studies in the Psychology of Music*, pages 306–345. University Press, Iowa City, 1936.

Gerhard Widmer. A machine learning analysis of expressive timing in pianists' performances of Schumann's "Träumerei". In Anders Friberg and Johan Sundberg, editors, *Proceedings of the KTH Symposion on Grammars for Music Performance*, pages 69–81. Department of Speech Communication and Music Acoustics, Stockholm, Sweden, 1995a.

Gerhard Widmer. Modeling rational basis for musical expression. *Computer Music Journal*, 19(2): 76–96, 1995b.

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

Gerhard Widmer. Large-scale induction of expressive performance rules: first quantitative results. In Ioannis Zannos, editor, *Proceedings of the 2000 International Computer Music Conference,*

Berlin, Germany, pages 344–347. International Computer Music Association, San Francisco, 2000.

Gerhard Widmer. Using AI and machine learning to study expressive music performance: Project survey and first report. *AI Communications*, 14(3):149–162, 2001.

Gerhard Widmer. In search of the Horowitz factor: Interim report on a musical discovery project. In *Proceedings of the 5th International Conference on Discovery Science (DS'02), Lübeck, Germany*. Springer, Berlin, 2002a.

Gerhard Widmer. Machine discoveries: A few simple, robust local expression principles. *Journal of New Music Research*, 31(1):37–50, 2002b.

Gerhard Widmer. Discovering simple rules in complex data: A meta-learning algorithm and some surprising musical discoveries. *Artificial Intelligence*, 146(2):129–148, 2003.

Gerhard Widmer. Studying a creative act with computers: Music performance studies with automated discovery methods. *Musicae Scientiae*, 9(1):11–30, 2005.

Gerhard Widmer, Simon Dixon, Werner Goebl, Elias Pampalk, and Asmir Tobudic. In search of the Horowitz factor. *AI Magazine*, 24(3):111–130, 2003.

Gerhard Widmer and Werner Goebl. Computational models of expressive music performance: The state of the art. *Journal of New Music Research*, 33(3):203–216, 2004.

Gerhard Widmer and Asmir Tobudic. Playing Mozart by analogy: Learning multi-level timing and dynamics strategies. *Journal of New Music Research*, 32(3):259–268, 2003.

Luke Windsor. Data collection, experimental design, and statistics in musical research. In Eric F. Clarke and Nicholas Cook, editors, *Empirical Musicology. Aims, Methods, and Prospects*, pages 197–222. University Press, Oxford, 2004.

Luke Windsor, Peter Desain, R. Aarts, Hank Heijink, and Renee Timmers. The timing of grace notes in skilled musical performance at different tempi. *Psychology of Music*, 29(2):149–169, 2001.

Chapter 5

Controlling Sound with Senses and Influencing Senses with Sound

Edited by:
Roberto Bresin[1], Gualtiero Volpe[2]

[1] KTH, Department of Speech, Music and Hearing, Sweden
[2] InfoMus Lab - DIST - University of Genova, Italy

5.1 Introduction

The problem of effectively controlling sound generation and processing has always been relevant for music research in general and Musical Informatics in particular. Moreover, an important aspect of research on control concerns the perceptual, cognitive, affective mechanisms affecting sound and music control from the study of the mechanisms involved in the control by musicians of traditional acoustic instruments to the novel opportunities offered by modern Digital Music Instruments. More recently, the problem of defining effective strategies for the real-time control of multimodal interactive systems, with particular reference to music but not limited to it, is receiving a growing interest from the scientific community since its relevance also for future research and applications in broader fields of human-computer interaction.

In this framework, research on control extended its scope to include for example analysis

of human movement and gesture (not only gestures of musicians playing an instruments but also gestures of subjects interacting with interactive 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-fertilization between scientific and technical knowledge on the one side, and art and humanities on the other side. Such need of cross-fertilization opens new perspectives to research in both fields: if from the one hand scientific and technological research can benefit from models and theories borrowed from psychology, social science, art, and humanities, on the other hand these disciplines can take advantage of the tools technology can provide for their own research, i.e., for investigating the hidden subtleties of human beings at a depth that was never reached 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 of the need and the importance of such cross-fertilization.

This work briefly surveys some relevant aspects of research on control putting into evidence research issues, achieved results, and problems that are still open for the future.

A first aspect concerns the development of a conceptual framework envisaging control at different levels, from the low-level analysis of audio signals, to feature extraction, to the identification and analysis of significant musical structures (note groups, phrases), to the high-level association of semantic descriptions including affective, emotional content. Moreover, such conceptual framework does not have to be limited to the music domain, but it needs to be applied to other modalities (e.g., movement and gesture) too. In particular, it has to enable multimodal and cross-modal processing and associations, i.e., it should include such a level of abstraction that features at that level do not belong to a given modality, rather they emerge from modalities and can be used for mapping between modalities. The definition of such high-level control spaces is still an open research issue deserving particular attention in the future. Chapter 5.2 presents a conceptual framework worked out in the EU-IST Project MEGA (Multisensory Expressive Gesture Applications) that can be considered as a starting point for research on this direction.

A second aspect is related to the definition of suitable scientific methodologies for investigated the subtleties involved in sound and music control. For example, an important topic for control research is gesture analysis of both performers and interacting subjects. Such analysis can be performed at different layers, from the tracking of the positions of given body parts, to the interpretation and classification of gestures in term of expressive, emotional content. Also different perspectives are possible. Chapter 5.3 provides an overview of some consolidated scientific methodologies for gesture analysis with particular focus on performing arts (dance and music performers) and presents different perspectives for analysis of music and dance.

Moving from these foundational issues, Chapter 5.4 and Chapter 5.5 address concrete examples of control problems. Chapter 5.4 focuses on control of music performance with a particular emphasis on the role of the affective, emotional information. It illustrates the problems involved with the analysis of expressive gestures of music performers and their mapping into synthesis of emotional expression in music. Chapter 5.5 deals with control issues related with sound production involving both control of traditional acoustic and digital musical instruments and control of sounding objects, i.e., the problem of effectively controlling (physical) sound models that are responsive to physical interactions and are easily matched to physical objects. These issues are illustrated with concrete examples of applications.

A common key issue for the research discussed in these Chapters is the development of suitable strategies for controlling and/or generating sound and music output in real-time. That is, even if algorithms able to correctly and reliably interpret information from performers/users including high-level expressive information from gesture were available, the problem of if and how to use such information especially in an artistic performance still remains very open. The problem is even more difficult to face since it often directly involves the artistic choices of the designer of the performance, i.e., how much degrees of freedom the designer wishes to leave to the automatic systems: in other words the role of technology in the artwork and, from a certain point of view, the concept of artwork.

Another key issue for the design of effective control strategies able to fully process and exploit possible high-level information is the development of multimodal and cross-modal algorithms and the identification of cross-modal features. Chapter 5.6 addresses this topic also discussing some concrete examples.

Finally some software often employed tools (e.g., the new version of the EyesWeb open platform for multimodal processing, Director Musices for expressive music performance), are briefly introduced along the book and some conclusions are drawn with a particular focus on the

most promising research topics that still need to be addressed in the future.

5.2 A conceptual framework for gestural control of interactive systems

A relevant foundational aspect for research in sound and music control concerns the development of a conceptual framework envisaging control at different levels, from the low-level analysis of audio signals, toward high-level semantic descriptions including affective, emotional content.

Such conceptual framework does not have to be limited to the music domain, but it needs to be applied to other modalities (e.g., movement and gesture) too. In particular, it has to enable multimodal and cross-modal processing and associations, i.e., it should include such a level of abstraction that features at that level do not belong to a given modality, rather they emerge from modalities and can be used for mapping between modalities.

This Chapter presents a conceptual framework worked out in the EU-IST Project MEGA (Multisensory Expressive Gesture Applications, 2000-2003) that can be considered as a starting point for research on this direction.

Research in the MEGA project moved from the assumption that the physical stimuli that make up an artistic environment contain information about expressiveness that can, to some extend, be extracted and communicated. With multiple modalities (music, video, computer animation) this allows the transmission of expressiveness parameters from one domain to another domain, for example from music to computer animation, or from dance to music. That is, expressive parameters are an example of parameters emerging from modalities and independent of them. In other words, expressive parameters define a cross-modal control space that is at a higher level with respect to modalities.

A main question in MEGA research thus relates to the nature of the physical cues that carry the 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 5.1 sketches the conceptual framework in terms

of (i) a syntactical layer that stands for the analysis and synthesis of physical properties (bottom), (ii) a semantic layer that contains descriptions of affects, emotions, and expressiveness (top), and (iii) 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 synthesize physical properties of that affect in another modality, such as animation. This path is followed, for example, when sadness is expressed in the music, and when an avatar is displaying this sadness in his posture.

5.2.1 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 Lesaffre et al. [2003] worked out a useful taxonomy of concepts that gives a structured understanding of this layer in terms of a number of justified distinctions. 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 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 (related to repetition in the energy of the signal), dynamics (intensity), roughness (or sensory dissonance), tonal tension (or the correlation between local pitch patterns and global or contextual pitch patterns), and so on.

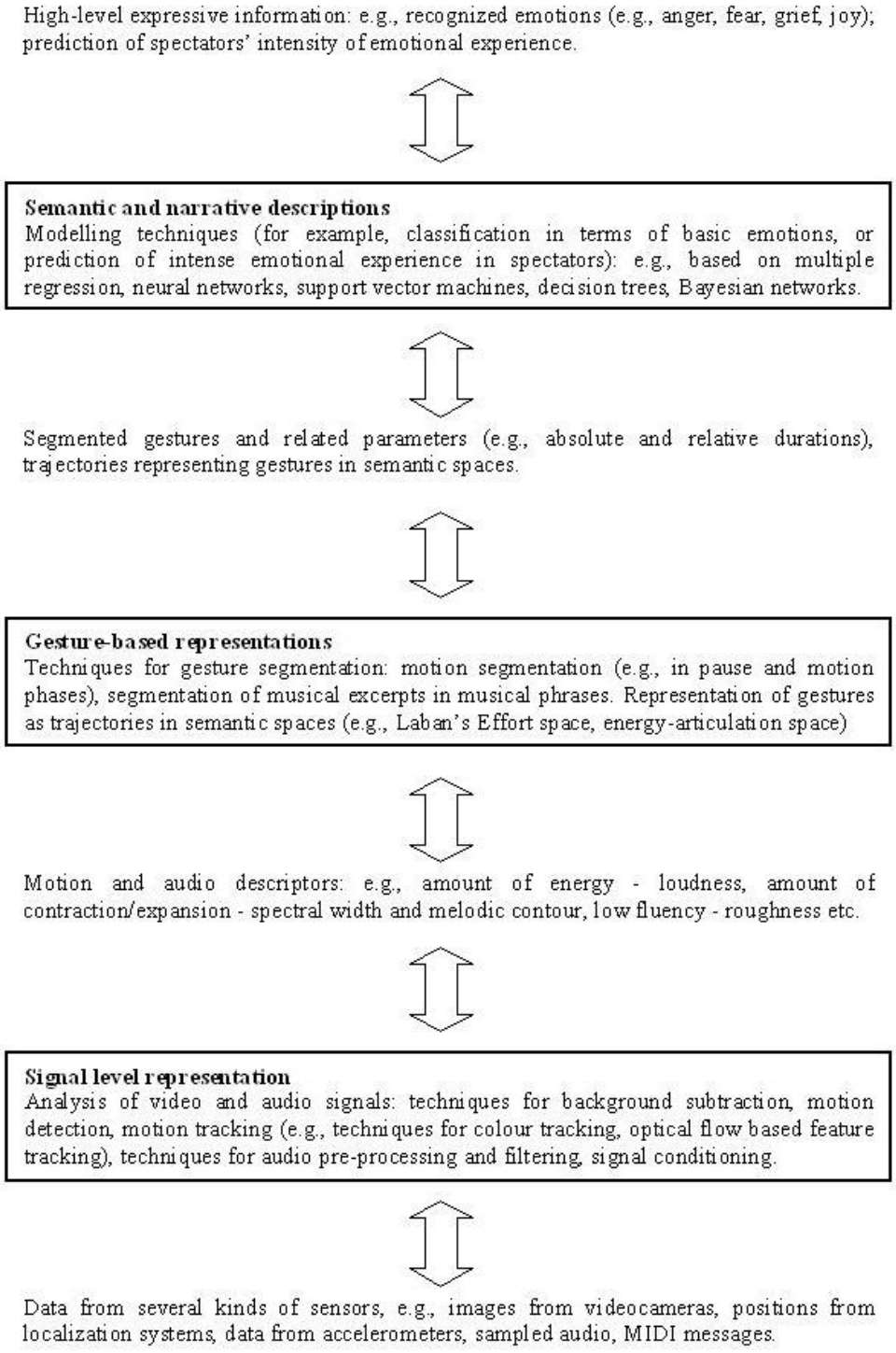

Figure 5.1: The layered conceptual framework makes a distinction between syntax and semantics, and in between, a connection layer that consists of affect / emotion / expressiveness (AEE) spaces and mappings.

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 5.1 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. Whether all these features are relevant in a context of affect processing and communication of expressiveness is another matter. The experiments discussed in the next sections tried to shed some light on this issue.

In the domain of movement (dance) analysis, a similar approach can be envisaged that draws 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"), 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 (see Camurri et al. [2003]), 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, kinematical 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 characterized 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).

The distinction between low-level, mid-level, and high-level descriptors will be further discussed in 5.3 as a possible perspective for gesture analysis.

5.2.2 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 be studied by allowing a large number of listeners to use adjectives (either on a completely free basis, or taken from an elaborate list) to specify the affective content of musical excerpts. Afterwards, the data is analyzed and clustered into categories. The early results of Hevner Hevner [1936], for example, showed that listeners tend to use 8 different categories of affect attribution. For a recent overview, see Sloboda and Juslin [2001]. 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.

5.2.3 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, dignity, 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, anger, fear and activity and low pitch may suggest sadness, dignity, excitement as well as boredom and pleasantness, and so on (see overviews in Juslin and Laukka [2003], Gabrielsson and Lindström [2001]). Leman and colleagues Leman et al. [2005] show that certain automatically extracted low-level features can be determinants of affect attribution and that maps can be designed that connect audio features with affect/emotion/expression descriptors. Bresin and Friberg Bresin and Friberg [2000b] synthesised music performances starting from a semantic map representing basic emotions. Using qualita-

tive cue descriptions from previous experiments, as listed above, each emotional expression was modeled in terms of a set of rule parameters in a performance rule system. This yielded a fine control of performance parameters relating to performance principles used by musicians such as phrasing and microtiming. A listening experiment was carried out confirming the ability of the synthesized performances to convey the different emotional expressions. Kinaesthetic 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 Canazza et al. [2003b]. 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, by using control based on timing and on dynamics of the notes. The MIDI parameters typically control tempo and key velocity. The audio-parameters control tempo, legato, loudness, brightness, attack time, vibrato, and envelope shape.

In human movement and dance the relationship between syntactical features and affect semantics has been investigated in several studies. For example, in the tradition of the work by Johansson [1973], it has been shown that it is possible for human observers to perceive emotions in dance from point light displays Walk and Homan [1984], Dittrich et al. [1996]. Pollick Pollick et al. [2001] analyzed recognition of emotion in everyday movements (e.g., drinking, knocking) and found significant correlations between motion kinematics (in particular speed) and the activation axis in the two-dimensional space having as axes activation and valence as described by Russell Russell [1980] with respect to his circumplex structure of affect. Wallbott Wallbott [2001] in his paper dealing with measurement of human expression after reviewing a collection of works concerning movement features related with expressiveness and techniques to extract them (either manually or automatically), classified these features by considering six different aspects: spatial aspects, temporal aspects, spatio-temporal aspects, aspects related to "force" of a movement, "gestalt" aspects, categorical approaches. Boone and Cunningham Boone and Cunningham [1998] starting from previous studies by De Meijer Meijer [1989] identified six expressive cues involved in the recognition of the four basic emotions anger, fear, grief, and happiness, and further tested the ability of children in recognizing emotions in expressive body movement through these cues. Such six cues are "frequency of upward arm movement, the duration of time arms were kept close to the body, the amount of muscle tension, the duration of time an individual leaned forward, the number of directional changes in face and torso, and the number of tempo changes an individual made in a given action sequence"

Boone and Cunningham [1998].

5.3 Methodologies, perspectives, and tools for gesture analysis

Antonio Camurri, Barbara Mazzarino, Gualtiero Volpe
InfoMus Lab - DIST - University of Genova

Discovering the key factors that characterize 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 characterization 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. One such scenario is dance (see for example Camurri et al. [2004c]). Another is music performance (see for example Dahl and Friberg [2004]). This chapter illustrates some consolidates approaches to gesture analysis and possible perspectives under which gesture analysis can be performed.

5.3.1 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 "microdance" 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 microdance 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 microdance performance. Spectators' ratings are used for evaluation and compared with the output of developed computational models (e.g., for the analysis of expressiveness). Moreover, microdances can also be used for testing feature extraction algorithms by comparing the outputs of the algorithms with spectators' rating of the same microdance performance (see for example Camurri et al. [2004b] for a work on spectators'

expectation with respect to the motion of the body center of gravity).

5.3.2 Subtractive approach

Microdances can be useful to isolate factors related to KANSEI and 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 microdance performed with different expressive intentions. Toward this aim, 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 utilized 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 feeling in the spectator but such the feeling is a negative one). The transitions between positive and negatives rates and a rate of zero (i.e. no expressiveness was found by the spectator in the analyzed 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 what information has to be subtracted) and the rating phase. This subtractive approach is different from the previous studies by Johansson [1973] and from more recent results Cowie et al. [2001] where it is demonstrated that a limited number of visible points on human joints allow an observer to recognize information on movement, including certain emotional content.

5.3.3 Space views

Gesture can be analyzed under different perspectives. Two of these aspects, space and time, are briefly discussed here.

A first aspect concerns the space under analysis, i.e. into which extent it is considered and which level of detail is assumed in the analysis. In his book "Modern Educational Dance" choreographer Rudolf Laban Laban [1963] introduces two relevant concepts: the Kinesphere,

also referred to as Personal Space, and the General Space, the whole space surrounding the Kinesphere. Laban says:

> Whenever the body moves or stands, it is surrounded by space. Around the body is the sphere of movement, or Kinesphere, the circumference of which can be reached by normally extended limbs without changing one's stance, that is, the place of support. (...) Outside this immediate sphere lies the wider or "general" space which man can enter only by moving away from their original stance. (p. 85).

A first distinction can thus be made between analysis in the Personal Space and analysis in the General Space. Further subdivisions can be made depending on the envisaged level of detail. For example, it is possible to consider the motion of only one person within the General Space or the motion of groups in order to analyze the behavior of the group as a whole.

In the Personal Space it is possible to consider global features, e.g. the global amount of detected motion, or local features, e.g. describing the motion of a given joint or of a given part of the body.

These subdivisions should not be considered as rigid and static ones, but rather as a continuum of possibilities through which the focus of attention dynamically moves. Many analyzes at each of the four levels of detail can be carried out in parallel and their results integrated toward a global interpretation of the detected movement. Moreover, the space view to be considered also depends on the kind of gesture under analysis. For example, in the analysis of a dance ensemble both the analysis in the General Space of the whole ensemble and the analysis in the Personal Space of each dancer will be important. In analysis of gesture of music performers movements in the Personal Space of each performer will usually be the focus of the analysis.

5.3.4 Time views

Time also plays an important role, mainly with respect to the time interval during which analyzes are carried out. This can vary from a few milliseconds (e.g., one frame from a videocamera) to several minutes (a whole performance) and it depends on the evolution of the performance and its narrative structure as well as on considerations about how movement/music is perceived by humans with respect to time.

As for analysis of music, a taxonomy of descriptors of musical audio has been worked out by Leman and colleagues (see for example Lesaffre et al. [2003] and Chapter 5.2) in the context of audio mining. A distinction is made among non-contextual "low level descriptors obtained from a frame-based analysis of the acoustical wave", mid-level descriptors "derived from musical context dependencies within time-scales of about 3 seconds" and allowing an event-based representation of musical objects, and high-level descriptors referring to time intervals longer than 3 seconds, and related to the cognitive and emotional/affective domains.

A similar approach can also be envisaged for motion descriptors. That is, it is possible to distinguish between descriptors calculated on different time scales. Low-level descriptors are calculated over a time interval of a few milliseconds (e.g. one or a few frames from a videocamera). For example the current amount of contraction/expansion can be calculated with just one frame. Mid-level descriptors are calculated over time durations of a few seconds. Examples of such descriptors are the overall direction of the movement in a gesture (e.g. upward or downward) or its directness. At this level it is also possible to segment movement in gestures and to develop an event-based representation of movement. High-level descriptors are related to the conveyed expressive content (but also to cognitive aspects) and refer to sequences of gestures. Time intervals in the case of dance performances range from a gesture (some seconds), to a microdance, to a whole dance performance (several minutes).

The time aspect is to a large extent complementary to the space aspect, i.e. it is possible to have low-level, mid-level, and high-level descriptors for the movement of a limb in the Personal Space, for the movement of the full-body in the Personal Space, and for the movement of individuals and groups in the General Space.

5.3.5 Examples of motion descriptors

Here we provide some examples of motion descriptors with details on how to extract them. With respect to the approaches discussed above these descriptors can be used in the bottom-up approach for characterizing motion (e.g., microdances). The top-down approach can be used for validating the descriptors with respect to their role and contribute in conveying expressive content.

Extraction of motion descriptors follows the layered conceptual framework described in Chapter 5.2. We also refer to dance (and to microdances) for presenting extraction of motion descriptors.

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 Figure 5.2 shows the extraction of a Silhouette Motion Image (SMI). A SMI is an image carrying information about variations of the shape and position of the dancer's silhouette in the last few frames. SMIs are inspired to MEI and MHI Bobick and Davis [2001]. We also use an extension of SMIs taking into account the internal motion in silhouettes.

Figure 5.2: The SMI is represented as the red area in the picture.

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 pixels) of a 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 Camurri et al. [2003] 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, e.g., Boone and Cunningham [1998], 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, e.g., length and overall direction of motion trajectories.
- Kinematical cues, e.g., velocity and acceleration on 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 5.3). 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 bells approach can also be applied also to sound signal analysis.

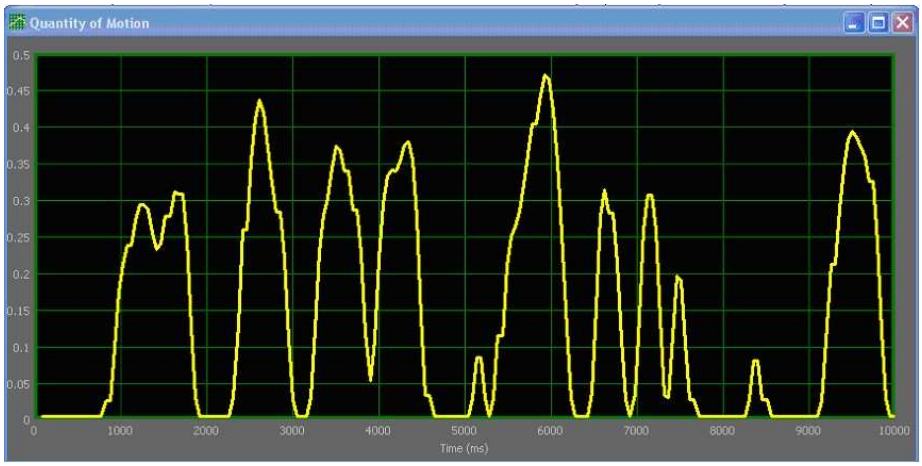

Figure 5.3: Motion bells and motion segmentation (Time on the x axis,QoM on the y axis).

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 characterized 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 characterized 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). In an experiment on analysis of dance performances carried out in collaboration with the Department of Psychology of Uppsala University (Sweden), a collection of 20 microdances (5 dancers per 4 basic emotions) was rated by subjects and classified by an automatic system based on decision trees. Five decision tree models were trained for classification on five training sets (85% of the available data) and tested on five test sets (15% of the available data). The samples for the training and test sets were randomly extracted from the data set and were uniformly distributed along the four classes and the five dancers. The data set included 18 variables extracted from the dance performances. The outcomes of the experiment shows a rate of correct classification for the automatic system (35.6%) in between chance level (25%) and spectators' rate of correct classification (56%): for further information see Camurri et al. [2004c].

5.3.6 Tools: the EyesWeb open platform

The EyesWeb open platform has been designed at DIST-InfoMus Lab with a special focus on the multimodal analysis and processing of non-verbal expressive gesture in human movement and music signals Camurri et al. [2000]. Since the starting of the EyesWeb project in 1997, the focus has been on the development of a system supporting on the one hand multimodal processing both in its conceptual and technical aspects, and allowing on the other hand fast development of

robust application prototypes for use in artistic performances and interactive multimedia installations. In 2001 the platform has been made freely available on the Internet (www.eyesweb.org) and the number of users has rapidly grown. In recent years, EyesWeb has been satisfactorily used by the DIST-InfoMus Lab both for research purposes and for several kinds of applications, e.g., in museum exhibits and in the field of performing arts. It has also been adopted as standard in several EU funded research projects (e.g., in the IST Program: projects MEGA, CARE-HERE, MEDIATE, TAI-CHI) and thousands of users currently employ it in universities, public and private research centers, and companies. Recently, the EyesWeb platform has been reconceived in order to fulfill new requirements coming from the continuously enlarging EyesWeb community. Such process led to the development of another platform (EyesWeb version 4.0) which is completely new with respect to its predecessors in the way it deals with the conceptual issues involved in multimodal processing, in how it supports and implements multimodality, in the additional features it provides to users Camurri et al. [2004a]. The first beta version of EyesWeb 4.0 has been publicly released in September 2004.

The EyesWeb open platform consists of a number of integrated hardware and software modules that can be easily interconnected and extended in a visual environment. The EyesWeb software includes a development environment and a set of libraries of reusable software components that can be assembled by the user in a visual language to build patches as in common computer music languages inspired to analog synthesizers. EyesWeb supports the user in experimenting computational models of non-verbal expressive communication and in mapping, at different levels, gestures from different modalities (e.g., human full-body movement, music) onto real-time generation of multimedia output (e.g., sound, music, visual media, mobile scenery). It allows fast development and experiment cycles of interactive performance setups. EyesWeb is a Win32 multi-thread application. At run-time, an original real-time patch scheduler supports several modalities of activation for modules in order to support and optimize management and integration of multimodal streams. A patch is automatically splitted by the scheduler according to its topology and possible synchronization needs. Asynchronous modules having an internal dynamics are also supported. They receive inputs as any other kind of modules but their outputs are asynchronous with respect to their inputs. For example, an "emotional resonator" able to react to the perceived expressive content of a dance performance, embedding an internal dynamics, may have a delay in activating its outputs due to its actual internal state, memory of past events. This is one of the mechanisms explicitly supported by the system to implement interaction metaphors beyond the "musical instrument" and to support interactive narrative structures. It should be noted that usually the user does not have to care about activation mechanisms and

scheduling of the modules, since EyesWeb directly manages these aspects. The user is therefore free to take care of higher-level tasks, e.g. the interactive narrative structure and dynamic evolution of patches in timelines or execution graphs. EyesWeb supports the integrated processing of different streams of (expressive) data, such as music audio, video, and, in general, gestural information.

A set of open libraries of basic modules is available including the following:

- Input and output modules: support for frame grabbers (from webcams to professional frame grabbers), wireless on-body sensors (e.g., accelerometers), live audio input, video and audio players (several different video and audio format supported), OSC (OpenSound-Control), Steinberg ASIO, MIDI, input from devices (e.g., mouse, keyboard, joystick, data gloves), audio, video, and numeric output both live and recorded on files (e.g., avi, wav, text files).
- Math and filters: e.g., modules for basic mathematical operations (both on scalars and matrices), pre-processing, signal conditioning, signal processing in the time and frequency domains.
- Imaging: processing and conversions of images, computer vision techniques, blob extraction and analysis, graphic primitives, support to FreeFrame plug-ins.
- Sound and MIDI libraries: audio processing, extraction of audio features in the time and frequency domains, extraction of features from MIDI, support to VST plug-ins.
- Communication: TCP/IP, serial, OSC, MIDI, Microsoft DCOM.

Users can also build new EyesWeb modules and use them in patches. In order to help programmers in developing blocks, the EyesWeb Wizard software tool has been developed and is available. Users can develop autonomously (i.e., possibly independently from EyesWeb) the algorithms and the basic software skeletons of their own modules. Then, the Wizard supports them in the process of transforming algorithms in integrated EyesWeb modules. Multiple versions of modules (versioning mechanism) are supported by the system, e.g., allowing the use in patches of different versions of the same data-type or module. The compatibility with future versions of the systems, in order to preserve the existing work (i.e., modules and patches) in the future is supported.

EyesWeb has been the basic platform of the MEGA EU IST project. In the EU V Framework Program it has also been adopted in the IST CARE HERE and IST MEDIATE projects on therapy and rehabilitation and by the MOSART network for training of young researchers. In the EU VI Framework Program EyesWeb has been adopted and extended to the new version 4.0 in the TAI-CHI project (Tangible Acoustic Interfaces for Computer-Human Interaction). Some partners in the EU Networks of Excellence ENACTIVE and HUMAINE adopted EyesWeb for research. EyesWeb is fully available at its website (www.eyesweb.org). Public newsgroups also exist and are daily managed to support the EyesWeb community.

5.3.7 Tools: the EyesWeb Expressive Gesture Processing Library

Many of the algorithms for extracting the motion descriptors illustrated above have been implemented as software modules for the EyesWeb open platform. Such modules are included in the EyesWeb Expressive Gesture Processing Library.

The EyesWeb Expressive Gesture Processing Library includes a collection of software modules and patches (interconnections of modules) contained in three main sub-libraries:

- The EyesWeb Motion Analysis Library: a collection of modules for real-time motion tracking and extraction of movement cues from human full-body movement. It is based on one or more videocameras and other sensor systems.
- The EyesWeb Space Analysis Library: a collection of modules for analysis of occupation of 2D (real as well as virtual) spaces. If from the one hand this sub-library can be used to extract low-level motion cues (e.g., how much time a given position in the space has been occupied), on the other hand it can also be used to carry out analyses of gesture in semantic, abstract spaces.
- The EyesWeb Trajectory Analysis Library: a collection of modules for extraction of features from trajectories in 2D (real as well as virtual) spaces. These spaces may again be either physical spaces or semantic and expressive spaces.

The EyesWeb Motion Analysis Library (some parts of this library can be downloaded for research and educational purposes from the EyesWeb website www.eyesweb.org) applies computer vision, statistical, and signal processing techniques to extract expressive motion features

(expressive cues) from human full-body movement. At the level of processing of incoming visual inputs the library provides modules including background subtraction techniques for segmenting the body silhouette, techniques for individuating and tracking motion in the images from one or more videocameras, algorithms based on searching for body centroids and on optical flow based techniques (e.g., the Lucas and Kanade tracking algorithm), algorithms for segmenting the body silhouette in sub-regions using spatio-temporal projection patterns, modules for extracting a silhouette's contour and computing its convex hull. At the level of extraction of motion descriptors a collection of parameters is available. They include the above mentioned Quantity of Motion, i.e., amount of detected movement, Contraction Index, Stability Index, Asymmetry Index, Silhouette shape, and direction of body parts. The EyesWeb Motion Analysis Library also includes blocks and patches extracting measures related to the temporal dynamics of movement. A main issue is the segmentation of movement in pause and motion phases. Several movement descriptors can be measured after segmenting motion in motion and pause phases: for example, blocks are available for calculating durations of pause and motion phases and inter-onset intervals as the time interval between the beginning of two subsequent motion phases.

The EyesWeb Space Analysis Library is based on a model considering a collection of discrete potential functions defined on a 2D space. The space is divided into active cells forming a grid. A point moving in the space is considered and tracked. Three main kinds of potential functions are considered: (i) potential functions not depending on the current position of the tracked point, (ii) potential functions depending on the current position of the tracked point, (iii) potential functions depending on the definition of regions inside the space. Objects and subjects in the space can be modeled by time-varying potentials. Regions in the space can also be defined. A certain number of "meaningful" regions (i.e., regions on which a particular focus is placed) can be defined and cues can be measured on them (e.g., how much time a tracked subject occupied a given region). The metaphor can be applied both to real spaces (e.g., scenery and actors on a stage, the dancer's General Space as described by Rudolf Laban) and to virtual, semantic, expressive spaces (e.g., a space of parameters where gestures are represented as trajectories): for example, if, from the one hand, the tracked point is a dancer on a stage, a measure of the time duration along which the dancer was in the scope of a given light can be obtained; on the other hand, if the tracked point represents a position in a semantic, expressive space where regions corresponds to basic emotions, the time duration along which a given emotion has been recognized can also be obtained. The EyesWeb Space Analysis Library implements the model and includes blocks allowing the definition of interacting discrete potentials on 2D spaces, the definition of regions, and the extraction of cues (such as, for example, the occupation rates of regions in the space).

The EyesWeb Trajectory Analysis Library contains a collection of blocks and patches for extraction of features from trajectories in 2D (real or virtual) spaces. It complements the EyesWeb Space Analysis Library and it can be used together with the EyesWeb Motion Analysis Library. Blocks can deal with many trajectories at the same time, for example trajectories of body joints (e.g., head, hands, and feet tracked by means of color tracking techniques - occlusions are not dealt with at this stage) or trajectories of points tracked using the Lucas-Kanade feature tracker available in the Motion Analysis Library. Features that can be extracted include geometric and kinematics measures. They include Directness index, Trajectory length, Trajectory local and average direction, Velocity, Acceleration, and Curvature. Descriptive statistic measures can also be computed both along time (for example, average and peak values of features calculated either on running windows or on all the samples between two subsequent commands such as the average velocity of the hand of a dancer during a given motion phase) and among trajectories (for example, average velocity of groups of trajectories available at the same time such as the average instantaneous velocity of all the tracked points located on the arm of a dancer). Trajectories can be real trajectories coming from tracking algorithms in the real world (e.g., the trajectory of the head of a dancer tracked using a tracker included in the EyesWeb Motion Analysis Library) or trajectories in virtual, semantic spaces (e.g., a trajectory representing a gesture in a semantic, expressive space).

5.4 Control of music performance

Anders Friberg and Roberto Bresin
KTH, Department of Speech, Music and Hearing

5.4.1 Introduction

Here we will look at the control of music on a higher level dealing with semantic/gestural descriptions rather than the control of each note as in a musical instrument. It is similar to the role of the conductor in a traditional orchestra. The conductor controls the overall interpretation of the piece but leaves the execution of the notes to the musicians. A music performance system typically consists of a human controller using gestures that are tracked and analysed by

a computer generating the performance. An alternative could be to use audio input. In this case the system would follow a musician or even computer-generated music. What do we mean by higher level control? The methods for controlling a music performance can be divided in three different categories: (1) Tempo/dynamics. A simple case is to control the instant values of tempo and dynamics of a performance. (2) Performance models. Using performance models for musical structure, such as the KTH rule system (see also Section 5.4.2), it is possible to control performance details such as how to perform phrasing, articulation, accents and other aspect of a musical performance. (3) Semantic descriptions. These descriptions can be an emotional expression such as aggressive, dreamy, melancholic or typical performance instructions (often referring to motion) such as andante or allegretto. The input gestures/audio can by analyzed in different ways roughly similar to the three control categories above. However, the level of detail obtained by using the performance models cannot in the general case be deduced from a gesture/audio input. Therefore, the analysis has to be based on average performance parameters. The overview of audio analysis including emotion descriptions is found in Section 5.4.1. The analysis of gesture cues is described in Chapter 5.3, above.

The fuzzy analyzer, described more in detail below, is a real time system for analyzing emotional expression from both audio and gestures. Several conductor systems using control of tempo and dynamics (thus mostly category 1) has been constructed in the past. The Radio Baton system, designed by Mathews [1989], was one the first systems and is still used both for conducting a score as well as a general controller. The Radio baton controller consists of two sticks and a rectangular plate. The 3D position of each stick above the plate is measured. Typically one stick is used for beating the time and the other stick is used for controlling dynamics. Using the conductor software, a symbolic score (a converted midi file) is played through a MIDI synthesizer. The system is very precise in the sense that the position of each beat is exactly given by the downbeat gesture of the stick. This allows for very accurate control of tempo but also requires practice - even for an experienced conductor! A more recent system controlling both audio and video is the *Personal Orchestra* developed by Borchers et al. [2004] and its further development in *You're the Conductor* [see Lee et al., 2004]. These systems are conducted using a wireless baton with infrared light for estimating baton position in two dimensions. The Personal Orchestra is an installation in House of Music in Vienna, Austria, where the user can conduct real recordings of the Vienna Philharmonic Orchestra. The tempo of both the audio and the video as well the dynamics of the audio can be controlled yielding a very realistic experience. The tempo is due to restrictions in the time manipulation model only controlled in discrete steps. The installation You're the conductor is also a museum exhibit but aimed for children rather

than adults. Therefore it was carefully designed to be intuitive and easily used. This time it is recordings of the Boston Pops orchestra that is conducted. A new time stretching algorithm was developed allowing any temporal changes of the original recording. From the experience with children users they found that the most efficient interface was a simple mapping of gesture speed to tempo and gesture size to volume. Several other conducting systems have been constructed. For example, the Conductor's jacket by Marrin Nakra [2000] senses several body parameters such as muscle tension and respiration that is translated to musical expression. The Virtual Orchestra is a graphical 3D simulation of an orchestra controlled by a baton interface developed by Ilmonen [2000]. In the following we will look more closely on the systems we have been developed using music performance models and semantic descriptions. It will start with an analyzer of emotional expression in gestures and music, present the KTH rule system and finally describe a "home conductor" system using these tools.

A fuzzy analyzer of emotional expression in music and gestures

An overview of the analysis of emotional expression is given in the contribution by OFAI in this CD. We will here[1] focus of one such analysis system aimed at real time applications. As mentioned, for basic emotions such as happiness, sadness or anger, there is a rather simple relationship between the emotional description and the cue values (i.e. measured parameters such as tempo, sound level or articulation). Since we are aiming at real-time playing applications we will focus here on performance cues such as tempo and dynamics. The emotional expression in body gestures has also been investigated but to a lesser extent than in music. Camurri et al. [2003] analyzed and modeled the emotional expression in dancing. Boone and Cunningham [1998] investigated children's movement patterns when they listened to music with different emotional expressions. Dahl and Friberg [2004] investigated movement patterns of a musician playing a piece with different emotional expressions. These studies all suggested particular movement cues related to the emotional expression, similar to how we decode the musical expression. We follow the idea that musical expression is intimately coupled to expression in body gestures and biological motion in general [see Friberg and Sundberg, 1999, Juslin et al., 2002]. Therefore, we try to apply similar analysis approach to both domains. Table 5.1 presents typical results from previous studies in terms of qualitative descriptions of cue values. As seen in the Table, there are several commonalities in terms of cue descriptions between motion and

[1]This section is a modification and shortening of the paper by Friberg [2005]

Emotion	Motion cues	Music performance cues
Anger	Large	Loud
	Fast	Fast
	Uneven	Staccato
	Jerky	Sharp timbre
Sadness	Small	Soft
	Slow	Slow
	Even soft	Legato
Happiness	Large	Loud
	Rather fast	Fast
		Staccato
		Small tempo variability

Table 5.1: A characterization of different emotional expressions in terms of cue values for body motion and music performance. Data taken from Dahl and Friberg (2004) and Juslin (2001).

music performance. For example, anger is characterized by both fast gestures and fast tempo. The research regarding emotional expression yielding the qualitative descriptions as given in Table 5.1 was the starting point for the development of current algorithms.

The analysis of emotional expression in music performance/gestures is realized in three steps:

1. *Cue extraction.* The first step is the extraction of basic cues. These cues are quite well defined for music input consisting of traditional tone parameters such as sound level, tempo, and articulation (staccato/legato). The audio cue extraction was designed for monophonic playing or singing. A previous prototype was described in [Friberg et al., 2002] and an improved version for non-real time use is found in [Friberg et al., in press]. The first part of the cue extraction segments the audio stream into tones by computing two different sound level envelopes using different time constants. The first sound level envelope follows roughly the shape of each tone and the second sound level envelope follows the general shape of the phrase. The crossings of two envelopes define the tone onsets and offsets. For each segmented tone five different cues are computed: sound level (dB), instant tempo (tones/second), articulation (relative pause duration), attack rate (dB/ms), and high-frequency content (high/low energy). In the body motion analysis we have been interested

in parameters describing general features of the movements rather than individual movements of each limb. A number of such general cues have been identified and algorithms have been developed for automatic extraction from video input using the EyesWeb platform [Camurri et al., 2004d, 2000], and Chapter 5.3. The current version of the cue analysis uses only a few basic tools within EyesWeb. In order to remove the background, the first step is to compute the difference signal between consecutive video frames. This means that the algorithm just "see" something when there is a movement. The use of difference signals makes the system quite robust and insensitive to e.g. light variations in the room. Three cues are computed from the difference signal. The total number of visible pixels constitutes the cue Quantity of Motion (QoM). The bounding rectangle defines the area in the picture that contains all non-zero pixels. The instant width and height of the bounding rectangle are computed and their peak-to-peak amplitude variations constitute the cues width-pp and height-pp.

2. *Calibration.* The second step is a semi-automatic calibration of cues. This is important for practical reasons in order to avoid otherwise lengthy optimization sessions trying to fine-tune internal parameters adapting to variations in musical content/body gestures or technical setup. Each cue is standardized, meaning that it is subtracted by its mean value and divided by its standard deviation. This results in cues with a mean of 0 and a standard deviation of 1. This is an important step implying that the following mapping does not need to be adjusted if the input conditions change, such as change of instrument, dancer, or artistic material. This requires a calibration phase before the recognition system is used in which the user is asked to move or play in, for example, a happy, sad and angry way thus defining the space of all possible variations.

3. *Expression mapping.* The third step is the mapping from calibrated cues to emotion description. Instead of using the more common data-driven methods, such as Neural Networks or Hidden Markov Models, we suggest here fuzzy set functions, allowing the direct use of qualitative data obtained from previous studies. For example Juslin and Laukka [2003] present a meta-analysis of about 40 studies regarding emotional expression in music. Most often, cues have been characterized mainly in terms of being either high or low in relation to different emotional expressions. One interesting extension in the meta-analysis was to classify some cues in terms of three levels. It indicated that an intermediate cue level might be important in some cases. This was used in the mapping algorithm. Following these ideas, we suggest here a method that uses the qualitative cue descriptions divided in three

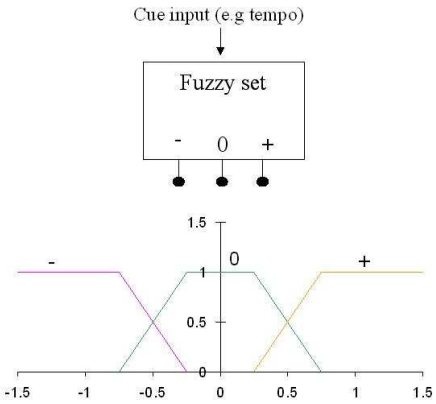

Figure 5.4: Qualitative classification of cues in terms of the three regions high (+), medium (0), and low (-), each with a separate output.

levels in order to predict the intended emotion. The same method is used both for musical and body motion cues. It uses fuzzy set functions to go from continuous cues to qualitative cues [Niskanen, 2004, Zimmerman, 1996, Seif El-Nasr et al., 2000, Bresin et al., 1995]. Each cue is divided into a three overlapping region functions, see Figure 5.4. Within a region the corresponding output is one and outside the region it is zero with an overlapping linear area at the boundaries. The final emotion prediction output is computed by taking an average of a selected set of fuzzy set functions. This selection is done according to previous qualitative descriptions. This results in a continuous output for each emotion with a range 0-1. If the value is 1 the emotion is completely predicted if it is 0 it is not at all predicted. Using an average of a set of region functions makes the output smoothly changing between emotions depending on the number of "right" cues. This averaging method was selected with the final application in mind. For example, if this algorithm is used for controlling the musical expressivity in computer-played performance, the transitions between different emotions need to be smooth. Other applications might ask for a discrete classification of emotional expression. This could be easily modeled using fuzzy logic.

The complete system is shown in Figure 5.5 using three cues extracted from audio input. The same setup is used for the body motion analysis using the body motion cues. The use of three cues with the mapping configuration indicated by the colored arrows in this example, has the advantage that emotion outputs are mutually exclusive, that is, if one output is 1, the other outputs

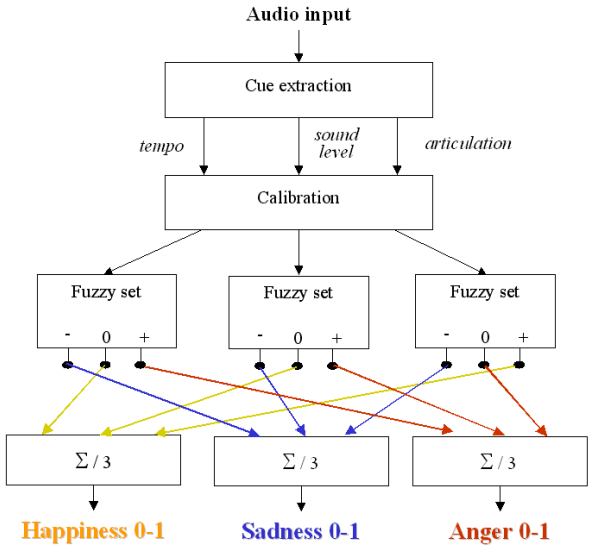

Figure 5.5: The complete system for estimating emotional expression in music performance using three cues. An audio input is analyzed in terms of tempo, sound level and articulation and the resulting prediction of emotional expression is output in terms of three functions ranging from 0 to 1.

are 0. All parts of the fuzzy analyzer except the motion cue analysis have been implemented using the program Pddeveloped by Puckette [1996]. Pdis a modular graphic environment for processing primarily audio and control information in real time. The fuzzy analyzer was implemented using preexisting blocks in the release Pd-extended 0.37, complemented with the library SMLib made by Johannes Taelman. The video cue analysis was implemented as a patch in EyesWeb, a similar graphic programming environment primarily for video processing [Camurri et al., 2000]. The audio analyzer makes modest claims on processing power and typically uses only a few percent of a Pentium 4 processor running Windows. Due to the cross-platform compatibility of Pd, the audio cue analysis could easily be ported to MacOS or Linux. The video cue analysis is currently restricted to the Windows platform.

Applications using the fuzzy analyser

The first prototype that included an early version of the fuzzy analyzer was a system that allowed a dancer to control the music by changing dancing style. It was called The Groove Machine and

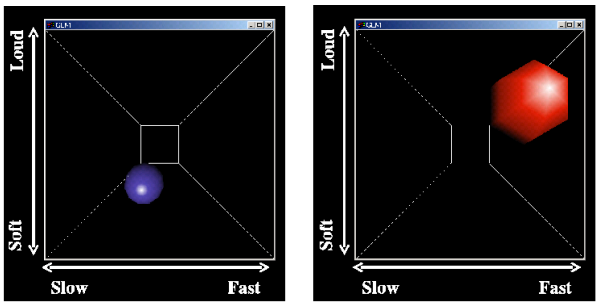

Figure 5.6: Two different examples of the Expressiball giving visual feedback of musical performance. Dimensions used in the interface are: X = tempo, Y =sound pressure level, Z =spectrum (attack time and spectrum energy), Shape = articulation, Colour = emotion. Left figure shows the feedback for a sad performance. Right figure shows the feedback for an angry performance.

was presented in a performance at Kulturhuset, Stockholm 2002. Three motion cues were used, QoM, maximum velocity of gestures in the horizontal plane, and the time between gestures in the horizontal plane, thus slightly different from the description above. The emotions analyzed were (as in all applications here) anger, happiness, and sadness. The mixing of three corresponding audio loops was directly controlled by the fuzzy analyzer output. For a more detailed description [Lindström et al., 2005]. The ExpressiBall, developed by Roberto Bresin, is a way to visualize a music performance in terms of a ball on a computer screen [Friberg et al., 2002]. A microphone is connected to the computer and the output of the fuzzy analyzer as well as the basic cue values are used for controlling the appearance of the ball. The position of the ball is controlled by tempo, sound level and a combination of attack velocity and spectral energy, the shape of the ball is controlled by the articulation (rounded-legato, polygon-staccato) and the color of the ball is controlled by the emotion analysis (red-angry, blue-sad, yellow-happy), see Figure 5.6. The choice of color mapping was motivated by recent studies relating color to musical expression [Bresin, 2005, Bresin and Juslin, 2005]. The ExpressiBall can be used as a pedagogical tool for music students or the general public. It may give an enhanced feedback helping to understand the musical expression. A future display is planned at Tekniska museet, Stockholm.

The latest application using the fuzzy analyzer has been the collaborative game Ghost in the Cave [Rinman et al., 2004]. It uses as its main input control either body motion or voice. One of the tasks of the game is to express different emotions either with the body or the voice; thus, both modalities are analyzed using the fuzzy analyzer described above. The game is played in

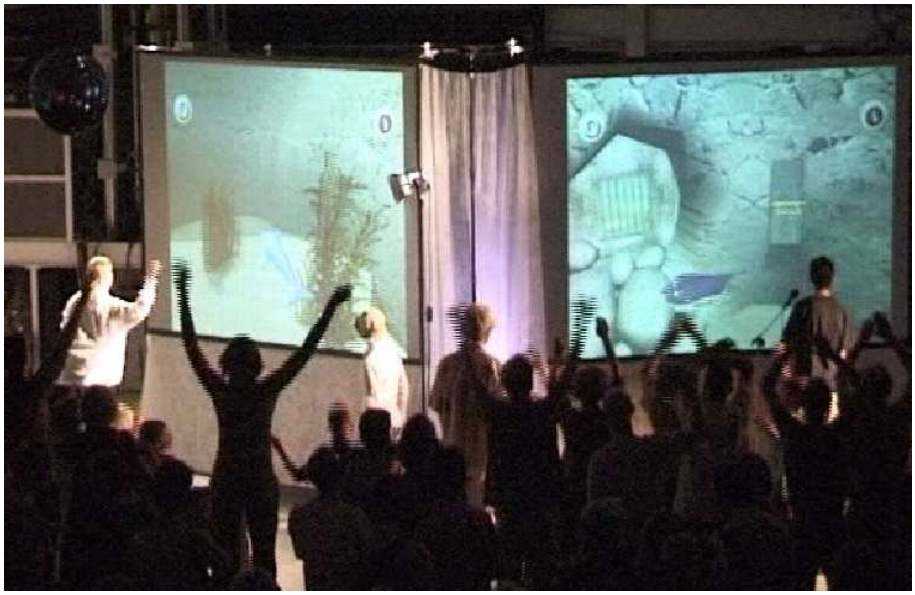

Figure 5.7: Picture from the first realization of the game Ghost in the Cave. Motion player to the left (in white) and voice player to the right (in front of the microphones).

two teams each with a main player, see Figure 5.7. The task for each team is to control a fish avatar in an underwater environment and to go to three different caves. In the caves there is a ghost appearing expressing different emotions. Now the main players have to express the same emotion, causing their fish to change accordingly. Points are given for the fastest navigation and the fastest expression of emotions in each subtask. The whole team controls the speed of the fish as well as the music by their motion activity. The body motion and the voice of the main players are measured with a video camera and a microphone, respectively, connected to two computers running two different fuzzy analyzers described above. The team motion is estimated by small video cameras (webcams) measuring the Quantity of Motion (QoM). QoM for the team motion was categorized in three levels (high, medium, low) using fuzzy set functions as shown in Figure 5.4. The music consisted of pre-composed audio sequences, all with the same tempo and key, corresponding to the three motion levels. The sequences were faded in and out directly by control of the fuzzy set functions. One team controlled the drums and one team controlled the accompaniment. The Game has been set up five times since the first realization summer at the Stockholm Music Acoustics Conference 2003, including the Stockholm Art and Science festival, Konserthuset, Stockholm, 2004, and Oslo University, 2004.

Summary and discussion

We describe a system that can be used for analyzing emotional expression both in music and body motion. The use of fuzzy mapping was a way of directly using previous results summarized in various investigations and turned out to be a robust mapping also in practical testing during the development of the applications.

The advantages of the model can be summarized as:
Generic - it is the same model for music performance and body motion.
Robust - The fuzzy set functions always stay between 0 and 1 implying that the emotion output is always between 0 and 1 as well. A collection of cues lowers the error influence from one cue proportionally.
Handle nonlinearities - This is not possible in e.g. a linear regression model.
Smooth transitions between emotions - This is achieved by the overlapping fuzzy set functions each with transition range.
Flexibility - It is easy to change the mapping using for example more cues since there is no need recalibrate the system.

The use of higher-level expression descriptions such as emotions has the advantage that it can provide a natural coherence between the controller's expression (visual or auditive) and the expression of the control device (could be a synthesizer, visual effects etc.) in a public performance. Take an example with a dancer controlling the music performance with a one-to-one correspondence between high-level description in the body motion and music performance. When the movements of the dancer are aggressive - the music also sounds aggressive. For a discussion of evaluation issues see [Friberg, 2005].

5.4.2 The KTH rule system for music performance

The KTH rule system is a result of an on-going long-term research project about music performance initiated by Johan Sundberg [e.g. Sundberg et al., 1983, Sundberg, 1993, Friberg, 1991, Friberg and Battel, 2002]. The idea of the rule system is to model the variations introduced by the musician when playing a score. The rule system contains currently about 30 rules modeling many performance aspects such as different types of phrasing, accents, timing patterns and intonation, see Table 5.2. Each rule introduce variations in one or several of the performance variables IOI (Inter-Onset Interval), articulation, tempo, sound level, vibrato rate, vibrato extent

as well a modifications of sound level and vibrato envelopes. Most rules operate on the "raw" score using only note values as input. However, some of the rules for phrasing as well as for harmonic, melodic charge need a phrase analysis and a harmonic analysis provided in the score. This means that the rule system does not in general contain analysis models. This is a separate and complicated research issue. One exception is the punctuation rule which includes a melodic grouping analysis [Friberg et al., 1998].

Table 5.2: Most of the rules in Director Musices, showing the affected performance variables (sl = sound level, dr = interonset duration, dro = offset to onset duration, va = vibrato amplitude, dc = cent deviation from equal temperament in cents)

M P C

Rule Name	**Performance Variables**	**Short Description**
High-loud	sl	The higher the pitch, the louder
Melodic-charge	sl dr va	Emphasis on notes remote from current chord
Harmonic-charge	sl dr	Emphasis on chords remote from current key
Chromatic-charge	dr sl	Emphasis on notes closer in pitch; primarily used for atonal music
Faster-uphill	dr	Decrease duration for notes in uphill motion
Leap-tone-duration	dr	Shorten first note of an up-leap and lengthen first note of a down-leap
Leap-articulation-dro	dro	Micropauses in leaps
Repetition-articulation-dro	dro	Micropauses in tone repetitions

Continued on next page

Table 5.2: (continued)

M	D	M	C
Rule Name	**Performance Variables**	**Short Description**	
Duration-contrast	dr sl	The longer the note, the longer and louder; and the shorter the note, the shorter and softer	
Duration-contrast-art	dro	The shorter the note, the longer the micropause	
Score-legato-art	dro	Notes marked legato in scores are played with duration overlapping with interonset duration of next note; resulting onset to offset duration is dr+dro	
Score-staccato-art	dro	Notes marked staccato in scores are played with micropause; resulting onset to offset duration is dr-dro	
Double-duration	dr	Decrease duration contrast for two notes with duration relation 2:1	
Social-duration-care	dr	Increase duration for extremely short notes	
Inegales	dr	Long-short patterns of consecutive eighth notes; also called swing eighth notes	

Continued on next page

Table 5.2: (continued)

Ensemble-swing	dr	Model different timing and swing ratios in an ensemble proportional to tempo
Offbeat-sl	sl	Increase sound level at offbeats

I

Rule Name	**Performance Variables**	**Short Description**
High-sharp	dc	The higher the pitch, the sharper
Mixed-intonation	dc	Ensemble intonation combining both melodic and harmonic intonation
Harmonic-intonation	dc	Beat-free intonation of chords relative to root
Melodic-intonation	dc	Close to Pythagorean tuning, e.g., with sharp leading tones

P

Rule Name	**Performance Variables**	**Short Description**
Punctuation	dr dro	Automatically locates small tone groups and marks them with lengthening of last note and a following micropause

Continued on next page

Table 5.2: (continued)

Phrase-articulation	dro dr	Micropauses after phrase and subphrase boundaries, and lengthening of last note in phrases
Phrase-arch	dr sl	Each phrase performed with arch-like tempo curve: starting slow, faster in middle, and ritardando towards end; sound level is coupled so that slow tempo corresponds to low sound level
Final-ritard	dr	Ritardando at end of piece, modeled from stopping runners

Continued on next page

Table 5.2: (continued)

S

Rule Name	**Performance Variables**	**Short Description**
Melodic-sync	dr	Generates new track consisting of all tone onsets in all tracks; at simultaneous onsets, note with maximum melodic charge is selected; all rules applied on this sync track, and resulting durations are transferred back to original tracks
Bar-sync	dr	Synchronize tracks on each bar line

The rules are designed using two methods, (1) the analysis-by-synthesis method, and (2) the analysis-by-measurements method. In the first method, the musical expert, Lars Frydén in the case of the KTH performance rules, tells the scientist how a particular performance principle functions. The scientist implements it, e.g. by implementing a function in lisp code. The expert musician tests the new rules by listening to its effect produced on a musical score. Eventually the expert asks the scientist to change or calibrate the functioning of the rule. This process is iterated until the expert is sutisfied with the results. An example of a rule obtained by applying the analysis-by-synthesis method is the Duration Contrast rule in which shorter notes are shortened and longer notes are lengthened [Friberg, 1991]. The analysis-by-measurements method consists of extracting new rules by analyzing databases of performances. For example two databases have been used for the design of the articulation rules. One database consisted in the same piece of music[2] performed by five pianists with nine different expressive intentions. The second database was made by 13 WA Mozart piano sonatas performed by a professional pianist. The performances of both databases were all made on computer-monitored grand pianos, a Yamaha Diskavier for the first database, and a Bösendorfer SE for the second one [Bresin and Battel, 2000,

[2]Andante movement of Mozart's sonata in G major, K 545.

Bresin and Widmer, 2000].

For each rule there is one main parameter k which controls the overall rule amount. When $k = 0$ there is no effect of the rule and when $k = 1$ the effect of the rule is considered normal. However, this "normal" value is selected arbitrarily by the researchers and should be used only for the guidance of parameter selection. By making a selection of rules and k values different performance styles and performer variations can be simulated. Therefore, the rule system should be considered as a musician's toolbox rather than providing a fixed interpretation (see 5.8).

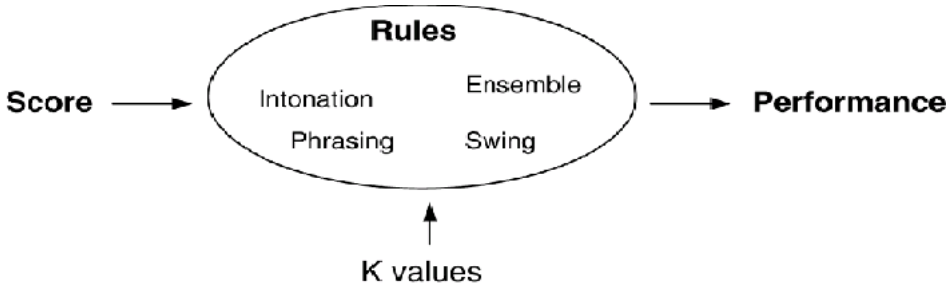

Figure 5.8: Functioning scheme of the KTH performance rule system.

A main feature of the rule system is that it most rules are related to the performance of different structural elements in the music (Friberg and Battel, 2002). Thus, for example, the phrasing rules enhance the division in phrases already apparent in the score. This indicates an interesting limitation for the freedom of expressive control: it is not possible to violate the inherent musical structure. One example would be to make ritardandi and accelerandi in the middle of a phrase. From our experience with the rule system such a violation will inevitably not be perceived as musical. However, this toolbox for marking structural elements in the music can also be used for modeling musical expression on the higher semantic level. Performances of emotional expressions can easily be modeled using different selections of rules and rule parameters as demonstrated by Bresin and Friberg [2000a]. Table 5.3 shows a possible organization of rules and their k parameters for obtaining performances with different emotional expressions.

Table 5.3: Cue profiles for emotions Anger, Happiness and Sadness, as outlined by Juslin [2001], and compared with the rule set-up utilized for the synthesis of expressive performances with Director Musices(DM)

A

Expressive Cue	**Juslin**	**Macro-Rule in DM**
Tempo	Fast	Tone IOI is shortened by 20%
Sound level	High	Sound level is increased by 8 dB
	Abrupt tone attacks	Phrase arch rule applied on phrase level and on sub-phrase level
Articulation	Staccato	Duration contrast articulation rule
Time deviations	Sharp duration contrasts	Duration contrast rule
	Small tempo variability	Punctuation rule

H

Expressive Cue	**Juslin**	**Macro-Rule in DM**
Tempo	Fast	Tone IOI is shortened by 15%
Sound level	High	Sound level is increased by 3 dB
Articulation	Staccato	Duration contrast articulation rule
	Large articulation variability	Score articulation rules
Time deviations	Sharp duration contrasts	Duration contrast rule

Continued on next page

Table 5.3: (continued)

	Small timing variations	Punctuation rule

S

Expressive Cue	**Juslin**	**Macro-Rule in DM**
Tempo	Slow	Tone IOI is lengthened by 30%
Sound level	Low	Sound level is decreased by 6 dB
Articulation	Legato	Duration contrast articulation rule
Articulation	Small articulation vaiability	Score legato articulation rule
Time deviations	Soft duration contrasts	Duration contrast rule
	Large timing variations	Phrase arch rule applied on phrase level and sub-phrase level
		Phrase arch rule applied on sub-phrase level
Final ritardando		Obtained from the Phrase rule with the *next* parameter

Director Musices[3] (DM) is the main implementation of the rule system and is a stand-alone lisp program available for Windows, MacOS, and GNU/Linux documented in [Friberg et al., 2000] and [Bresin et al., 2002].

[3]http://www.speech.kth.se/music/performance/download/dm-download.html

pDM - Real time control of the KTH rule system

In DM and its lisp environment there is no support for real time music processing. Therefore, in order implement a real time control of the rules we needed a new application. We selected two-step approach using a combination of DM and a player. The rules are first applied to the score in DM producing an enhanced score containing all the possible rule-induced variations of performance parameters. This new score is then played by an application written in `Pd`. There were several advantages using this approach. First of all, it was not necessary to reimplement the rule system - a long and tedious process since each rule needs to be verified on several musical examples. Also, it avoids the splitting into two different systems to support. The first prototype following this approach was made by Canazza et al. [2003a] using the EyesWeb platform. In the following we give an overview of the procedure including a description of the pDM player. For a more detailed description, see [Friberg, in press].

Rule application In the first step, the rules are applied to the score using DM. Most of the relevant rules defined in DM are applied with their default quantities. Each rule is applied on the original score and normalized with respect to overall length and dynamics. The deviations in the performance parameters tempo, sound level, and articulation are collected for each note. The original score together with all the individual rule deviations are stored in a custom pDM file format.

pDM player In the second step, the produced score is loaded into pDM, which is essentially an extended sequencer. Since all rule knowledge is kept in DM the structure of pDM is quite simple and is written in `Pd`-extended v. 0.37 using only the provided libraries. The sequencer is centered around the qlist object in `Pd`. qlist is a text-based sequencer allowing any data provided in an input file to be executed in time order. During playing, each of the three basic score parameters tempo (T_{nom}), sound level (SL_{nom}) and duration (DUR_{nom}) are modified using a weighting factor k_i for each rule:

$$T = T_{nom} \cdot (1 + \sum_{i=1}^{14} k_i \Delta T_i) \tag{5.1}$$

$$SL = SL_{nom} \cdot (1 + \sum_{i=1}^{11} k_i \Delta SL_i) \tag{5.2}$$

$$DUR = DUR_{nom} \cdot (1 + \sum_{i=1}^{5} k_i \Delta ART_i) \tag{5.3}$$

where T, SL, DUR stands for the resulting tempo, sound level, and duration; i is the rule parameter number, k_i is the weighting factor for the corresponding rule, and T_i, SL_i, DUR_i are the rule deviations given in the pDM file. According to the formulas above, the effect of several simultaneous rules acting on the same note is additive. This might lead to that the same note receives too much or contrary amount of deviations. This is in reality not a big problem and some rule interaction effects are already compensated for in the DM rule application. For example, the Duration contrast rule (shortening of relatively short notes) is not applied where the Double duration rule would be applied and lengthen relatively short notes. These performance parameters are computed just before a note is played. In this way, there is no perceived delay from a real-time input since all control changes will appear at the next played note. Figure 5.9 shows the window for individual rule control in which the rule weights can be manipulated.

pDM Expression mappers pDM contains a set of mappers that translate high-level expression descriptions into rule parameters. We have mainly used emotion descriptions (happy, sad, angry, tender) but also other descriptions such as hard, light, heavy or soft has been implemented. The emotion descriptions have the advantages that there has been substantial research made describing the relation between emotions and musical parameters [Sloboda and Juslin, 2001, Bresin and Friberg, 2000a]. Also, these basic emotions are easily understood by laymen. Typically, these kinds of mappers have to be adapted to the intended application as well as considering the function of the controller being another computer algorithm or a gesture interface. Usually there is a need for interpolation between the descriptions. One option implemented in pDM is to use a 2D plane in which each corner is specified in terms of a set of rule weightings corresponding to a certain description. When moving in the plane the rule weightings are interpolated in a semi-linear fashion. This 2D interface can easily be controlled directly with the mouse. In this way, the well-known Activity-Valence space for describing emotional expression can be implemented [Juslin, 2001]. Activity is related to high or low energy and Valence is related to positive or

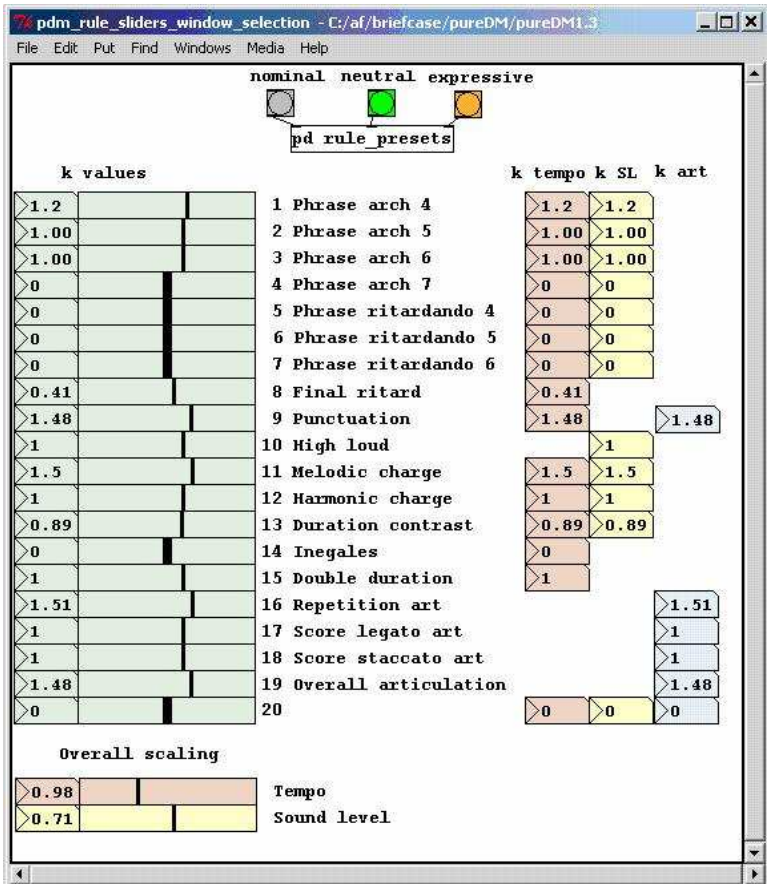

Figure 5.9: pDM window for controlling individual rule parameters. The sliders to the left control the overall amount of each rule (k_i values).

negative emotions. The quadrants of the space can be characterized as happy (high activity, positive valence), angry (high activity, negative valence), tender (low activity, positive valence), and sad (low activity, negative valence). An installation using pDM in which the user can change the emotional expression of the music while it is playing is currently part of the exhibition "Se Hjärnan" (Swedish for "See the Brain") touring Sweden for two years.

5.4.3 A home conducting system

Typically the conductor express by gestures overall aspects of the performance and the musician interpret these gestures and fill in the musical details. However, previous conductor systems

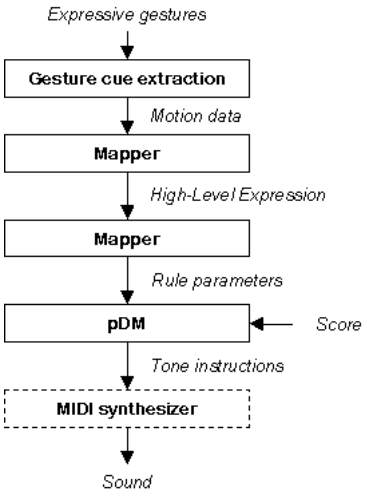

Figure 5.10: Overall schematic view of a home conductor system.

have often been restricted to the control of tempo and dynamics. This means that the finer details will be static and out of control. An example would be the control of articulation. The articulation is important for setting the gestural and motional quality of the performance but cannot be applied on an average basis. Amount of articulation (*staccato*) is set on a note-by-note basis dependent on melodic line and grouping, as reported by Bresin and Battel [2000] and Bresin and Widmer [2000]. This makes it too difficult for a conductor to control directly. By using the KTH rule system with pDM described above these finer details of the performance can be controlled on a higher level without the necessity to shape each individual note. Still the rule system is quite complex with a large number of parameters. Therefore, the important issue when making such a conductor system is the mapping of gesture parameters to music parameters. Tools and models for doing gesture analysis in terms of semantic descriptions of expression has recently been developed, see above. Thus, by connecting such a gesture analyzer to pDM we have a complete system for controlling the overall expressive features of a score. An overview of general system is given in Figure 5.10.

Recognition of emotional expression in music has been shown to be an easy task for most listeners including children from about 6 years of age even without any musical training [Peretz, 2001, see e.g.]. Therefore, by using simple high-level emotions descriptions such as (happy, sad, angry) the system have the potential of being intuitive and easily understood by most users including children. Thus, we envision a system that can be used by the listeners in their homes

rather than a system used for the performers on the stage. Our main design goals have been a system that is (1) easy and fun to use for novices as well as experts, (2) realized on standard equipment using modest computer power. In the following we will describe the system more in detail starting with the gesture analysis followed by different mapping strategies.

Gesture cue extraction We use a small video camera (webcam) as input device analysed by a robust and simple motion detection algorithm. The video signal is analyzed with the EyesWeb tools for gesture recognition [Camurri et al., 2000, 2004d]. The first step is to compute the difference signal between video frames. This is a simple and convenient way of removing all background (static) information in the picture. Thus, there is no need to worry about special lightning, clothes or background content. For simplicity, we have been using a limited set of tools within EyesWeb such as the overall quantity of motion (QoM), x y position of the overall motion, size and velocity of horizontal and vertical gestures.

Mapping gesture cues to rule parameters Depending on the desired application and user ability the mapping strategies can be divided in three categories:

Level 1 (listener level) The musical expression is controlled in terms of basic emotions (happy, sad, angry). This creates an intuitive and simple music feedback comprehensible without the need for any particular musical knowledge.

Level 2 (simple conductor level) Basic overall musical features are controlled using for example the energy-kinematics space previously found relevant for describing the musical expression as in [Canazza et al., 2003b].

Level 3 (advanced conductor level) Overall expressive musical features or emotional expressions in level 1 and 2 are combined with the explicit control of each beat similar to the Radio-Baton system.

Using several interaction levels makes the system suitable both for novices, children and expert users. Contrary to traditional instruments, this system may "sound good" even for a beginner when using a lower interaction level. It can also challenge the user to practice in order to master higher levels similar to the challenge provided in computer games. A few complete prototypes for level 1 and 2 have been assembled and tested using different mappings. One direct way of a simple interface on level 1 is to extract the semantic expressive descriptions from the motion cues using the fuzzy analyzer described above and connect that to the emotion control

in pDM. However, the mapping that we have used the most is a simpler but effective gesture interface mapping. It uses two cues from the video analysis: (1) overall quantity of motion (QoM) computed as the total number of visible pixels in the difference image, (2) the vertical position computed as the center of gravity for the visible pixels in the difference image. These two cues are directly mapped to the Activity-Valence emotion space included in pDM with QoM connected to Activity (high QoM - high Activity) and vertical position connected to Valence (high position-positive Valence). This interface has been demonstrated several times with very positive responses. However, formal testing in form of usability studies is planned in future work.

5.5 Controlling sound production

Roberto Bresin and Kjetil Falkenberg Hansen
KTH, Dept. of Speech, Music and Hearing

Matti Karjalainen, Teemu Mäki-Patola, Aki Kanerva, and Antti Huovilainen
Helsinki University of Technology

Sergi Jordà, Martin Kaltenbrunner, Günter Geiger, Ross Bencina
Universitat Pompeu Fabra, Music Technology Group/IUA

Amalia de Götzen and Davide Rocchesso
University of Verona, Dept. of Computer Science

The structure of this chapter is a particular one, it is made of five contributions by different authors. Each contribution is self-contained and can be read separately. The main reason for this structure is that the field of sound control is a relative new one and it is open to different approaches and applications. In this chapter we try to summarize what it is the state-of-the-art in the field within the Consortium partners of the S2S^2 Coordinating Action. The chapter starts with an introductory part in which we present general concepts and problems related to sound control. In particular it is outlined the important issue of the choice of sound models that can give suitable and responsive feedback in continuous control, as in the cases of sound generated by body motion and sound as feedback in interaction. The introductory part is followed by four sections in which

we present state-of-the-art applications. The first two sections, ”DJ Scratching” and ”Virtual Air Guitar”, focus on the control of musical instruments, and in particular on control that makes sense of sound production. The next two sections, ”The reacTable*” and ”The Interactive Book”, focus on the control of sounding objects, and are characterized by applications that control sounds that make sense. The chapters sequence can also be seen as ordered after increasing abstraction of sound models (from sampled sounds to cartoonized sounds) and decreasing complexity of gesture control (from DJ scratching to simple sliding).

Authors of the different sessions are:

5.5.1 Introduction
Roberto Bresin and Davide Rocchesso

5.5.2 DJ scratching with Skipproof
Kjetil Falkenberg Hansen and Roberto Bresin

5.5.3 Virtual air guitar
Matti Karjalainen, Teemu Mäki-Patola, Aki Kanerva, and Antti Huovilainen

5.5.4 The reacTable*
Sergi Jordà, Martin Kaltenbrunner, Günter Geiger, Ross Bencina

5.5.5 The interactive book
Amalia de Götzen and Davide Rocchesso

5.5.1 Introduction

[4] With the new millennium there are a few emerging facts that are conditioning our present approaches to the study of sound. Sensors of many different kinds are available at low cost and they can be organized into networks. Computing power is generously available even in tiny and low-power processors that can be easily embedded into artefacts of different nature and size. New design strategies that take advantage of these technological opportunities are emerging: physical computing, natural interaction, calm technologies are some of the many buzzwords that are being proposed as labels for these new trends in design. For the purpose of sound-based communication, the concept of embodied interaction [Dourish, 2001] is particularly

[4] Parts of this section are extracted and modified from a recent work by Rocchesso and Bresin [2005]

significant. Embodiment is considered a property of how actions are performed with or through artefacts, thus embracing the position that treats meanings as inextricably present in the actions between people, objects, and the environment. A key observation that emerges from embodied interaction examples is that human interaction in the world is essentially continuous and it relies on a complex network of continuous feedback signals. This is significantly important if one considers that most interfaces to technological artefacts that are currently being produced are developed around switches, menus, buttons, and other discrete devices. The design of graphical user interfaces has been largely inspired by ecological psychology and concepts such as direct perception and affordances [Gibson, 1979]. When designing embodied interfaces, we call for a reconciliation of ecological psychology and phenomenology that looks, with equal emphasis, at the objects and at the experiences. By means of physical modeling we can represent and understand the objects. By direct observation we can tell what are the relevant phenomena, which physical components are crucial for perception, what degree of simplification can be perceptually tolerated when modeling the physical reality. Specifically, sound designers are shifting their attention from sound objects to sounding objects, in some way getting back to the sources of acoustic vibrations, in a sort of ideal continuity with the experimenters of the early twentieth century, especially futurists such as Luigi Russolo and his intonarumori. In the contemporary world, sounding objects should be defined as sounds in action, intimately attached to artefacts, and dynamically responsive to continuous manipulations. As opposed to this embodied notion of sound, consider an instrument that came shortly after the intonarumori, the theremin invented 1919 by Lev Termen. It is played by moving the hands in space, near two antennae controlling amplitude and frequency of an oscillator. Its sound is ethereal and seems to come from the outer space. This is probably why it has been chosen in the soundtracks of some science-fiction movies [see the documentary by Martin, 2001]. Even though relying on continuous control and display, the lack of physical contact may still qualify the theremin as a schizophonic artefact, and it is not by coincidence that it is the only musical instrument invented in the twentieth century (the schizophonic age) that was used by several composers and virtuosi. Indeed, nephews of the theremin can be found in several recent works of art and technology making use of sophisticated sensors and displays, where physical causality is not mediated by physical objects, and the resulting interaction is pervaded by a sense of disembodiment.

Sound and motion

Sounds are intimately related to motion, as they are usually the result of actions, such as body gestures (e.g. the singing voice) or mechanical movements (e.g. the sound of train wheels on rails). In the same way as we are very accurate in recognizing the animate character of visual motion only from a few light points corresponding to the head and the major limb-joints of a moving person [Johansson, 1973], we are very sensitive to the fluctuations of auditory events in the time-frequency plane, so that we can easily discriminate walking from running [Bresin and Dahl, 2003] or even successfully guess gender of a person walking [Li et al., 1991]. It is not a surprise that gestures are so tightly related with sound and music communication. A paradigmatic case is that of the singing voice, which is directly produced by body movements (see also Chapters 5.3 and 5.4 for overviews on gestures in music performance). In general, gestures allow expressive control in sound production. Another example is DJ scratching, where complex gestures on the vinyl and on the cross-fader are used for achieving expressive transformation of prerecorded sounds [Hansen and Bresin, 2003]. In the context of embodied interfaces, where manipulation is mostly continuous, it is therefore important to build a gesture interpretation layer, capable to extract the expressive content of human continuous actions, such as those occurring as preparatory movements for strokes [see Dahl, 2004]. Body movements preceding the sound production give information about the intentions of the user, smother and slower movements produce softer sounds, while faster and sudden movements are associated to louder sounds. Gestures and their corresponding sounds usually occur in time sequences, and it is their particular time organization that helps in classifying their nature. Indeed, if properly organized in time, sound events can communicate a particular meaning. Let us consider the case of walking sounds [Bresin and Giordano, submitted]. The sound of a step in isolation is difficult to identify, while it gives the idea of walking if repeated a number of times. If the time sequence is organized according to equations resembling biological motion, then walking sounds can be perceived as more natural [Bresin and Dahl, 2003]. In addition, if sound level and timing are varied, it is possible to communicate different emotional intentions with walking sounds. In fact, the organization in time and sound level of structurally organized events, such as notes in music performance or phonemes in speech, can be controlled for communicating different emotional expressions. For instance in hyper- and hypoarticulated speech [Lindblom, 1990] and in enhanced performance of musical structure [Bresin and Friberg, 2000a] the listener recognizes the meaning being conveyed as well as the expressive intention on top of it. Research results show that not only we are able to recognize different emotional intentions used by musicians or

speakers [Juslin and Laukka, 2003] but also we feel these emotions. It has been demonstrated by psychophysical experiments that people listening to music evoking emotions experience a change in biophysical cues (such as blood pressure, etc.) that correspond to the feeling of that specific emotion and not only to the recognition. Krumhansl [1997] observed that sad music produced largest changes in heart rate, blood pressure, skin conductance and temperature, while happy music produced largest changes in measures of respiration. Music and sound in general have therefore the power to effect the variation of many physiological parameters in our body. These results could be taken into account in the design of more engaging applications where sound plays an active role.

Sound and interaction

Important role in any controlling action is played by the feedback received by the user, which in our case is the sound resulting from the user's gestures on an object or a musical instrument. Therefore sound carries information about the user's actions. If we extend this concept and consider sounds produced by any object in the environment we can say that sound is a multidimensional information carrier and as such can be used by humans for controlling their actions and reactions relative the environmental situation. In particular, humans are able to extract size, shape, material, distance, speed, and emotional expression from sonic information. These capabilities can be exploited to use sound as a powerful channel of communication for displaying complex data. Interactive sonification[5] is a new emerging field where sound feedback is used in a variety of applications including sport, medicine, manufacturing, and computer games. There are many issues that have been raised in such applications, and answers are expected to come from interaction design, perception, aesthetics, and sound modeling. For instance, how do we achieve pleasant and effective navigation, browsing, or sorting of large amount of data with sounds? In the framework of the Sounding Object project , the concept of sound cartoonification has been embraced in its wider sense and applied to the construction of engaging everyday sound models. Simplified and exaggerated models have been proved to be efficient in communicating the properties of objects in actions, thus being excellent vehicles for informative feedback in human-artefact communication. For instance, it has been shown that temporal control of sound events helps in communicating the nature of the sound source (e.g. a footstep) and the action that is being performed (walking/running). The possibility of using continuous interaction with sounding objects allows for expressive control of the sound production and, as a result, to higher

[5]See [Hunt and Hermann, 2005] for a recent overview of the field

engagement, deeper sense of presence, and experiential satisfaction. Low-cost sensors and recent studies in artificial emotions enable new forms of interaction using previously under-exploited human abilities and sensibilities. For instance, a cheap webcam is sufficient to capture expressive gesture nuances that, if appropriately interpreted, can be converted into non-visual emotional cues. These new systems, albeit inexpensive and simple in their components, provide new challenges to the designer who is called to handle a palette of technologies spanning diverse interaction modalities. In the future, the field of interaction design is expected to provide some guidelines and evaluation methods that will be applicable to artefacts and experiences in all their facets. It is likely that the classic methods of human-computer interaction will be expanded with both fine-grained and coarse-grained analysis methods. On a small scale, it is often necessary to consider detailed trajectories of physical variables in order to make sense of different strategies used with different interaction modalities. On a large scale, it is necessary to measure and analyze the global aesthetic quality of experiences.

Examples

Examples In the next four Sections we present state-of-the-art applications that can be grouped into two classes of sound control. Paraphrasing the S2S^2 Coordinating Action name, the one class, focusing on the control of musical instruments, represents applications that control the sense of sound production; the other class, focusing on the control of sounding objects, is characterized by applications that control sounds that make sense. The chapters sequence can also be seen as ordered after increasing abstraction of sound models (from sampled sounds to cartoonized sounds) and decreasing complexity of gesture control (from DJ scratching to simple sliding).

Control of musical instruments As far as continuous control is concerned, sampled sounds offer limited possibilities, unless a very complicated control is constructed, as it is found in sample-based musical keyboards. A successful example of such control is DJ scratching. In the next Section 5.5.2, Hansen and Bresin present a program for the simulation of typical DJ scratching gestures. This interface allows the novice musician to emulate DJs' sound "molding" techniques. This can be achieved by using gesture controllers instead of the traditional turntable, mixer and vinyl record set-up.

In the nineties it became clear that interfaces need sound models, i.e. audio synthesis procedures whose parameters humans can make sense. These are sound models that can directly

react to gestures. Section 5.5.3 is dedicated to a brief presentation of the Virtual Air Guitar (VAG) developed at the Helsinki Technical University by Matti Karjalainen and co-workers. It is a model for the control of guitar playing without a guitar. The sound model used is a physics-based one and it is control with the player's hands using a variety of devices such as web camera, and gesture sensor. Both Skipproof and VAG enable non-expert musicians to control virtual versions of real instruments. These interfaces enhance the expressive possibilities for non-expert musicians and yet allow the idiosyncrasies of a real instrument in ways that could appeal to professional musicians. The interfaces accomplish this by enabling the same or similar gesture controls both directly and at a higher level, i.e. by simulating them.

Control of sounding objects Research on the control of acoustic musical instrument offers insights for the design of new interaction models. In particular, in models making use of multisensory embodied interaction, sounds should be embedded in the artefacts and controlled by continuous interaction, such as in violin playing. In some cases the actual sound diffusion could be displaced elsewhere, but the tight coupling between gestures on the artefacts and sound models should give the illusion of embodied sounds. Two Sections are dedicated to interfaces manipulating sounding objects using typical everyday gesture interaction such as grasping, pulling, pushing, scratching, and hitting.

In Section 5.5.4, Sergi Jordà and co-workers present the reacTable*, a musical instrument based on a tabletop tangible user interface that allows cooperative and distributed multi-user music performance and composition. This instrument can be played by manipulating a set of objects that are distributed on top of a table surface.

In Section 5.5.5 Amalia de Götzen and Davide Rocchesso give an overview of book interfaces focusing on future developments of children books. This class of books, with their typical interactive gesture interface based on tabs and flaps that can be pulled and lifted, is ideal for the design of innovative applications associating sounds to interaction and therefore enhancing both narrative of the story and immersion of the reader.

5.5.2 DJ scratching with Skipproof

Scratching is a very popular way of making music with a turntable and a mixer. It is considered to be one of various DJ playing styles. Scratching and the lesser known, related playing style *beat juggling* both require much training for mastering complex gesture control of turntable and

mixer.

Skipproof, a patch written for `Pd` [Puckette, 1996], is both a virtual turntable and audio mixer, and an application for playing synthesized scratch techniques. The main purpose is to "scratch" sound files using gesture controllers of different kinds. Skipproof has been used by a DJ in two live concert situations, using a number of sensors.

Overview and background

Scratch performances are normally built up by the sequential executions of well-defined hand movements. Combinations of a gesture with the hand controlling the record and a gesture with the hand controlling the crossfader on the mixer are called scratch techniques. These have become common language for DJs, and they refer to them by name (*baby*, *crab*, *flare*) or by its characteristics (1-click *flare*, 2-click *flare*, reversed *tear*). About 100 techniques were recognized and more than 20 analysed in previous studies [e.g. Hansen, 2002, Hansen and Bresin, 2003]. The measurements focussed on the movement of the record and the crossfader, and based on the analysis, models with synthesized scratch techniques were collected.

Originally the software was intended to be a tool for reproducing and exploring the modelled scratch techniques with different characteristics. For instance we could change speed and extent of the record movements. The graphical user interface allows for easy experimenting with the models. We decided to develop the patch so it could be used also for controlling sound files in a turntable-like manner. In the end Skipproof was combining features from turntable, audio mixer and scratch vinyl records.

State of the art in scratch tools There is an increasing market for turntable-like interfaces used with digital music players such as CD and MP3 players, and for controlling music files on a computer. Many DJs nowadays feel comfortable with these interfaces and prefer CDs or MP3s to vinyl. This is not yet the case for scratch DJs, who still mainly perform with an ordinary turntable and mixer.

Even though learning to scratch properly is very demanding, no commercial products aims to help the aspiring DJ with easier ways to play the instrument. Only the crossfader has been experimented with to some degree (*Samurai* mixers from Vestax [Hansen, 2002]) making it easy to perform several crossfader clicks with only one simple gesture. On ordinary turntables

the pitch (speed change) has been extended from ± 8% to ± 100%, and with options to reverse the motor in order to play backwards.

A disadvantage of having record players and a mixer for scratching is the large size. Following a design idea by the world's leading scratch artist, DJ Q-Bert, Vestax has manufactured the first turntable with a built-in crossfader, the *QFO*. This small, round turntable represents a significant step towards a self-contained musical instrument for scratching.

One of very few attempts at improving the turntable as controller is a haptic force feedback turntable called *D'Groove*, developed by Beamish et al. [2004]. D'Groove controls a computer sound file like other systems, but has a motor that guides the DJ with various haptic feedback (like modern joysticks in a computer game). It can for instance re-create what kind of music is currently playing by mapping motor strength to intensity in the sound file[6], or mark each beat in the music with a bump in the vinyl. This can be highly useful in localizing the right spot on the album, and for beat juggling and mixing. Test results from D'Groove show that scratch DJs are eager to modify and experiment with their techniques to take full advantage of the new possibilities.

Skipproof's features

With Skipproof the user can play sound files with a virtual turntable and mixer, and apply models of scratch techniques. The most important feature is to be able to control the execution of the models (to play techniques with varying gestures). Another feature is easy customization of gesture controllers and sensors for operating the virtual turntable, mixer and techniques.

The graphical user interface (GUI), made with GrIPD [Sarlo, 2003], serves as a collection of controllers a DJ normally would expect to find. Skipproof's GrIPD window contains (see Fig. 5.11):

an area for scratching with the mouse

volume slider (simulating master/channel volume)

switch for muting the sound (simulating crossfader break-in point or channel select)[7]

[6] For a relatively short, isolated sound it will feel like pushing the record uphill as the resistance grows with the tone attack, and downhill in the decay.

[7] The crossfader normally has a very short span from silent to full sound. *Break-in point* is the place where sound

progress bar for showing position in the sound

a selection of samples to play with (simulating moving the needle)

slider for changing the effect of a movement (similar to having hand positions on the vinyl towards the edge or towards the middle)

pitch (tempo) change

on/off button for turntable

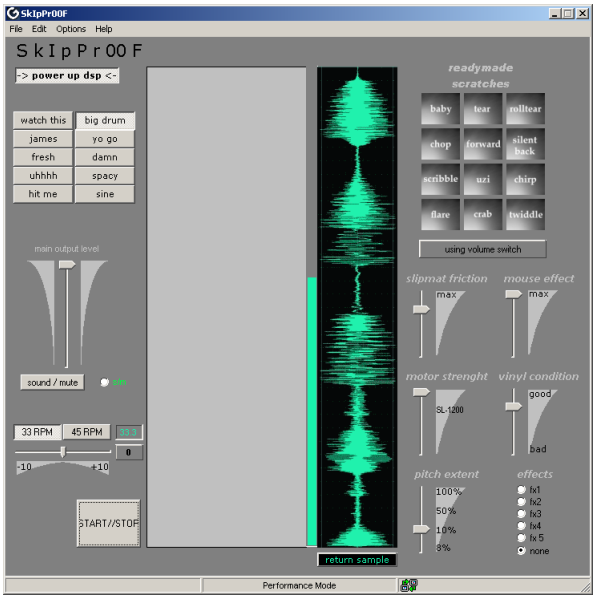

Figure 5.11: Skipproof GUI.

Some functions are not found in the real turntable and mixer, such as:

graphical representation of the sound file

triggers for modelled techniques

sound effects (reverberation, delay etc.)

is turned on, often a span of not more than a millimeter. It is more realistic to consider the crossfader as a switch than a fader. The channel selector is used as such switch.

slider for adjusting turntable motor strength

slider for adjusting friction between vinyl and platter

Some functions are not found in the model, but in the real turntable and mixer, such as:

equalizer/tone control

crossfader (changing between two sound sources)

two turntables, allowing to do beat juggling

Some of the sound effects implemented control frequency filters and stereo panning, much like the equalizer and panning knobs on a regular mixer.

All sound samples that are used with the patch are 1.6 seconds long, equivalent to one rotation with 33⅓ RPM (rounds per minute) on a vinyl record. This is reflecting the idea of a *skip proof* record[8]. The sounds are taken from a DJ tool record and are all common to perform scratching with.

Implementation of synthesized techniques There are 12 different scratch techniques ready to use in the patch. Each of these are models based on analysis of recordings made by a professional DJ, see Hansen [2002] for an overview. The typical scratch technique consists of a forward–backward movement of the record and one or more synchronized "clicks" with the crossfader. All the implemented techniques start and stop at the same position on the virtual record, with the exceptions of a slow forward motion and a silent backward motion.

To change the way a technique is performed depends on the controller. In most cases large gestures will make longer scratch movements than small gestures would within the same time span. As a result pitch will be higher. Even the speed is typically determined by the gesture as a sudden gesture will trigger a fast scratch and a slow gesture a slow scratch. These mappings may be changed in any way, and when using several controllers, it is possible to have many-to-many, one-to-many and many-to-one mappings [Hunt and Kirk, 2000]. For instance, with three

[8] The name Skipproof is taken from a feature found on DJ-tools records called a *skip proof* section, where a sound (or set of sounds) are exactly one rotation long and repeated for a couple of minutes. If the needle should happen to jump during a performance, the chances are quite good it will land on the same spot on the sound, but in a different groove.

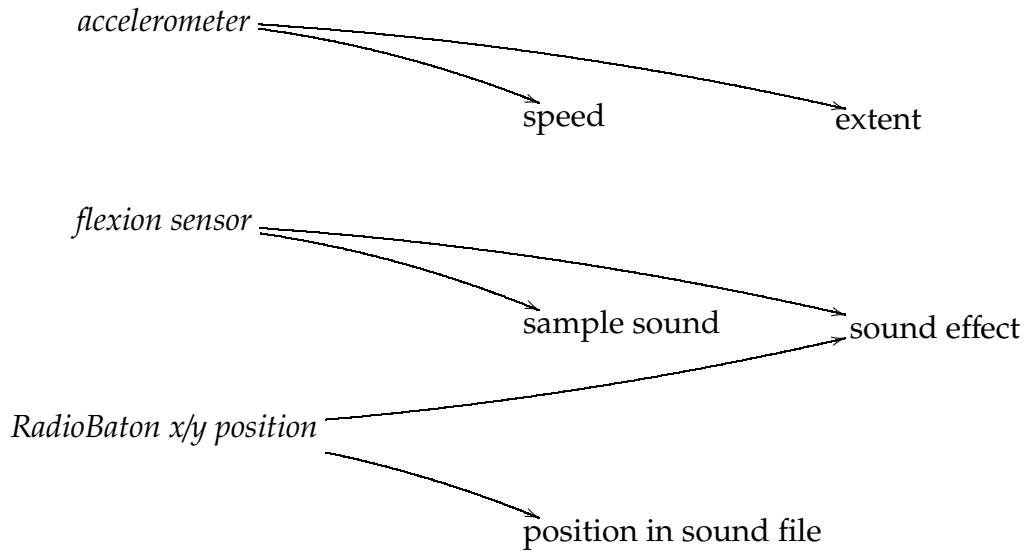

Figure 5.12: Possible mappings between three gesture sensors and five Skipproof parameters.

sensors and five actions, the mapping could look as in Figure 5.12, which shows the one-to-many mappings with all three gesture sensors and a many-to-one mapping to the sound effect action. For the set-ups described in the following section, one sensor often only controlled one action.

In the first version of Skipproof the techniques can be altered in extent and speed of the movement, the two most important parameters in scratching [Hansen and Bresin, 2003]. Both extent and speed can be exaggerated in the sense that scratch techniques can be performed faster and with larger movements than on a real turntable. Even slow and small movements can be exaggerated, but with a less interesting effect.

Controlling the patch

All the elements in the GUI can be controlled by either computer hardware, MIDI instruments or other sensors. An interface for mapping sensors to actions was developed, and it is also easy to calibrate sensors using this interface. It is possible to customize keyboard and mouse events to the actions. Until now Skipproof has been played by a number of devices:

Traditional input device (including mouse, keyboard, joystick, touch pad and tablet)

MIDI devices such as keyboard, sliders and knobs

Max Mathews' *RadioBaton* ([Boulanger and Mathews, 1997])

Pico D/A converter with custom sensors (including light sensor, potentiometers, buttons)

LaKitchen's Kroonde and Toaster sensor interfaces (including pressure sensor, magnetic field sensor, accelerometers, flexion sensor)

Of these controllers, the RadioBaton in combination with sliders, buttons, magnetic field, flexion and light sensors has proven to be most successful. Even a computer mouse in combination with other sensors can be quite effective.

Skipproof used in concerts A professional DJ has been using Skipproof in two live concert situations, see Fig. 5.13. First time it was with RadioBaton controlling the turntable and technique models, and some foot switches for other tasks. Sound level was controlled with the crossfader on a standard audio mixer. The RadioBaton sticks were replaced with treated gloves, so the musician could control the turntable speed and the techniques easily, moving his hand in a 3-D space over the antennae. The technique models were triggered with a foot switch, and the hand placement over the antennae determined how they were played.

In a second concert the DJ used again the RadioBaton, but now he could trigger the models more directly based on the gesture. The approach toward a defined area of the antennae was measured in speed and distance, and this gesture determined scratch speed and extent. For this performance a *light switch* was used to replace the crossfader. A light sensor was placed pointing directly toward a lamp, and the DJ could break the light beam by placing his hand close to the sensor, or by waving an open hand as a comb. In that way, controlled and rapid sound on–off events was possible, just like with a crossfader.

Feedback from the DJ and test persons The DJ that performed with the RadioBaton commented that although there was a lack of mechanical feedback from the interface, it opened up to new possibilities. Controlling recorded techniques was considered to be hard, especially to get the correct tempo. A scratch DJ uses physical markings on the vinyl (stickers, label) to see where in a sound the pick-up is, and this feature is moved from the controller to the GUI in Skipproof. This takes time getting comfortable with and is not at all optimal.

Persons without DJ experience have found the set-up with RadioBaton as turntable and light switch as crossfader to be intuitive and exciting, and quite fast they could perform with it

Figure 5.13: DJ 1210 Jazz scratching with Skipproof in a concert. The RadioBaton is operated with his right hand. Beside the computer screen is the lamp for the light sensor, on the floor is a rack of foot switches. His left hand is on the crossfader.

in a simple fashion.

Conclusions

It is realizable to build a system that enhances DJ performances of scratching, and it is desirable to experiment with the equipment currently preferred by DJs. Performing with models of scratch techniques is still a novel approach that needs to be tested more.

Commercial products are currently striving towards giving accurate control of digital sound files by means of turntable-like devices. Evolution of the crossfader is mostly focussed on the mechanical perspectives such as what kind of fader is used (optical, magnetic etc.).

DJs seem overall to be interested and intrigued by new technology and possibilities. Different playing styles demand different controllers and instruments, therefore the tools for mixing, scratching and beat juggling might become more specialized in the future.

5.5.3 Virtual air guitar

A combination of hand-held controllers and a guitar synthesizer with audio effects is called here the ”Virtual Air Guitar” (VAG). The name refers to playing an ”air guitar”, i.e., just acting the playing with music playback, and the term virtual refers to making a playable synthetic instrument. Sensing of the distance of hands is used for pitch control, the right hand movements for plucking, and the finger positions may in some cases be used for other features of sound production. The synthetic guitar algorithm supports electric guitar sounds, augmented with sound effects and intelligent mapping from playing gestures to synthesis parameters. This section describes the main principles of the instrument, particularly its control and how it is played.

Introduction

Electronic and computer-based musical instruments are typically developed to be played from keyboard, possibly augmented by foot, breath or other controllers. In the present study we have explored the possibility to make an intuitive yet simple user interface for playing a particular virtual (synthetic) instrument, the electric guitar. In addition to the synthetic instrument and related audio effects (amplifier distortion and loudspeaker cabinet simulation) we have explored three different controllers for player interface: data gloves in a virtual room environment, webcam-based camera tracking of player’s hands, and special hand-held controller sticks. The first one (data glove control) is for flexible experimentation of possible control features, while the two others are intended for maximally simplified guitar playing, designed for wide audience visiting a science center exhibition.

The main effort has been to study how the hand positions and movements can be mapped to control typical playing of the electric guitar. The two most important parameters needed are the pitch control ’ (corresponding to fretting position) and the string plucking action. In all three cases of controller design the pitch-related information is taken by measuring the distance of the two hands, which was found easier to use than the distance of left hand to a reference such as the players body. This distance information can be mapped also to other forms of control, such as selecting pre-programmed chords to be played. The string plucking action is most easily captured by the downward stroke of the right hand.

This means a highly simplified user interface, which sets strict limits to what can be played

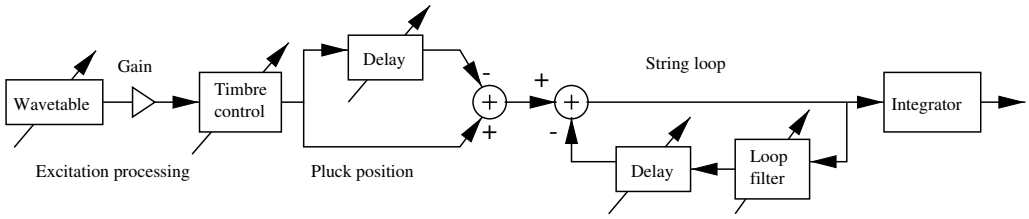

Figure 5.14: Basic structure of the single-delay loop filter and excitation mechanism for implementing the Extended Karplus-Strong string model.

by such a virtual guitar, certainly not satisfactory for a professional guitarist. It is, however, an interesting case of studying what can be done to demonstrate the basic features of playing a particular instrument with a given style, or to make inexpensive 'toy' instruments for fun. This can then be augmented by extra functions in more complex controllers, for example using finger movements, foot pedals, etc. In such cases the challenge is to find expressive and intuitive forms of controls that may be useful also for professional musicians. Another way, to get rich sound from minimalistic controllers, is to use complex rule-based mappings from simple control signals to more advanced control signals for playing a virtual instrument.

In this section we present an overview of the virtual air guitar design, paying most attention to the user interface aspects from the players point of view. A more detailed description[9] is found for example in Karjalainen et al. [2004]. Publications on physics-based modeling and control of the synthetic guitar have been described earlier for example in Jaffe and Smith [1983], Sullivan [1990], Karjalainen and Laine [1991], Jánosy et al. [1994], Karjalainen et al. [1998], and Mäki-Patola et al. [2005].

Synthesizing the electric guitar

The virtual instrument used in this study is a simulation of an electric guitar tuned to sound like the Fender Stratocaster. In this section we briefly characterize the development of the guitar model and of the sound effects used with it.

[9]Web documents on the project, including videos of playing the VAG, are available at: http://www.tml.hut.fi/~tmakipat/airguitar/virtualAirGuitar.html, and http://www.acoustics.hut.fi/demos/VAG/

Virtual Stratocaster The virtual electric guitar is realized using the Extended Karplus-Strong modeling technique described in Karjalainen et al. [1998] for the acoustic guitar except that the electric guitar does not need a body but instead a magnetic pickup. Figure 5.14 depicts the principle of a single string simulation as a signal processing diagram. An excitation wavetable signal is fed through spectral shaping (timbre control) to the string model, which consists of a pluck position filter and a string loop filter. This loop includes a delay to adjust the pitch and a loop filter to adjust the decay of vibration corresponding to string losses. The integrator converts the signal to string velocity. The magnetic pickup model is not shown, but it is similar to the pluck position filter.

Two such substring models need to be used in parallel to simulate the two polarizations of string vibration and possible beating in signal envelope due to the difference of string length in these two polarizations. Six such strings are needed for a full guitar model.

The string models need to be calibrated to yield sounds that are close to the original electric guitar sound. This is accomplished in the following way. Each string is plucked sharply very close to the bridge for the first and the 12th fret fingering, and the resulted signal is analyzed. The decay rate at each harmonic frequency is used to design the loop filter parameters for the two fingering positions. These parameters are then interpolated linearly for other fingering positions. The loop delay is also tuned as a function of fret position to get correct pitch. The two polarizations are tuned to have slightly different delays in order to obtain minor beating in string sound envelope for increased naturalness. The magnetic pickup filter (not shown) is designed to include a lowpass filter around 8 kHz to match with the recorded signals. Figure 5.15 shows a real measured spectrum and the corresponding modeled signal spectrum for string 1, fret 1, using bridge pickup, and when plucked 1 cm from the bridge.

Simulation of tube amplifier and loudspeaker Electric guitar players prefer to play through a tube amplifier that adds more or less distortion to the sound. A loudspeaker cabinet also adds distortion and shapes the signal spectrum. The distortion makes the spectrum richer, and for solo playing and simple chords the distortion is often preferred perceptually when compared to clean electric sound.

For our virtual electric guitar the preamplifier stages of a tube amplifier and loudspeaker cabinet were simulated digitally. Figure 5.16(a) presents a tube amplifier simulator as a signal processing diagram. In addition to lowpass (LPF), highpass (HPF), and delay (z^{-1}) blocks, the most essential part is a nonlinear block (F_{tube}) to simulate the triode tube stage characteristics,

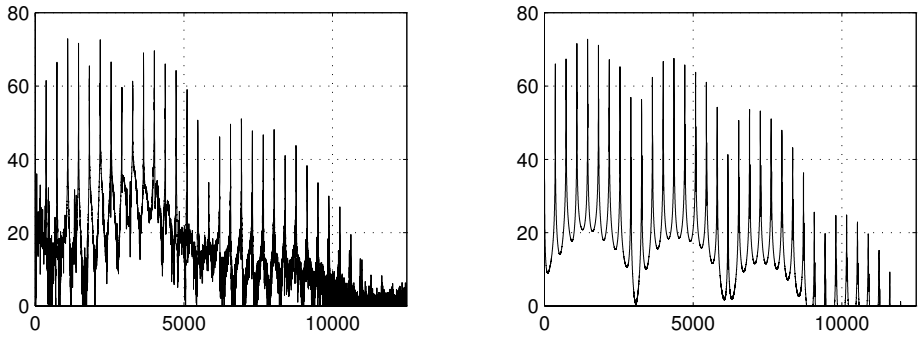

Figure 5.15: Comparison of (a) measured and (b) modeled spectrum response for string 1, fret 1, bridge pickup, plucked 1 cm from the bridge. The model is excited by an impulse. X-axis is frequency in Hz. Y-axis is level in dB.

which makes the distortion.

Figure 5.16(b) plots the frequency response of a loudspeaker cabinet simulator implemented simply as an FIR filter that has some ripple in the response and cuts high frequencies beyond 5 kHz. No nonlinearities of the speaker are simulated.

Controllers and user interfacing

The playability of a real electric guitar is based on right hand plucking/picking by a plectrum or by fingers and left hand fingering of strings at frets or letting strings to vibrate open. Each hand can be used to damp string vibration by touching as a soft termination. In slide guitar playing the effective string length is varied by terminating the string by a relatively hard object that can slide on the string.

A synthetic electric guitar can be played using any controllers that properly support the main control features of the instrument. As mentioned above, pitch and plucking are the minimum requirements, while many other controls are needed in professional playing. In this section we discuss three alternatives that have been used in our experimental VAG designs. Two of them were designed for a science centre exhibition, which required them to run without too special hardware, to be easy to grasp, and hard to break. Rule-based mappings from the physical parameters to instrument control parameters are also discussed.

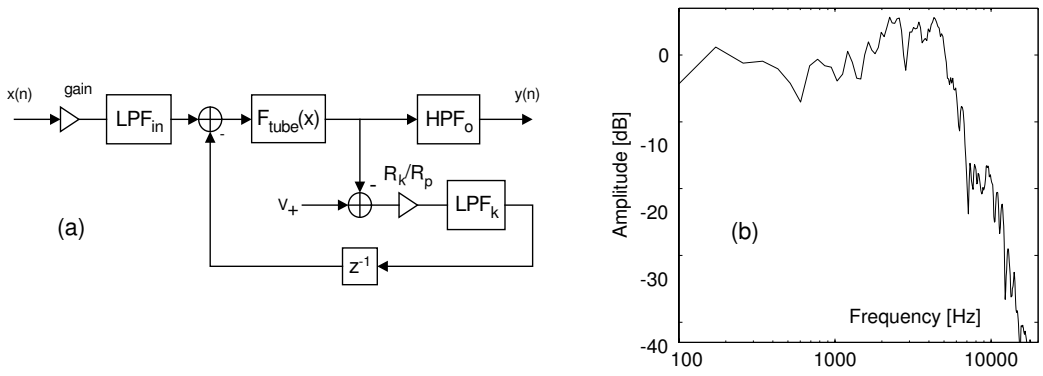

Figure 5.16: (a) Digital model of the tube amplifier stage and (b) Frequency response of Celestion Vintage 30 speaker in 4x12″ cabinet.

Data gloves Data gloves with 3-D tracking of hand positions as well as finger flexure parameters are used in a virtual room environment for experimentation of virtual guitar playing as shown in Figure 5.17. The distance between hands is measured for pitch control, right hand down-stroke makes the plucking, and finger information can be used for other effects, such as switching between continuous sliding of pitch or holding the pitch sampled during a pluck until the next pluck. In the latter case the pitch can be quantized to discrete fretted pitches. Among other parameters that can make playing expressive are string damping for long or short sustain, selection of string or chords, or controlling different audio effects such as distortion, echo, and reverberation.

So far we have applied the data glove control primarily to simple control strategies, because the studies have been targeted to the design and use of the simpler interfaces described below. The disadvantage of using data gloves is that they are expensive if constructed for accurate tracking. Thus they are useful primarily for research purposes.

Control sticks A version of user interface for the VAG has been developed that supports pitch and pluck control variables using special hand-held devices. Figure 5.18 illustrates a prototype version of such control sticks.

The pitch information to control the string length is measured as the distance of a sound transmitter and receiver in the sticks. The right-hand stick includes a tiny loudspeaker that sends high-frequency pulses and the left-hand stick has an electret microphone to receive the pulses.

Figure 5.17: Soulful playing of a VAG in a virtual room using data gloves.

The acoustic delay of the pulses due to sound propagation is converted to string length control. The frequency range of the transmitted pulses is around 20 kHz so that standard audio interfaces with sampling rate of 44.1 kHz can be applied, yet the pulses are not audible by the user due to insensitivity of the human ear at those frequencies.

The plucking information is captured by an acceleration sensor microchip inside the right-hand stick. The acceleration pulse triggers a transient signal (noise burst), proportional in amplitude to the acceleration peak value. This signal is fed to the string model(s) as an excitation to start a sound.

In addition to the hand-held sticks, a foot pedal can be used to control the sustain time or some audio effects, such as a wah-wah effect.

The electric guitar model of this case is controlled typically so that a pluck triggers two strings tuned to the interval of fifth, corresponding to what guitar players call the "power chord". When played through a tube amplifier distortion simulator, the VAG makes sounds that are typical in blues and hard rock music.

Figure 5.18: Control sticks for playing the VAG. Right-hand stick (on the left) includes an acceleration sensor as well as a small loudspeaker to send the distance measurement pulse. Left-hand stick (on the right) receives the pulse by an electret microphone.

Hand-tracking by a webcam camera A camera interface, being non-obstructive, gives the performer freedom to move intensely without worrying about breaking the controller or tying wires. By these properties it is the perfect interface for an intensive playing experience, especially for a science centre exhibition where tens of thousands of people, many of them children, will play the instrument.

Our camera interface for the guitar, see Figure 5.19, uses a common webcam camera for tracking the user. The user wears orange gardening gloves, the locations of which are extracted by finding the two largest blobs of orange color from the camera frames. This location data is then fed into a gesture extractor to detect when certain playing gestures take place.

A pluck is detected when the right hand passes through the imaginary guitar centerline. The guitar moves with the player and allows him to play even near the feet of behind the neck. Yet, as the location of the right hand is smoothed (lowpass filtered) for the centerline, the guitar does not shake along with the plucks.

The distance between the hands is transferred into fretting positions. The maximum width of the user's grip calibrates the "size" of the guitar. Thus, it can be played equally well by adults

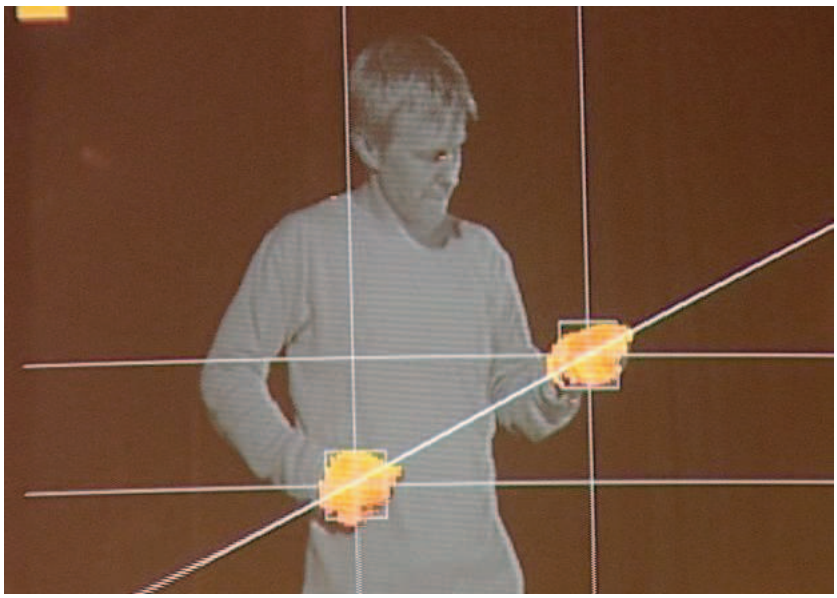

Figure 5.19: Camera tracking of hand positions (orange gloves), as seen by the computer.

and children. In addition to these basic controls a vibrato can be switched on by shaking the left hand along the imaginary guitar neck. Moving the hand more continuously along the neck translates into a fret-based slide. It is also possible to mute the strings by occluding one of the hands, for instance, behind one's back. The webcam interface also contains a foot pedal for changing between different playing modes.

Musical intelligence Instead of attempting to replicate every aspect of a real guitar, the Virtual Air Guitar's interface can use a complex mapping system to produce the sounds of guitar playing. The result is an entertainment device that is more in line with the way of playing air guitar - with showmanship, intensity and fun. In the webcam interface, the user is presented with two playing modes: rhythm guitar and solo guitar.

The rhythm guitar mode allows the user to strum four different power chords. These produce a heavy, thick sound, allowing the user to play "riffs", or rhythmic passages. Using these four chords, it is also possible to play the introduction to the song "Smoke on the Water" by Deep Purple, a well-known piece that is often used for air guitar playing.

The solo mode, on the other hand, allows the user to freely play a guitar solo on a pentatonic minor scale, with additional techniques such as fret sliding and vibrato. This mode also takes

into account the intensity of playing, producing satisfying distortion when the user plays very fast and hard.

To achieve this control, the mapping from gesture to sound model contains rules and procedures called the musical intelligence. The user's gestures are first interpreted into a meta language that describes guitar playing techniques on an abstract, musical level. For example, moving the right hand over the imaginary guitar strings in a strumming motion is interpreted as a pluck event. These events are in turn converted into control parameters for the sound model. The musical intelligence thus contains both the rules by which gestures are converted into these musical events, and the implementations of the events for a certain sound model.

Summary and conclusion

In this section we have described experiments on a playable virtual instrument called Virtual Air Guitar (VAG). It consists of hand-held controllers and a guitar synthesizer with sound effects. Three different user interfaces for controlling the instrument are experimented: data gloves used in a virtual room, optical tracking of hand movements, and special control sticks using acoustic and acceleration sensing of hand movements. The control parameters are mapped to synthesis control parameters in various ways: from direct control of pitch and plucking to artificial intelligence based advanced control.

The VAG is a good example of a virtual instrument that requires special controllers and playing strategies, different from keyboard oriented control. The simple versions described above are intended for toy-like applications, such as games, or instructional devices to characterize the most essential features of plucked string instrument playing.

Among future challenges are studies on more expressive control interfaces, which could be useful also by professional musicians. In contrary to real guitar playing, there is much more freedom to apply different gesture parameters for virtual instrument control.

5.5.4 The reacTable*

[10]The reacTable* is a state-of-the-art interactive music instrument, which seeks to be collaborative (local and remote), intuitive (zero manual, zero instructions), sonically challenging and

[10]This Section is a longer version of a paper published by same the authors [Jordà et al., 2005]

interesting, learnable and masterable [Wessel and Wright, 2002], suitable for complete novices (in installations) and for advanced electronic musicians (in concerts) and completely controllable (no random, no hidden presets). The reac-Table* uses no mouse, no keyboard, no cables, no wearables. In other words, the technology it involves is transparent to the user. It allows a flexible number of users, and these can enter or leave the instrument-installation without previous announcements.

Antecedents

FMOL and Visual feedback The reacTable* can be considered the successor of FMOL, a previous project developed by Jordà [Jordà, 1999, Jordà and Wüst, 2001, Jordà, 2002]. FMOL is a sonigraphical musical instrument [Jordà and Wüst, 2003] that has been used by hundreds of performers, be tween 1998 and 2002 in several on-line collective composition calls, and which is still used by the author in live performance and improvisational contexts. One of the main features of FMOL is its visual feedback capacity, which intuitively helps the understanding and the mastery of the interface, enabling the simultaneous control of a high number of parameters that could not be possible without this visual feedback. Like the abacus, in which beads, rods, and frame serve as manipulable physical representations of numerical values and operations [Ullmer and Ishii, 2001], FMOL makes no conceptual distinction between input and output. In FMOL, the mental model suggested by the interface reflects the synthesis engine conceptual model, while the resulting geometric "dance" of all of these elements, tightly reflects the temporal activity and intensity of the piece and gives multidimensional cues to the player. Looking at a screen like figure 5.21, which is taken from a quite dense FMOL fragment, the player can intuitively see the loudness, the dominating frequencies and the timbrical content of every channel, the amount of different applied effects, and the activity of more than thirty LFOs.

What is even more important, is that no indirection is needed to modify any of these parameters, since anything in the screen behaves simultaneously as an output and as an input. However, FMOL drags a very severe limitation: FMOL still is a conventional GUI Model-View-Controller application. There are limits to what can be efficiently achieved in real-time by means of a mouse and a computer keyboard. Building a tangible FMOL interface for a faster and more precise multi-parametric control seemed therefore a tempting idea. Designing a video detection or ultrasound system that would allow musicians to interact on a big projection screen, grabbing and moving strings with their hands, was the first idea we had. This could surely

Figure 5.20: The reacTable*

add a lot of visual impact to live concerts, although we soon discovered that musical control and performance did not improve but rather worsen with these additional technologies. These and other considerations took us to a completely new path, which should profit the knowledge gained during this years and bring it to a much more ambitious project: The reacTable*, first though as a new tangible controller for the FMOL synthesizer, finally evolved into a completely new instrument.

Tangible User Interfaces As the Tangible Media Group[11] directed by Professor Hiroshi Ishii at the MIT Media Lab states, people have developed sophisticated skills for sensing and manipulating our physical environments. However, most of these skills are not employed by traditional GUI. The goal is to change the painted bits of GUIs to tangible bits, taking advantage of the richness of multimodal human senses and skills developed through our lifetime of interaction with the physical world.[Fitzmaurice et al., 1995, Ishii and Ullmer, 1997, Ullmer and Ishii, 2001]. Several tangible systems have been constructed based on this philosophy. Some for musical applications, like SmallFish[12], the Jam-O-Drum [Blaine and Perkis, 2000, Blaine and Forlines,

[11]Tangible Media Group: http://tangible.media.mit.edu
[12]SmallFish: http://hosting.zkm.de/wmuench/small_fish

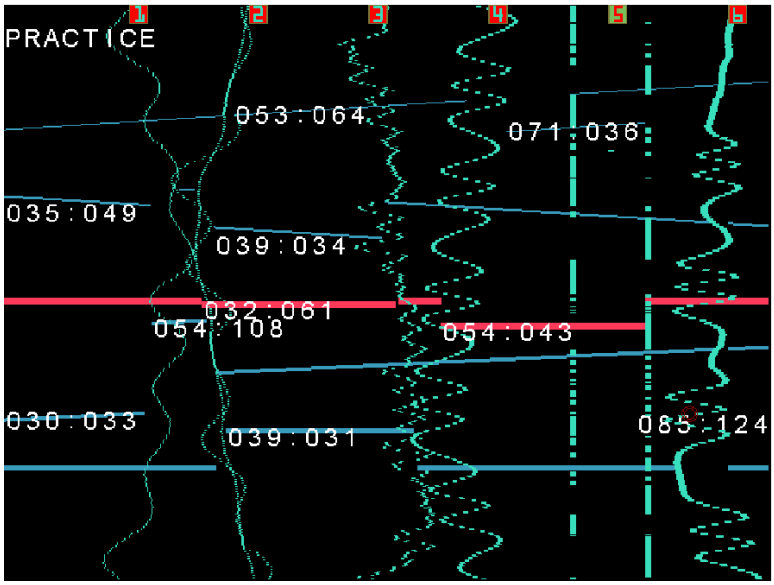

Figure 5.21: FMOL in full action

2002], the Musical Trinkets [Paradiso and Hsiao, 2000], Augmented Groove[13] or the Audiopad[14] [Patten et al., 2002], but we believe that no one attempts the level of integration, power and flexibility we propose: a table-based collaborative music instrument that uses computer vision and tangible user interface technologies, within a Max-like architecture and scheduler, and with FMOL-inspired HCI models and visual feedback.

Conception and design

Everything is possible The first step is to believe everything is feasible: we assume that we have access to a universal sensor which provides all the necessary information about the instrument and the player state, enabling the conception and design of the instrument without being driven by technology constraints. The reacTable* is a musical instrument based on a round table with a transparent surface, a video camera situated beneath the table continuously analyzes the table surface, tracking the nature, position and orientation of the objects that are distributed on its surface and the hand movements over the table. A projector draws a dynamic and interactive interface on it, while a spatial audio system provides the actual sonic feedback. The objects are

[13] Augmented Groove: http://www.csl.sony.co.jp/~poup/research/agroove/ [Poupyrev, 2000]
[14] Audiopad: http://tangible.media.mit.edu/projects/

mostly passive , without any sensors or actuators and are made out various materials of different shapes. Users interact with them by moving them, changing their orientation on the table or changing their faces (in the case of volumetric objects). For future versions of the reacTable* we are planning more complex objects such as flexible plastic tubes for continuous multi-parametric control, little wooden dummy 1-octave keyboards, combs (for comb-filters), or other everyday objects.

Modular synthesis and visual programming In the reacTable*, two additional concerns are no less important than visual feedback and tangibility: modular synthesis and visual programming. The concept of modular synthesis goes back to the first sound synthesizers, both in the digital [Mathews, 1963, Mathews and Moore, 1969] as in the analog domains, with Robert Moog's or Donald Buchla's Voltage-controlled synthesizers [Moog, 1965, Chadabe, 1975]. Modular synthesis has largely proved its unlimited sound potential and can be considered indeed as the starting point of all the visual programming environments for sound and music, which started with Max in the late 1980s [Puckette, 1988], and have developed into Pd[Puckette, 1996], jMax [Déchelle et al., 1999] or AudioMulch [Bencina, 1998], to mention a few. As shown by all of these healthy environments, visual programming constitutes nowadays one of the more flexible and widespread paradigms for interactive music making. The reacTable* is probably the first system that seeks to incorporate all of these paradigms, in order to build a flexible, powerful and intuitive new music instrument; a table-based instrument which can be played by manipulating a set of objects that are distributed on top of the table surface.

Objects, connections and visual feedback The set of objects (as depicted in Figure 5.22), which are made available on the table, can be manipulated by the players in various ways: once put onto the table the objects are activated, the objects can be moved on the table surface - enabling the objects to relate to each other. The rotation angle of the object is tracked as well. ReacTable* objects are plain and passive, meaning that they do not come with any cables, switches buttons or whatsoever. The user also does not have to wear any special sensor or controller equipment for the object handling; the players plain hands are the only necessary controller. This of course, should not rule out the possibility of "smart" objects that incorporate additional internal electronics in order to retrieve some additional sensor data, coming from squeezing, bending or bouncing them, like in the case of the Squeezables [Weinberg and Gan, 2001]. A rubber hose or a wooden toy snake, whose state could be either determined by the computer vision or by using some

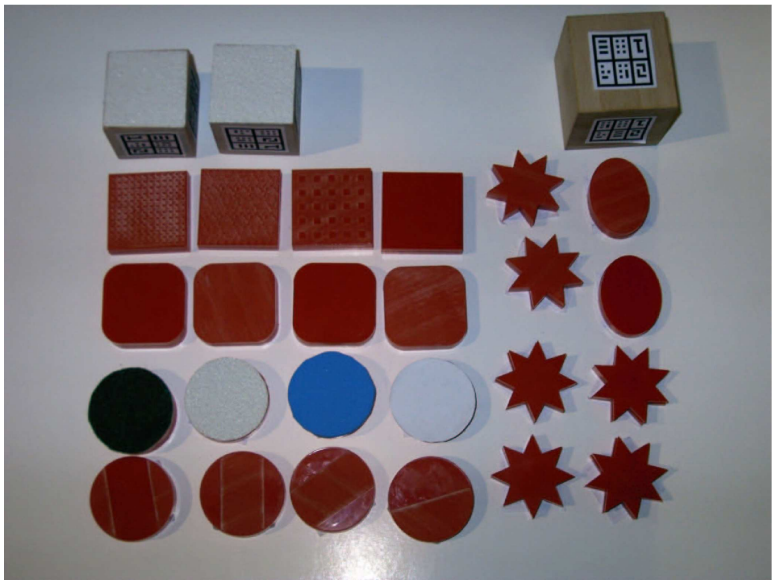

Figure 5.22: Several reacTable* objects (photography by Martin Kaltenbrunner)

bending sensors like in the Sonic Banana [Singer, 2003], can serve as an advanced controller producing multi-dimensional control data. In any case, this would have to be achieved in a completely transparent way, using wireless technology for example, so that the performer can treat all objects in an equal way.

Each of these objects has its dedicated function for the generation, modification or control of sound. Like Max and its cousins, the reacTable* distinguishes between control and sound objects, and between control and sound connections. Bringing these objects into proximity with each other constructs and plays the instrument at the same time. When a control flow is established between two objects, a thick straight line is drawn between them, showing by means of dynamic animations, the flux direction, its rate and its intensity. Audio flows, on their turn, are represented like those in FMOL, by means of instantaneous waveforms.

Moreover, the reacTable* projection follows the objects on the table, wrapping them with auras. An LFO, for example, will be wrapped by a blinking animation that will keep showing the frequency, the amplitude and the shape (e.g. square vs. sinusoidal) of the oscillation. These visualizations never shows text, buttons, sliders or widgets of any kind. What is shown at every moment, is only the instrument activity and behaviour, the objects' types and positions and the relations between them all, in an abstract but direct and non-symbolic way. Figure 5.20 illustrates this visual feedback.

Figure 5.23: The first working reacTable* prototype

The hands play an important role: not only can they manipulate reacTable* objects, they are treated as superobjects themselves. We also track the position and state of the hands in order to retrieve additional control data. A simple gesture to illustrate this principle is the cut or muting of a sound stream, which is done with a karate-style hand gesture. Another nice example are the waveform and the envelopes objects, which can be simply programmed by finger-painting respectively a waveform or an envelope close to them. This painted waveform is "absorbed" by the object that starts playing it immediately. As already mentioned above, the hand can also control the visual sound representation by cutting or redirecting the sound flow.

First informal tests indeed show that this visual feedback is actually crucial for the playability of the instrument. The central feature of this visual feedback is the visualization of the sound and control flows between the processing objects, which are basically a representation of the waveform state between two objects.

The reacTable* Architecture

Figure 5.23 shows the first prototype from 2003, which demonstrates the camera setup from below. Figure 5.24 illustrates all the reacTable* system components. In this section we will discuss each of them.

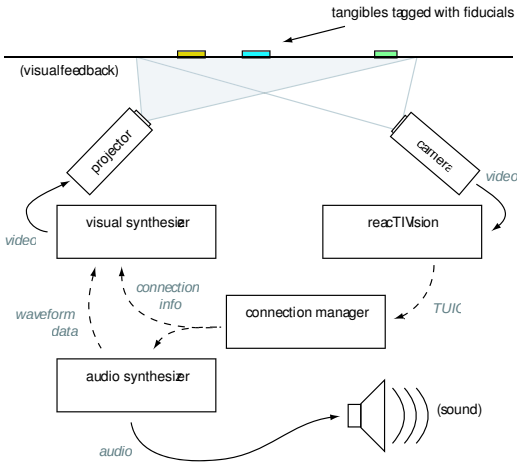

Figure 5.24: The reacTable* architecture

Vision The reacTable* sensor component reacTIVision is an open source system for tracking the type, location and orientation of visual markers in a real-time video stream. The system was developed within the Music Technology Group by Ross Bencina [Bencina et al., submitted] after we had developed an initial prototype using Costanza and Robinson's d-touch system [Costanza and Robinson, 2003] that is based on the recognition of similar fiducial markers. The decision to develop a new tracking system stemmed primarily from the relatively low frame rates achieved with d-touch, we sought to support high frame rates, while maintaining its overall robustness and accuracy. Later, additional requirements emerged such as reducing the dimensions of the fiducials and increasing the number of uniquely identifiable fiducials.

reacTIVision is a multi-platform standalone application, which sends the acquired sensor data via TUIO [Kaltenbrunner et al., submitted] a protocol based on OpenSound control [Wright, 2003] to a client application. Objects have to be tagged with simple black and white markers, with unique topologies to allow their distinction. The marker size usually depends on the camera resolution and distance; we are currently working with markers of around 4 cm^2. A vision engine has to be sufficiently fast for the needs of an expressive musical instrument, thus providing a high temporal resolution for the tracking of fast movements. The reacTIVision engine in our current setup processes 60 frames at a resolution of 640x80 pixels in real-time on a 2GHz Athlon system, and scales without any problems to higher frame rates or resolutions. We are using a set of 128 unique marker symbols, although the amount of symbols can be easily extended up to several thousands at the cost of slightly larger marker sizes.

Connection manager: dynamic patching Provided the raw sensor data of the objects' type, position and orientation a central connection manager is calculating the actual patch network similar to traditional visual programming languages. The resulting connections are then sent to the actual sound and graphics synthesizer components.

Dynamic patching does not require the user to explicitly connect the objects. We defined a simple set of rules, which automatically connect and disconnect objects. All objects have a certain number of different in-output connectors: sound, control, sync etc. Based on a simple distance rule, each object checks its neighborhood for objects, which can provide both compatible and available ports. It will therefore always choose the closest available object to connect to. The reacTable* connection paradigm produces a highly dynamic environment. Moving an object around the table surface permanently interferes and alters existing connections, creating extremely variable synthesizer morphologies. This behavior might be disturbing in certain conditions; therefore we introduced an additional hard-link gesture, which establishes a permanent link between objects that can't be broken by the conventional rules. Other objects such as the synchronizers emit a kind of pseudo-physical force fields around them, thus synchronizing only objects, which are under their current influence. Again this connections are calculated based on the relative positions of all objects and only the resulting connection or disconnection commands are sent to the synthesizer.

Along with the actual patch network information the connection manager also sends a series of raw control data values for each object. These values include distances and angles of connected objects as well as their movement speed and acceleration values. It is decided within the synthesizer component how to map this raw environment control data onto the synthesis process. Please note that this data is retrieved directly from the table setup and is not related to the control connections between synthesis objects.

Audio synthesizer Currently, the reacTable* objects can be generally categorized into seven different functional groups: Generators (1 audio out and a varied number of control in), Audio Filters (1 audio in, 1 audio out and a variable number of control in), Controllers (1 control out), Control Filters (1 control in and 1 control out), Mixers (several audio in, 1 audio out), Clock synchronizers and Containers. There are also some exceptions that do not fit within any of these categories. Figure 5.25 shows a topology with three independent audio threads, as taken from the reacTable* software simulator.

The synthesizer is controlled by the connection manager, which means that information

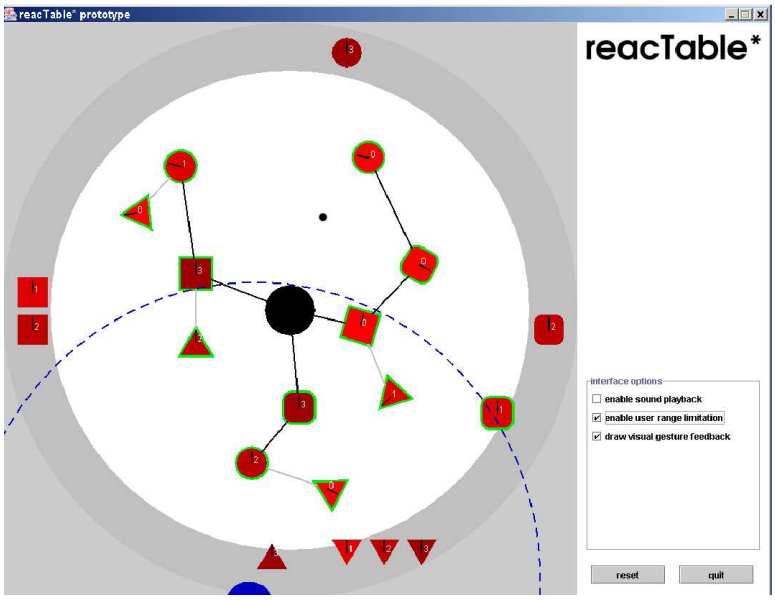

Figure 5.25: A snapshot of the reacTable* software simulator

about which objects are on the table and what is connected is sent via Open Sound Control (OSC) [Wright, 1997] to the synthesizer. The current implementation of the synthesizer uses the Pure Data computer music system, but it is possible to write synthesis modules in any system that supports OSC, as long as it is possible to implement the mapping logic and to instantiate, connect and delete objects dynamically. Using a higher level synthesis language allows to easily adapt and expand the audio synthesis by changing existing objects and adding new ones.

In order to avoid clicks, connections and disconnections of audio data are done with a fadeout/fadein. The general set of OSC messages that are understood by the synthesizer module are:

new/delete of objects (parameter is the name of an object class)

connect/disconnect between the different object ports.

activate/deactivate

messages for parameter passing, always referenced by a numerical ID.

several special messages (e.g additional info from hand gestures)

The synthesizer also takes care of the mapping of parameters from the table. The implementation of the mapping on the synthesizer side allows for very flexible mapping algorithms and presets.

A set of a dozen precomputed parameters, ranging from 0 to 1 is sent by the connection manager. These parameters are position of the object in polar coordinates, angle of the object, angle and distance to connected objects. The synthesizer engine can map these according to object types. e.g. the angle of an oscillator influences the frequency, and the angle of the oscillator regarding to the next object might influence the waveform. For a sample player, the angle might influence the playback speed.

Visual synthesizer The visual feedback is implemented in another standalone application in a similar way as the audio synthesizer, thus constituting a "visual synthesizer". The connection manager sends the information about the objects' connection state and their parameters to this engine, which interprets this information in a straightforward way; It draws the objects at their correct positions and draws lines for connecting them.

For the correct visualization of these connections, the graphics engine has an additional connection to the audio synthesizer, from where it receives the information about the data flows in the synthesizer. Audio connections are visualized by their waveforms in the time-domain and are sent as a OSC blob. Control connection data is sent directly as floating point values over the same OSC protocol. Additional information such as the internal object state is also visualized with data provided by the sound synthesizer. The "Aura" of a LFO for example, is pulsating in order to reflect the LFO's oscillation frequency in a visual way.

To draw the visual feedback the graphics engine uses the platform independent accelerated OpenGL library. Drawing styles can be changed by selecting different graphical schemes, which commonly are called skins. It is also possible to create new skins, which makes the look-and-feel of a reacTable* performance easily customizable. Figure 5.20 shows an example of the visualization.

reacTable* hardware As in the current prototype the complete reacTable* hardware is installed within a round wooden table with a semitransparent plastic surface. This surface allows the projection of the visual feedback and is sufficiently transparent to clearly detect the object symbols while they are in contact with the table surface. This property has the additional advantage that

objects are immediately lost by the vision sensor when they are lifted from the table, From beneath a standard projector and a firewire camera are covering the complete table surface of around one meter in diameter. This is achieved within a considerable small distance by using mirrors and wide-angle lenses. In order to separate the camera acquisition from the projection image the table's optical system is working within two separate spectra. The camera is operating completely within the near infrared spectrum, illuminating the table's interior with an array of infrared LEDs and filtering the visible light using an IR photography filter. While common CCD cameras are sensitive to the near IR illumination it is completely invisible to the human player. The projection of course is operating within the visible spectrum, while its infrared component is filtered away to avoid light flares in the camera image.

Performing with the reacTable*

At the time of this writing, the reacTable* is not yet completely finished. It is still more a prototype than a real musical instrument. While we have yet to learn to play it, some models of playing can already be anticipated.

Novel and occasional users: discovering the reacTable* The reacTable* has been conceived for a wide spectrum of users. From the absolute novice in an interactive installation setup, to professional performers in concert venues. This is attempted by designing an instrument as intuitive as possible, and at the same time, capable of the maximum complexities.

The reacTable* was conceived to be played from the first minute, without a user manual or a single line of instructions. The approach taken towards novice users could be summarized in the following sentence: avoid user's frustration at any cost. To avoid frustrations, a system does not necessarily have to be completely understandable, but it has to be coherent and responsible. The reacTable* has to work "by default" and any gesture has to produce audible results, so for example if on start-up, a user activates an object that does not sound (i.e. a control object) the closest audio object is automatically linked to it (and the link is visualized).

The reacTable* wants to be user-proof. For instance, it seems natural that in an installation context, after some minutes exploring the instrument some visitors may start stressing the system in different ways, like placing personal objects onto the table. Although it is no possible to anticipate all objects that users may use, some of the more common could be detected (mobile phones, keys, pens, lipsticks, cigarette packets, lighters) and a "funny" functionality could be

added to them (e.g. mobiles could be used as cheesy mobile-like melody generators).

Advanced reactablists: performing and mastering the instrument The reacTable* can bring increasing and compelling complexities for the advanced users, allowing them to combine simultaneously diverse and complementary performance approaches and techniques. We describe here four of them.

Towards the luthier-improviser continuum Within traditional modular visual programming synthesizers, there is a clear separation between building and playing the patch (or instrument): There is an editing and an execution mode. The editing is usually a lengthy development process, which leads to a final and stable instrument patch, which then during the execution mode is controlled on screen or via any available controller device. The reacTable* has to be built and played at the same time. Each piece has to be constructed from scratch starting from an empty table (or from a single snapshot which has been re-constructed on the table before the actual performance). This is a fundamental characteristic of this instrument, which therefore always has to evolve and change its setup. Building the instrument is equivalent to playing it and vice-versa. Remembering and repeating the construction of a building process can be compared to the reproduction of a musical score. The reacTable* establishes thus a real continuum not only between composition and performance, but between lutherie, composition and performance.

Whereas FMOL has proved to be quite flexible, its architecture is predefined by a grid of 6x4 generators and processors, which have to be previously selected from a palette of presets, so that the process of building an orchestra is not done in realtime while playing. Moreover, in FMOL all the macro-control of form is done like in traditional analog synthesizers, by means of simple and relatively low-level LFOs and arpeggiators. The reacTable* overcomes all these limitations. Its open structure favors (a) the existence of all types of higher level objects, such as the ones we could imagine within an environment such as Max or `Pd`(e.g. sophisticated rhythm and melody generators, chaotic generators, pitch quantizers and harmonizers, etc.), and (b) the construction of all kind of sound synthesis and processing nets and topologies.

Moreover, the reacTable* connection paradigm in which by moving an object around the table surface, the performer is able to permanently interfere and alter existing connections, creates extremely variable and yet because of visual feedback easily understandable and predictable synthesizer morphologies, just at the reach of one hand.

The bongosero-karateka model The reacTable* also permits physically intense playing. Grabbing, lifting and dropping objects with both hands, cutting flows with karate-like gestures and reactivating them by touching the objects again, will only be limited by the computer vision engine speed. The current implementation predicts that the input frame rate will not go below 30 Hz, while 60 Hz is attainable with an adequate camera.

The caresser-masseur-violinist model Objects on the table permanently sense at least three parameters, depending on their relative position within the net topology and of their angle orientation. This allows for very subtle and intimate control. Moving and twisting delicately two objects allows to precisely control six parameters, without scarifying voluntarily brusque and sudden morphology changes and discontinuities.

The painter model The reacTable* allows free finger-drawing directly on all of the table's surface. This functionality includes drawing envelopes, wavetables or spectra, depending on which objects are situated nearby.

The reacTable* as a collaborative multi-user instrument The reacTable* supports a flexible number of users with no predefined roles, and allows simultaneously additive (users working on independent audio threads) as well as multiplicative (users sharing control of audio threads) behaviors. Because of the way physical objects are visually and virtually augmented, the reacTable* also constitutes a perfect example of the local and remote all-at-once multi-user instrument. In a local collaboration scenario two or more players can share the same physical objects and their space. This collaborative space is only limited by the diameter of the table and by the players' claustrophobia, but a normal situation we would support between two and four players. This amount can be extended when two or more reacTables* are connected through the net. Sharing the same virtual space only, performers can only move the physical objects on their local table, while these are only projected onto the remote table, Their movement may modify the shared audio threads, thus provoking interactions between displaced objects, so that one filter controlled in Barcelona may process the output of a generator in Linz. In a third collaboration scenario, remote users could join a reacTable* session with a software simulation, where the virtual table would have the same impact as remote tables, without the tangible interaction.

The reacTable*: Conclusion

In the reacTable*, performers share complete access to all the musical threads by moving physical objects (representing generators, filters, etc.) on a table surface and constructing different audio topologies in a sort of tangible modular synthesizer or graspable flow-controlled programming Max-like language. Unlike many new designed instruments, its origin does not come from approaching its creation by exploring the possibilities of a specific technology, nor from the perspective of mimicking a known instrumental model. The reacTable* comes from our experience designing instruments, making music with them, and listening and watching the way others have played them. The reacTable* team[15] is currently constituted by Sergi Jordà, Martin Kaltenbrunner, Günter Geiger, Ross Bencina, Hugo Solis, Marcos Alonso and Alvaro Barbosa. It is an ambitious project, in which, needless to say, we have placed great hope and expectation. We expect it to leave the laboratory soon, and for people to start creating wonderful music with it.

5.5.5 The interactive book

A book is a very well known object which everyone has used at least once in his life. It plays an important role in children education: most of us learned colors, names of animals and numbers just 'reading' or better interacting with some nice, colored pull-the-tab and lift-the-flap books. In the last decades children's books have been modified in order to use new interaction channels, inserting technology inside this old medium or using the book metaphor to develop new interfaces. It is quite clear that technology did not change too much for the book in thousand years: the history of books has seen new printing and composition techniques but the users are still basically dealing with the same artifact. Thus, the book as an object guarantees a high level of functionality. This section[16] is a review of book interfaces with particular attention to future

[15]We would like to thank the former interns within the reacTable* team Ignasi Casasnovas, José Lozano and Gerda Strobl for their valuable contributions to the project. We are also thankful for Enrico Costanza's contribution of his d-touch vision engine and Sile O'Modhrain's suggestions to our initial tangible prototype. This work was partially supported by the European Commission Cost287-ConGAS action on Gesture controlled Audio Systems.

[16]The conception and realization of an early prototype of a sound-augmented book were carried on by the second author as part of the Sounding Object project[17]. Later on, students Damiano Battaglia (Univ. of Verona) and Josep Villadomat Arro (Univ. Pompeu Fabra, Barcelona, visiting Verona in 2004) realized the sketches that are described in this paper as part of graduation projects, under the guidance of the authors. At the moment, the first author is pursuing her PhD on fundamental issues in sound-mediated interaction, such as causality and Fitts' Law, and she

developments of children's books, since the latter feature most of the possible applications related to sound and interaction.

Introduction

Current commercial interactive books for children are very often similar to conventional colored stories with the addition of some pre-recorded sounds which can be triggered by the reader. The limitations of these books are evident: the sounds available are limited in number and diversity and they are played using a discrete control (typically a button). This means that sounds are irritating rather than being a stimulus to interact with the toy-book or allowing for learning by interaction.

Pull-the-tab and lift-the-flap books play a central role in the education and entertainment of most children all over the world. Most of these books are inherently cross-cultural and highly relevant in diverse social contexts. For instance, Lucy Cousins, the acclaimed creator of Maisy (Pina in Italy), has currently more than twelve million books in print in many different languages. Through these books, small children learn to name objects and characters, they understand the relations between objects, and develop a sense of causality by direct manipulation [Hutchins et al., 1986, Schneiderman, 2002] and feedback. The importance of sound as a powerful medium has been largely recognized and there are books on the market that reproduce prerecorded sounds upon pushing certain buttons or touching certain areas. However, such triggered sounds are extremely unnatural, repetitive, and annoying. The key for a successful exploitation of sounds in books is to have models that respond continuously to continuous action, just in the same way as the children do when manipulating rattles or other physical sounding objects. In other words, books have to become an embodied interface [Dourish, 2001] in all respects, including sound.

The history of interactive books

In the nineties, the introduction of the *e-book* was supposed to trigger a revolution similar to that of CDs versus vinyl recordings. Many companies tried to produce hardware and software devoted to digital reading: an important example is the Gemstar GEB 2150 [Negroponte, 1996], which was a portable e-book reader. Unfortunately, in 2003 Gemstar had to stop the development and production of its products (Softbook, Nuvomedia, GEB, etc.). The main problem that was

is looking at children books as a possible application scenario.

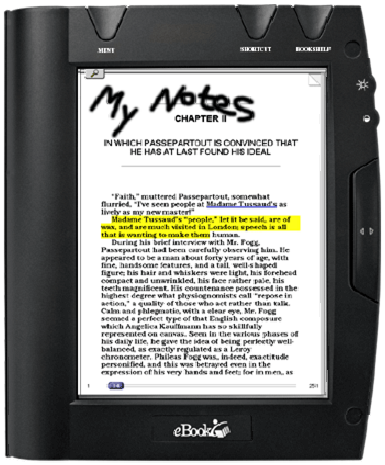

Figure 5.26: The Gemstar GEB 2150 e-book portable reader

immediately apparent in this commercial flop (beyond the enthusiastic expectations [Negroponte, 1996]) was that researchers thought to identify the book object with just one informative medium: its text. Such an electronic book cannot win when compared with the traditional book: the thickness of a book, the possibility of quickly browsing its pages, etc. are all important features. Some experiments about the rendering of such features have been carried out [Chu et al., 2003, 2004, Sheridon and Berkovitz, 1977].

In 1968 Kay Goldberg [Kay and Goldberg, 1977] started to work on the *Dynabook*. The *Dynabook* was thought to be a portable instrument which guaranteed simultaneous memory storage and access to all the necessary information using different user–selectable modes. The same information (texts, images or sounds) could be managed in a different way each time around according to user needs. The Dynabook in Fig.5.27 can be seen as an ancestor of the modern notebooks or tablet-PCs.

The research aim was mostly focused on translating in electronic form the book object and all the information that could be derived from the medium. The technology was still the main focus and the main problem, while user interaction was still the last priority. The user had to bend over to technology and the opposite was hardly taken into consideration. Forty years ago, when the first Dynabook was realized, the main problem was to realize portable and effective instruments with the available technology: nowadays the computation power and the costs

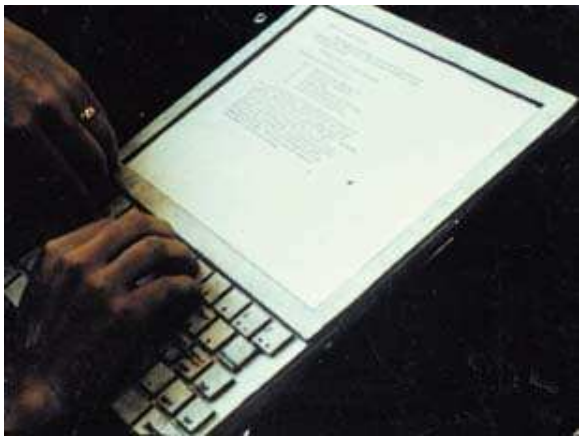

Figure 5.27: The Dynabook

are not a problem any longer while creativity is the new frontier. The Gemstar and Dynabook experiences have shown that the book interface is really strong and that a successful approach could involve the hiding of technology leaving the original interface similar to the old winning one as much as possible.

Focusing on augmented books for children, it is worthwhile to mention the work done by researchers of the Xerox PARC (Palo Alto Research Center) Laboratories. In the SIT (Sound-Image-Text) book prototype [Back et al., 2001] the idea was to use sensors in order to use the reader hands' speed as control parameter for some sounds effects. Listen Reader [Back et al., 1999] is the natural prosecution of the SIT Book project: while in the SIT Book the sounds effects are used to create a soundscape for the narrative content, here the sound part is a foreground element in itself, conveying information about what the child is reading. Gestures are used to control the sound synthesis, and the electronic part is completely hidden to enhance the naturalness of the interaction.

The Listen Reader [Back et al., 2001] was the natural development of the SIT Book project. While in the SIT book the audio effects are a sort of background for the story, this interface uses music as an important element of the storyboard. Sounds are informative about the environment and about what is happening in the story. The prototype is a seat with a stand where the interactive book is placed: user gestures control sound synthesis and a Radio-Frequency Identifier (RFID) sends information about the precise page which the user is currently looking at. The technology is hidden from the user who is interacting with some very usual objects such as a seat and a book (see fig. 5.28).

Figure 5.28: Children playing with the Listen Reader

Future perspectives

In recent years, the European project "The Sounding Object"[18] was entirely devoted to the design, development, and evaluation of sound models based on a cartoon description of physical phenomena. In these models the salient features of sounding objects are represented by variables whose interpretation is straightforward because based on physical properties. As a result, the models can be easily embedded into artefacts and their variables coupled with sensors without the need of complex mapping strategies.

Pop-up and lift-the-flap books for children were indicated as ideal applications for sounding objects [Rocchesso et al., 2003], as interaction with these books is direct, physical, and essentially continuous. Even though a few interactive plates were prototyped and demonstrated, in-depth exploitation of continuous interactive sounds in children books remains to be done.

Everyday sounds can be very useful because of the familiar control metaphor: no explanation nor learning is necessary [Brewster, 2002]. Moreover, it is clear that the continuous audio feedback affects the quality of the interaction and that the user makes continuous use of the information provided by sounds to adopt a more precise behavior: the continuously varying sound of a car engine tells us when we have to shift gears. In this perspective sound is the key

[18] http://www.soundobject.org

for paradigmatic shifts in consumer products. In the same way as spatial audio has become the characterizing ingredient for home theatres (as opposed to traditional TV-sets), continuous interactive sounds will become the skeleton of electronically-augmented children books of the future. The book-prototype is designed as a set of scenarios where narration develops through sonic narratives, and where exploration is stimulated through continuous interaction and auditory feedback. Through the development of the book, the class of models of sounding objects has been deeply used and verified[19]. The physical models of impacts and friction have been used to synthesize a variety of sounds: the steps of a walking character, the noise of a fly, the engine of a motor bike, and the sound of an inflatable ball.

Scenarios Design

The integration and combination of the sound models available from the Sounding Object project in an engaging tale has been studied. The first step was to create demonstration examples of interaction using different kinds of sensors and algorithms. During this phase the most effective interactions (i.e. easier to learn and most natural) have been chosen, and several different scenarios were prepared with the goal of integrating them in a common story. The scenarios use embedded sensors, which are connected to a central control unit. Data is sent to the main computer using UDP messages through a local network from sensors and the sound part is synthesized using custom designed Pure Data (`Pd`)[20] patches. These `Pd`patches implement a set of physical models of everyday sounds such as friction, impacts, bubbles, etc. and the data coming from sensors is used to control the sound object model in real time. In the following subsections an investigation scenario will be described.

Steps The *steps* scenario shows a rural landscape with a road; an embedded slider allows the user to move the main character along the road, and all movement data are sent to the computer, where the velocity of the character is calculated and a sound of footsteps is synthesized in real-time. The timing, distance, and force of the sound of each step is modified as a function of the control velocity. Fig. 5.29 shows a preliminary sketch, while fig. 5.30 shows the final prototype with the embodied sensor.

[19] http://www.soundobject.org/articles.html
[20] http://www.pure-data.info

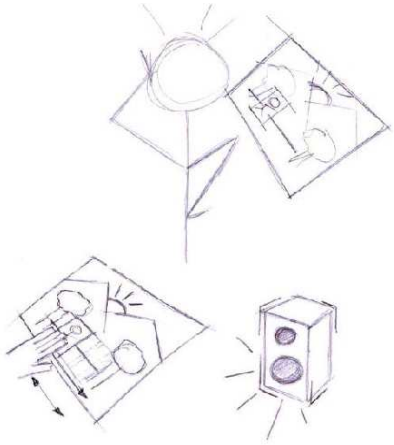

Figure 5.29: The user is looking at the scene, identifies the moving part and tries to move the character generating sound

Conclusions

Our investigation shows that in a near future lift-the-flap books for children will be augmented by sounds that respond continuously and consistently to control gestures. The sample scenario shown in the previous paragraph demonstrates the effectiveness of sound as an engaging form of feedback and the possibility of embedding real-time physics-based models of everyday sounds in small inexpensive stand-alone systems. A relevant part of future work will concentrate on real-world tests with children that will enhance the playability/usability of prototype books. Another aspect which will be further developed is the embedding and the sophistication of the technologies used.

5.6 Multimodal and cross-modal control of interactive systems

The development of multimodal and cross-modal algorithms for integrated analysis of multimedia streams offers an interesting challenge and opens novel perspectives for research on multimedia content analysis, multimodal interactive systems, innovative natural and expressive interfaces Camurri et al. [2004e], especially in the framework of multimodal and expressive control of interactive systems.

Multimodal processing enables the integrated analysis of information coming from dif-

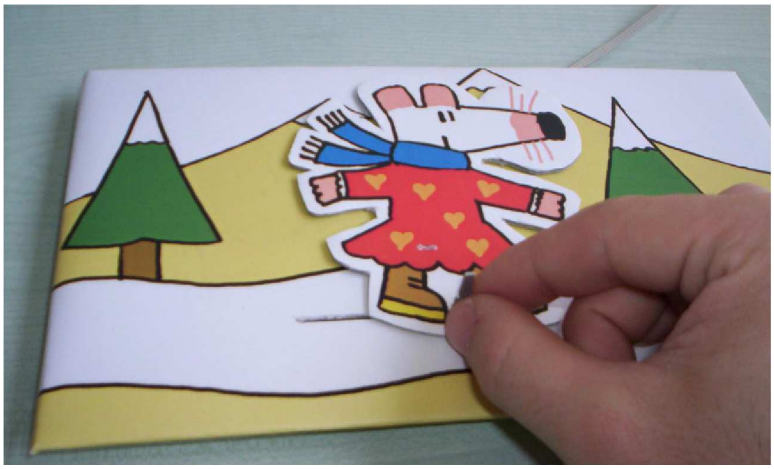

Figure 5.30: Interaction through slider: the footsteps scenario prototype

ferent multimedia streams (audio, video) and affecting different sensorial modalities (auditory, visual). Cross-modal processing enables exploiting potential similarities in the approach for analyzing different multimedia streams so that algorithms developed for analysis in a given modality (e.g., audio) can be also employed for analysis in another modality (e.g., video).

This chapter presents and discusses some concrete examples of multimodal and cross-modal algorithms we used for analysis of expressive gesture Camurri et al. [2004c] and for real-time control of interactive systems. The algorithms have been implemented as software modules (blocks) or applications (patches) for the new EyesWeb 4 platform Camurri et al. [2005b] (www.eyesweb.org) which differently from its predecessors directly and explicitly supports multimodal and cross-modal processing.

5.6.1 Cross-modal processing: visual analysis of acoustic patterns

A first application of cross-modal processing consists in the analysis by means of computer vision techniques of acoustic patterns extracted from an audio signal by means of a collection of EyesWeb 4 modules for auditory modeling.

Such modules are included in an EyesWeb library providing the whole auditory processing chain, i.e., cochlear filter banks, hair cell models, and auditory representations including excitation pattern, cochleogram, and correlogram Camurri et al. [2005a]. The design of the cochlear filter banks relies on the Matlab Auditory Toolbox Slaney [1994]. To date, a filter bank configura-

tion can be exported in XML format and loaded into the EyesWeb plug in (see Figure 5.31). For example the cochleogram of voice sound is depicted in Figure 5.32.

The cochleogram images can be analyzed by image processing techniques to extract information that is not so directly accessible through audio analysis (e.g., activation of particular regions in the image, pattern matching with template images).

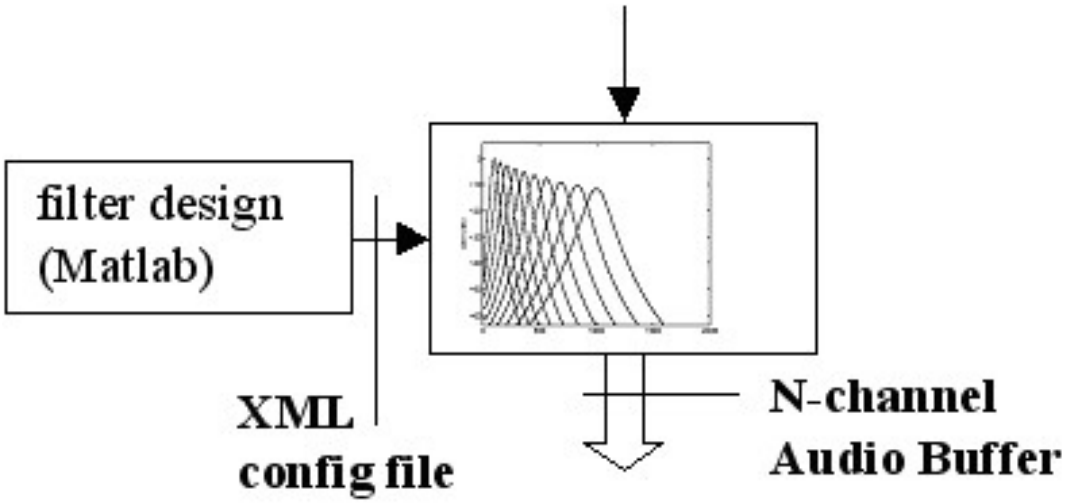

Figure 5.31: Design of the auditory filter bank through the Matlab Auditory Toolbox

In this first example of cross-modal techniques the cochleogram images are analyzed by applying to them the techniques for motion analysis included in the EyesWeb Gesture Processing Library Camurri et al. [2004d]. For example, in order to quantify the variation of the cochleogram, i.e., the variance over time of the spectral components in the audio signal, Silhouette Motion Images (SMIs) and Quantity of Motion (QoM) Camurri et al. [2003] are used. Figure 5.33 shows the SMI of a cochleogram (red shadow). It represents the combined variation of the audio signal over time and frequency in the last 200 ms. The area (i.e., number of pixels) of the SMI (that in motion analysis is usually referred to as Quantity of Motion, i.e., the amount of detected overall motion) summarizes such variation of the audio signal, i.e., it can be considered as the detected amount of variation of the audio signal both along time and along frequency in the time interval over which the corresponding SMI is computed (200 ms in this example).

From a first analysis of the data obtained with this approach it seems that the QoM obtained from the SMIs of the cochleograms can be employed for onset detection especially at the phrase level, i.e., it can be used for detection of phrase boundaries. In speech analysis

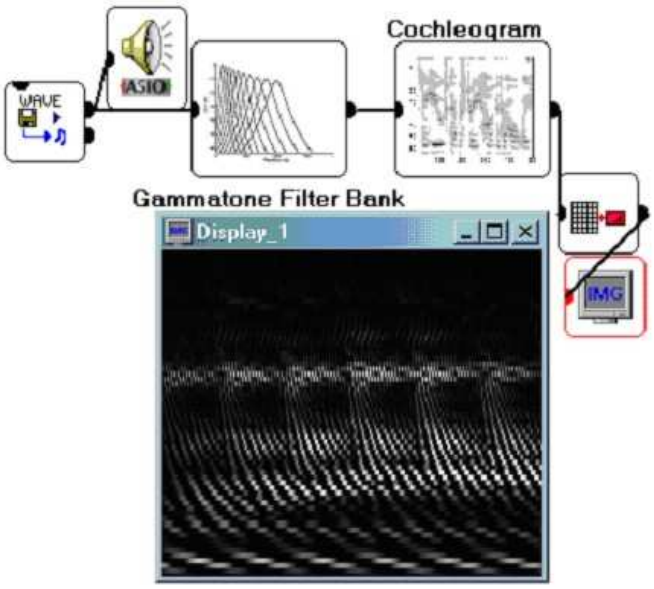

Figure 5.32: Cochleogram of a voice sound obtained through the auditory model blocks

the same technique can be used for segmenting words. Current research includes performance analysis and comparison with state-of-the-art standard techniques.

5.6.2 Cross-modal processing: auditory-based algorithms for motion analysis

Cross-modal processing applications can also be designed in which the analysis of movement and gestures is inspired by audio analysis algorithms. An example is the patch shown in Figure 5.34, in which a pitch detector is used to measure the frequency of periodic patterns in human gestures: the vertical displacement of a moving hand, measured from the video input signal and rescaled, is converted into the audio domain through an interpolation block, and then analyzed through a pitch detector based on the autocorrelation function.

Motion-derived signals and audio signals differ in terms of sampling rate and band characteristics. The conversion from a motion-derived signal to one in the audio domain can be performed in principle by upsampling and interpolating the input signal, and a dedicated conversion block is available to perform this operation. If m_{i-1} and m_i are the previous and present input values respectively, and t_i is the initial time of the audio frame in seconds, the audio-rate

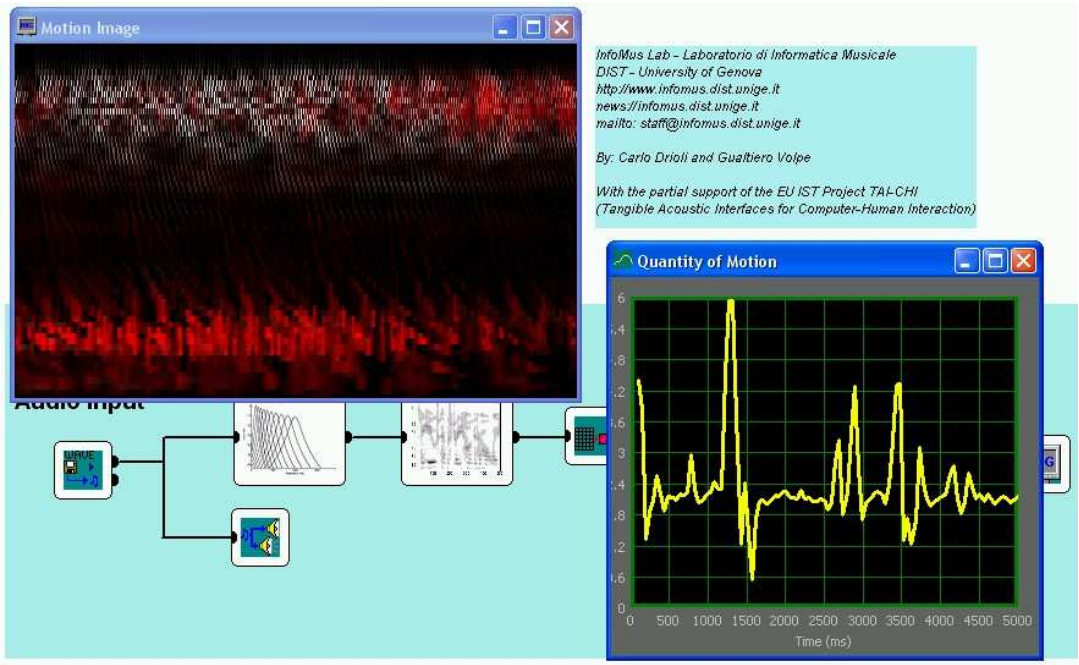

Figure 5.33: SMI of a cochleogram (red shadow) and graph of the corresponding QoM

samples are computed by linear interpolation as

$$s(t_i + \frac{n}{F_s}) = m_{i-1} + n\frac{(m_i - m_{i-1})}{N_s}, n = 1 \ldots N_s$$

where N_s is a selected audio frame length at a given audio sampling rate F_s. However, often sound analysis algorithms are designed to operate in frequency ranges that are much higher if compared to those related to the velocity of body movements. For this reason, the conversion block also provides amplitude modulation (AM) and frequency modulation (FM) functions to shift the original signal band along the frequency axis. If

$$c(t) = A_c \cos(2\pi f_c t)$$

is a sinusoidal carrier wave with carrier amplitude A_c and carrier frequency f_c, an AM audio-rate signal can be computed as

$$s_m(t) = A_c s(t) \cos(2\pi f_c t),$$

and an FM signal as

$$s_m(t) = A_c \cos(2\pi f_c t + 2\pi \int_0^t s(t)dt).$$

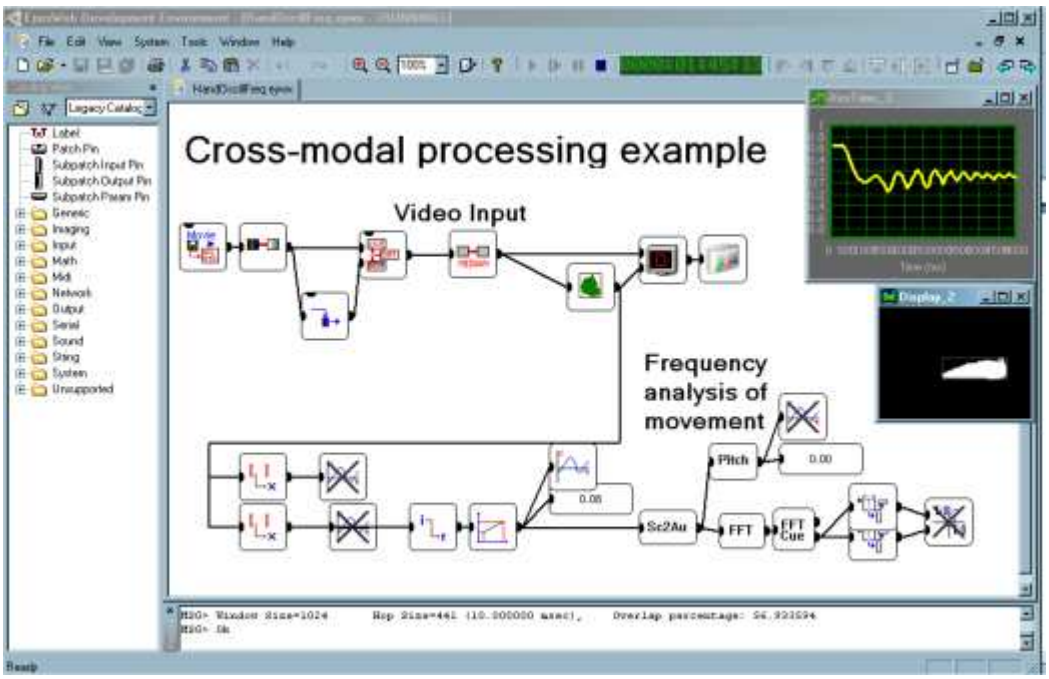

Figure 5.34: An example of EyesWeb application for cross-modal analysis of movement: the hand vertical displacement, measured from the video signal, is converted into the audio domain and analyzed through a pitch detector.

The approach to motion analysis by algorithms inspired to acoustic and/or musical cues extraction can be explored further. A possible application is, for example, the control of a digital score reproduction (e.g., a MIDI file) through the detection of tempo, onset, IOI, and other similar musical parameters from the arm and hand movements.

5.6.3 Multimodal processing for analysis of touch gestures

As an example of multimodal analysis of gestural information let us consider an experimental application for the analysis of touch gesture based on Tangible Acoustic Interfaces (TAIs).

Designing and developing TAIs consists of exploring how physical objects, augmented surfaces, and spaces can be transformed into tangible-acoustic embodiments of natural seamless unrestricted interfaces. TAIs can employ physical objects and space as media to bridge the gap between the virtual and physical worlds and to make information accessible through large size touchable objects as well as through ambient media. Research on TAI is carried out for example in

the framework of the EU-IST project TAI-CHI (Tangible Acoustic Interfaces for Computer-Human Interaction).

The aim of the sample application here described is twofold: (i) locate where on a TAI the touch gesture takes place, and (ii) analyze how touching is performed (i.e., individuating the expressive qualities of the touching action, such as for example whether the touching action is light and delicate or heavy and impulsive).

The approach to analysis is multimodal since both the information extracted from the acoustic signal generated by the touching action on the TAI and the information extracted from a video-camera toward the touching position are used.

Localization is based on two algorithms for in-solid localization of touching positions developed by the partners in the TAI-CHI project. The first algorithm, developed by the Image and Sound Processing Group at Politecnico di Milano employs 4 sensors and is based on the computation of the Time Delay of Arrival (TDOA) of the acoustical waves to the sensors Polotti et al. [2005]. The second algorithm developed by the Laboratoire Ondes et Acoustique at the Institut pour le Developement de la Science, l'Education et la Technologie, Paris, France, employs just 1 sensor and is based on pattern matching of the sound patterns generated by the touching action against a collection of stored patterns. In order to increase the reliability of the detected touching position we developed an EyesWeb application integrating the two methods and compensating the possible weakness of one method with the outcomes of the other one.

The position and time of contact information obtained from audio analysis can be employed to trigger and control in a more precise way the video-based gesture analysis process: e.g., we are testing hi-speed and hi-res videocameras in EyesWeb 4 in which it is also possible to select the portion of the active ccd area using (x,y) information from a TAI interface.

Video-based analysis (possibly combined with information extracted from the sound generated by the touching action, e.g., the sound level) is then used for extraction of expressive qualities. Gesture analysis is based on hand detection and tracking and builds upon the extraction of information concerning both static and dynamic aspects. As for the static aspects we developed a collection of EyesWeb modules for real-time classification of hand postures. Classification employs machine learning techniques (namely, Support Vector Machines). As for the dynamic aspects we used the expressive features currently available in the EyesWeb Expressive Gesture Processing Library (e.g., Quantity of Motion, Contraction/Expansion, Directness Index etc.). Figure 5.35 shows for example the output of an EyesWeb module for the extraction of the

hand skeleton.

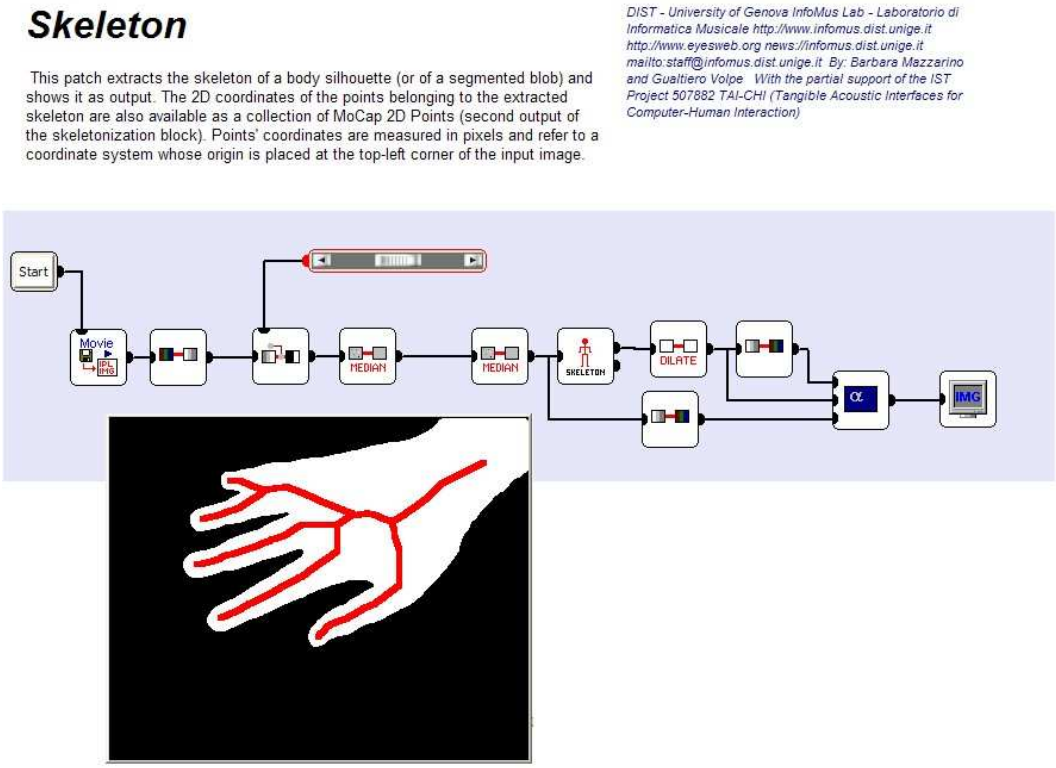

Figure 5.35: An EyesWeb 4 patch extracting the skeleton of a hand touching a TAI

In other words, while the contact position is detected through an acoustic based localization system, visual information is employed to get information on how the hand approaches and touches the interface (e.g., with a fluent movement, or in a hesitating way, or in a direct and quick way etc.).

5.6.4 Future perspectives for cross-modal analysis

This chapter presented some concrete examples of cross-modal and multimodal analysis techniques. The preliminary results from these sample applications indicate the potentialities of a multimodal and cross-modal approach to expressive gesture processing: 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. Multimodality 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 (especially in expressive gesture processing). 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 work described in this paper concerns 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.

5.7 Acknowledgements

This work has been partially supported by the EU-IST Coordinated Action S2S^2 ("Sound to Sense, Sense to Sound",IST-2004-03773, www.s2s2.org).

The editors and the authors thank all the people involved in the project with particular reference to the staffs of the Labs of the S2S^2 partners.

Bibliography

Maribeth Back, Jonathan Cohen, Rich Gold, Steve Harrison, and Scott Minneman. Listen reader: An electronically augmented paper-based book. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 23–29. ACM Press, New York, NY, USA, 2001.

Maribeth Back, Rich Gold, and Dana Kirsh. The sit book: Audio as affective imagery for interactive storybooks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 202–203. ACM Press, New York, NY, USA, 1999.

Timothy Beamish, Karon Maclean, and Sidney Fels. Manipulating music: Multimodal interaction for djs. In *Conference for Human-Computer Interaction 2004*, April 2004.

R. Bencina. Oasis rose, the composition: Real-time dsp with audiomulch. In *Proceedings of the Australasian Computer Music Conference*, 1998.

R. Bencina, M. Kaltenbrunner, and S. Jordà. Improved topological fiducial tracking in the reactivision system. In *PROCAMS 2005–IEEE International Workshop on Projector-Camera Systems*, submitted.

T. Blaine and C. Forlines. Jam-o-world: Evolution of the jam-o-drum multi-player musical controller into the jam-o-whirl gaming interface. In *Proceedings of the 2002 Conference on New Interfaces for Musical Expression (NIME-02*, pages 17–22, Dublin, 2002.

T. Blaine and T. Perkis. Jam-o-drum, a study interaction design. In *Proceedings of the ACM DIS 2000 Conference*, NY, 2000. ACM Press.

A.F Bobick and J. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001.

R.T. Boone and J.G. Cunningham. Children's decoding of emotion in expressive body movement: The development of cue attunement. 34:1007–1016, 1998.

J. Borchers, E. Lee, and W. Samminger. Personal orchestra: a real-time audio/video system for interactive conducting. *Multimedia Systems*, 9(5):458–465, 2004.

Richard Boulanger and Max Mathews. The 1997 mathews radio-baton and improvisation modes. In *Proceedings International Computer Music Conference*, Thessaloniki, Greece, 1997.

R. Bresin. What color is that music performance? In *International Computer Music Conference - ICMC 2005*, Barcelona, 2005.

R. Bresin and G. U. Battel. Articulation strategies in expressive piano performance. analysis of legato, staccato, and repeated notes in performances of the andante movement of Mozart's sonata in G major (K 545). *Journal of New Music Research*, 29(3):211–224, 2000.

R. Bresin and S. Dahl. Experiments on gestures: walking, running, and hitting. In D. Rocchesso and F. Fontana, editors, *The Sounding Object*, pages 111–136. Mondo Estremo, Florence, Italy, 2003.

R. Bresin, G. De Poli, and R. Ghetta. Fuzzy performance rules. In J. Sundberg, editor, *KTH Symposium on "Grammars for music performance"*, pages 15–36, Stockholm, 1995.

R. Bresin and A. Friberg. Emotional coloring of computer-controlled music performances. *Computer Music Journal*, 24(4):44–63, 2000a.

R. Bresin, A. Friberg, and J. Sundberg. Director musices: The KTH performance rules system. In *SIGMUS-46*, pages 43–48, Kyoto, 2002.

R. Bresin and P. N. Juslin. Rating expressive music performance with colours. *Manuscript submitted for publication*, 2005.

R. Bresin and G. Widmer. Production of staccato articulation in Mozart sonatas played on a grand piano. preliminary results. *TMH-QPSR, Speech Music and Hearing Quarterly Progress and Status Report*, 2000(4):1–6, 2000.

Roberto Bresin and Anders Friberg. Emotional coloring of computer-controlled music performances. *Computer Music Journal*, 24(4):44–6, 2000b.

Roberto Bresin and Bruno Giordano. Do we play how we walk? an experiment on expressive walking. In *Proc. of Sound and Music Computing 2005*, Salerno, Italy, submitted.

Steven Brewster. Non-speech auditory output. In Jacko J. and Sears A., editors, *The Human-Computer Interaction Handbook*. Lawrence Erlbaum, 2002.

Antonio Camurri, Paolo Coletta, Carlo Drioli, Alberto Massari, and Gualtiero Volpe. Audio processing in a multimodal framework. Barcelona, Spain, May 2005a.

Antonio Camurri, Paolo Coletta, Alberto Massari, Barbara Mazzarino, Massimiliano Peri, Matteo Ricchetti, Andrea Ricci, and Gualtiero Volpe. Toward real-time multimodal processing: Eyesweb 4.0. In *Proceedings AISB 2004 Convention: Motion, Emotion and Cognition*, pages 22–26, Leeds, UK, March 2004a.

Antonio Camurri, Shuji Hashimoto, Matteo Ricchetti, Riccardo Trocca, Kenji Suzuki, and Gualtiero Volpe. Eyesweb: Toward gesture and affect recognition in interactive dance and music systems. *Computer Music Journal*, 24(1):941–952, Spring 2000.

Antonio Camurri, Carol L. Krumhansl, Barbara Mazzarino, and Gualtiero Volpe. An exploratory study of anticipating human movement in dance. Genova, Italy, June 2004b.

Antonio Camurri, Ingrid Lagerlöf, and Gualtiero Volpe. Recognizing emotion from dance movement: Comparison of spectator recognition and automated techniques. *International Journal of Human-Computer Studies*, 59(1):213–225, July 2003.

Antonio Camurri, Barabara Mazzarino, Matteo Ricchetti, Renee Timmers, and Gualtiero Volpe. Multimodal analysis of expressive gesture in music and dance performances. In Antonio Camurri and Gualtiero Volpe, editors, *Gesture-based Communication in Human-Computer Interaction, LNAI 2915*, pages 20–39. Springer Verlag, February 2004c.

Antonio Camurri, Barabara Mazzarino, and Gualtiero Volpe. Analysis of expressive gesture: The eyesweb expressive gesture processing library. In Antonio Camurri and Gualtiero Volpe, editors, *Gesture-based Communication in Human-Computer Interaction, LNAI 2915*, pages 460–467. Springer Verlag, February 2004d.

Antonio Camurri, Barbara Mazzarino, and Gualtiero Volpe. Expressive interfaces. *Cognition, Technology, and Work*, 6(1):15–22, February 2004e.

Antonio Camurri, Giovanni De Poli, Marc Leman, and Gualtiero Volpe. Toward communicating expressiveness and affect in multimodal interactive systems for performing art and cultural applications. *IEEE Multimedia Magazine*, 12(1):43–53, January 2005b.

S. Canazza, A. Friberg, A. Rodà, and P. Zanon. Expressive Director: a system for the real-time control of music performance synthesis. In R. Bresin, editor, *Stockholm Music Acoustics Conference – SMAC 2003*, volume 2, pages 521–524, Stockholm, 2003a.

Sergio Canazza, Giovanni De Poli, Antonio Rodà, and Alvise Vidolin. An abstract control space for communication of sensory expressive intentions in music performance. *Journal of New Music Research*, 32(3):281–294, 2003b.

J. Chadabe. The voltage-controlled synthesizer. In John Appleton, editor, *The development and practice of electronic music*. Prentice-Hall, New Jersey, 1975.

Yi-Chun Chu, David Bainbridge, Matt Jones, and Ian H. Witten. Realistic books: a bizarre homage to an obsolete medium? In *Proceedings of the 2004 joint ACM/IEEE Conference on Digital Libraries*, pages 78–86. ACM Press, New York, NY, USA, giugno 2004.

Yi-Chun Chu, Ian H. Witten, Richard Lobb, and David Bainbridge. How to turn a page. In *Proceedings of the third joint ACM/IEEE-CS Conference on Digital Libraries*, pages 186–188. IEEE Computer Society, Washington, DC, USA, 2003.

E. Costanza and J.A. Robinson. A region adjacency tree approach to the detection and design of fiducials. In *Vision, Video and Graphics (VVG) 2003*, 2003.

R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor. Emotion recognition in human-computer interaction. 18(1):32–80, January 2001.

Sofia Dahl. Playing the accent - comparing striking velocity and timing in an ostinato rhythm performed by four drummers. *Acta Acoustica United with Acoustica*, 90(4):762–776, 2004.

Sofia Dahl and Anders Friberg. Expressiveness of musician's body movements in performances on marimba. In Antonio Camurri and Gualtiero Volpe, editors, *Gesture-based Communication in Human-Computer Interaction, LNAI 2915*. Springer Verlag, February 2004.

F. Déchelle, R. Borghesi, M. De Cecco, E. Maggi, B. Rovan, and N. M. Schnell. jMax: An environment for real time musical applications. *Computer Music Journal*, 23(3):50–58, 1999.

W.H. Dittrich, T. Troscianko, S.E.G. Lea, and D. Morgan. Perception of emotion from dynamic point-light displays represented in dance. 25:727–738, 1996.

Paul Dourish. *Where the Action Is: the foundations of embodied interaction*. MIT Press, Cambridge, MA, 2001.

G. Fitzmaurice, H. Ishii, and W. Buxton. Bricks: Laying the foundations of graspable user interfaces. In *Proceedings of CHI'95 Conference on Human Factors in Computing systems*, pages 442–449, 1995.

A. Friberg. Generative rules for music performance: A formal description of a rule system. *Computer Music Journal*, 15(2):56–71, 1991.

A. Friberg. A fuzzy analyzer of emotional expression in music performance and body motion. In J. Sundberg and B. Brunson, editors, *Proceedings of Music and Music Science, October 28-30, 2004*, Stockholm: Royal College of Music, 2005.

A. Friberg. pDM: an expressive sequencer with real-time control of the KTH music performance rules. *Computer Music Journal*, in press.

A. Friberg and G. U. Battel. Structural communication. In R. Parncutt and G. E. McPherson, editors, *The Science and Psychology of Music Performance: Creative Strategies for Teaching and Learning*, pages 199–218. Oxford University Press, New York and Oxford, 2002.

A. Friberg, R. Bresin, L. Frydén, and J. Sundberg. Musical punctuation on the microlevel: Automatic identification and performance of small melodic units. *Journal of New Music Research*, 27(3):271–292, 1998.

A. Friberg, V. Colombo, L. Frydén, and J. Sundberg. Generating musical performances with Director Musices. *Computer Music Journal*, 24(3):23–29, 2000.

A. Friberg, E. Schoonderwaldt, and P. N. Juslin. Cuex: An algorithm for extracting expressive tone variables from audio recordings. *Acoustica united with Acta Acoustica*, in press.

A. Friberg, E. Schoonderwaldt, P. N. Juslin, and R. Bresin. Automatic real-time extraction of musical expression. In *International Computer Music Conference - ICMC 2002*, pages 365–367, Göteborg, 2002.

A. Friberg and J. Sundberg. Does music performance allude to locomotion? A model of final ritardandi derived from measurements of stopping runners. *Journal of the Acoustical Society of America*, 105(3):1469–1484, 1999.

A. Gabrielsson and E. Lindström. The influence of musical structure on emotion. 2001.

J.J. Gibson. *The Ecological Approach to Visual Perception*. Lawrence Erlbaum Ass., Cambridge, MA, 1979.

Kjetil Falkenberg Hansen. The basics of scrathing. *Journal of New Music Research*, 31(4):357–365, 2002.

Kjetil Falkenberg Hansen and Roberto Bresin. Analysis of a genuine scratch performance. In *Gesture Workshop*, pages 519–528, 2003.

K. Hevner. Experimental studies of the elements of expression in music. 48:246–268, 1936.

A. Hunt and R Kirk. Multiparametric control of real-time systems. In M. Battier, J. Rovan, and M. Wanderley, editors, *Trends in Gestural Control of Music*. FSU - Florida, April 2000.

Andy Hunt and Thomas Hermann. Special issue on Interactive Sonification. *IEEE Multimedia*, 12(2), 2005.

Edwin L. Hutchins, James D. Hollan, and Donald A. Norman. Direct manipulation interfaces. In D. Norman and S. W Draper, editors, *User-Centred System Design*, pages 87–124. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1986.

T. Ilmonen. The virtual orchestra performance. In *Proceedings of the CHI 2000 Conference on Human Factors in Computing Systems, Haag, Netherlands*, pages 203–204. Springer Verlag, 2000.

H Ishii and B. Ullmer. Tangible bits: Towards seamless interfaces between people, bits and atoms. In *Proceedings of CHI 97 Conference on Human Factors in Computing systems*, pages 22–27, Atlanta, Georgia USA, 1997.

D. Jaffe and J. O. Smith. Extensions of the Karplus-Strong plucked-string algorithm. *Computer Music Journal*, 7(2):76–87, 1983.

Z. Jánosy, M. Karjalainen, and V. Välimäki. Intelligent synthesis control with applications to a physical model of the acoustic guitar. In *Proc. Int. Computer Music Conference (ICMC'94)*, Aarhus, Denmark, 1994.

G. Johansson. Visual perception of biological motion and a model for its analysis. 14:201–211, 1973.

S. Jordà. Faust Music On Line (FMOL): An approach to real-time collective composition on the internet. *Leonardo Music Journal*, 9(1):5–12, 1999.

S. Jordà. FMOL: Toward user-friendly, sophisticated new musical instruments. *Computer Music Journal*, 26(3):23–39, 2002.

S. Jordà and O. Wüst. FMOL: A system for collaborative music composition over the web. In *Proceedings of the 12th International Workshop on Database and Expert Systems Applications*, pages 537–542, 2001.

S. Jordà and O. Wüst. Sonigraphical instruments: From FMOL to the reacTable*. In *Proceedings of the 2003 International Conference on New Interfaces for Musical Expression (NIME-03)*, pages 70–76, Montreal, 2003.

Sergi Jordà, Martin Kaltenbrunner, Günter Geiger, and Ross Bencina. the reacTable*. In *Proceedings International Computer Music Conference*, San Francisco, 2005. International Computer Music Association.

P. N. Juslin. Communicating emotion in music performance: A review and a theoretical framework. In P. N. Juslin and J. A. Sloboda, editors, *Music and emotion: Theory and research*, pages 305–333. Oxford University Press, New York, 2001.

P. N. Juslin, A. Friberg, and R. Bresin. Toward a computational model of expression in performance: The GERM model. *Musicae Scientiae*, Special issue 2001-2002:63–122, 2002.

P.N. Juslin and J. Laukka. Communication of emotions in vocal expression and music performance: Different channels, same code? 129(5):770–814, 2003.

M. Kaltenbrunner, T. Bovermann, R. Bencina, and E. Costanza. TUIO: A protocol for table-top tangible user interfaces. In *6th International Gesture Workshop*, Vannes 2005, submitted.

M. Karjalainen and Unto K. Laine. A model for real-time sound synthesis of guitar on a floating-point signal processor. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'91)*, Toronto, Canada, 1991.

M. Karjalainen, T. Mäki-Patola, Aki Kanerva, Antti Huovilainen, and Pekka Jänis. Virtual air guitar. In *Proc. Audio Eng. Soc. 117th Convention*, San Francisco, 2004.

M. Karjalainen, Vesa Välimäki, and Tero Tolonen. Plucked-string models: From the Karplus-Strong algorithm to digital waveguides and beyond. *Computer Music Journal*, 22(3):17–32, 1998.

Alan Kay and Adele Goldberg. Personal dynamic media. *Computer*, 10(3):31–44, marzo 1977.

Carol L. Krumhansl. Am exploratory study of musical emotions and psychophysiology. *Canadian Journal pf Experimental Psychology*, 51(4):336–352, 1997.

Rudolf Laban. *Modern Educational Dance*. 1963.

E. Lee, T.M. Nakra, and J. Borchers. You're the conductor: A realistic interactive conducting system for children. In *Proc. of NIME 2004*, pages 68–73, 2004.

M. Leman, V. Vermeulen, L. De Voogdt, and D. Moelants. Prediction of musical affect attribution using a combination of structural cues extracted from musical audio. *Journal of New Music Research*, 34(1):39–67, January 2005.

M. Lesaffre, M. Leman, K. Tanghe, B. De Baets, H. De Meyer, and J.P. Martens. User-dependent taxonomy of musical features as a conceptual framework for musical audio-mining technology. pages 635–638, August 2003.

Xiaofeng Li, Robert J. Logan, and Richard E. Pastore. Perception of acoustic source characteristics: Walking sounds. *The Journal of the Acoustical Society of America*, 90(6):3036–3049, 1991.

Björn Lindblom. Explaining phonetic variation: a sketch of the H&H theory. In Hardcastle and Marchal, editors, *Speech production and speech modeling*, pages 403–439. Kluwer, Dordrecht, 1990.

Erik Lindström, Antonio Camurri, Anders Friberg, Gualtiero Volpe, and Marie-Louise Rinmann. Affect, attitude and evaluation of multi-sensory performances. *Journal of New Music Research*, 34(1), January 2005.

T. Mäki-Patola, A. Kanerva, J. Laitinen, and T. Takala. Experiments with virtual reality instruments. In *Proc. Int. Conference on New Interfaces for Musical Expression (NIME05)*, Vancouver, Canada, 2005.

T. Marrin Nakra. *Inside the conductor's jacket: analysis, interpretation and musical synthesis of expressive gesture*. PhD thesis, MIT, 2000.

Steven Martin. *Theremin, an electronic odissey*. MGM, 2001. Documentary on DVD.

M. V. Mathews. The digital computer as a musical instrument. *Science*, 142(11):553–557, 1963.

M. V. Mathews. The conductor program and the mechanical baton. In M. Mathews and J. Pierce, editors, *Current Directions in Computer Music Research*, pages 263–282. The MIT Press, Cambridge, Mass, 1989.

M. V. Mathews and R. Moore. GROOVE, a program for realtime control of a sound synthesizer by a computer. In *Proceedings of the 4th Annual Conference of the American Society of University Composers*, pages 22–31, NY: ASUC, Columbia University, 1969.

M. De Meijer. The contribution of general features of body movement to the attribution of emotions. 13:247–268, 1989.

R. Moog. Voltage-controlled electronic music modules. *Journal of the Audio Engineering Society*, 13(3):200–206, 1965.

Nicholas Negroponte. Books without pages. *ACM SIGDOC Asterisk Journal of Computer Documentation*, 20(3):2–2, agosto 1996.

V. A. Niskanen. *Soft Computing Methods in Human Sciences*. Springer Verlag, Berlin, 2004.

J. A. Paradiso and K.-Y. Hsiao. Musical trinkets: New pieces to play. In *SIGGRAPH 2000 Conference Abstracts and Applications*, NY, 2000. ACM Press.

J. Patten, B. Recht, and H. Ishii. Audiopad: A tag-based interface for musical performance. In *Proceedings of the 2003 International Conference on New Interfaces for Musical Expression (NIME-02)*, pages 11–16, Dublin, 2002.

I. Peretz. Listen to the brain: a biological perspective on musical emotions. In P. N. Juslin and J. A. Sloboda, editors, *Music and emotion: Theory and research*, pages 105–134. Oxford University Press, New York, 2001.

F.E. Pollick, A. Bruderlin, and A.J. Sanford. Perceiving affect from arm movement. 82:B51–B61, 2001.

P. Polotti, M. Sampietro, A. Sarti, S. Tubaro, and A. Crevoisier A. Acoustic localization of tactile interactions for the development of novel tangible interfaces. September 2005.

I. Poupyrev. Augmented groove: Collaborative jamming in augmented reality. In *ACM SIGGRAPH 2000 Conference Abstracts and Applications*, NY, 2000. ACM Press.

Miller Puckette. The patcher. In *Proceedings International Computer Music Conference*, pages 420–429, San Francisco, 1988. International Computer Music Association.

Miller Puckette. Pure data. In *Proceedings International Computer Music Conference*, pages 269–272, Hong Kong, August 1996.

M.-L. Rinman, A. Friberg, B. Bendiksen, D. Cirotteau, S. Dahl, I. Kjellmo, B. Mazzarino, and A. Camurri. Ghost in the cave - an interactive collaborative game using non-verbal communication. In A. Camurri and G. Volpe, editors, *Gesture-based Communication in Human-Computer Interaction, LNAI 2915*, volume LNAI 2915, pages 549–556, Berlin Heidelberg, 2004. Springer-Verlag.

Davide Rocchesso and Roberto Bresin. Emerging sounds for disappearing computers. in press, 2005.

Davide Rocchesso, Roberto Bresin, and Michael Fernström. Sounding objects. *IEEE Multimedia*, 10(2):42–52, April 2003.

J.A. Russell. A circumplex model of affect. 39:1161–1178, 1980.

J. A. Sarlo. Gripd: A graphical interface editing tool and run-time environment for pure data. In *Proceedings International Computer Music Conference*, pages 305–307, Singapore, August 2003.

B. Schneiderman. *Leonardo's laptop: human needs and the new computing technologies*. MIT press, Cambridge, 2002.

M. Seif El-Nasr, J. Yen, and T. R. Iorger. Flame - fuzzy logic adaptive mode of emotions. *Autonomous Agents and Multi-Agent Systems*, 3:219–257, 2000.

N.K. Sheridon and M.A. Berkovitz. The gyricon–a twisting ball display. In *Prooceedings Of The Society For Information Display*, pages 289–293. Boston, MA, maggio 1977.

E. Singer. Sonic banana: A novel bend-sensor-based MIDI controller. In *Proceedings of the 2003 International Conference on New Interfaces for Musical Expression (NIME-03)*, pages 220–221, Montreal, 2003.

M. Slaney. Auditory toolbox documentation. technical report 45. Technical report, Apple Computers Inc., 1994.

J.A. Sloboda and P.N. Juslin, editors. *Music and Emotion: Theory and Research*, 2001.

C. S. Sullivan. Extending the Karplus-Strong algorithm to synthesize electric guitar timbres with distortion and feedback. *Computer Music Journal*, 14(3):26–37, 1990.

J. Sundberg. How can music be expressive? *Speech Communication*, 13:239–253, 1993.

J. Sundberg, A. Askenfelt, and L. Frydén. Musical performance: A synthesis-by-rule approach. *Computer Music Journal*, 7:37–43, 1983.

B. Ullmer and H. Ishii. Emerging frameworks for tangible user interfaces. In John M. Carnoll, editor, *Human Computer Interaction in the New Millenium*, pages 579–601. Addison-Wesley, Reading, MA, 2001.

R.D. Walk and C.P. Homan. Emotion and dance in dynamic light displays. 22:437–440, 1984.

H.G. Wallbott. The measurement of human expressions. pages 203–228. 2001.

G. Weinberg and S. Gan. The squeezables: Toward an expressive and interdependent multiplayer musical instrument. *Computer Music Journal*, 25(2):37–45, 2001.

D. Wessel and M. Wright. Problems and prospects for intimate musical control of computers. *Computer Music Journal*, 26(3):11–22, 2002.

M. Wright. Implementation and performance issues with OpenSound control. In *Proceedings International Computer Music Conference*, pages 224–227, University of Michigan - Ann Arbor, 1997. International Computer Music Association.

M. Wright. OpenSound control: State of the art 2003. In *Proceedings of the 2003 International Conference on New Interfaces for Musical Expression (NIME-03)*, pages 153–159, Montreal, 2003.

H.-J. Zimmerman. *Fuzzy set theory - and its applications*, volume 3rd ed. Kluwer, Boston, 1996.

Chapter 6

Physics-based Sound Synthesis

Cumhur Erkut, Vesa Välimäki, Matti Karjalainen, and Henri Penttinen

Helsinki University of Technology, Lab. Acoustics and Audio Signal Processing, Espoo, Finland

Abstract: This chapter provides the current status and open problems in the field of physics-based sound synthesis. Important concepts and methods of the field are discussed, the state of the art in each technique is presented. The focus is then shifted towards the current directions of the field. The future paths are derived and problems that deserve detailed collaborative research are indicated.

6.1 Introduction

Physics-based sound synthesis focuses on developing efficient digital audio processing algorithms built upon the essential physical behavior of various sound production mechanisms. The model-based representation of audio can be used in many digital audio applications, including digital sound synthesis, structural analysis of sounds, automatic transcription of musical signals, and parametric audio coding.

Physics-based sound synthesis is currently one of the most active research areas in audio signal processing Välimäki et al. [2004b], Välimäki [2004], Smith [2004c]. Many refinements to

existing algorithms, as well as several novel techniques are emerging. The aim of this chapter is to provide the current status in physics-based sound synthesis by summarizing various approaches and methodologies within the field, capture the current directions, and indicate open problems that deserve further research. A comprehensive review of physics-based sound synthesis methods is underway Välimäki et al. [2005], and other excellent reviews and tutorials are readily available Smith [2004c, 1996], Borin et al. [1992], De Poli and Rocchesso [1998], Välimäki and Takala [1996], Välimäki [2004]. Our aim is not duplicate these efforts; we rather focus on selective aspects related to each method. Sec. 6.2 presents background information about these aspects. An important point is that we structurally classify the physics-based sound synthesis methods into two main groups according to their variables used in computation.

In Sec. 6.3, without going into technical details (the reader is referred to Välimäki et al. [2005] for a detailed discussion of each method), we briefly outline the basics, indicate recent research, and enlist available implementations. We then consider some current directions in physics-based sound synthesis in Sec. 6.3.3, including the discussion on recent systematic efforts to combine the two structural groups of physics-based sound synthesis.

A unified modular modeling framework, in our opinion, is one of the most important open problems in the field of physics-based sound synthesis. There are, however, other problems, which provide the content of Sec. 6.4.

6.2 General Concepts

A number of physical, and signal processing concepts are of paramount importance in physics-based sound synthesis. The background provided in this section is crucial for understanding problem definition in Sec. 6.2.3, as well as the state of the art and the open problems discussed in the subsequent sections.

6.2.1 Different flavors of modeling Tasks

Physical mechanisms are generally complex, and those related to the sound production mechanisms are no exceptions. A useful approach for dealing with complexity is to use a *model*, which typically is based on an abstraction that suppress the non-essential details of the original

problem and allows selective examination with the essential aspects[1]. Yet, an abstraction is task-dependent and it is used for a particular purpose, which in turn determines what is important and what can be left out.

One level of abstraction allows us to derive mathematical models (i.e., differential equations) of physical phenomena. Differential equations summarize larger-scale temporal or spatio-temporal relationships of the original phenomena on an infinitesimally small basis. *Musical acoustics*, a branch of physics, relies on simplified mathematical models for a better understanding of the sound production in musical instruments Benade [1990], Fletcher and Rossing [1998]. Similar models are used to study the biological sound sources Fletcher [1992].

Computational models is for long a standard tool in various disciplines. At this level, the differential equations of the mathematical models are discretized and solved by computers, one small step at a time. Computational models inherit the abstractions of mathematical models, and add one more level of abstraction by imposing an *algorithm* for solving them Press et al. [2002]. Among many possible choices, *digital signal processing* (DSP) provides an advanced theory and tools that emphasize computational issues, particularly maximal efficiency.

Computational models are the core of physics-based sound synthesis (hence the aliases *physical modeling* [Smith, 1992, 1996] or *model-based sound synthesis* [Karjalainen et al., 2001]). In addition, physics-based sound synthesis inherits constraints from the task of sound synthesis Smith [1991], Tolonen et al. [1998], Cook [2002b], i.e., representing huge amount of audio data preferably by a small number of meaningful parameters. Among a wide variety of synthesis and processing techniques, physically-based methods have several advantages with respect to their parameters, control, efficiency, implementation, and sound quality Jaffe [1995].

6.2.2 Physical domains, systems, variables, and parameters

Physical phenomena occur in different *physical domains*: string instruments operate in *mechanical*, wind instruments in *acoustical*, and electro-acoustic instruments (such as the analog synthesizers) operate in *electrical* domains. The domains may interact, as in the electro-mechanical Fender Rhodes, or they can be used as *analogies* (equivalent models) of each other. Analogies make unfamiliar phenomena familiar to us. It is therefore not surprising to find many electrical circuits as analogies to describe phenomena of other physical domains in a musical acoustics textbook

[1] As in Einstein's famous dictum: everything should be made as simple as possible, but no simpler.

Fletcher and Rossing [1998].

A physical system is a collection of objects united by some form of interaction or interdependence. A mathematical model of a physical system is obtained through rules (typically differential equations) relating measurable quantities that come in pairs of *variables*, such as force and velocity in the mechanical domain, pressure and volume velocity in the acoustical domain, or voltage and current in the electrical domain. If there is a linear relationship between the dual variables, this relation can be expressed as a *parameter*, such as impedance $Z = U/I$ being the ratio of voltage U and current I, or by its inverse, admittance $Y = I/U$. An example from the mechanical domain is mobility (mechanical admittance) as the ratio of velocity and force. When using such parameters, only one of the dual variables is needed explicitly, because the other one is achieved through the constraint rule.

The physics-based sound synthesis methods use two types of variables for computation, *K-variables* and *wave variables*. K-variables refer to the Kirchhoff continuity rules of dual quantities mentioned above, in contrast to wave components of physical variables. Instead of pairs of K-variables, the wave variables come in pairs of *incident* and *reflected* wave components. The decomposition into wave components is clear in such wave propagation phenomena, where opposite-traveling waves add up to the actual observable K-quantities. A wave quantity is directly observable only when there is no other counterpart. It is, however, a highly useful abstraction to apply wave components to any physical cases, since this helps in solving computability (causality) problems in discrete-time modeling.

6.2.3 Dichotomies, problem definition, and schemes

Similar to the wave and K-variables, important concepts in physics-based sound synthesis form dichotomies that come into the play in the structure, design, implementation, and execution of the physics-based sound synthesis techniques. These dichotomies are enlisted in Table. 6.1. In general, the properties in the first column are easier to handle compared to those in the second. Thus, the properties in the second column readily point out open research problems. We will elaborate these problems in Sec.6.4.

For the purposes of this chapter, the main problem of physics-based sound synthesis is to derive efficient, causal, and explicit computational models for high-quality, natural-sounding synthetic audio, which are optimally balancing accuracy, efficiency, and ease of control. These models should operate in the widest range of physical domains and handle the nonlinearities and

causal	non-causal
explicit	implicit
lumped	distributed
linear	nonlinear
time-invariant	time-varying
DSP-based	not DSP-based
terminals	ports
passive	active
stability guarantee	no stability guarantee
monolithic	modular

Table 6.1: Dichotomies in physics-based sound synthesis

parameter updates in a robust and predictable manner. In this respect, a DSP-based formulation and stability guarantee are desirable features. Port-based formulations and modular schemes have certain advantages when attempting to design a general, unified framework for physics-based sound synthesis.

Based on the dichotomies and the problem definition, two general schemes for physics-based sound synthesis emerge. One way of decomposing a physics-based sound synthesis system is highlighting the functional elements *exciter* and *resonator* Borin et al. [1992] (abbreviated in Fig. 6.1 as E-object and R-object, respectively). In this generic scheme, the exciter and the resonator are connected through ports. The exciter is usually nonlinear, whereas resonator is usually linear, and can be decomposed in sub-models. The interaction between the objects is usually handled implicitly within the system.

Alternatively, a modular system with explicit local interactions is schematically illustrated in Fig. 6.2. This scheme was first proposed in Borin et al. [1992], but only recently it is being used for implementing physics-based sound synthesis systems. In Fig. 6.2, an *S-object* represents a synthesis module that can correspond to both the exciter and the resonator of Fig. 6.1. An *I-object* is an explicit interconnection object (connector). Each synthesis module has *internal* and *external* parameters, with a reference of their accessibility from the connector. Internal parameters of a synthesis module (such as port admittances) are used by a connector for distributing the outgoing signals; they are only meaningful if the objects are linked. The external parameters

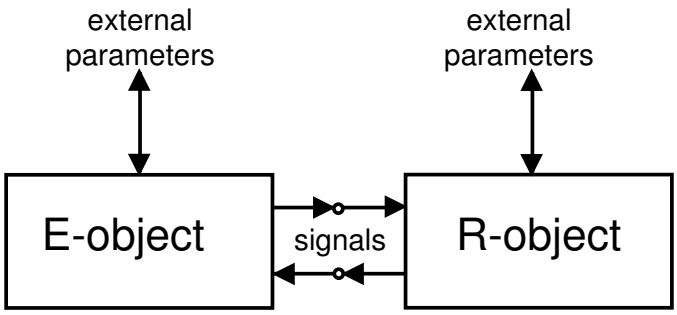

Figure 6.1: Excitation plus resonator paradigm of physics-based sound synthesis.

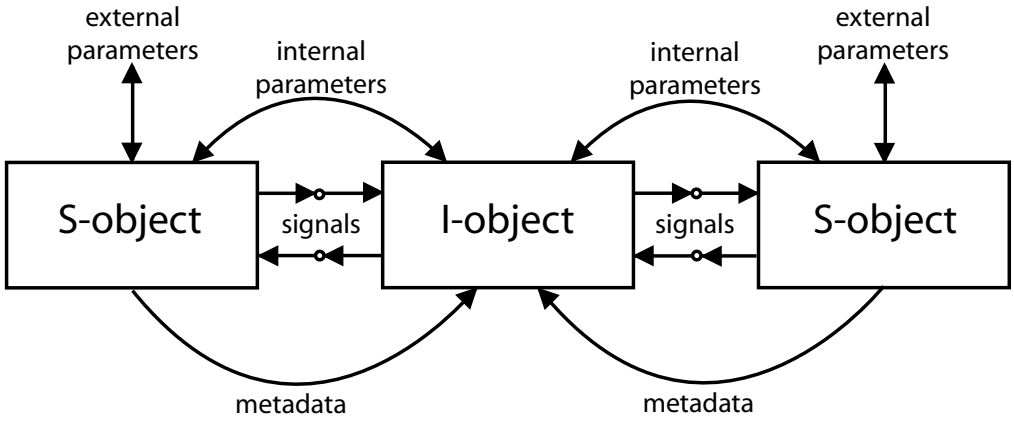

Figure 6.2: Modular interaction diagram.

are specific attributes of a synthesis module. Finally, metadata contains descriptors such as the domain or the type of the synthesis module. Note that locality implies that only neighboring synthesis modules are connected to a connector.

A reader who is already familiar with the concepts mentioned so far may want to proceed to Sec. 6.3, and read how the available methods relate to the general schemes presented here, and what is the current status in each of them. For others, these concepts are explained in the rest of this section.

6.2.4 Important concepts explained

Physical structure and interaction

Physical phenomena are observed as structures and processes in space and time. As a universal property in physics, the interaction of entities in space always propagates with a finite velocity. *Causality* is a fundamental physical property that follows from the finite velocity of interaction from a cause to the corresponding effect. The requirement of causality introduces special computability problems in discrete-time simulation, because two-way interaction with no delay leads to the *delay-free loop problem*. An evident solution is to insert a unit delay into the delay-free loop. However, this arbitrary delay has serious side effects (see Borin et al. [2000], Avanzini [2001]). The use of wave variables is advantageous, since the incident and reflected waves have a causal relationship.

Taking the finite propagation speed into account requires using a spatially *distributed* model. Depending on the case at hand, this can be a full 3-D model such as that used for room acoustics, a 2-D model such as for a drum membrane, or a 1-D model such as for a vibrating string. If the object to be modeled behaves homogeneously as a whole, for example due to its small size compared to the wavelength of wave propagation, it can be considered a *lumped* system that does not need spatial dimensions.

Signals, signal processing, and discrete-time modeling

The word *signal* typically means the value of a measurable or observable quantity as a function of time and possibly as a function of place. In signal processing, signal relationships typically represent one-directional cause-effect chains. Modification of signals can be achieved technically by active electronic components in analog signal processing or by numeric computation in DSP. This simplifies the design of circuits and algorithms compared to two-way interaction that is common in (passive) physical systems, for example in systems where the reciprocity principle is valid. In true physics-based modeling, the two-way interactions must be taken into account. This means that, from the signal processing viewpoint, such models are full of feedback loops, which further implicates that the concepts of computability (causality) and stability become crucial, as will be discussed later.

We favor the discrete-time signal processing approach to physics-based modeling when-

ever possible. The motivation for this is that digital signal processing is an advanced theory and tool that emphasizes computational issues, particularly maximal efficiency. This efficiency is crucial for real-time simulation and sound synthesis. Signal flow diagrams are also a good graphical means to illustrate the algorithms underlying the simulations.

The sampling rate and the spatial sampling resolution need more focus in this context. According to the sampling theorem Shannon [1948], signals must be sampled so that at least two samples must be taken per period or wavelength for sinusoidal signal components or their combinations, in order to make the perfect reconstruction of a continuous-time signal possible. This limit frequency, one half of the sampling rate, is called the Nyquist frequency. If a signal component higher in frequency f_x is sampled by rate f_s, it will be *aliased*, i.e. mirrored by the Nyquist frequency back to the *base band* by $f_a = f_s - f_x$. In audio signals, this will be perceived as very disturbing distortion, and should be avoided. In linear systems, if the inputs are bandlimited properly, the aliasing is not a problem because no new frequency components are created, but in nonlinear systems aliasing is problematic. In modeling physical systems, it is also important to remember that *spatial aliasing* can be a problem if the spatial sampling grid is not dense enough.

Linearity and time invariance

Linearity of a system means that the superposition principle is valid, i.e., quantities and signals in a system behave additively 'without disturbing' each other. Mathematically, this is expressed so that if the responses $\{y_1(t), y_2(t)\}$ of the system to two arbitrary input signals $\{x_1(t), x_2(t)\}$, respectively, are $x_1(t) \rightarrow y_1(t)$ and $x_2(t) \rightarrow y_2(t)$, then the response to $Ax_1(t)+Bx_2(t) \rightarrow Ay_1(t)+By_2(t)$ is the same as the sum of the responses to $Ax_1(t)$ and $Bx_2(t)$, i.e., $Ay_1(t) + By_2(t)$, for any constants A and B.

A linear system cannot create any signal components with new frequencies. If a system is nonlinear, it typically creates harmonic (integer multiples) or intermodulation (sums and differences) frequency components. This is particularly problematic in discrete-time computation because of the aliasing of new signal frequencies beyond the Nyquist frequency.

If a system is both *linear and time invariant* (LTI), there are constant-valued parameters that effectively characterize its behavior. We may think that in a time-varying system its characteristics (parameter values) change according to some external influence, while in a nonlinear system the characteristics change according to the signal values in the system.

Linear systems or models have many desirable properties. In digital signal processing, LTI systems are not only easier to design but also are typically more efficient computationally. A linear system can be mapped to transform domains where the behavior can be analyzed by algebraic equations Oppenheim et al. [1996]. For continuous-time systems, the Laplace and Fourier transforms can be applied to map between the time and frequency domains, and the Sturm-Liouville transform Trautmann and Rabenstein [2003] applies similarly to the spatial dimension[2] . For discrete-time systems, the Z-transform and the discrete Fourier transform (DFT and its fast algorithm, FFT) are used.

For nonlinear systems, there is no such elegant theory as for the linear ones; rather, there are many forms of nonlinearity, which require different methods for example, depending on which effect is desired. In discrete-time modeling, nonlinearities bring problems that are difficult to solve. In addition to aliasing, the delay-free loop problem and stability problems can become worse than they are in linear systems. If the nonlinearities in a system to be modeled are spatially distributed, the modeling task is even more difficult than with a localized nonlinearity.

Energetic behavior and stability

The product of dual variables such as voltage and current gives power, which, when integrated in time, yields energy. Conservation of energy in a closed system is a fundamental law of physics that should also be obeyed in physics-based modeling.

A physical system can be considered *passive* in the energetic sense if it does not produce energy, i.e., if it preserves its energy or dissipates it into another energy form, such as thermal energy. In musical instruments, the resonators are typically passive, while excitation (plucking, bowing, blowing, etc.) is an *active* process that injects energy to the passive resonators.

The *stability* of a physical system is closely related to its energetic behavior. Stability can be defined so that the energy of the system remains finite for finite-energy excitations. In this sense, a passive system always remains stable. From the signal processing viewpoint, stability may also be meaningful if it is defined so that the variables, such as voltages, remain within a linear operating range for possible inputs in order to avoid signal clipping and distortion. For system transfer functions, stability is typically defined so that the system poles (roots of the denominator polynomial) in a Laplace transform remain in the left half plane, or that the poles in a Z-transform

[2] A technical detail: unlike Laplace transform, Sturm-Liouville transform utilizes a non-unique kernel that depends on the boundary conditions.

in a discrete-time system remain inside the unit circle Oppenheim et al. [1996]. This guarantees that there are no responses growing without bounds for finite excitations.

In signal processing systems with one-directional interaction between stable subblocks, an instability can appear only if there are feedback loops. In general, it is impossible to analyze such a system's stability without knowing its whole feedback structure. Contrary to this, in models with physical two-way interaction the passivity rule is a sufficient condition of stability, i.e., if each element is passive, then any arbitrary network of such elements remains stable.

Modularity and locality of computation

For a computational realization, it is desirable to decompose a model systematically into blocks and their interconnections. Such an object-oriented approach helps manage complex models through the use of the modularity principle. The basic modules can be formulated to correspond to elementary objects or functions in the physical domain at hand. Abstractions of new macro blocks on the basis of more elementary ones helps hiding details when building excessively complex models.

For one-directional interactions in signal processing, it is enough to have input and output terminals for connecting the blocks. For physical interaction, the connections need to be done through ports, with each port having a pair of K- or wave variables depending on the modeling method used. This follows the mathematical principles used for electrical networks Nilsson and Riedel [1999].

Locality of interaction is a desirable modeling feature, which is also related to the concept of causality. In a physical system with a single propagation speed of waves, it is enough that a block interacts only with its nearest neighbors; it does not need global connections to compute its task. If the properties of one block in such a localized model vary, the effect automatically propagates throughout the system. On the other hand, if some effects propagate for example at the speed of light but others with the speed of sound in air, the light waves are practically simultaneously everywhere. If the sampling rate in a discrete-time model is tuned to audio bandwidth (typically 44.1 or 48 kHz sample rate), the unit delay between samples is too long to represent light wave propagation between blocks. Two-way interaction with zero delay means a delay-free loop, the problem that we often face in physics-based sound synthesis. Technically it is possible to realize fractional delays Laakso et al. [1996], but delays shorter than the unit delay contain a delay-free component, so the problem is hard to avoid. There are ways to make such

systems computable, but the cost in time (or accuracy) may become prohibitive for real-time processing.

Types of complexity in physics-based modeling

Models are always just an approximation of real physical phenomena. Therefore, they reduce the complexity of the target system. This may be desired for a number of reasons, such as keeping the computational cost manageable, or more generally forcing some cost function below an allowed limit. These constraints are Particularly important in real-time sound synthesis and simulation.

A model's complexity often is the result of the fact that the target system is conceptually overcomplex to a scientist or engineer developing the model, and thus cannot be improved by the competence or effort available. An overcomplex system may be deterministic and modelable in principle but not in practice: It may be stochastic due to noise-like signal components, or it may be chaotic so that infinitesimally small disturbances lead to unpredictable states.

A particularly important form of complexity is perceptual overcomplexity. For example, in sound synthesis there may be no need to make the model more precise, because listeners cannot hear the difference. Phenomena that are physically prominent but do not have any audible effect can be excluded in such cases.

6.3 State-of-the-Art

This section starts with an analytical overview of physics-based *methods and techniques* for modeling and synthesizing musical instruments with an emphasis on the state of the art in each technique. The methods are grouped according to their variables. Wherever possible, we indicate their relation to the concepts and general schemes discussed in Sec.6.2.

Although some basic methods are commonly used in acoustics, we have excluded them because they do not easily solve the task of discrete-time modeling and simulation. For example, methods to solve the underlying partial differential equations are theoretically important but do not directly help in simulation or synthesis. Finite element and boundary element methods are generic and powerful for solving system behavior numerically, particularly for linear systems in the frequency domain, but we focus on inherently time-domain methods. Three-dimensional spaces, such as rooms and enclosures, can be modeled by the image source and ray

tracing techniques combined with late reverberation algorithms, but only the last one is useful in approximating resonators in musical instruments.

The second part of this section is devoted to a discussion of the current status of the *field*, as indicated by recent publications. Our approach there is holistic and synthetic. This part also helps us to extrapolate the current trends into the future paths and indicate the open problems of the field.

Our discussion so far has (indirectly) pointed out many fields related to the physics-based sound synthesis, including physics (esp. musical acoustics), mathematics, computer science, electrical engineering, digital signal processing, computer music, perception, human-computer interaction, and control. A novel result in these fields surely effects our field. However in order to keep our focus directed and the size of this chapter manageable, we have excluded these fields in our discussion.

6.3.1 K-models

Finite difference models

A finite difference scheme is a generic tool for numerically integrating differential equations Strikwerda [1989]. In this technique, the mathematical model, which is typically distributed on a bounded spatio-temporal domain, corresponds to the excitation plus resonator paradigm (see Fig.6.1). This mathematical model is discretized with the help of grid functions and difference operators. The numerical model can be *explicit* or *implicit* (in this case, iteration may be needed) Strikwerda [1989]. In either case, the operations are local. Typically one physical K-variable is directly observable, and the other is hidden in the states of the system.

In general, finite differences can be applied to a broad range of physical domains, such as electro-magnetic Taflove [1995], acoustic Botteldooren [1994], or mechanical Chaigne [1992]. An early example of using finite differences in physics-based sound synthesis can be found in Hiller and Ruiz [1971a,b]. Since then, finite differences have been applied successfully to multi-dimensional structures Chaigne [1992], Chaigne and Askenfelt [1994a,b], Chaigne and Doutaut [1997]. Currently, Chaigne systematically extends this line of research Chaigne [2002]. Among similar lines, a full-scale finite-difference piano model has been recently proposed in Giordano and Jiang [2004].

The finite difference model parameters are typically derived from the physical material properties, although the loss terms in most cases are simplified due to the lack of a general theory. Recently, a mixed derivative term is shown to have superior numerical properties for modeling frequency dependent losses compared to higher-order temporal differences Bensa et al. [2003b].

Standard DSP tools for analysis and design cannot be facilitated for finite difference models, as they do not follow regular DSP formulations. Recent DSP-oriented finite difference (re)formulations attempt to fill this gap Smith [2004b,a], Pakarinen [2004], Karjalainen and Erkut [2004]. In particular, Smith [2004a] is based on the *state-space* formalism Zadeh and Desoer [1963] to relate the finite differences to digital waveguides (see Sec. 6.3.2) or to any other LTI model.

The starting point in Karjalainen and Erkut [2004] is the duality between the ideal *waveguide mesh* (see Sec. 6.3.2) and 2-D finite difference model Savioja et al. [194]. This duality has been generalized to non-homogeneous media in higher dimensions, resulting in a modular local interaction scheme based on two K-variables (see Fig.6.2). Another advantage of this formulation is its efficiency in higher dimensions, which has been nicely exploited in Kelloniemi et al. [2005] for designing an efficient artificial reverberator with a dense modal pattern. From a different perspective, the relation of finite difference models and other modular techniques has been tackled in Barjau and Gibiat [2002] in case of wind instrument models.

The DSP tools aside, *von Neumann* analysis provides a standard technique for investigating the stability of an LTI finite-difference structure Strikwerda [1989], Press et al. [2002], Savioja [1999], Bilbao [2001]. Although finite difference schemes have provisions for modeling nonlinear and time-variant systems, it has been difficult to analyze their stability and passivity. Recent efforts of Bilbao provided a time-domain energetic analysis technique that is applicable to nonlinear and time-varying cases Bilbao [2005a]. In addition, a finite difference model that successfully simulates the distributed nonlinearities is presented in Pakarinen et al. [2005], Pakarinen [2004].

Although the locality of the finite difference structures have been exploited for parallel processing in general applications, in sound synthesis a parallel implementation has been rarely addressed. An exception is Motuk et al. [2005], which reports a parallel hardware implementation of a 2-D plate equation. Despite the large number of publications in the field, available sound synthesis software consists of a few Matlab toolboxes that focus on 1-D structures Kurz and Feiten [1996], Kurz [1995], Karjalainen and Erkut [2004]. The DSP-oriented finite difference structures have been implemented in BlockCompiler[3] Karjalainen [2003b,a], Karjalainen et al. [2003].

[3] `http://www.acoustics.hut.fi/software/BlockCompiler/`. Distribution and license undetermined.

Mass-spring networks

This group of techniques (also refereed to as *mass-interaction, cellular* or *particle* systems) decompose the original physical system in its structural atoms Cadoz et al. [1983]. These structural atoms are masses, springs, and dash-pots in the mechanical domain, although domain analogies may also be used. The interactions between the atoms are managed via explicit interconnection elements that handle the transfer of the K-variables between the synthesis objects. By imposing a constraint on the causality of action and reaction, and by using finite-difference formalism, modularity is also achieved Cadoz et al. [1983]. Thus, it is possible to construct complex modular cellular networks that are in full compliance with the diagram in Fig. 6.2. Mass-spring systems typically include special interaction objects for implementing time-varying or nonlinear interactions Florens and Cadoz [1991]. However, the energetic behavior and stability analysis of the resulting network is hard to estimate, since the existing analysis tools apply only to LTI cases.

The principles of mass-spring networks for physics-based sound synthesis were introduced by Cadoz and his colleagues within their system CORDIS-ANIMA Cadoz et al. [1983, 1993], which is a comprehensive audio-visual-tactile system. Their developments, achievements, and results in a large time-span are outlined in a recent review article Cadoz et al. [2003]. The most advanced sound synthesis model by mass-spring networks (and probably by any physics-based algorithm) is the model that creates the musical piece "pico..TERA". In this model, thousands of particles and many aggregate geometrical objects interact with each other to create 290 seconds of music without any external interaction or post-processing. Despite the successful examples, constructing a detailed mass-spring network is still a hard task, since the synthesis objects and their interaction topology require a large number of parameters. To address this issue, Cadoz and his coworkers developed helper systems for support, authoring, analysis, and parameter estimation of mass-spring networks Castagné and Cadoz [2002], Cadoz et al. [2003], Szilas and Cadoz [1998].

A renewed interest (probably due to the intuitiveness of the mass-spring metaphor) in cellular networks resulted in other systems and implementations, which are built upon the basic idea of the modular interactions but placing additional constraints on computation, sound generation, or control. These systems are PMPD Henry [2004b,a], TAO Pearson [1995, 1996], and CYMATIC Howard and Rimell [2004].

PMPD[4] closely follows the CORDIS-ANIMA formulation for visualization of mass-spring

[4]PMPD has multi-platform support and it is released as a free software under the GNU Public License (GPL). It

networks within the pd-GEM environment Puckette [1997], and defines higher-level aggregate geometrical objects such as squares and circles in 2-D or cubes or spheres in 3-D. Although the package is a very valuable tool for understanding the basic principles or mass-spring networks, it has limited support for audio synthesis.

TAO[5] specifically addresses the difficulty of model construction and introduces a scripting language. It uses a static topology of masses and springs, and provides pre-constructed 1-D (string) or 2-D (triangle, rectangle, circle, and ellipse) modules, but 3-D modules are not supported. Operations such as deleting the mass objects for constructing shapes with holes and joining the shapes are defined. For efficiency and reduction in the number of parameters, TAO constrains the spring objects by using a fixed spring constant. The system is driven by a score; the audio output is picked-up by virtual microphones and streamed to a file, which is normalized when the stream finishes.

The synthesis engine of CYMATIC is based on TAO, but it introduces two important improvements. The first improvement is the replacement of the forward differences common in all previous systems by central differences. The central differences result in a more stable model and reduce the frequency warping. The second improvement over TAO is the support for 3-D structures.

Modal synthesis

Linear resonators can also be described in terms of their vibrational modes in the frequency domain. This representation is particularly useful for sound sources that have a small number of relatively sharp resonances (such as the xylophone or the marimba Bork et al. [1999]), and may be obtained by experimental modal analysis Ewins [1986], Bissinger [2003]. The modal description, which essentially a frequency-domain concept, was successfully applied to discrete-time sound synthesis by Adrien Adrien [1989, 1991]. In his formulation, the linear resonators (implemented as a parallel filterbank) are described in terms of their modal characteristics (frequency, damping factor, and mode shape for each mode), whereas connections (representing all non-linear aspects) describe the mode of interaction between objects (e.g. strike, pluck, or bow). These ideas were implemented in MOSAIC software platform Morrison and Adrien [1993], which is later ported

can be downloaded from `http://drpichon.free.fr/pmpd/`.

[5]TAO is an active software development project and it is released as a free software under the GPL. It resides at `http://sourceforge.net/projects/taopm/`

and extended to Modalys[6].

Modal synthesis is best suited for mechanical domain and uses K-variables; it is modular and supports the bi-directional interaction scheme of Fig.6.2 (usually by iteration). The resonator filterbank is essentially a lumped model, however a matrix block brings back the spatial characteristics of a distributed system by transforming the input force to modal coordinates for weighting the individual resonances. An excellent DSP formulation of modal synthesis, based on the state-space formalism, can be found in Smith [2004b].

If the modal density of a sound source is high (such as a string instrument body), or if there are many particles contained in a model (such as a maracas) modal synthesis becomes computationally demanding. If the accuracy is of paramount importance (for example, in understanding the string-body coupling mechanism in a guitar Woodhouse [2004a,b]), then simplifications are not preferred. However, in sound synthesis, instead a detailed bookkeeping of each mode or particle, using stochastic methods significantly reduce the computational cost without sacrificing the perceived sound quality Cook [1997], Lukkari and Välimäki [2004b]. The basic building blocks of modal synthesis, as well as stochastic extensions are included in STK Cook [2002b], Cook and Scavone [1999][7].

Two linear modal resonators linked by an interaction element a la Fig.6.2 has been reported in Rocchesso and Fontana [2003]. The interaction element simulates nonlinear impact or friction iteratively and provides energy to the modal resonators. These impact and friction models are implemented as pd plugins[8].

The functional transform method (FTM) is a recent development closely related to the modal synthesis Trautmann and Rabenstein [2003]. In FTM, the modal description of a resonator is obtained directly from the governing PDEs by applying two consecutive integral transforms (Laplace and Sturm-Liouville) to remove the temporal and spatial partial derivatives, respectively. The advantage of this approach is that while traditional modal synthesis parameters are bound to the measured modal patterns of complex resonators, FTM can more densely explore the parameter space, if the problem geometry is simple enough and physical parameters are available. More recently, nonlinear extensions of the method, as well multirate implementations to reduce the computational load have been reported Trautmann and Rabenstein [2004],

[6]Proprietary software of IRCAM, see `http://www.ircam.fr/logiciels.html`

[7]STK has multiplatform support and it is released as open source without any specific license. It can be downloaded from `http://ccrma.stanford.edu/software/stk/`

[8]Available from `http://www.soundobject.org/software.html`, license unspecified.

Petrausch and Rabenstein [2005b].

Source-filter models

When an exciter in Fig.6.1 is represented by a signal generator, a resonator by a time-varying filter, and the bi-directional signal exchange between them is reduced to unidirectional signal flow from the exciter towards the resonator, we obtain a *source-filter model*. In some cases, these reductions can be physically justified (see Erkut [2002] for discussion concerning plucked string instruments), however in general they are mere simplifications, especially when the source is extremely complex, such as in the human voice production Kob [2004], Titze [2004], Arroabarren and Carlosena [2004] or in biosonar signal formation of marine mammals Erkut [2004].

Since the signal flow is strictly unidirectional, this technique does not provide good means for interactions. However, the resonators may be decomposed to arbitrary number of subblocks, and outputs of several exciters may be added. Thus, to a certain degree, the modularity is provided. The exciter is usually implemented as a switching wavetable, and resonators are simple time-varying filters. An advantage here is that these components are included in every computer music and audio signal processing platform. Thus, source-filter models can be used as early prototypes of more advanced physical models.

This is for example the case in *virtual analog synthesis*. This term became popular when the Nord Lead 1 synthesizer was introduced to the market as "an analog-sounding digital synthesizer that uses no sampled sounds[9]". Instead, a source-filter based technique was used. More physically oriented sound synthesis models of analog electric circuits have been recently reported in Huovilainen [2004]Karjalainen et al. [2004a].

The main reason of our focus on the source-filter models here is the *commuted synthesis* technique Smith [1993], Karjalainen et al. [1993b]. Recent references include Laurson et al. [2004], Välimäki et al. [2004a, 2003], Laurson et al. [2002, 2001].

[9] `http://www.clavia.com/`

6.3.2 Wave models

Wave digital filters

The *wave digital filter* (WDF) theory is originally formulated for conversion of analog filters into digital filters Fettweis [1986]. In phyics-based sound synthesis, a physical system is first converted to an equivalent electrical circuit using the domain analogies, then each circuit element is discretized (usually by the bilinear transform). Each object is assigned a port impedance and the energy transfer between objects is carried out by explicit interconnection objects (*adaptors*), which implement Kirchhoff laws and eliminate the delay-free loops. WDF models are mostly used as exciters, but are also applicable to resonators.

Recent references include Bilbao [2001], Bensa et al. [2003a], Bilbao et al. [2003], Bilbao [2004, 2005b] de Sanctis et al. [2005], Sarti and Tubaro [2002], Sarti and Poli [1999] van Walstijn and Scavo [2000], van Walstijn and Campbell [2003]

Digital waveguides

Digital waveguides (DWGs) are the most popular physics-based method for 1-D structures, such as strings and wind instruments Smith [2004b]. The reason for this is their extreme computational efficiency. They have been used also in 2-D and 3-D modeling, but in such cases they are not superior in efficiency.

A DWG is a bi-directional delay line pair with an assigned port admittance Y and it accommodates the wave variables of any physical domain. The change in Y across a junction of the waveguide sections causes *scattering*, and the scattering junctions of interconnected ports have to be formulated. Since DWGs are based on the wave components, this is not a difficult task, as the reflected waves can be causally formulated as a function of incoming waves. DWGs are mostly compatible with wave digital filters, but in order to be compatible with K-modeling techniques, special conversion algorithms must be applied to construct hybrid models.

Recent references include Shelly and Murphy [2005] Bensa et al. [2005] Essl et al. [2004b]Essl et al. [2004a] Smith [2004b] Bank et al. [2003] Esquef and Välimäki [2003] Rocchesso and Smith [2003] de la Cuadra et al. [2001] Serafin and Smith [2001]

6.3.3 Current directions in physics-based sound synthesis

Around the mid-1990s, the research reached the point, where most Western orchestral instruments could be synthesized based on a physical model. A comprehensive summary of this line of research is provided in Smith [2004c]. More recently, many papers have been published on the modeling of ethnic and historical musical instruments. These include, for example, the Finnish kantele Karjalainen et al. [1993a], Erkut et al. [2002], the Turkish tanbur Erkut and Välimäki [2000], the bowed bar Essl and Cook [2000], ancient Chinese flutes de la Cuadra et al. [2001], an African flute de la Cuadra et al. [2002], the Tibetan praying bowl Essl and Cook [2002], Essl et al. [2004a], and the Japanese sho Hikichi et al. [2003], among others.

We have also recently seen applications of physical modeling techniques to non-musical sound sources. Some examples of this are physical modeling of bird song Kahrs and Avanzini [2001], Smyth and Smith [2002], Smyth et al. [2003], various everyday sounds, such as those generated by wind chimes Cook [1997], Lukkari and Välimäki [2004a], footsteps Cook [2002a], and beach balls Rocchesso and Dutilleux [2001], and friction models that can be applied in many cases Avanzini et al. [2002]. An interesting aspect in this line of research, especially in Cook [1997] and Rocchesso and Fontana [2003] is the stochastic higher-level control blocks that govern the dynamics of simplistic (”cartoonified”) low-level resonator structures.

Another direction is the subjective evaluation of perceptual features and parameter changes in physics-based synthesis, see, Rocchesso and Scalcon [1999], Lakatos et al. [2000], Järveläinen et al. [2001], Järveläinen and Tolonen [2001]. This line of research provides musically relevant information on the relation of timbre and the properties of human hearing [REF:ENS CHAPTER]. These results help in reducing the complexity of synthesis models, because details that are inaudible need not be modeled.

The first attempts at audio restoration based on physical models were conducted recently Esquef et al. [2002]. While this can be successful for single tones, the practical application of such methods for recordings including a mix of several instruments is a challenge for future research. The main problem is high-quality source separation, which is required before this kind of restoration process. Sophisticated algorithms have been devised for this task, but generally speaking, separation of a musical signal into individual source signals is still a difficult research problem (see e.g. Klapuri [2003] and [REF:UPF CHAPTER]).

Using hybrid approaches in sound synthesis to maximize strengths and minimize weaknesses of each technique, has been previously addressed in Jaffe [1995]. It has been pointed

out that hybridization typically shows up after a technique has been around for some time and its characteristics have been extensively explored. A basic question, with increasing research interest, is to understand how different discrete-time modeling paradigms are interrelated and can be combined, whereby K-models and wave models can be understood in the same theoretical framework. In Karjalainen et al. [2004b] recent results are indicated, both in the form of theoretical discussions and by examples. Several other examples are given in Smith [2004c] andVälimäki et al. [2005]. Here, we focus on systematic approaches for constructing hybrid models, which address computability, accuracy, stability, and efficiency of the resulting structure.

A general method for constructing K-hybrids has been reported in Borin et al. [2000]. Based on the state-space formulation, this method performs a geometrical transformation on the nonlinearities to cut instantaneous dependencies. This can be done in two different ways; either as a table lookup, or iteratively by the Newton-Raphson method. The second approach has been successfully applied to construction of K-hybrids in Avanzini [2001]. Note that this approach essentially converts exciter-resonator type of K-models (see Fig. 6.1) to the modular, interactive scheme of Fig. 6.2.

The wave methods are more apt for hybrid modeling, as the WDF and DWG structures are mostly compatible. Moreover, the adaptors in WDFs and the isomorphic scattering junctions eliminate the delay-free loops and support the modular interaction scheme of Fig. 6.2. Because of these desirable properties, many wave-hybrids are reported, see Smith [2004c,b]. In addition, a recent generalized formulation of multivariable waveguides and scattering allows DWG networks to be constructed in a more compact way Rocchesso and Smith [2003].

Two systematic ways of interconnecting the wave models and K-models to construct KW-hybrids have been recently reported. One way of constructing KW-hybrids is to formulate a particular modular K-model with explicit instantaneous interaction elements a la Fig.6.2, and then to use a special KW-converter. The advantage of this approach is that the full dynamics of the K-model is preserved and its scheduling is made similar to that of the wave model. The disadvantage of this approach is that it is not general, as each K-model should be formulated separately for instantaneous modular interactions. Such a formulation is carried out in Karjalainen and Erkut [2004] for finite-difference structures.

Another way of constructing KW-hybrids is to formulate the K-models within the state-space formalism (as a black-box with added ports), and choose the port resistance to break instantaneous input-output path to avoid delay-free loops. The advantage of this approach is its generality, as any LTI K-model can be formulated as a state-space structure Smith [2004b]. The

disadvantage of this approach is that the dynamics of the K-model is hidden and its scheduling has to be separately authorized. A KW-hybrid modeling formulation based on the state-space formalism is presented in Petrausch and Rabenstein [2005a].

An example might better indicate why the hybrid modeling is currently a hot-topic. *RoomWeaver* Beeson and Murphy [2004] is a software based integrated development environment for the design, modeling and rendering of virtual acoustic spaces, based on the multidimensional DWG mesh. Since multidimensional finite difference structures Savioja et al. [194] are computationally more efficient Karjalainen and Erkut [2004], Kelloniemi et al. [2005] but poor in implementing the boundary conditions as DWG meshes do, a KW-hybrid (constructed by using the KW-converters Karjalainen and Erkut [2004]) is used in Room Weaver for an efficient solution. A fan-shaped room was simulated at the sampling rate of 44.1 kHz using 260330 nodes. The results show that both the computation time and the memory usage halves in the hybrid mesh compared to the conventional one.

6.4 Open Problems and Future Paths

6.4.1 Sound sources and modeling algorithms

There are many ethnic and historical musical instruments yet to be studied. An acoustical study may be combined with the physics-based sound synthesis in order to verify the acoustical characteristics of the instrument in focus. Moreover, there is a vast amount of performance characteristics to be explored. Ideally, these characteristics should be extracted from recordings rather than isolated experiments.

Physical parameter extraction techniques need to be extended. For best sound quality, computational methods that automatically calibrate all the parameter values of a physical model according to the sound of a good instrument should exist. This is very challenging and almost hopeless for some methods, and relatively easy only for some special cases.

Physical virtual analog synthesis is an important future path. Building an analog circuit and comparing the measured physical variables with the synthetic ones may improve the tuning of the virtual analog model parameters, and thus the quality of the audio output. Since many analog electrical, mechanical, and acoustical systems can be decomposed into elementary components, it is desirable to build a library of such components. The theory of wave digital filters

[Fettweis, 1986] may be facilitated for this purpose.

Important future directions in hybrid modeling include analysis of the dynamic behavior of parametrically varying hybrid models, as well as benchmark tests for computational costs of the proposed structures.

Motivated by the possible immersive and virtual reality applications [REF: DEI CHAPTER, VIPS-CHAPTER], the directivity and distributed radiation research and modeling are expected to be major challenging problems in physics-based sound synthesis in the next decade.

6.4.2 Control

The use of nonlinear dynamics Strogatz [1994] to control the simplistic low-level sound source models (as opposed to the stochastic control blocks of Cook [1997] and Rocchesso and Fontana [2003]) is surprisingly under-researched (the most mature applications are mentioned in Cadoz et al. [2003]). If carefully designed, the discrete-time nonlinear blocks can successfully modify the characteristics of simplistic synthesis objects in a dynamical fashion. This way, coherent and plausible sonic behavior (including synchronization and flocking Strogatz [2003]) of a large group of animate/inanimate objects may be efficiently modeled. The research in this direction is underway, and preliminary results are reported in Peltola [2004].

Going back to less exotic sound sources, the user control (or "playing") of physical models of musical instruments is another problem area where general solutions are unavailable. The piano is one of the easiest cases, because the player only controls the fundamental frequency and dynamic level of tones. In the cases of string and wind instruments, the control issue requires clever technical solutions. The control of virtual musical instruments is currently a lively research field Paradiso [1997], Cook [1992], Howard and Rimell [2004], Karjalainen and Mäki-Patola [2004]. Physics-based modeling for real-time sound synthesis of musical instruments is well-suited for interactive virtual reality. The parameters for model control are intuitive and closely related to the parameters used in controlling real instruments. These issues are further elaborated in [REF: DEI SECTION] and [REF:ControlChapter].

6.4.3 Applications

An ultimate dream of physical modeling researchers and instrument builders is virtual prototyping of musical instruments. This application will preeminently require physical models with excellent precision in the simulation of sound production. A musical instrument designer should have the possibility to modify a computer model of a musical instrument and then play it to verify that the design is successful. Only after this would the designed instrument be manufactured. Naturally, fine details affecting the timbre of the instrument should be faithfully simulated, since otherwise this chain of events would be fruitless. Current research is still far away from this goal and more research work is required.

The concept of Structured Audio introduced as part of the MPEG-4 international multimedia standard has opened a new application field for physical models Vercoe et al. [1998]: parametric coding of music, where a program for sound generation of the instruments and control data for playing the instrument are transmitted. The practical use of this idea remains a dream for the future.

In addition to synthesizing musical sounds, in the future, the physical modeling techniques are expected to be applied to numerous everyday sound sources for human-computer interfaces, computer games, electronic toys, sound effects for films and animations, and virtual reality applications.

Despite a long development history, significant recent advances, and premises such as control, efficiency, and sound quality, the physics-based sound synthesis still lacks a wide-spread use in music as a compositional tool for a composer/user (as opposed to a performance tool for an enthusiast), although several case studies have been reported in Chafe [2004]. We believe that the most important factor behind this is the lack of a unified modular modeling framework in full compliance with the scheme in Fig.6.2. Such a framework should optimally balance accuracy, efficiency, and ease of control, and operate in the widest range of physical domains. It should also and handle the parameter updates in a robust and predictable manner in real-time. Useful tools and metaphors should minimize the time devoted to instrument making and maximize the time devoted to music making. Designing such a framework may require a holistic approach spanning the domain from the sound to sense, and bringing the expertise in audio, control, and music together. In this respect, only the surface is scratched so far de Sanctis et al. [2005], Karjalainen [2003b], Karjalainen et al. [2003] (see also `http://www-dsp.elet.polimi.it/alma/`) and there is a vast amount of opportunities for further research and development.

6.5 Conclusions

To be written.

Bibliography

J.-M. Adrien. The missing link: modal synthesis. In G. De Poli, A. Piccialli, and C. Roads, editors, *Representations of Musical Signals*, pages 269–297. The MIT Press, Cambridge, Massachusetts, USA, 1991. URL `PAPERS/NeedHomepage.txt`.

Jean-Marie Adrien. Dynamic modeling of vibrating structures for sound synthesis, modal synthesis. In *Proc. AES 7th Int. Conf.*, pages 291–300, Toronto, Canada, May 1989. Audio Engineering Society. URL `PAPERS/NeedHomepage.txt`.

Ixone Arroabarren and Alfonso Carlosena. Vibrato in singing voice: The link between source-filter and sinusoidal models. *EURASIP Journal on Applied Signal Processing*, 2004(7):1007–1020, July 2004. URL `PAPERS/NeedHomepage.txt`. Special issue on model-based sound synthesis.

F. Avanzini, S. Serafin, and D. Rocchesso. Modeling interactions between rubbed dry surfaces using an elasto-plastic friction model. In *Proc. COST G6 Conf. Digital Audio Effects*, pages 111–116, Hamburg, Germany, Sep. 2002. URL `http://www.dei.unipd.it/~avanzini/`.

Federico Avanzini. *Computational Issues in Physically-based Sound Models*. PhD thesis, Dept. of Computer Science and Electronics, University of Padova, Italy, 2001. URL `http://www.dei.unipd.it/~avanzini/`.

B. Bank, F. Avanzini, G. Borin, G. De Poli, F. Fontana, and D. Rocchesso. Physically informed signal-processing methods for piano sound synthesis: a research overview. *EURASIP Journal on Applied Signal Processing*, 2003(10):941–952, Sep. 2003. URL `http://www.mit.bme.hu/~bank/`. Special Issue on Digital Audio for Multimeda Communications.

Ana Barjau and Vincent Gibiat. Delay lines, finite differences and cellular automata: Three close but different schemes for simulating acoustical propagation in 1D systems. *Acta Acustica united with Acustica*, 88(4):554–566, 2002. URL `http://www.upc.es/em/personal/barjau/`.

M. J. Beeson and D. T. Murphy. Roomweaver: A digital waveguide mesh based room acoustics research tool. In *Proc. COST G6 Conf. Digital Audio Effects*, pages 268–273, Naples, Italy, Oct. 2004. URL `PAPERS/NeedHomepage.txt`.

Arthur H. Benade. *Fundamentals of musical acoustics*. Dover Publications, Inc., New York, NY, USA, 2nd edition, 1990. URL `http://ccrma-www.stanford.edu/marl/Benade/`.

J. Bensa, S. Bilbao, R. Kronland-Martinet, J. O. Smith III, and T. Voinier. Computational modeling of stiff piano strings using digital waveguides and finite differences. *Acta Acustica united with Acustica*, 91(2):289–298, 2005.

Julien Bensa, Stefan Bilbao, Richard Kronland-Martinet, and Julius O. Smith III. A power normalized nonlinear lossy piano hammer. In *Proc. Stockholm Musical Acoustics Conf.*, pages 365–368, Stockholm, Sweden, Aug. 2003a. URL `http://www.lma.cnrs-mrs.fr/bensa.htm`.

Julien Bensa, Stefan Bilbao, Richard Kronland-Martinet, and Julius O. Smith. The simulation of piano string vibration: From physical models to finite difference schemes and digital waveguides. *J. Acoust. Soc. Am.*, 114(2):1095–1107, Aug. 2003b. URL `http://www.lma.cnrs-mrs.fr/bensa.htm`.

S. Bilbao. An energy-conserving difference scheme for nonlinear coupled transverse/longitudinal string vibration. In *Proc. European Sig. Proc. Conf.*, Antalya, Turkey, Sep. 2005a. URL `http://www.music.qub.ac.uk/school/staff/stefan.htm`. Accepted for publication.

S. Bilbao. Time-varying generalizations of allpass filters. *IEEE Signal Processing Letters*, 2005b. accepted for publication.

Stefan Bilbao. *Wave and Scattering Methods for Numerical Simulation*. John Wiley and Sons, Chichester, UK, 2004. ISBN 0-470-87017-6. URL `http://www.music.qub.ac.uk/school/staff/stefan.htm`.

Stefan Bilbao, Julien Bensa, and Richard Kronland-Martinet. The wave digital reed: A passive formulation. In *Proc. COST G6 Conf. Digital Audio Effects*, London, England, Sep. 2003. URL `http://www.music.qub.ac.uk/school/staff/stefan.htm`.

Stefan Damian Bilbao. *Wave and scattering methods for the numerical integration of partial differential equations*. PhD thesis, Stanford University, California, USA, May 2001. URL `http://www.music.qub.ac.uk/school/staff/stefan.htm`.

George Bissinger. Modal analysis of a violin octet. *J. Acoust. Soc. Am.*, 113(4):2105–2113, Apr. 2003. URL `http://www.ecu.edu/physics/George.htm`.

G. Borin, G. De Poli, and D. Rocchesso. Elimination of delay-free loops in discrete-time models of nonlinear acoustic systems. *IEEE Trans. Speech and Audio Processing*, 8(5):597–605, Sep. 2000. URL `PAPERS/NeedHomepage.txt`.

Gianpaolo Borin, Giovanni De Poli, and Augusto Sarti. Algorithms and structures for synthesis using physical models. *Computer Music J.*, 16(4):30–42, 1992.

Ingolf Bork, Antoine Chaigne, Louis-Cyrille Trebuchet, Markus Kosfelder, and David Pillot. Comparison between modal analysis and finite element modeling of a marimba bar. *Acta Acustica united with Acustica*, 85:258–266, 1999. URL `http://www.ptb.de/en/org/1/14/1401/index.htm`.

Dick Botteldooren. Acoustical finite-difference time-domain simulation in a quasi-cartesian grid. *J. Acoust. Soc. Am.*, 95(5):2313–2319, May 1994. URL `http://www.intec.rug.ac.be/data/default.html`.

C. Cadoz, A. Luciani, and J.-L. Florens. Responsive input devices and sound synthesis by simulation of instrumental mechanisms: The CORDIS system. *Computer Music J.*, 8(3):60–73, 1983. URL `http://www-acroe.imag.fr/`.

C. Cadoz, A. Luciani, and J.-L. Florens. Artistic creation and computer interactive multisensory simulation force feedback gesture transducers. In *Proc. Conf. New Interfaces for Musical Expression NIME*, pages 235–246, Montreal, Canada, May 2003. URL `http://www-acroe.imag.fr/`.

C. Cadoz, Annie Luciani, and Jean-Loup Florens. CORDIS-ANIMA. a modeling and simulation system for sound and image synthesis. the general formalism. *Computer Music J.*, 17(1):19–29, Spring 1993. URL `http://www-acroe.imag.fr/`.

N. Castagné and C. Cadoz. Creating music by means of ´physical thinking‘: The musician oriented Genesis environment. In *Proc. COST G6 Conf. Digital Audio Effects*, pages 169–174, Hamburg, Germany, Sep. 2002. URL `http://www-acroe.imag.fr/`.

C. Chafe. Case studies of physical models in music composition. In *Proc. 18th International Congress on Acoustics (ICA)*, pages 297 – 300, Kyoto, Japan, April 2004.

A. Chaigne. On the use of finite differences for musical synthesis. Application to plucked stringed instruments. *J. Acoustique*, 5(2):181–211, Apr. 1992. URL `http://wwwy.ensta.fr/~chaigne`.

A. Chaigne and A. Askenfelt. Numerical simulations of piano strings. I. A physical model for a struck string using finite difference methods. *J. Acoust. Soc. Am.*, 95(2):1112–1118, Feb. 1994a. URL `http://wwwy.ensta.fr/~chaigne`.

A. Chaigne and A. Askenfelt. Numerical simulations of piano strings. II. Comparisons with measurements and systematic exploration of some hammer-string parameters. *J. Acoust. Soc. Am.*, 95(3):1631–1640, March 1994b.

A. Chaigne and V. Doutaut. Numerical simulations of xylophones. I. Time-domain modeling of the vibrating bars. *J. Acoust. Soc. Am.*, 101(1):539–557, Jan. 1997. URL `http://wwwy.ensta.fr/~chaigne`.

Antoine Chaigne. Numerical simulations of stringed instruments – today's situation and trends for the future. *Catgut Acoustical Society Journal*, 4(5):12–20, May 2002. URL `http://wwwy.ensta.fr/~chaigne/`.

P. R. Cook. A meta-wind-instrument physical model, and a meta-controller for real-time performance control. In *Proc. Int. Computer Music Conf.*, pages 273–276, San Jose, California, October 14-18 1992.

P. R. Cook. Physically informed sonic modeling (PhISM): Synthesis of percussive sounds. *Computer Music J.*, 21(3):38–49, 1997. URL `http://www.cs.princeton.edu/~prc/`.

P. R. Cook. Modeling Bill's gait: Analysis and parametric synthesis of walking sounds. In *Proc. AES 22nd Int. Conf. Virtual, Synthetic, and Entertainment Audio*, pages 73–78, Espoo, Finland, June 2002a. URL `http://www.cs.princeton.edu/~prc/`.

Perry R. Cook. *Real sound synthesis for interactive applications*. A. K. Peters, Natick, MA, USA, 2002b. URL `http://www.cs.princeton.edu/~prc/`.

Perry R. Cook and Gary P. Scavone. The Synthesis ToolKit STK. In *Proc. Int. Computer Music Conf.*, Beijing, China, Oct. 1999. URL `http://www.cs.princeton.edu/~prc/`. For an updated version of STK see `http://ccrma-www.stanford.edu/software/stk/`.

P. de la Cuadra, T. Smyth, C. Chafe, and H. Baoqiang. Waveguide simulation of neolithic Chinese flutes. In *Proc. Int. Symposium on Musical Acoustics*, pages 181–184, Perugia, Italy, Sep. 2001. URL `http://ccrma.stanford.edu/~pdelac`.

P. de la Cuadra, C. Vergez, and R. Caussé. Use of physical-model synthesis for developing experimental techniques in ethnomusicology — The case of the ouldeme flute. In *Proc. Int. Computer Music Conf.*, pages 53–56, Gothenburg, Sweden, 2002. URL `http://ccrma.stanford.edu/~pdelac`.

G. De Poli and D. Rocchesso. Physically based sound modelling. *Organised Sound*, 3(1):61–76, 1998.

G. de Sanctis, A. Sarti, G. Scarparo, and S. Tubaro. An integrated system for the automatic block-wise synthesis of sounds through physical modeling. In *Proc. European Sig. Proc. Conf.*, Antalya, Turkey, Sep. 2005. URL `PAPERS/NeedHomepage.txt`. Accepted for publication.

C. Erkut, M. Karjalainen, P. Huang, and V. Välimäki. Acoustical analysis and model-based sound synthesis of the kantele. *J. Acoust. Soc. Am.*, 112(4):1681–1691, Oct. 2002. URL `http://www.acoustics.hut.fi/~cerkut`.

C. Erkut and V. Välimäki. Model-based sound synthesis of tanbur, a Turkish long-necked lute. In *Proc. ICASSP*, pages 769–772, Istanbul, Turkey, June 2000. URL `http://www.acoustics.hut.fi/~cerkut`.

Cumhur Erkut. *Aspects in analysis and model-based sound synthesis of plucked string instruments*. PhD thesis, Helsinki University of Technology, Espoo, Finland, Nov. 2002. URL `http://www.acoustics.hut.fi/~cerkut`. Available at http://lib.hut.fi/Diss/2002/isbn9512261901/.

Cumhur Erkut. Bioacoustical modeling for sound synthesis: a case study of odontoceti clicks. In *Proc. Baltic-Nordic Acoustics Meeting*, Mariehamn, Åland, June 2004. URL `http://www.acoustics.hut.fi/~cerkut`. The proceedings published electronically online at `http://www.acoustics.hut.fi/asf/bnam04/webprosari/onlineproc.html` and the papers do not have page numbers.

P. Esquef, V. Välimäki, and M. Karjalainen. Restoration and enhancement of solo guitar recordings based on sound source modeling. *J. Audio Eng. Soc.*, 50(4):227–236, Apr. 2002. URL `http://www.acoustics.hut.fi`.

Paulo A. A. Esquef and Vesa Välimäki. Design of an efficient inharmonic digital waveguide filter for synthesis of hand-bell sounds. In *Proc. FINSIG*, pages 49–53, Tampere, Finland, May 2003. URL `http://www.acoustics.hut.fi`.

G. Essl and P. R. Cook. Measurements and efficient simulations of bowed bars. *J. Acoust. Soc. Am.*, 108(1):379–388, July 2000.

G. Essl and P. R. Cook. Banded waveguides on circular topologies and of beating modes: Tibetan singing bowls and glass harmonicas. In *International Computer Music Association*, pages 49–52, Gothenburg, Sweden, 2002. URL `http://www.mle.media.mit.edu/~georg`.

G. Essl, S. Serafin, P. R. Cook, and J. O. Smith. Musical applications of banded waveguides. *Computer Music J.*, 28(1):51–63, 2004a. URL `http://www.mle.media.mit.edu/~georg`.

G. Essl, S. Serafin, P. R. Cook, and J. O. Smith. Theory of banded waveguides. *Computer Music J.*, 28(1):37–50, 2004b. URL `http://www.mle.media.mit.edu/~georg`.

D. J. Ewins. *Modal Testing: Theory and practice*. Research Studies Press Ltd., Herts, UK, 3rd edition, 1986.

Alfred Fettweis. Wave digital filters: Theory and practice. *Proc. IEEE*, 74(2):270–327, Feb. 1986. URL `http://www.nt.ruhr-uni-bochum.de/Extern/en/Group/Who/Prof/Fettweis/person.htm`.

Neville H. Fletcher. *Acoustic systems in biology*. Oxford University Press, New York, USA, 1992. URL `http://wwwrsphysse.anu.edu.au/eme/profile.php/34`.

Neville H. Fletcher and Thomas D. Rossing. *The Physics of Musical Instruments*. Springer-Verlag, New York, NY, USA, 2nd edition, 1998. URL `http://wwwrsphysse.anu.edu.au/eme/profile.php/34`.

J.-L. Florens and C. Cadoz. The physical model: Modeling and simulating the instrumental universe. In G. De Poli, A. Piccialli, and C. Roads, editors, *Representations of Musical Signals*, pages 227–268. The MIT Press, Cambridge, Massachusetts, 1991. URL `http://www-acroe.imag.fr/`.

N. Giordano and M. Jiang. Physical modeling of the piano. *EURASIP Journal on Applied Signal Processing*, 2004(7):926–933, July 2004. URL `http://www.physics.purdue.edu/faculty/ng`. Special issue on model-based sound synthesis.

C. Henry. Physical modeling for pure data (PMPD) and real time interaction with an audio synthesis. In *Proc. Sound and Music Computing*, Paris, France, Oct. 2004a. URL `http://drpichon.free.fr/pmpd/`.

C. Henry. PMPD : Physical modelling for pure data. In *Proc. Int. Computer Music Conf.*, Coral Gables, Florida, USA, Nov. 2004b. International Computer Music Association. URL `http://drpichon.free.fr/pmpd/`.

T. Hikichi, N. Osaka, and F. Itakura. Time-domain simulation of sound production of the sho. *J. Acoust. Soc. Am.*, 113(2):1092–1101, February 2003.

L. Hiller and P. Ruiz. Synthesizing musical sounds by solving the wave equation for vibrating objects: Part 1. *J. Audio Eng. Soc.*, 19(6):462–470, June 1971a. URL `PAPERS/NeedHomepage.txt`.

L. Hiller and P. Ruiz. Synthesizing musical sounds by solving the wave equation for vibrating objects: Part 2. *J. Audio Eng. Soc.*, 19(7):542–551, 1971b. URL `PAPERS/NeedHomepage.txt`.

D. M. Howard and S. Rimell. Real-time gesture-controlled physical modelling music synthesis with tactile feedback. *EURASIP Journal on Applied Signal Processing*, 2004(7):1001–1006, July 2004. URL `http://www-users.york.ac.uk/~dmh8/`. Special issue on model-based sound synthesis.

Antti Huovilainen. Non-linear digital implementation of the moog ladder filter. In *Proc. COST G6 Conf. Digital Audio Effects*, pages 61–64, Naples, Italy 2004. URL `http://www.acoustics.hut.fi/contact/ajhuovil`. Additional material and audio examples are available at `http://www.acoustics.hut.fi/publications/papers/dafx2004-moog/`.

D. A. Jaffe. Ten criteria for evaluating synthesis techniques. *Computer Music J.*, 19(1):76–87, 1995. URL `http://www.jaffe.com/`.

H. Järveläinen and T. Tolonen. Perceptual tolerances for decay parameters in plucked string synthesis. *J. Audio Eng. Soc.*, 49(11):1049–1059, Nov. 2001. URL `http://www.acoustics.hut.fi/~hjarvela`.

H. Järveläinen, V. Välimäki, and M. Karjalainen. Audibility of the timbral effects of inharmonicity in stringed instrument tones. *Acoustics Research Letters Online*, 2(3):79–84, July 2001. URL `http://www.acoustics.hut.fi/~hjarvela`.

M. Kahrs and F. Avanzini. Computer synthesis of bird songs and calls. In *Proc. COST G6 Conf. Digital Audio Effects*, pages 23–27, Limerick, Ireland, Dec. 2001. URL `PAPERS/NeedHomepage.txt`.

M. Karjalainen. Blockcompiler: A research tool for physical modeling and DSP. In *Proc. COST G6 Conf. Digital Audio Effects*, pages 264–269, London, UK, Sep. 2003a. URL `http://www.acoustics.hut.fi`.

M. Karjalainen. BlockCompiler: Efficient simulation of acoustic and audio systems. In *Proc. 114th AES Convention*, Amsterdam, Netherlands, Mar. 2003b. URL `http://www.acoustics.hut.fi`. Preprint 5756.

M. Karjalainen, J. Backman, and J. Pölkki. Analysis, modeling, and real-time sound synthesis of the kantele, a traditional Finnish string instrument. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pages 229–232, Minneapolis, MN, 1993a.

M. Karjalainen and C. Erkut. Digital waveguides versus finite difference structures: Equivalence and mixed modeling. *EURASIP Journal on Applied Signal Processing*, 2004(7):978–989, July 2004. URL `http://www.acoustics.hut.fi/~mak`.

M. Karjalainen, C. Erkut, and L. Savioja. Compilation of unified physical models for efficient sound synthesis. In *Proc. ICASSP*, volume 5, pages 433–436, Hong Kong, Apr. 2003. URL `http://www.acoustics.hut.fi/~mak`.

M. Karjalainen and T. Mäki-Patola. Physics-based modeling of musical instruments for interactive virtual reality. In *Proc. Int. Workshop on Multimedia Signal Processing*, pages 223–226, Siena, Italy, September 2004.

Matti Karjalainen, Teemu Mäki-Patola, Aki Kanerva, Antti Huovilainen, and P. Jänis. Virtual air guitar. In *AES 117th Convention*, San Francisco, CA, USA, Oct. 2004a. URL `http://www.acoustics.hut.fi`. NeedPreprint nr.

Matti Karjalainen, Jyri Pakarinen, Cumhur Erkut, Paulo A. A. Esquef, and Vesa Välimäki. Recent advances in physical modeling with k- and w-techniques. In *Proc. COST G6 Conf. Digital Audio Effects*, pages 107–112, Naples, Italy, Oct. 2004b. URL `http://www.acoustics.hut.fi`.

Matti Karjalainen, Tero Tolonen, Vesa Välimäki, Cumhur Erkut, Mikael Laurson, and Jarmo Hiipakka. An overview of new techniques and effects in model-based sound synthesis. *J. New Music Research*, 30(3):203–212, 2001. URL `http://www.acoustics.hut.fi`.

Matti Karjalainen, Vesa Välimäki, and Zoltán Jánosy. Towards high-quality sound synthesis of the guitar and string instruments. In *Proc. Int. Computer Music Conf.*, pages 56–63, Tokyo, Japan, Sep. 1993b.

A. Kelloniemi, V. Välimäki, P. Huang, and L. Savioja. Artificial reverberation using a hyper-dimensional FDTD mesh. In *Proc. European Sig. Proc. Conf.*, Antalya, Turkey, Sep. 2005. URL `http://www.tml.hut.fi/~kellonie`. Accepted for publication.

A. Klapuri. Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *IEEE Trans. Speech and Audio Processing*, 11(6):804–816, Nov. 2003. URL `PAPERS/NeedHomepage.txt`.

Malte Kob. Singing voice modeling as we know it today. *Acta Acustica united with Acustica*, 90 (4):649–661, July/Aug. 2004. URL `http://www.akustik.rwth-aachen.de/~malte/index.html.en`. Selection of papers presented at SMAC 2003.

Michael Kurz. Klangsynthese mittels physicalisher Modellierung einer schwingenden Saite durch numerische Integration der Differentialgleichung. Master's thesis, Technical University Berlin, 1995.

Michael Kurz and Bernhard Feiten. Physical modelling of a stiff string by numerical integration. In *Proc. Int. Computer Music Conf.*, pages 361–364, Hong Kong, Aug. 1996. International Computer Music Association.

T. I. Laakso, V. Välimäki, M. Karjalainen, and U. K. Laine. Splitting the unit delay — Tools for fractional delay filter design. *IEEE Signal Processing Mag.*, 13(1):30–60, Jan. 1996. URL `http://www.acoustics.hut.fi`.

S. Lakatos, P. R. Cook, and G. P. Scavone. Selective attention to the parameters of a physically informed sonic model. *Acoustics Research Letters Online*, Mar. 2000. URL `PAPERS/NeedHomepage.txt`. Published in J. Acoust. Soc. Am. 107, L31-L36.

M. Laurson, C. Erkut, V. Välimäki, and M. Kuuskankare. Methods for modeling realistic playing in acoustic guitar synthesis. *Computer Music J.*, 25(3):38–49, 2001. URL `http://www2.siba.fi/soundingscore/MikaelsHomePage/MikaelsHomePage.html`. Available at http://lib.hut.fi/Diss/2002/isbn9512261901/.

Mikael Laurson, Vesa Norillo, and Mika Kuuskankare. PWGLSynth, a visual synthesis language for virtual instrument design and control. *Computer Music J.*, 2004. URL `PAPERS/NeedHomepage.txt`. Submitted.

Mikael Laurson, Vesa Välimäki, and Cumhur Erkut. Production of virtual acoustic guitar music. In *Proc. AES 22nd International Conference*, pages 249–255, Espoo, Finland, June 2002. URL `PAPERS/NeedHomepage.txt`.

T. Lukkari and V. Välimäki. Modal synthesis of wind chime sounds with stochastic event triggering. In *Proc. Sixth Nordic Signal Processing Symposium*, pages 212–215, Espoo, Finland, June 2004a. URL `http://www.acoustics.hut.fi/publications/papers/norsig04-wind/`.

Teemu Lukkari and Vesa Välimäki. Modal synthesis of wind chime sounds with stochastic event triggering. In *Proc. NORSIG 2004*, Espoo, Finland, June 2004b. URL `http://www.acoustics.hut.fi/publications/papers/norsig04-wind/`.

J. Morrison and J. M. Adrien. MOSAIC: A framework for modal synthesis. *Computer Music Journal*, 17(1):45–56, 1993.

E. Motuk, R. Woods, and S. Bilbao. Parallel implementation of finite difference schemes for the plate equation on a FPGA-based multi-processor array. In *Proc. European Sig. Proc. Conf.*, Antalya, Turkey, Sep. 2005. URL `http://www.music.qub.ac.uk`. Accepted for publication.

J. W. Nilsson and S. A. Riedel. *Electric Circuits*. Prentice-Hall, 6th edition, 1999.

A. V. Oppenheim, A. S. Willsky, and S. H. Navab. *Signals and Systems*. Prentice-Hall, second edition edition, 1996.

J. Pakarinen, V. Välimäki, and M. Karjalainen. Physics-based methods for modeling nonlinear vibrating strings. *Acta Acustica united with Acustica*, 91(2):312–325, March/April 2005. URL `http://www.acoustics.hut.fi`.

Jyri Pakarinen. Spatially distributed computational modeling of a nonlinear vibrating string. Master's thesis, Helsinki University of Technology, Espoo, Finland, June 2004. URL `http://www.acoustics.hut.fi`.

J. Paradiso. Electronic music interfaces: New ways to play. *IEEE Spectrum*, 34(12):18–30, December 1997.

M. D. Pearson. Tao: a physical modelling system and related issues. *Organised Sound*, 1(1):43–50, Apr. 1995. URL `http://sourceforge.net/projects/taopm`.

M. D. Pearson. *Synthesis of Organic Sounds for Electroacoustic Music: Cellular Models and the TAO Computer Music Program*. PhD thesis, Department of Electronics, University of York, May 1996. URL `http://sourceforge.net/projects/taopm`.

Leevi Peltola. Analysis, parametric synthesis, and control of hand clapping sounds. Master's thesis, Helsinki University of Technology, Espoo, Finland, Dec. 2004. URL `http://www.acoustics.hut.fi/contact/leevi`.

S. Petrausch and R. Rabenstein. Interconnection of state space structures and wave digital filters. *IEEE Transactions on Circuits and Systems, Part II (Express Briefs)*, 52(2):90–93, Feb. 2005a.

Stefan Petrausch and Rudolf Rabenstein. Tension modulated nonlinear 2D models for digital sound synthesis with the functional transformation method. In *Proc. European Sig. Proc. Conf.*, Antalya, Turkey, Sep. 2005b. URL `http://www.lnt.de/lms`. Accepted for publication.

William H. Press, Saul A. Teukolsky, William A. Vetterling, and Brian P. Flannery. *Numerical recipes in C++*. Cambridge Univestiy Press, Cambridge, UK, 2nd edition, 2002. URL `http://www.nr.com`.

M. Puckette. Pure data. In *Proc. Int. Computer Music Conf.*, pages 224–227, Thessaloniki, Greece, Sep. 1997. URL `http://www.crca.ucsd.edu/~msp`.

D. Rocchesso and P. Dutilleux. Generalization of a 3-D acoustic resonator model for the simulation of spherical enclosures. *EURASIP Journal on Applied Signal Processing*, 2001(1):15–26, Mar. 2001. URL `http://profs.sci.univr.it/~rocchess/`.

D. Rocchesso and F. Fontana, editors. *The Sounding Object*. Edizioni di Mondo Estremo, Firenze, Italy, 2003. URL `http://profs.sci.univr.it/~rocchess/`.

D. Rocchesso and F. Scalcon. Bandwidth of perceived inharmonicity for musical modeling of dispersive strings. *IEEE Trans. Speech and Audio Processing*, 7(5):597–601, Sep. 1999. URL `http://profs.sci.univr.it/~rocchess/`.

D. Rocchesso and J. O. Smith. Generalized digital waveguide networks. *IEEE Trans. Speech and Audio Processing*, 11(3):242–254, May 2003. URL `http://profs.sci.univr.it/~rocchess/`.

Augusto Sarti and Giovanni De Poli. Toward nonlinear wave digital filters. *IEEE Transactions on Signal Processing*, 47(6):1654–1668, June 1999.

Augusto Sarti and Stefano Tubaro. Dynamical modeling of physics-based interactions in musical acoustics. In *Proc. AES 22nd International Conference*, pages 324–330, Espoo, Finland, June 2002.

L. Savioja, T. J. Rinne, and T. Takala. Simulation of room acoustics with a 3-D finite difference mesh. In *Proc. Int. Computer Music Conf.*, pages 463–466, Aarhus, Denmark, Sep. 194. URL `http://www.tml.hut.fi/~las/`.

Lauri Savioja. *Modeling Techniques for Virtual Acoustics*. PhD thesis, Helsinki University of Technology, Espoo, Finland, 1999. URL `http://www.tml.hut.fi/~las/`.

S. Serafin and J. O. Smith. Impact of string stiffness on digital waveguide models of bowed strings. *Catgut Acoustical Society Journal*, 4(4):49–52, 2001.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, July and October 1948.

S. Shelly and D. T. Murphy. Diffusion modelling at the boundary of a digital waveguide mesh. In *Proc. European Sig. Proc. Conf.*, Antalya, Turkey, Sep. 2005. URL `http://www.elec.york.ac.uk/ME/`. Accepted for publication.

J. O. Smith. Physical modeling using digital waveguides. *Computer Music J.*, 16(4):74–91, 1992. URL `http://ccrma.stanford.edu/~jos`.

J. O. Smith. On the equivalence of the digital waveguide and finite difference time domain schemes. URL `http://ccrma.stanford.edu/~jos`. Submitted for publication, July 2004a.

J. O. Smith. *Physical Audio Signal Processing: Digital Waveguide Modeling of Musical Instruments and Audio Effects, Draft*. http://ccrma.stanford.edu/~jos/pasp04/, August 2004b. URL `http://ccrma.stanford.edu/~jos`.

J. O. Smith. Virtual acoustic musical instruments: Review and update. *J. New Music Research*, 33 (3):283–304, Sep. 2004c. URL `http://ccrma.stanford.edu/~jos`.

Julius O. Smith. Viewpoints on the history of digital synthesis. In *Proc. Int. Computer Music Conf.*, pages 1–10, Montreal, Canada, Oct. 1991.

Julius O. Smith. Efficient synthesis of stringed musical instruments. In *Proc. Int. Computer Music Conf.*, pages 64–71, Tokyo, Japan, Sep. 1993.

Julius O. Smith. Physical modeling synthesis update. *Computer Music J.*, 20(2):44–56, 1996.

T. Smyth, J. S. Abel, and J. O. Smith. The estimation of birdsong control parameters using maximum likelihood and minimum action. In *Proc. Stockholm Musical Acoustics Conference*, pages 413–416, Stockholm, Sweden, August 2003.

T. Smyth and J. O. Smith. The sounds of the avian syrinx – are the really flute-like? In *Proc. International Conference on Digital Audio Effects*, pages 199–202, Hamburg, Germany, September 2002.

John C. Strikwerda. *Finite difference schemes and partial differential equations*. Wadsworth, Brooks & Cole, California, 1989. URL `PAPERS/NeedHomepage.txt`.

Steven Strogatz. *Nonlinear Dynamics and Chaos*. Studies in nonlinearity. Westview Press, 1994. URL `http://www.tam.cornell.edu/Strogatz1.html`.

Steven Strogatz. *Sync*. Allen Lane, The Penguin Press, London, UK, 2003. URL `http://www.tam.cornell.edu/Strogatz1.html`.

N. Szilas and C. Cadoz. Analysis techniques for physical modeling networks. *Computer Music J.*, 22(3):33–48, 1998. URL `http://www-acroe.imag.fr/`.

Allen Taflove. *Computational electrodynamics: The finite-difference time-domain method*. Artech House, Boston, MA, USA, 1995.

I. R. Titze. Theory of glottal airflow and source-filter interaction in speaking and singing. *Acta Acustica united with Acustica*, 90(4):641–648, July/Aug. 2004. URL `http://www.shc.uiowa.edu/wjshc/facultyandstaff/titze.html`.

Tero Tolonen, Vesa Välimäki, and Matti Karjalainen. Evaluation of modern sound synthesis methods. Technical Report 48, Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, Espoo, Finland, Mar. 1998.

L. Trautmann and R. Rabenstein. *Digital Sound Synthesis by Physical Modeling Using the Functional Transformation Method*. Kluwer Academic/Plenum Publishers, New York, NY, 2003. URL `http://www.lnt.de/lms/staff/index.php?lang=en&function=1&person=19`.

L. Trautmann and R. Rabenstein. Multirate simulations of string vibrations including nonlinear fret-string interactions using the functional transformation method. *EURASIP Journal on Applied Signal Processing*, 2004(7):949–963, June 2004.

V. Välimäki, M. Laurson, and C. Erkut. Commuted waveguide synthesis of the clavichord. *Computer Music J.*, 27(1):71–82, 2003. URL `http://www.acoustics.hut.fi`.

V. Välimäki and T. Takala. Virtual musical instruments - natural sound using physical models. *Organised Sound*, 1(2):75–86, 1996.

Vesa Välimäki. Physics-based modeling of musical instruments. *Acta Acustica united with Acustica*, 90(4):611–617, July/Aug. 2004. URL `http://www.acoustics.hut.fi/~vpv`.

Vesa Välimäki, Jyri Pakarinen, Cumhur Erkut, and Matti Karjalainen. Discrete-time modelling of musical instruments. *Rep. Prog. Phys.*, 2005. URL `PAPERS/NeedHomepage.txt`. Submitted for publication.

Vesa Välimäki, Henri Penttinen, Jonte Knif, Mikael Laurson, and Cumhur Erkut. Sound synthesis of the harpsichord using a computationally efficient physical model. *EURASIP Journal on Applied Signal Processing*, 2004(7):934–948, July 2004a. URL `http://www.acoustics.hut.fi/~vpv`. Special issue on model-based sound synthesis, companion page at `http://www.acoustics.hut.fi/publications/papers/jasp-harpsy/`.

Vesa Välimäki, Augusto Sarti, Matti Karjalainen, Rudolf Rabenstein, and Lauri Savioja. Editorial. *EURASIP Journal on Applied Signal Processing*, 2004(7):923–925, July 2004b. URL `http://www.acoustics.hut.fi`. Special issue on model-based sound synthesis.

M. van Walstijn and M. Campbell. Discrete-time modeling of woodwind instrument bores using wave variables. *J. Acoust. Soc. Am.*, 113(1):575–585, Jan. 2003. URL `http://www.ph.ed.ac.uk/~maarten`.

M. van Walstijn and G. P. Scavone. The wave digital tonehole model. In *Proc. Int. Computer Music Conf.*, pages 465–468, Berlin, Germany, NeedMonth 2000. URL `http://www.ph.ed.ac.uk/~maarten`.

B. L. Vercoe, W. G. Gardner, and E. D. Scheirer. Structured audio: creation, transmission, and rendering of parametric sound representations. *Proc. IEEE*, 86(3):922–940, Nov. 1998. URL `PAPERS/NeedHomepage.txt`.

Jim Woodhouse. On the synthesis of guitar plucks. *Acta Acustica united with Acustica*, 90(5): 928–944, 2004a. URL `http://www2.eng.cam.ac.uk/~jw12/`.

Jim Woodhouse. Plucked guitar transients: Comparison of measurements and synthesis. *Acta Acustica united with Acustica*, 90(5):945–965, 2004b. URL `http://www2.eng.cam.ac.uk/~jw12/`.

L. A. Zadeh and C. A. Desoer. *Linear System Theory: The State Space Approach*. Series in system science. McGraw-Hill, New York, NY, 1963. URL `http://www.cs.berkeley.edu/~zadeh`.

Chapter 7

Interactive sound

Federico Avanzini
University of Padova, Dep. of Information Engineering, Padova, Italy

Abstract

The research status in sound modeling for interactive computer animation and virtual reality applications is reviewed. It is argued that research in this field needs to become more aware of studies in ecological perception and multimodal perception. A number of relevant studies in these fields, that address auditory perception, are reviewed.

7.1 Introduction

This chapter tries to trace a route that, starting from studies in ecological perception and action-perception loop theories, goes down to sound modeling and design techniques for interactive computer animation and virtual reality applications.

We do not intend to (and most of all we am not able to) provide an in-depth discussion about different theories of perception. We rather review a number of studies from experimental psychology that can have relevant implications for the design of auditory feedback in interactive settings.

We start from the analysis of relevant related literature in perception and end with sound modeling. The technically inclined reader may turn the chapter upside-down and start reading the last sections, referring to the initial material when needed.

7.2 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. General introductions to the ecological approach to perception are to be found in Gibson Gibson [1986] and in Michaels and Carello Michaels and Carello [1981]. Carello and Turvey Carello and Turvey [2002] also provide a synthetic overview of the main concepts of the ecological approach.

The label "ecological" reflects two main themes that distinguish this approach from the establishment view. First, perception is an achievement of animal-environment systems, not simply animals (or their brains). What makes up the environment of a particular animal –cliffs or caves or crowds– is part of this theory of perception. Second, perception's main purpose is guiding activity, so a theory of perception cannot ignore what animals do. The kinds of activities that a particular animal does how it eats and moves and mates are part of this theory of perception.

7.2.1 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 and cochlea. 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 bear 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 part called the retina) has lost and distorted a lot of detail. The job of perception, then, becomes one of fixing the input and adding meaningful interpretations to it so that the brain can make an inference (or educated guess) 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 filling in requires inferential cognitive processing.

A theory of *direct* perception, in contrast, argues that sensory stimulation is lawfully determined in such a way that there exists a 1:1 correspondence between patterns of sensory stimulation and the underlying aspects of physical reality Gibson [1986]. This is a very strong assumption, since it basically says that reality is specified in the available sensory stimulation. Intermediary steps are only needed if the scientist has described the input incorrectly.

Gibson Gibson [1986] provides an example in the domain of visual perception, which supports, in his opinion, the direct perception theory. For centuries, scientists believed that distance is not perceivable by eye alone. Indeed, if the objects are treated as isolated points in otherwise empty space, then their distances on a line projecting to the eye are indistinct: each stimulates the same retinal location. Gibson argues that this formulation is inappropriate for addressing how we see. Instead he emphasizes the contribution of a continuous background

surface to providing rich visual structure. The simple step of acknowledging that points do not float in the air but are attached to a surface such as the ground introduces what might be considered a higher-order property, the gradient.

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. For Gibson, this realization opens up the new possibility that perception might be veridical, that is, about facts of the world. A relevant consequence of the direct perception approach is that sensory reference frames are unnecessary: if perception is direct, then perceivables can be measured relative to physical referents.

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 Gibson [1986] 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 having to do with whether any muscle commands had been issued to move the eyes or the head or the legs. In the absence of counter-acting motor commands, object motion can be concluded; in the presence of such commands, the retinal signals would be counteracted, allowing the alternative conclusion of self-motion. Overall (global) change in the pattern of light is specific to self-motion; local change against a stationary background is specific to object motion.

This simple insight opened a new field of research devoted to uncovering the structure in changing patterns of light: *optic flow*. Optic flow refers to the patterns of light, structured by particular animal-environment settings, available to a point of observation. The goal of optic flow research is to discover particular reliable patterns of optical structure, called invariants, relevant to guiding activity. Outflow and inflow are distinct forms of optic flow distinct flow morphologies that tell the perceiver whether she is moving forward or backward. As scientists consider how that flow is structured by the variety of clutter that we encounter as we move around doorways and hillsides and the like they discover invariants specific to those facts as well.

Perceivers exploits *invariants* in the optic flow, in order to effectively guide their activities. Carello and Turvey Carello and Turvey [2002] provides the following instructive example: as

a busy waiter rushes towards the swinging door of the restaurant kitchen, he makes subtle adjustments to his behavior in order to control his collision. He needs to maintain enough speed to push through the door but not so much that he crashes into it. Effective behavior requires that he knows when a collision will happen (so he does not slow down too early) and how hard the collision will be (so that he slows down enough). Optical structure relevant to these facts can be identified, and provides examples of quantitative invariants.

The above considerations apply not only to visual perception but also to other senses, including audition (see Section 7.2.2 next). Moreover, recent research has introduced the concept of *global array* Stoffregen and Bardy [2001]. 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 multiple forms of ambient 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.

Stoffregen and Bardy Stoffregen and Bardy [2001] exemplify this concept by examining the well known McGurk effect McGurk and MacDonald [1976], which is widely interpreted as reflecting general principles of intersensory interaction. In studies of this effect the visual portion of a videotape shows a speaker saying one syllable, while on the audio track a different syllable is presented. Observers are instructed to report the syllable on the audio track, and perceptual reports are strongly influenced by the nominally ignored visible speaker. One of the most consistent and dramatic findings is that perceptual reports frequently are not consistent with either the visible or the audible event. Rather, observers often report "a syllable that has not been presented to either modality and that represents a combination of both". The sustained interest in the McGurk effect arises in part from the need to explain how it is that the final percept differs qualitatively from the patterns in the optic and acoustic arrays.

Stoffregen and Bardy Stoffregen and Bardy [2001] claim that the McGurk effect is consistent with the general premise that perceptual systems do not function independently, but work in a cooperative manner to pick up higher-order patterns in the global array. If speech perception is based on information in the global array, then it must be unnatural (or at least uncommon) for observers who can both see and hear the speaker to be asked to report only what is audible; the global array provides information about what is being said, rather than about what is visible or what is audible: multiple perceptual systems are stimulated simultaneously and the stimulation has a single source (i. e., a speaker). In research on the McGurk effect the discrepancy between

the visible and audible consequences of speech is commonly interpreted as a conflict between the two modalities, but it could also be interpreted as creating information in the global array that specifies the experimental manipulation, that is, the global array may specify that what is seen and what is heard arise from two different speech acts.

Affordances

The most radical contribution of Gibson's theory is probably the notion of *affordance*. Gibson [Gibson, 1986, p. 127] 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. If a stair is a certain proportion of a person's leg length, it affords climbing ; if a surface is rigid relative to the weight of an animal, it affords stance and perhaps traversal; if a ball is falling with a certain velocity relative to the speed that a person can generate in running toward it, it affords catching, and so on.

Therefore, affordances are the possibilities for action of a particular animal-environment setting; they are usually described as -ables, as in catch-able, pass-through-able, climbable, and so on. What is important is that affordances are not determined by absolute properties of objects and environment, but depend on how these relate to a particular animal, including that animal's size, agility, style of locomotion, and so on Stoffregen [2000].

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 Stoffregen [2000]. 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.

7.2.2 Everyday sounds and the acoustic array

Ecological psychology has 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 environ-

ment, 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 students of audition investigate the identification of sound source properties, such as material, size, shape, and so on.

Two companion papers by Gaver Gaver [1993b,a] have greatly contributed to the build-up of a solid framework for ecological acoustics. Specifically, Gaver [1993b] deals with foundational issues, addresses such concepts as the acoustic array and acoustic invariants, and proposes a sort of "ecological taxonomy" of sounds.

Musical listening versus everyday listening

In Gaver [1993b] Gaver formulates an interestin example: you are walking along a road at night when you hear a sound. On the one hand, you might pay attention to its pitch and loudness and the ways they change with time. You might attend to the sound's timbre, whether it is rough or smooth, bright or dull. You might even notice that it masks other sounds, rendering them inaudible. These are all examples of *musical listening*, in which the perceptual dimensions and attributes of concern have to do with the sound itself, and are those used in the creation of music. These are the sorts of perceptual phenomena of concern to most traditional psychologists interested in sound and hearing.

On the other hand, as you stand there in the road, it is likely that you will notice that the sound is made by an automobile with a large and powerful engine. Your attention is likely to be drawn to the fact that it is approaching quickly from behind. And you might even attend to the environment, hearing that the road you are on is actually a narrow alley, with echoing walls on each side. This is an example of *everyday listening*, the experience of listening to events rather than sounds. Most of our experience of hearing the day-to-day world is one of everyday listening: we are concerned with listening to the things going on around us, with hearing which are important to avoid and which might offer possibilities for action. The perceptual dimensions and attributes of concern correspond to those of the sound-producing event and its environment, not to those of the sound itself. This sort of experience is not well understood by traditional approaches to audition.

The experience of everyday listening may serve as the foundation for a new explanatory framework for understanding sound and listening. Such a framework would allow us to understand listening and manipulate sounds along dimensions of sources rather than sounds. Studies

of audition have been further constrained by sensation-based theories of perception and the supposed primitives of sound they suggest. Physical descriptions of sound are dominated by those suggested by the Fourier transform: frequency, amplitude, phase, and duration. Traditional explanations of psychophysics take these "primitive" physical dimensions as their elemental stimuli and use them to motivate the identification of corresponding "elemental" sensations. questions concerning auditory event perception – if recognised at all – are left to higher-level cognitive accounts.

A theoretical framework for everyday listening must answer two simple but fundamental questions. First, in expanding upon traditional accounts of elemental sensations, we must develop an account of ecologically relevant perceptual entities: the dimensions and features of events that we actually obtain through listening. Thus the first question to be answered is: "What do we hear?". Similarly, in expanding traditional accounts of the primitive physical features of sound, we must seek to develop an ecological acoustics, which describes the acoustic properties of sounds that convey information about the things we hear. Thus the second question to be answered is: "How do we hear it?"

Acoustic flow and acoustic invariants

Back to the example of hearing an approaching automobile: there is continuum of energy between the source event (the automobile) and the experience. There are several landmarks along the way, each with its own characteristics as a medium for structured patterns of energy.

The first landmark is that of a sound-producing event, or source of sound. In the case of the automobile, some proportion of the energy produced by burning petrol causes vibrations in the material of the car itself (instead of contributing to its gross movement). Things tap, scrape, slosh, rub, roll, and flutter. These mechanical vibrations, in turn, produce waves of alternating high and low pressure in the air surrounding the car. The pattern of these pressure waves 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 spreading pressure waves, then, may serve as information about the vibrations that caused them, and thus for the event itself. When they change with sufficient amplitude and within a range of frequencies to which the auditory system is sensitive, the result is a sound wave from which a listener might obtain such information.

Each source of sound involves an interaction of materials. For instance, when two gears rub against each other, the patterns of their vibration depend both on the force, duration, and

changes over time of their interaction as well as the size, shape, material, and texture of the gears themselves. The pressure waves produced by the resulting vibration are determined by these attributes, and thus may serve as information about them.

Sound is also structured by and informative about the environment in which the event occurs. Much of the sound that reaches us from a source has reflected off various other objects in the environment, which colour the spectrum of reflected sound. In addition, the medium itself shapes the sounds it conveys: sound waves lose energy, especially high-frequency energy, as they travel through the air, and thus provide information about the distance of their sources. As sources move with respect to a potential observation point, their frequencies shift, producing the Doppler effect. Finally, changes in loudness caused by changes in distance from a source may provide information about time-to-contact in an analogous fashion to changes in visual texture.

The result is an auditory array, analogous to the optical array. In sum, then, sound provides information about an interaction of materials at a location in an environment. We can hear an approaching automobile, its size and speed. We can hear where it is, and how fast it is approaching. And we can hear the narrow, echoing walls of the alley it's driving along. Traditional psychologists only study part of the continuum from source to experience. Typical research focuses on the sound itself, analysing it in terms of properties such as amplitude and perceived loudness, or frequency and perceived pitch. Such research misses the higher-level structures that are informative about events.

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 likely that many parameters that change frequency also change other attributes of a sound. For example, changing the size of an object will change the frequencies of the sound it produces, but not their pattern. Changing the shape, on the other hand, changes both the frequencies and their relationships. These complex patterns of change may serve as information distinguishing the physical parameters responsible: These are the acoustic invariants that an ecological acoustics is concerned with discovering.

Maps of everyday sounds

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 causing the vibration, those of the vibrating object's material, and those of the object's configuration.

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, such as the hissing of a leaky pipe or the rush of wind from a fan, are caused by more continuous events. Another sort of aerodynamic event involves situations in which changes in pressure themselves impart energy to objects, causing them to vibrate. For example, when wind passes through a wire,

Sound-producing events involving liquids (e.g., dripping and splashing) are like those of vibrating solids in that they depend on an initial deformation that is countered by the material's restoring forces. But it seems the resulting vibration of the liquid does not usually affect the air in direct and audible ways. Instead, the resulting sounds are determined by the formation and change of resonant cavities in the surface of the liquid. As the object hits the liquid, it pushes it aside, forming a cavity that resonates to a characteristic frequency, amplifying and modifying the pressure wave formed by the impact itself.

Although all sound-producing events seem to involve vibrating solids, aerodynamic, or liquid interactions, many also depend on complex patterns of the simple events described above. So footsteps consist of temporal patterns of impact sounds, while door slams involve the squeak of scraping hinges and the impact of the door on its frame. Some of these involve timing of successive events, so that, for instance, successive footstep sounds probably must occur within a range of rates and regularities to be heard as walking. Others are likely to involve mutual constraints on the objects that participate in related events. For instance, concatenating the creak of a heavy door closing slowly with the slap of a light door slammed shut would be likely to sound quite unnatural.

Starting from these considerations, Gaver derives a tentative map of everyday sounds (see also figure 7.1).

Basic Level Sources: consider, for example, the region describing sounds made by vibrating solids. Four fundamentally 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, breaking,

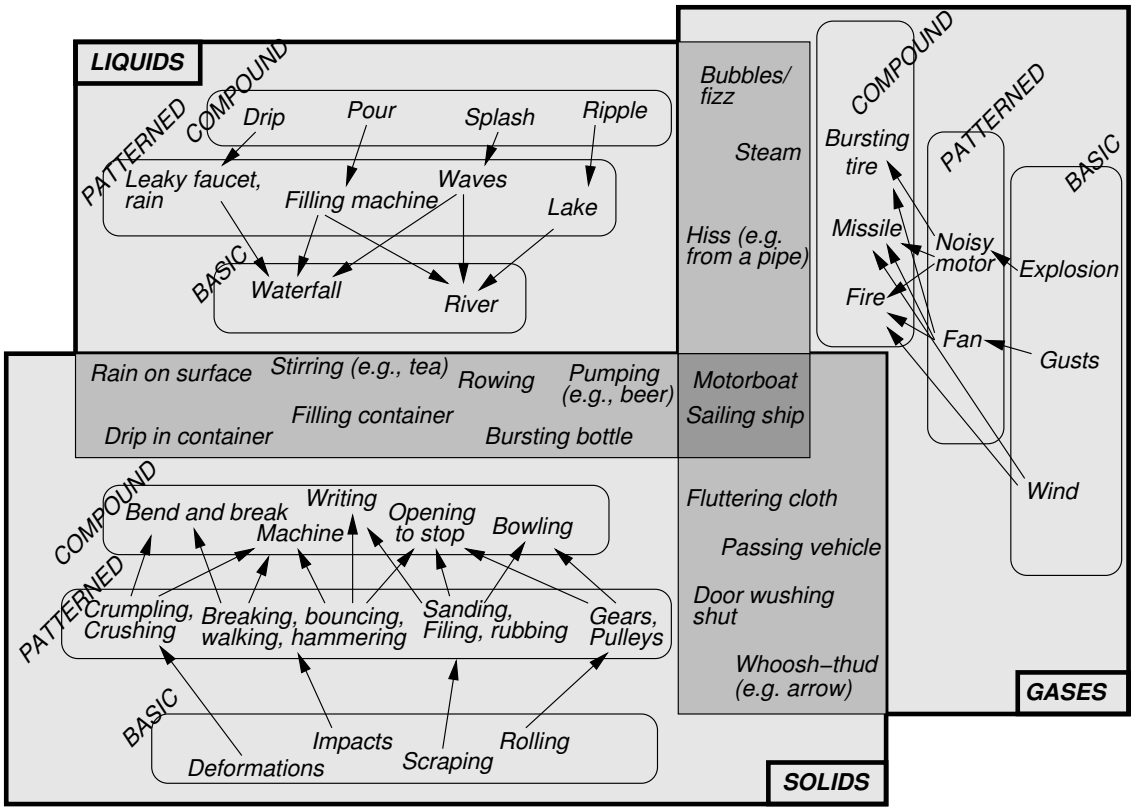

Figure 7.1: A map of everyday sounds, complexity increases towards the center. Figure based on Gaver [1993b].

spilling, walking and hammering 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 sort of basic level event. For instance, the sounds made by writing 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 sort of material is involved. For instance, when water drips on a reverberant surface, the resulting sounds 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.

7.2.3 Relevant studies

Most of the papers reviewed below present results on "solids" (see figure 7.1, while there seems to be a lack of studies on sound-producing events that involve liquids and aerodynamic interactions. Anyway sounds from solids are especially interesting when talking about interaction: auditory cues frequently occur when we touch or interact with objects, and these sounds often convey potentially useful information regarding the nature of the objects with which we are interacting.

Basic level sources

Wildes and Richards Wildes and Richards [1988] have studied material perception. The purpose of the authors was to find an acoustical parameter that could characterize material type uniquely, i.e. despite variations in objects features such as size or shape. Materials can be characterized using the coefficient of internal friction, which is a measure of anelasticity (in descending order of anelasticity we have steel, glass, wood and rubber). In the acoustical domain the coefficient of internal friction was found to be measurable using both the quality factor Q and the decay time of vibration t_e, this latter measured as the time required for amplitude to decrease to $1/e$ of its initial value. For decreasing anelasticity we have an increase in Q and a decrease in t_e.

Lutfi and Oh Lutfi and Oh [1997] have also performed a study on material perception. They studied material discrimination in synthetic struck clamped bar sounds. Stimulus synthesis was based on variations in the elasticity and density of the bars, whose values were perturbed about those found in iron, silver, steel, copper, glass, crystal, quartz, and aluminium. Perturbations were applied either to all the frequency components together (lawful covariation) or independently to each of them (independent perturbation). On half of the trials participants had to tell which of two presented stimuli was an iron sound, silver, steel, and copper being the alternatives, while on the other half of the trials the target was glass, and the alternatives were crystal, quartz, and aluminium. Participants were given feedback on the correctness of the response after each trial. Participants performance was analyzed in terms of the weights given to three different acoustical parameters: frequency, decay, and amplitude. Data revealed that discrimination was mainly based on frequency in all conditions, with amplitude and decay rate being of secondary importance.

Klatzky and coworkers Klatzky et al. [2000] investigated material discrimination in stimuli with variable frequency and modulus of internal friction $tan\phi$. In the first two experiments subjects had to judge on a continuous scale the perceived difference in the material of an object. Stimuli had the same values of frequency and $tan\phi$, but in the seconds experiment they were equalized by overall energy. As results did not differ significantly in the two experiments, it could be concluded that intensity is not relevant in the judgment of material difference. Experiments 3 and 4 were conducted on the same set of stimuli used in experiment 2. In the former subjects had to judge the difference in the perceived length of the objects, in the latter they had to categorize the material of the objects using four response alternatives: rubber, wood, glass and steel. Results indicated that judgments of material difference and of length difference were significantly influenced by both the $tan\phi$ coefficient and the fundamental frequency, even though the contribution of the decay parameter to length difference was smaller than that to material difference. An effect of both these variables was found in a categorization task: for lower decay factors steel and glass were chosen over rubber and plexiglass. Glass and wood were chosen for higher frequencies than steel and plexiglass.

Freed Freed [1990] addressed hardness perception in impact sounds. Freed's study is oriented toward measuring an attack related timbral dimension using a sound source oriented judgment scale: hardness. Investigated stimuli were generated by percussing four cooking pans, with variable diameter, with six mallets of variable hardness. Mallet hardness ratings were found to be independent of the size of the pans, thus revealing the ability to judge properties of the percussor independently of properties of the sounding object. The main goal of this study was to derive a psychophysical function for mallet hardness ratings, based on the properties of the acoustical signal. Preliminary experiments pointed out that the useful information for mallet hardness rating was contained in the first 300 ms of the signals. For this reason acoustical analyses focused on this portion of the signals. Four acoustical indices were measured: average spectral level, spectral level slope (i.e., rate of change in spectral level, a measure of damping), average spectral centroid ,and spectral centroid TWA (i.e., time weighted average). These acoustical indices were used as predictor in a multiple regression analysis. Altogether they accounted for 75% of the variance of the ratings.

Carello *et al.* Carello et al. [1998] have investigated the recognition of the length of wood rods dropped on the floor. In both the experiments, the latter focusing on shorter lengths than the ones used in the first experiment, subjects judged the perceived length by adjusting the distance of a visible surface in front of them. Subjects where found to be able to scale length of the rods

consistently. Physical length was found to correlate strongly with estimated length (r = .95 in both cases), although the second experiment showed a grater compression of the length estimates (the slope of the linear regression function equaled .78 in the first experiment, .44 in the second one). Analysis of the relationship between the acoustical and perceptual levels was carried on using three acoustical features: signal duration, amplitude and spectral centroid. None of the considered acoustical variables, apart from log amplitude in the second experiment, predicted length estimates better than actual length. Length estimates were finally explained by means of an analysis of the moments of inertia of a falling rod. Results of these latter analysis show potential analogies between the auditory and the tactile domain.

Lederman Lederman [1979] compared the effectiveness of tactile and auditory information in judging the roughness of a surface (i.e., texture). Roughness of aluminium plates was manipulated by varying the distance between adjacent grooves of fixed width, or by varying the width of the grooves. Subjects task was to rate numerically the roughness of the surface. In one condition (auditory) participants were presented the sounds generated by the experimenter who moved his fingertips along the grooved plate. In the second two conditions subjects were asked to move their fingertips onto the plate. In the tactile condition they whore cotton plugs and earphones while touching the plate. In the auditory+tactile condition they were able to ear the sounds they generated when touching the plate. Roughness estimates were not different between the auditory+tactile and tactile conditions, but differed in the auditory condition. In other words when both kinds of information were present, the tactile one played the strongest role in determining experimental performance. Roughness estimates were shown to increase as both the distance between grooves and the width of the grooves decreased. Additionally roughness estimates increased as the force exerted by the finger on the surface increased, and as the speed of the movement of the fingers decreased. The effect of the force on roughness estimates in the auditory condition was however not constant across subjects. A speculative discussion concerning the relative role of pitch and loudness in determining the estimates is provided by the author, although no acoustical analyses of the experimental stimuli are provided. More recent research by Lederman and coworkers Lederman et al. [2002] has focused on surface roughness perception when the surface is explored using a rigid probe rather than with the bare skin. It is known that in this case the role of vibratory texture coding is different, because the probe provides a rigid link between the skin and the surface. When people feel a surface with a rigid probe, *vibratory* roughness perception occurs. This study investigates relative contributions of tactile and auditory information to judgments of surface roughness. The haptic and auditory stimuli were obtained by asking subjects to use a probe to explore a set of plates with periodic

textures of varying interelement spacings. Experiments were run using three modality conditions (touch-only, audition-only, and touch+audition), and results showed that, although tactual dominance is still found, sound plays a more relevant role when using a probe than in the case of direct contact with bare fingers. The authors argue that this may be due not only to the different interaction conditions, but also to the simple fact that the amplitude of the accompanying sounds is considerably greater for probe-based exploration than for bare skin contact.

Patterned sources

Warren and coworkers Warren and Verbrugge [1988] have investigated acoustic invariants in bouncing and breaking events. According to the terminology adopted by the ecological approach to perception, we distinguish between two classes of invariants (i.e., higher-order acoustical properties) that specify the sound generating event. Structural invariants specify the properties of the objects, while transformational invariants specify their style of change. Warren and Verbrugge investigated the nature of the structural invariant that allow identification of breaking and bouncing events. On the basis of a physical analysis of these two kinds of events the authors hypothesized that the nature of these invariants was essentially temporal, static spectral properties having no role in the identification of breaking and bouncing events. Experimental stimuli were generated by dropping one of three different glass objects on the floor from different heights, so that for each of the objects a bouncing event and a breaking one were recorded. Once the ability of participants to correctly identify these two types of events was assessed with the original stimuli, two further experiments were conducted using synthetic stimuli. The bouncing event was synthesized by superimposing four trains of damped quasi-periodic pulses generated using, for each one, a recording from one of four different bouncing glass tokens. These four sequences had the same damping. The breaking event was synthesized by superimposing the same four damped quasi-periodic sequences, but using a different damping coefficient for each of them (in the second experiment the breaking stimuli were preceded by a 50 ms noise burst of the original breaking sound). Identification performance was extremely accurate in all cases, despite the strong simplifications of the spectral and temporal profile of the acoustical signal. Therefore the transformational invariants for bouncing was identified as a single damped quasi-periodic sequence of pulses, while that for breaking was identified as a multiple damped quasi-periodic sequence of pulses.

Repp Repp [1987] reports a study on auditory perception of hands clapping. Repp's

work is an extension of the so called motor theory of speech perception to the investigation of a non-speech communicative sound: claps. In particular Repp hypothesized the subjects' ability to recognize the size and the configuration of clapping hands on the basis of the auditory information. Recognition of hands size was also related to recognition of the gender of the clapper, given that male have in general bigger hands than females. Several clapping sounds were recorded from different clappers. In the first experiment clapper gender and hand size recognition were investigated indirectly, as participants were asked to recognize the identity of the clapper. Overall clapper recognition was not good, although listeners performance in the identification of their own claps was much better. Gender recognition was barely above chance. Gender identification appeared to be guided by misconceptions: faster, higher-pitched and fainter claps were judged to be produced by females and vice-versa. In the second experiment, subjects had to recognize the configuration of the clapping hands. Overall subjects were found to be able to recover correctly the hand configuration from sound. Although hands configuration was a determinant of the clapping sound spectrum, the best predictor of performance was found to be clapping rate, spectral variables having only a secondary role.

Li *et al.* Li et al. [1991] studied gender recognition in walking sounds. Walking sounds of seven females and seven males were recorded. Subjects were asked to categorize the gender of the walker on the basis of a four step walking sequence. Results show recognition levels well above chance. Several anthropometric measures were collected on the walkers. Male and female walkers were found to differ in height, weight and shoe size. Spectral and duration analyses were performed on the recorded walking excerpts. Duration analysis indicated that female and male walkers differed in respect to the relative duration of the stance and swing phases, but not inrespect to the walking speed. Nonetheless judged maleness was significantly correlated with the latter of these two variables, but not with the former. Several spectral measures were derived from the experimental stimuli: spectral centroid, skewness, and kurtosis, spectral mode, average spectral level, and low and high spectral slopes. Two components where then derived by applying a principal components analysis on the spectral predictors. These components were used as predictors for both physical and judged gender. Overall male walkers were characterized by a lower spectral centroid, mode and high frequency energy than females, and by higher values for skewness, kurtosis and low-frequency slope. The same tendencies were found when the two components were used as predictors for judged gender. Results gathered from the analysis of the relationship between the acoustical and the perceptual levels were then tested in a further experiment. Stimuli were generated by manipulating the spectral mode of the two most ambiguous walking excerpts (spectral slopes too were altered, but manipulation of this

feature was not completely independent of the manipulation of the spectral mode). Consistently with previous analyses, the probability of choosing the response "male" was found to decrease as spectral mode increased. A final experiment showed that judged gender could be altered by having a walker wear shoes of the opposite gender.

Unlike other studies on the perception of environmental sounds, the work of Gygi *et al.* Gygi et al. [2004] does not focus on a specific event or feature. Instead the authors use for their experiments a large (70) and varied catalog of sounds, which covers "nonverbal human sounds, animal vocalizations, machine sounds, the sounds of various weather conditions, and sounds generated by human activities". Patterned, compound, and hybrid sounds (according to the terminology used by Gaver Gaver [1993b]) are included, e.g., beer can opening, bowling, bubbling, toilet flushing, etc. The experiments apply to non-verbal sound an approach used in early studies of speech perception, namely the use of low-, high-, and bandpass filtered speech to assess the importance of various frequency regions for speech identification. The third experiment (see abstract) is perhaps the most interesting one. The authors seem to follow an approach already suggested by Gaver: "[. . .] if one supposes that the temporal features of a sound are responsible for the perception of some event, but that its frequency makeup is irrelevant, one might use the amplitude contour from the original sound to modify a noise burst." Gaver [1993a].

The results show that identifiability is heavily affected by experience and that it has a strong variability between sounds. The authors try to quantify the relevance of tem temporal structures through a selection of time- and frequency-domain parameters, including statistics of the envelope (a quantitative meaasure of the envelope "roughness"), autocorrelation statistics (to reveal periodicities in the waveform), and moments of the longterm spectrum (to see if some spectral characteristics were preserved when the spectral information was drastically reduced.). Correlation of these parameters with the EMN identification results shows that three variables are mainly used by every group of listeners in every experimental condition: number of auto-correlation peaks, ratio of burst duration to total duration, cross-channel correlation. These are all temporal features, reflecting periodicity, amount of silence, and coherence of envelope across channels.

7.3 Multimodal perception and interaction

7.3.1 Combining and integrating auditory information

Humans achieve robust perception through the combination and integration of information from multiple sensory modalities. According to the intersensory integration view of perceptual development, multisensory perception emerges gradually during the first months of life, and experience significantly shapes multisensory functions. By contrast, according to the intersensory differentiation view, 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 Ernst and Bülthoff [2004]: 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. That is, they may be in different units, coordinate systems, or about complementary aspects of the same environmental property. Disambiguation and cooperation are examples for two such interactions: if a single modality is not enough to come up with a robust estimate, information from several modalities can be combined. For example, for object recognition different modalities complement each other with the effect of increasing the information content.

By contrast, sensory integration describes interactions between redundant signals. That is, to be integrated, the sensory estimates must be in the same units, the same coordinates and about the same aspect of the environmental property. Ernst and Bülthoff Ernst and Bülthoff [2004] illustrate this concept with an example: when knocking on wood at least three sensory estimates about the location (L) of the knocking event can be derived: visual (V), auditory (A) and proprioceptive (P). 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. The *modality precision* or *modality appropriateness* hypothesis by Welch and Warren Welch and Warren [1986] is often cited when trying to explain which modality dominates under what circumstances. These hypotheses state that discrepancies are always resolved in favour of the more precise or more appropriate modality. In spatial tasks, for example, the visual modality usually dominates, because it is the most precise at determining spatial information. For temporal judgments,however, the situation is reversed and audition, being the more appropriate modality, usually dominates over vision. In texture perception tasks haptics dominates on other modalities, and so on.

With regard to this concept, Ernst and Bülthoff Ernst and Bülthoff [2004] note thatthe terminology *modality precision* and *modality appropriateness* is misleading because it is not the modality itself or the stimulus that dominates. Rather, the dominance is determined by the estimate and how reliably it can be derived within a specific modality from a given stimulus. Therefore, the term *estimate precision* would probably be more appropriate. The authors also list a series of questions for future research, among which one can find "What are the temporal aspects of sensory integration?". This is a particularly interesting question, since, as already noted, audition provides salient temporal information.

Auditory capture and illusions

In psychology, there is a long history of studying intermodal conflict and illusions to uncover the principles of perception. Studying multisensory illusions is a promising approach for investigating multisensory integration. is frequently studied using intermodal conflict, as in the ventriloquist effect whereby the perceived location of a sound shifts towards a visual stimulus presented at a different position. Much of the multisensory literature has focused on spatial interactions, as in the ventriloquist effect whereby the perceived location of a sound shifts towards

a visual stimulus presented at a different position. Identity interactions are also studied, such as the already mentioned McGurk effect McGurk and MacDonald [1976]. In the McGurk effect, 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, it is also the case that the senses can interact in time, that is, not where or what 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 to spatial ventriloquism, as the effect 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. For example, Shams *et. al* Shams et al. [2002] presented subjects with a briefly flashed visual stimulus that was accompanied by one, two or more auditory beeps. There was a clear influence of the number of auditory beeps on the perceived number of visual flashes. That is, if there were two beeps subjects frequently reported seeing two flashes when only one was presented. Maintaining the terminology above, this effect may be called auditory capture.

Another recent study Morein-Zamir et al. [2003] has tested a related hypothesis: that auditory events can alter the perceived timing of target lights. Specifically, four experiments reported in Morein-Zamir et al. [2003] investigated whether irrelevant sounds can influence the perception of lights in a visual temporal order judgment task, where participants judged which of two lights appeared first. The results show that presenting one sound before the first light and another one after the second light improves performance relative to baseline (sounds appearing simultaneously with the lights), as if the sounds pulled the perception of lights further apart in time. More precisely, the performance improvement results from the second sound trailing the second light. On the other hand, two sounds intervening between the two lights lead to a decline in performance, as if the sounds pulled the lights closer together. These results demonstrate a temporal analogue of the spatial ventriloquist effect, where visual events can alter the perceived location of target sounds.

These capture effects, or broadly speaking, these integration effects, are of course not only limited to vision and audition. In principle they can occur between any modalities (even within modalities).

Some authors have investigated whether audition can influence tactile perception similarly to what Shams *et al.* have done for vision and audition, and found that an incompatible number of auditory beeps can increase or decrease the number of simultaneously felt taps.

Hötting and Röder Hötting and Röder [2004] report upon a series of experiments where a single tactile stimulus was delivered to the right index finger of subjects, accompanied by one to four task-irrelevant tones. Participants (both sighted and congenitally blind) had to judge the number of tactile stimuli. As a test of whether possible differences between sighted and blind people were due to the availability of visual input during the experiment, half of the sighted participants were run with eyes open (sighted seeing) and the other half were blindfolded (sighted blindfolded). The first tone always preceded the first tactile stimulus by 25 ms and the time between the onsets of consecutive tones was 100 ms. Participants were presented with trials made of a single tactile stimulus accompanied by no, one, two, three, or four tones. All participants reported significantly more tactile stimuli when two tones were presented than when no or only one tone was presented. Sighted participants showed a reliable illusion for three and four tones as well, while blind participants reported a lower number of perceived tactile stimuli than sighted seeing or sighted blindfolded participants. These results extend the finding of the auditory-visual illusion established by Shams *et al.* Shams et al. [2002] to the auditory-tactile domain. Moreover, the results (especially the discrepancies between sighted and congenitally blind participants) suggest that interference by a task-irrelevant modality is reduced if processing accuracy of the task-relevant modality is high.

Bresciani *et al.* Bresciani et al. [2005] have conducted a very similar study, and investigated whether the perception of tactile sequences of two to four taps delivered to the index fingertip can be modulated by simultaneously presented sequences of auditory beeps when the number of beeps differs (less or more) from the number of taps. This design allowed to systematically test whether task-irrelevant auditory signals can really modulate (influence in both directions) the perception of tactile taps, or whether the results of Hötting and Röder Hötting and Röder [2004] merely reflected an original but very specific illusion. In the first experiment, the auditory and tactile sequences were always presented simultaneously. Results demonstrate that tactile tap perception can be systematically modulated by task-irrelevant auditory inputs. Another interesting point is the fact that subjects responses were significantly less variable when redundant tactile and auditory signals were presented rather than tactile signals alone. This suggests that even though auditory signals were irrelevant to the task, tactile and auditory signals were probably integrated. A second experiment tested whether the auditory and tactile stimuli are integrated

when the timing between auditory and tactile sequences is manipulated. The sequences of beeps and taps were always similar, but for some timing conditions they were presented in temporal asynchrony. Results show that the auditory modulation of tactile perception was weaker when the auditory stimuli were presented immediately before the onset or after the end of the tactile sequences. This modulation completely vanished with a 200 ms gap between the auditory and tactile sequences. Shams *et al.* Shams et al. [2002] found that the temporal window in which audition can bias the perceived number of visual flashes is about 100 ms. These results suggest that the temporal window of auditory-tactile integration might be wider than for auditory-visual integration. One more observed effect was that the auditory sequence biased tactile perception when presented immediately *after* the tactile sequence, but not when presented immediately *before*. The authors regard this as a reasonable result: since the tactile sensory signals from the fingertip take longer to reach the brain than the auditory signals coming from the ears, the tactile stimulus has to precede the auditory one in order for the two to be centrally represented as simultaneous.

These studies provide evidence of the fact that the more salient (or reliable) a signal is, the less susceptible to bias this signal should be. In the same way, the more reliable a biasing signal is, the more bias it should induce. Therefore, the fact that auditory signals can bias both visual and tactile perception probably indicates that, when counting the number of events presented in a sequence, auditory signals are more reliable than both visual and tactile signals. When compared to the studies by Shams et al. Shams et al. [2002], the effects observed on tactile perception are relatively small. This difference in the magnitude of the auditory-evoked effects likely reflects a higher saliency of tactile than visual signals in this kind of non-spatial task.

Other studies have concentrated on auditory-tactile integration in surface texture perception. Studies by Lederman and coworkers Lederman [1979], Lederman et al. [2002], already mentioned in the previous section, have shown that audition had little influence on texture perception when participants touched the stimulus with their fingers Lederman [1979]. However, when the contact was made via a rigid probe, resulting in an increase of touch-related sound and a degradation of tactile information, auditory and tactile cues were integrated Lederman et al. [2002]. These results suggest that although touch is mostly dominant in texture perception, the degree of auditory-tactile integration can be modulated by the reliability of the single-modality information

In a related study, Guest *et al.* Guest et al. [2002] have focused on audio-tactile interactions in roughness perception. In their experimental setup, participants were required to make forced-

choice discrimination responses regarding the roughness of abrasive surfaces which they touched briefly. Texture sounds were captured by a microphone located close to the manipulated surface and subsequently presented through headphones to the participants in three different conditions: veridical (no processing), amplified (12dB boost on the $2 - 20$kHz band), and attenuated (12dB attenuation in the same band). The authors investigated two different perceptual scales: smooth-rough, and moist-dry. Analysis of discrimination errors verified that attenuating high frequencies led to a bias towards an increased perception of tactile smoothness (or moistness), and conversely the boosted sounds led to a bias towards an increased perception of tactile roughness (or dryness). This work is particuarly interesting from a sound-design perspective, since it investigate the effects of a *non-veridical* auditory feedback (not only the spectral envelope is manipulated, but sounds are picked up in the vicinity of the surface and are therefore much louder than in natural listening conditions).

7.3.2 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 simple 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 what takes place in the brain, and that very likely there are internal representations in the brain (e.g. content-bearing internal states), more recent theories have questioned such a modular decomposition in which cognition interfaces between perception and action. The ecological approach discussed in section 7.2 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

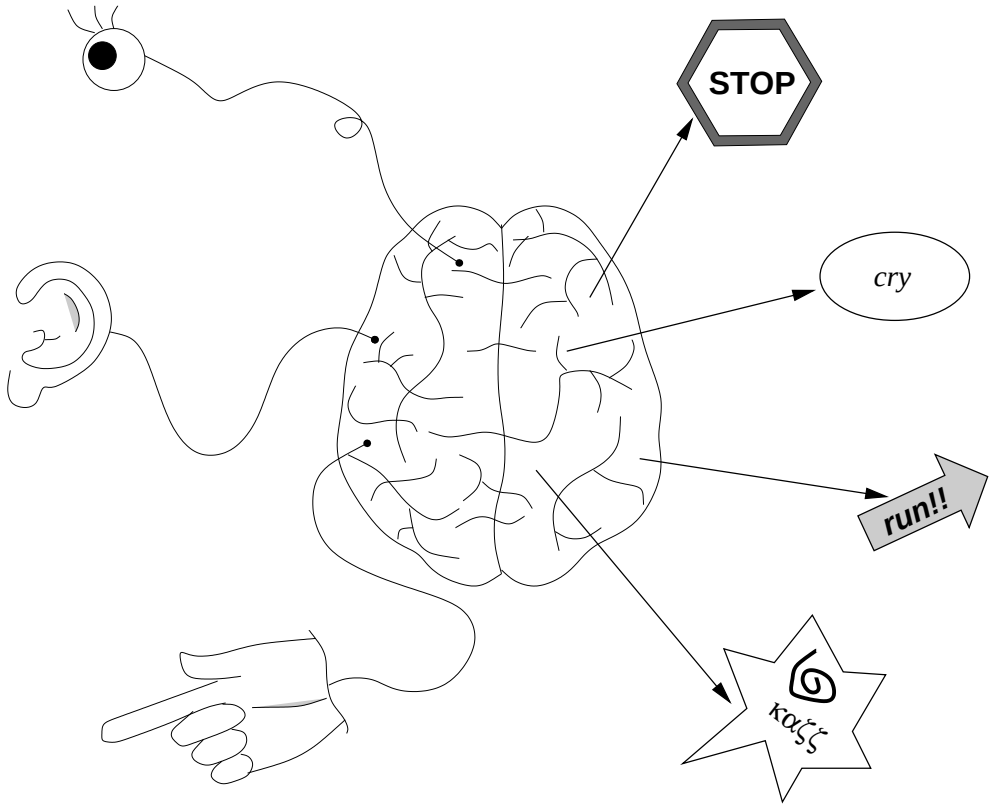

Figure 7.2: A cartoon representation of traditional views of the perception-action functions as a causal one-way flow.

(not even truly conceivable) 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 skillful activity on the part of the animal as a whole. As stated by Nöe Noë [2005], blind creatures may be capable of thought, but thoughtless creatures could never be capable of sight, or of any genuine content-bearing perceptual experience. In other words, 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 and coworkers (see O'Regan and Noë [2001] for a detailed review of other works based on similar ideas). Varela, Thompson and Rosch Varela et al. [1991] presented an "enactive conception" of experience according to which experience is not something that occurs inside the animal, but is something the animal *enacts* as it explores the environment in which it is situated. In this view, the subject

of mental states is taken to be 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 come from having a body with various 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 live cognition.

O'Regan and Nöe O'Regan and Noë [2001] have proposed a closely related approach, according to which perception consists in exercising an exploratory skill. The authors illustrate their approach with an example: the sensation of softness that one might experience in holding a sponge consists in being aware that one can exercise certain practical skills with respect to the sponge: one can for example press it, and it will yield under the pressure. The experience of softness of the sponge is characterized by a variety of such possible patterns of interaction with the sponge. O'Regan and Nöe term *sensorimotor contingencies* the laws that describe these interactions. When a perceiver knows, in an implicit, practical way, that at a given moment he is exercising the sensorimotor contingencies associated with softness, then he is in the process of experiencing the sensation of softness.

O'Regan and Nöe O'Regan and Noë [2001] then classify sensory inputs according to two criteria, i.e. *corporality* and *alerting capacity*. *Corporality* is the extent to which activation in a neural channel systematically depends on movements of the body. Sensory input from sensory receptors like the retina, the cochlea, and mechanoreceptors in the skin possesses corporality, because any body motion will generally create changes in the way sensory organs are positioned in space, and consequently in the incoming sensory signals (the situation is less clear for the sense of smell, but sniffing, blocking the nose, moving the head, do affect olfactory stimulation). Proprioceptive input from muscles also possesses corporality, because there is proprioceptive input when muscle movements produce body movements. The authors argue that corporality is one important factor that explains the extent to which a sensory experience will appear to an observer as being truly sensory, rather than non-sensory, like a thought, or a memory. The *alerting capacity* of sensory input as the extent to which that input can cause automatic orienting behaviors that peremptorily capture the organism s cognitive processing resources. Vision, touch, hearing, and smell have not only high corporality but also high alerting capacity, provided by the fact that sudden changes in visual, tactile, auditory or olfactory stimulation provoke immediate

orienting behaviors that peremptorily modify cognitive processing. With high corporality and high alerting capacity, vision, touch, hearing and smell have strong phenomenal presence. This is in accordance with the usual assumption that they are the prototypical sensory modalities.

A possible objection to the definitions of perception and action given above is that most sensations can be perceived without any exploratory skill being engaged. For example, having the sensation of red or of a bell ringing does not seem to involve the exercising of skills. Such an objection can be overcome by realizing that sensations are never instantaneous, but are always extended over time, and that at least potentially, they always involve some form of activity. O'Regan and Nöe refer in O'Regan and Noë [2001] to a number of experiments, especially in the domain of visual perception, that support this idea Experiments on "change blindness" present observers with displays of natural scenes and ask them to detect cyclically repeated changes (e.g., large object shifting, changing colors, and so on). Under normal circumstances a change of this type would create a transient signal in the visual system that would be detected by low-level visual mechanisms and would attract attention to the location of the change. However in the change blindness experiments conditions were arranged such that these transients were hidden by superimposing a brief global flicker over the whole visual field at the moment of the change. The results of the experiments showed that in many cases observers have great difficulty seeing changes, even when the changes are extremely large (and are perfectly visible to someone who knows what they are). Such results contrast with the subjective impression of "seeing everything" in an observed scene or picture. O'Regan and Nöe regards them as a support to the view that an observer sees the aspects of a scene which he/she is currently "visually manipulating", which makes it reasonable that only a subset of scene elements that share a particular scene location can be perceived at a given moment.

Again in the domain of visual perception, Nöe Noë [2005] discuss the concept of "experiential blindness" and reports upon cases where this phenomenon has been observed. According to Nöe there are, broadly speaking, two different kinds of blindness: blindness due to damage or disruption of the sensitive apparatus (e.g., caused by cataracts, by retinal disease or injury, or by brain lesion in the visual cortex), and blindness that is not due to the absence of sensation or sensitivity, but rather to the person's inability to integrate sensory stimulation with patterns of movement and thought. The latter is termed experiential blindness because it occurs despite the presence of normal visual sensation.

As an example of the occurrence of experiential blindness, Nöe considers attempts to restore sight in congenitally blind individuals whose blindness is due to cataracts impairing

the eye's sensitivity by obstructing light on its passage to the retina. The medical literature reports that surgery restores visual sensation, at least to a significant degree, but that it does not restore sight. In the period immediately after the operation, patients suffer blindness despite rich visual sensations. This clearly contrasts with the traditional input-output picture described at the beginning of this section, according to which removing the cataract and letting in the light should enable normal vision.

A related phenomenon is that of blindness caused by paralysis. Normally the eyes are in nearly constant motion, engaging in sharp movements several times a second. If the eyes cease moving, they lose their receptive power. A number of studiies are reported in Noë [2005], which show that images stabilized on the retina fade from view. This is probably an instance of the more general phenomenon of sensory fatigue thanks to which we do not continuously feel our clothing on our skin, the glasses resting on the bridge of our nose, or a ring on our finger. This suggests that some minimal amount of eye and body movement is necessary for perceptual sensation.

Audition and sensory substitution

According to the theories discussed above, the quality of a sensory modality does not derive from the particular sensory input channel or neural activity involved in that specific modality, but from the laws of sensorimotor skills that are exercised. The difference between hearing and seeing amounts to the fact that, among other things, one is seeing if there is a large change in sensory input when blinking; on the other hand, onee is hearing if nothing happens when one blinks but there is a left/right difference when one turns the head, and so on. This line of reasoning implies that it is possible to obtain a visual experience from auditory or tactile input, provided the sensorimotor laws that are being obeyed are the laws of vision.

The phenomenon of *sensory substitution* is coherent with this view. Perhaps the first studies on sensory substitution are due to Bach-y-Rita who, starting from 1967 has been experimenting with devices to allow blind people to "see" via tactile stimulation provided by a matrix of vibrators connected to a video camera. A comprehensive review of this research stream can be found in Kaczmarek et al. [1991].

The tactile visual substitution systems developed by Bach-y-Rita and coworkers use matrices of vibratory or electrical cutaneous stimulators to represent the luminance distribution captured by a camera on a skin area (the back, the abdomen, the forehead or the fingertip). Note

that due to technical reasons and to bandwidth limitations of tactile acuity, these devices have a rather poor spatial resolution, being generally matrices of not more than 20×20 stimulators. One interesting result from early studies was that blind subjects were generally unsuccessful in trying to identify objects placed in front of a fixed camera. It was only when the observer was allowed to actively manipulate the camera that identification became possible. Although subjects initially located the stimulation on the skin area being stimulated, with practice they started to locate objects in space (although they were still able to feel local tactile sensation). This point support the idea that the experience associated with a sensory modality is not wired into the neural hardware, but is rather a question of exercising sensorimotor skills: seeing constitutes the ability to actively modify sensory impressions in certain law-obeying ways.

There is a certain amount of studies that investigates sensory substitution phenomena in which audition is involved. One research stream investigates the use of echolocation devices to provide auditory signals to a user, depending on the direction, distance, size, and surface texture of nearby objects. Such devices have been extensively studied as prostheses for the blind As an example, Ifukube *et al.* Ifukube et al. [1991] designed an apparatus in which a frequency-modulated ultrasound signal (with carrier and modulating frequencies in a similar range as that produced by bats for echolocation) is emitted from a transmitting array with broad directional characteristics in order to detect obstacles. Reflections from obstacles are picked up by a two-channel receiver and subsequently digitally downconverted by a 50:1 factor, resulting in signals that are in the audible frequency range and can be presented binaurally through earphones. The authors evaluated the device through psychophysical experiments in order to establish whether obstacles may be perceived as localized sound images corresponding to the direction and the size of the obstacles. Results showed that the auditory feedback was succesfully used for the recognition of small obstacles, and also for discriminating between several obstacles at the same time without any virtual images.

While such devices cannot provide a truly visual experience, they nevertheless provide users with the clear impression of things being "out in front of them". In this sense, these devices can be thought as variants of the blind person's cane. Blind persons using a cane sense the external environment that is being explored through the cane, rather than the cane itself. The tactile sensations provided by the cane are "relocated" onto the environment, and the cane itself is forgotten or ignored. O'Regan and Nöe O'Regan and Noë [2001] prefer to say that sensations in themselves are situated nowhere, and that the location of a sensation is an abstraction constructed in order to account for the invariance structure of the available sensorimotor contingencies.

A related research was conducted by Meijer Meijer [1992], who developed an experimental system for the conversion of a video stream into sound patterns, and investigated possible applications of such a device as a vision substitution device for the blind. According to the image-to-sound mapping chosed by Meijer, a $N \times M$ pixel image is sampled from the video stream at a given rate, and converted into a spectrogram in which grey level of the image corresponds to partial amplitude. Therefore the device potentially conveys more detailed information than the one developed by Ifukube *et al.* Ifukube et al. [1991],since it provides a representation of the entire scene rather than simply detecting obstacles and isolated objets. In this sense, the approach followed by Mejer resembles closely the work by Bach-y-Rita, except that audition instead of tactile stimulation is used to substitute for vision.

Although from a purely mathematical standpoint the chosen image-to-sound mapping ensures the preservation of visual information to a certain extent, it is clear that perceptually such a mapping is highly abstract and *a priori* completely unintuitive. Accordingly, Meijer remarks in Meijer [1992] that the actual perception of these sound representations remains to be evaluated. However, it must also be noted that users of such devices sometimes testify that a transfer of modalities indeed takes place[1]. Again, this finding is consistent with the sensorimotor theories presented above, since the key ingredient is the possibility for the user to actively manipulate the device.

7.4 Sound modeling for multimodal interfaces

This section focuses on applications that involve direct interaction of an operator with virtual objects and environments: interactive computer animation and virtual reality applications. Musical interfaces are an interesting special case to which we devote some attention. The general topic of the use of sound in interfaces is addressed in chapter 9.

7.4.1 Interactive computer animation and VR applications

Various applications of virtual reality and teleoperation:

[1]Pat Fletcher reported her experience at the Tucson 2002 Consciousness Conference, and explicitly described herself as seeing with the visual-to-auditory substitution device. The presentation is available at http://www.seeingwithsound.com/tucson2002f.ram

Medicine; surgical simulators for medical training; manipulating micro and macro robots for mini- mally invasive surgery; remote diagnosis for telemediclne; aids for the disabled such as haptic interfaces for the blind.

Entertainment: video games and simulators that enable the user to feel and manipulate virtual solids, fluids. tools, and avatars.

Education., giving students the feel of phenomena at nano, macro, or astronomical scales; what if scenarios for non-terrestrial physics; experiencing complex data sets.

Industry: integration of haptics into CAD systems such that a designer can freely manipulate the mechanical components of an assembly in an immersive environment.

Graphic arts: virtual art exhibits, concert rooms, and museums in which the user can log in remotely to play the musical instruments, and to touch and feel the haptic attributes of the displays; individual or co-operative virtual sculpturing across the Internet.

The need for multisensory feedback

Most of the virtual environments (VEs) built to date contain visual displays, primitive haptic devices such as trackers or gloves to monitor hand position, and spatialized sound displays. To realize the full promise of VEs, accurate auditory and haptic displays are essential. Being able to hear, touch, and manipulate objects in an environment, in addition to seeing them, provides a sense of immersion in the environment that is otherwise not possible. It is quite likely that much greater immersion in a VE can be achieved by the synchronous operation of even simple haptic and auditory displays with a visual one, than by large improvements in the fidelity of the visual display alone.

Multimodal VEs that combine the visual, haptic, and auditory sensory information are essential for designing immersive virtual worlds. It is known that an individual's perceptual experience can be influenced by interactions among various sensory modalities. For example, in real environments, visual information has been shown to alter the haptic perception of object size, orientation, and shape Welch and Warren [1986]. An important implication to virtual environments is that 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. Based on these considerations, many authors (see e.g. Hahn *et al.* Hahn et al. [1998]

and Srinivasan and Basdogan Srinivasan and Basdogan [1997]) emphasize the need to make a more concerted effort to bring the three modalities together in VEs.

According to Hahn *et al.* Hahn et al. [1998], the problem of generating effective sounds in VE can be divided into three sub-problems: sound *modeling*, sound *synchronization*, and sound *rendering*. The first problem has long been studied in the field of computer music (see also Chapter 6). However, the primary consideration in VE is the effective parameterization of sound models so that the parameters being generated from the motion can be mapped and synchronized to them. Finally, sound rendering refers to the process of generating sound signals from their models within a given environment, which is in principle very much equivalent to the process of generating images from their geometric models: the sound energy being emitted needs to be traced within the environment, and the sound reaching the listener then needs to be processed to take into account the listener effects (e.g. via filtering with Head Related Transfer Functions). The whole process of rendering sounds can be seen as a rendering pipeline analogous to image rendering pipeline.

It has to be noted that until recently the primary focus for sound generation in VEs has been in spatial localization of sounds. On the contrary, research about the links between object motion/interaction and the resulting sounds has been minimal. In section 7.4.2 we will concentrate on this latter topic.

Learning the lessons from perception studies

Given the needs and the requirements addressed in the previous section, many lessons can be learned from the studies in direct (ecological) perception and in the action-perception loop that we have reviewed in the first part of this chapter.

The concept of "global array" proposed by Stoffregen Stoffregen and Bardy [2001] is a very powerful one: the global array provides information that can optimize perception and performance, and that is not available in any other form of sensory stimulation. Humans may detect informative global array patterns, and they may routinely use this information for perception and control, in both VE and daily life. According to Stoffregen and Bardy [2001], in a sense VE designers do not need to make special efforts to make the global array available to users: the global array already is available to users. Rather than attempting to create the global array, designers need to become aware of the global array that already exists, and begin to understand how multisensory displays structure the global array. The essential aspect is the

initial identification of the relevant global array parameters, which makes it possible to construct laboratory situations in which this parameter can be manipulated, and in which their perceptual salience and utility for performance in virtual environments can be evaluated.

For the specific case of auditory information, the description of sound producing events by Gaver Gaver [1993a]— provides a framework. Gaver emphasizes that, since it is often difficult to identify the acoustic information for events from acoustic analysis alone, it is useful to supplement acoustic analyses with physical analyses of the event itself. Studying the physics of sound-producing events is useful both in suggesting relevant source attributes that might be heard and in indicating the acoustic information for them. Resynthesis, then, can be driven by the resulting physical simulations of the event.

Gygi *et al.* Gygi et al. [2004] also suggest that the results reported in their work may be useful to investigators and applications developers in VEs

The effects of multimodal interactions on human perception need to be investigated in more detail through a study of normal and altered relationships among haptic, visual, and auditory displays. This will then lead to a rational basis upon which multimodal VEs can be designed and implemented.

7.4.2 Sound modeling approaches

Most sounds currently used in VE are sampled from real sounds or synthesized using "traditional" sound synthesis techniques (e.g. additive, subtractive, FM), which are based on signal theoretic parameters and tools. While deep research has established a close connection to conventional musical terms, such as pitch, timbre or loudness, the research in ecological acoustics reported in section 7.2 points out that the nature of everyday listening is rather different and that auditory perception delivers information which goes beyond attributes of musical listening.

A second problem with these approaches is that the sounds cannot be easily parameterized so that they may be correlated to motions. Parameterizing real recorded sounds by their attributes such as amplitude and pitch is not a trivial task, since it corresponds to a sort of "reverse engineering" problem where one tries to determine how the sounds were generated starting from the sounds themselves.

Finally, physically-based models allow for a high degree of interactivity, since the physical parameters of the sound models can be naturally controlled by the gestures and the actions of a

user. This remark establishes a strong link with the studies in action-perception loop described in section 7.3.2.

For all these reasons, it would be desirable to have at disposal sound modeling techniques that incorporate complex responsive acoustic behaviors and can reproduce complex invariants of primitive features: physically-based models offer a viable way to synthesize naturally behaving sounds from computational structures that respond to physical input parameters.

Physical models are widely developed in the computer music community, especially using the waveguide simulation paradigm, but their main application has been the faithful simulation of existing musical instruments. The literature on physically-based sound modeling is reviewed in chapter 6.

Contact sounds

As already remarked in section 7.2 an important class of sound events is that of *contact* sounds between solids, i.e. sounds generated when objects come in contact with each other (collision, rubbing, etc.: see also figure 7.1). Various modeling approaches have been proposed in the literature.

Van den Doel *et al.* van den Doel and Pai [1998], van den Doel et al. [2001] proposed modal synthesis Adrien [1991] as an efficient yet accurate framework for describing the acoustic properties of objects. Contact forces are used to drive the modal synthesizer, under the assumption that the sound-producing phenomena are linear, thus being representable as source-filter systems.

$$y_k(t) = \sum_{n=1}^{N} a_{nk} e^{-d_n t} \sin(2\pi f_n t) \tag{7.1}$$

The modal representation of a resonating object is naturally linked to many *ecological* dimensions of the corresponding sounds. The frequencies and the amount of excitation of the modes of a struck object depend on the *shape* and the geometry of the object. The *material* determines to a large extent the decay characteristics of the sound. The amplitudes of the frequency components depend on where the object is struck (as an example, a table struck at the edges makes a different sound than when struck at the center). The amplitude of the emitted sound is proportional to the square root of the energy of the impact.

The possibility of linking the physical model parameter to ecological dimensions of the sound has been demonstrated in Klatzky et al. [2000], already discussed in Section 7.2. In this

work, the modal representation proposed in van den Doel and Pai [1998] has been applied to the synthesis of impact sounds with material information.

An analogous modal representation of resoating objects was also adopted by Avanzini *et al.* Avanzini et al. [2003]. The main difference with the above mentioned works lies in the approach to contact force modeling. While van den Doel and coworkers adopt a feed-forward scheme in which the interacting resonators are set into oscillation with driving forces that are externally computed or recorded, the models proposed in Avanzini et al. [2003] embed direct computation of non-linear contact forces. Despite the complications that arise in the sound models, this approach provides access to (and control over) other ecological dimensions of the sound events. As an example, the impact model used in Avanzini et al. [2003], and originally proposed by Hunt and Crossley Hunt and Crossley [1975], describe the non-linear contact force as

$$f(x(t), v(t)) = \begin{cases} kx(t)^{\alpha} + \lambda x(t)^{\alpha} \cdot v(t) & x > 0, \\ 0 & x \leq 0, \end{cases} \tag{7.2}$$

where x is the interpenetration of the two colliding objects and $v = \dot{x}$. Then force parameters such as the stiffness k can be related to the perceived stiffness of the impact.

Furthermore, modeling the interaction forces explicitly improves the interactivity of the models themselves. This is particularly true for continuous contact, such as stick-slip friction Avanzini et al. [2005].

Finally, this approach allows for a natural translation of the map of everyday sounds proposed by Gaver into a hierarchical structure in which "patterned" and "compound" sounds models are built upon low-level, "basic" models of impact and friction (see 7.1). Models for bouncing, breaking, rolling, crumpling sounds are described in Rath and Fontana [2003], Rath and Rocchesso [2005].

A different phisically-based approach has been proposed by O'Brien and coworkers O'Brien et al. [2001, 2002]. Rather than making use of heuristic methods that are specific to particular objects, their approach amounts to employing finite-element simulations for generating both animated video and audio. This task is accomplished by analyzing the surface motions of objects that are animated using a deformable body simulator, and isolating vibrational components that correspond to audible frequencies. The system then determines how these surface motions will generate acoustic pressure waves in the surrounding medium and models the propagation of those waves to the listener. In this way, sounds arising from complex nonlinear phenomena can

be simulated, but the heavy computational load prevents real-time sound generation and the use of the method in interactive applications.

An important aspect in using physically-based sound models is that of synchronization with other modalities. The parameters that are needed to characterize the sounds resulting from mechanical contact (e.g. impulsive sounds due to collision), come directly from the simulation. In other cases where only simple kinematic information like trajectory is present, needed information like velocity and acceleration can be calculated. Examples of synchronization of physically-based models for audio and graphics have been given in the above referenced papers Avanzini et al. [2005], O'Brien et al. [2001, 2002], van den Doel et al. [2001].

Audio-haptic rendering

One interesting application of the contact sound models described in the previous section is in simultaneous audio-haptic rendering. There is a significant amount of literature that deals with the design and the evaluation of interfaces that involve auditory feedback in conjunction with haptic/tactile feedback.

Realistic auditory and haptic cues should be synchronized so that they appear perceptually simultaneous. They should also be perceptually similar – a rough surface would both sound and feel rough. This type of interface could improve the amount of control a user could exert on their virtual environment and also increase the overall aesthetic experience of using the interface.

Therefore there is the need for tight synchronization of the auditory mode and the haptic mode. User interaction with the simulated environment generates contact forces, these forces are rendered to the hand by a haptic force-feedback device, and to the ear as contact sounds: this is more than synchronizing two separate events. Rather than triggering a pre-recorded audio sample or tone, the audio and the haptics change together when the user applies different forces to the object.

Perceptual experiments on a platform that integrates haptic and sound displays to were reported in DiFranco et al. [1997]. Prerecorded sounds of contact between several pairs of objects were played to the user through the headphones to stimulate the auditory senses. The authors studied the influence of auditory information on the perception of object stiffness through a haptic interface. In particular, contact sounds influenced the perception of object stiffness during tapping of virtual objects through a haptic interface. These results suggest that, although the

range of object stiffnesses that can be displayed by a haptic interface is limited by the force-bandwidth of the interface, the range perceived by the subject can be effectively increased by the addition of properly designed impact sounds.

While the auditory display adopted in DiFranco et al. [1997] was rather poor (the authors used recorded sounds), a more sophisticated approach amounts to synthesize both auditory and haptic feedback using physically-based models. This approach was taken in the work of DiFilippo and Pai DiFilippo and Pai [2000]. In this work the modal synthesis techniques described in van den Doel and Pai [1998] were applied to audio-haptic rendering. Contact forces are computed at the rate of the haptic rendering routine (e.g., 1kHz), then the force signals are upsampled at the rate of the audio rendering routine (e.g., 44.1kHz) and filtered in order to remove spurious impulses at contact breaks and high frequency position jitter. The resulting audio force is used to drive the modal sound model. This architecture ensures a low latency between haptic and audio rendering (the latency is 1ms if the rate of the haptic rendering routine is 1kHz). Experimental results reported in DiFilippo and Pai [2000] suggest that a 2ms latency lies below the perceptual tolerance for detecting synchronization between auditory and haptic contact events.

We conclude this section by addressing some works that, although not specifically related with audio but rather with visual and haptic feedback, contain interesting ideas that may be applied to auditory rendering.

Lecuyer Lécuyer et al. [2000] developed interaction techniques for simulating contact without a haptic interface, but with a passive input device combined with the visual feedback of a basic computer screen. An example reported in Lécuyer et al. [2000]: let us assume that one manipulates a virtual cube in a 3D virtual environment. The cube must be inserted inside a narrow duct. As the cube penetrates the duct, its speed is reduced. Consequently, the user will instinctively increase its pressure on the ball which results in the feeding back of an increasing reaction force by the static device. The coupling between the slowing down of the object on the screen and the increasing reaction force coming from the device gives the user an "illusion" of force feedback as if a friction force between the cube and the duct was directly applied to him.

Very similar ideas have driven the work of van Mensvoort van Mensvoort [2002] who developed a cursor interface in which the cursor position is manipulated to give feedback to the user. The user has main control over the cursor movements, but the system is allowed to apply tiny displacements to the cursor position. These displacements are similar to those experienced when using force-feedback systems, but while in force-feedback systems the location of the cursor

is manipulated as a result of the force sent to the haptic display, in this case the cursor location is directly manipulated. These active cursor displacements result in interactive animations that induce haptic sensations like stickiness, stiffness, or mass.

Also in light of the remarks given in section 7.3.1, similar ideas may be experimented with auditory instead of visual feedback: audition indeed appears to be an ideal candidate modality to support *illusion of substance* in direct manipulation of virtual objects, while in many applications the visual display does not appear to be the best choice as a replacement of kinesthetic feedback. Touch and vision represent different priorities, with touch being more effective in conveying information about "intensive" properties (material, weight, texture, and so on) and vision emphasizing properties related to geometry and space (size, shape). Moreover, the auditory system tends to dominate in judgments of temporal events, and intensive properties strongly affect the temporal behavior of objects in motion, thus producing audible effects at different time scales.

Other classes of sounds

The map of everyday sounds developed by Gaver (see figure 7.1) comprises three main classes: solids, liquids, and gases. Research on sound modeling is clearly biased toward the first of these classes, while little has been done for the others.

A physically-based liquid sound synthesis methodology has been developed by van den Doel van den Doel [2004]. The fundamental mechanism for the production of liquid sounds is identified as the acoustic emission of bubbles. After reviewing the physics of vibrating bubbles as it is relevant to audio synthesis, the author has developed a sound model for isolated single bubbles and validated it with a small user study. A stochastic model for the real-time interactive synthesis of complex liquid sounds such as produced by streams, pouring water, rivers, rain, and breaking waves is based on the synthesis of single bubble sounds. It is shown in van den Doel [2004] how realistic complex high dimensional sound spaces can be synthesized in this manner.

Dobashi *et al.* Dobashi et al. [2003] have proposed a method for creating aerodynamic sounds. Examples of aerodynamic sound include sound generated by swinging swords or by wind blowing. A major source of aerodynamic sound is vortices generated in fluids such as air. The authors have proposed a method for creating sound textures for aerodynamic sound by making use of computational fluid dynamics. Next, they have developed a method using the sound textures for real-time rendering of aerodynamic sound according to the motion of objects or wind velocity.

7.4.3 A special case: musical interfaces

Bibliography

Jean-Marie Adrien. The missing link: Modal synthesis. In Giovanni De Poli, Aldo Piccialli, and Curtis Roads, editors, *Representations of Musical Signals*, pages 269–297. MIT Press, Cambridge, MA, 1991.

Federico Avanzini, Matthias Rath, Davide Rocchesso, and Laura Ottaviani. Low-level sound models: resonators, interactions, surface textures. In Davide Rocchesso and Federico Fontana, editors, *The Sounding Object*, pages 137–172. Mondo Estremo, Firenze, 2003. URL `http://www.dei.unipd.it/~avanzini/`.

Federico Avanzini, Stefania Serafin, and Davide Rocchesso. Interactive simulation of rigid body interaction with friction-induced sound generation. *IEEE Trans. Speech Audio Process.*, 13(6), Nov. 2005. URL `http://www.dei.unipd.it/~avanzini/`.

Jean-Pierre Bresciani, Marc O. Ernst, Knut Drewing, Guillaume Bouyer, Vincent Maury, and Abderrahmane Kheddar. Feeling what you hear: auditory signals can modulate tactile tap perception. *Exp. Brain Research*, In press, 2005. URL `http://www.kyb.tuebingen.mpg.de/~bresciani`.

Claudia Carello, Krista L. Anderson, and Andrew Kunkler-Peck. Perception of object length by sound. *Psychological Science*, 9(3):211–214, May 1998. URL `http://www.sp.uconn.edu/~wwwpsyc/Faculty/Carello/Carello.html`.

Claudia Carello and Michael T. Turvey. The ecological approach to perception. In Encyclopedia of cognitive science. London: Nature Publishing Group., 2002. URL `http://www.sp.uconn.edu/~wwwpsyc/Faculty/Carello/Carello.html`.

Derek DiFilippo and Dinesh K. Pai. The AHI: An audio and haptic interface for contact interactions. In *Proc. ACM Symp. on User Interface Software and Technology (UIST'00)*, San Diego, CA, Nov. 2000.

David E. DiFranco, G. Lee Beauregard, and Mandayam A. Srinivasan. The effect of auditory cues on the haptic perception of stiffness in virtual environments. Proceedings of the ASME Dynamic Systems and Control Division, (DSC-Vol.61), 1997.

Yoshinori Dobashi, Tsuyoshi Yamamoto, and Tomoyuki Nishita. Real-time rendering of aerodynamic sound using sound textures based on computational fluid dynamics. In *Proc. ACM SIGGRAPH 2003*, pages 732–740, San Diego, CA, July 2003.

Marc O. Ernst and Heinrich H. Bülthoff. Merging the senses into a robust percept. *TRENDS in Cognitive Sciences*, 8(4):162–169, Apr. 2004. URL `http://www.kyb.tuebingen.mpg.de/main/staff.php?user=marc`.

Daniel J. Freed. Auditory correlates of perceived mallet hardness for a set of recorded percussive events. *J. Acoust. Soc. Am.*, 87(1):311–322, Jan. 1990.

William W. Gaver. How do we hear in the world? explorations of ecological acoustics. *Ecological Psychology*, 5(4):285–313, 1993a. URL `www.interaction.rca.ac.uk/research/people/bill/1.html`.

William W. Gaver. What in the world do we hear? an ecological approach to auditory event perception. *Ecological Psychology*, 5(1):1–29, 1993b. URL `www.interaction.rca.ac.uk/research/people/bill/1.html`.

James J. Gibson. *The ecological approach to visual perception*. Lawrence Erlbaum Associates, Mahwah, NJ, 1986.

Steve Guest, Caroline Catmur, Donna Lloyd, and Charles Spence. Audiotactile interactions in roughness perception. *Exp. Brain Research*, 146(2):161–171, Sep. 2002.

Brian Gygi, Gary R. Kidd, and Charles S. Walson. Spectral-temporal factors in the identification of environmental sounds. *J. Acoust. Soc. Am.*, 115(3):1252–1265, Mar. 2004. URL `http://www.ebire.org/speechandhearing/`.

James K. Hahn, Hesham Fouad, Larry Gritz, and Jong Won Lee. Integrating sounds in virtual environments. *Presence: Teleoperators and Virtual Environment*, 7(1):67–77, Feb. 1998.

Kirsten Hötting and Brigitte Röder. Hearing Cheats Touch, but Less in Congenitally Blind Than in Sighted Individuals. *Psychological Science*, 15(1):60, Jan. 2004. URL http://www.bpn.uni-hamburg.de/pers/Ekirsteng.html.

K.... H. Hunt and F..... R. E. Crossley. Coefficient of restitution interpreted as damping in vibroimpact. *ASME J. Applied Mech.*, 42:440–445, June 1975.

Tohru Ifukube, Tadayuki Sasaki, and Chen Peng. A blind mobility aid modeled after echolocation of bats. *IEEE Trans. Biomedical Engineering*, 38(5):461–465, May 1991. URL http://www.human.rcast.u-tokyo.ac.jp.

Kurt A. Kaczmarek, John G. Webster, Paul Bach-y-Rita, and Willis J. Tompkins. Electrotactile and vibrotactile displays for sensory substitution systems. *IEEE Trans. Biomedical Engineering*, 38 (1):1–16, Jan. 1991. URL http://uwcreate.engr.wisc.edu.

Roberta L. Klatzky, Dinesh K. Pai, and Eric P. Krotkov. Perception of material from contact sounds. *Presence: Teleoperators and Virtual Environment*, 9(4):399–410, Aug. 2000. URL http://www.psy.cmu.edu/faculty/klatzky/.

Anatole Lécuyer, Sabine Coquillart, and Abderrahmane Kheddar. Pseudo-haptic feedback: Can isometric input devices simulate force feedback? In *IEEE Int. Conf. on Virtual Reality*, pages 83–90, New Brunswick, 2000. URL http://www.irisa.fr/siames/GENS/alecuyer/page1.html.

Susan J. Lederman. Auditory texture perception. *Perception*, 8(1):93–103, Jan. 1979. URL http://psyc.queensu.ca/faculty/lederman/lederman.html.

Susan J. Lederman, Roberta L. Klatzki, Timothy Morgan, and Cheril Hamilton. Integrating multimodal information about surface texture via a probe: Relative contribution of haptic and touch-produced sound sources. In *Proc. IEEE Symp. Haptic Interfaces for Virtual Environment and Teleoperator Systems (HAPTICS 2002)*, pages 97–104, Orlando, FL, 2002. URL http://psyc.queensu.ca/faculty/lederman/lederman.html.

Xiaofeng Li, Robert J. Logan, and Richard E. Pastore. Perception of acoustic source characteristics: Walking sounds. *J. Acoust. Soc. Am.*, 90(6):3036–3049, Dec. 1991.

Robert A. Lutfi and Eunmi L. Oh. Auditory discrimination of material changes in a struck-clamped bar. *J. Acoust. Soc. Am.*, 102(6):3647–3656, Dec. 1997. URL http://www.waisman.wisc.edu/abrl/.

Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, Dec. 1976.

Peter B.L. Meijer. An experimental system for auditory image representations. *IEEE Trans. Biomedical Engineering*, 39(2):112–121, Feb. 1992. URL `http://www.seeingwithsound.com`.

Claire F. Michaels and Claudia Carello. *Direct Perception*. Prentice-Hall, Englewood Cliffs, NJ, 1981. URL `http://www.sp.uconn.edu/~wwwpsyc/Faculty/Carello/Carello.html`.

Sharon Morein-Zamir, Salvador Soto-Faraco, and Alan Kingstone. Auditory capture of vision: examining temporal ventriloquism. *Cognitive Brain Research*, 17:154–163, 2003.

Alva Noë. *Action in perception*. MIT press, Cambridge, Mass., 2005. URL `http://ist-socrates.berkeley.edu/~noe/action.html`.

James F. O'Brien, Perry R. Cook, and Georg Essl. Synthesizing sounds from physically based motion. In *Proc. ACM SIGGRAPH 2001*, pages 529–536, Los Angeles, CA, Aug. 2001.

James F. O'Brien, Chen Shen, and Christine M. Gatchalian. Synthesizing sounds from rigid-body simulations. In *Proc. ACM SIGGRAPH 2002*, pages 175–181, San Antonio, TX, July 2002.

J. Kevin O'Regan and Alva Noë. A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24(5):883–917, 2001. URL `http://nivea.psycho.univ-paris5.fr/`.

Matthias Rath and Federico Fontana. High-level models: bouncing, breaking, rolling, crumpling, pouring. In Davide Rocchesso and Federico Fontana, editors, *The Sounding Object*, pages 173–204. Mondo Estremo, Firenze, 2003.

Matthias Rath and Davide Rocchesso. Continuous sonic feedback from a rolling ball. *IEEE Multimedia*, 12(2):60–69, Apr. 2005.

Bruno H. Repp. The sound of two hands clapping: an exploratory study. *J. Acoust. Soc. Am.*, 81 (4):1100–1109, Apr. 1987. URL `http://www.haskins.yale.edu/Haskins/STAFF/repp.html`.

Ladan Shams, Yukiyasu Kamitani, and Shinsuke Shimojo. Visual illusion induced by sound. *Cognitive Brain Research*, 14(1):147–152, June 2002. URL `http://vmpl.psych.ucla.edu/~lshams/`.

Mandayam A. Srinivasan and Cagatay Basdogan. Haptics in virtual environments: taxonomy, research status, and challenges. *Comput. & Graphics*, 21(4):393–404, July 1997. URL `http://touchlab.mit.edu/`.

Thomas A. Stoffregen. Affordances and events. *Ecological Psychology*, 12(1):1–28, Winter 2000. URL `http://education.umn.edu/kls/faculty/tas.htm`.

Thomas A. Stoffregen and Benoît G. Bardy. On specification and the senses. *Behavioral and Brain Sciences*, 24(2):195–213, Apr. 2001. URL `http://education.umn.edu/kls/faculty/tas.htm`.

Kees van den Doel. Physically-based models for liquid sounds. In *Proc. Int. Conf. Auditory Display (ICAD2004)*, Sydney, July 2004.

Kees van den Doel, Paul G. Kry, and Dinesh K. Pai. Foleyautomatic: Physically-based sound effects for interactive simulation and animation. In *Proc. ACM SIGGRAPH 2001*, pages 537–544, Los Angeles, CA, Aug. 2001.

Kees van den Doel and Dinesh K. Pai. The sounds of physical shapes. *Presence: Teleoperators and Virtual Environment*, 7(4):382–395, Aug. 1998.

Koert van Mensvoort. What you see is what you feel – exploiting the dominance of the visual over the haptic domain to simulate force-feedback with cursor displacements. In *Proc. ACM Conf. on Designing Interactive Systems (DIS2004)*, pages 345–348, London, June 2002.

Francisco Varela, Evan Thompson, and Eleanor Rosch. *The Embodied Mind*. MIT Press, Cambridge, MA, 1991.

William H. Warren and Robert R. Verbrugge. Auditory perception of breaking and bouncing events: Psychophysics. In Whitman A. Richards, editor, *Natural Computation*, pages 364–375. MIT Press, Cambridge, Mass., 1988. URL `http://www.cog.brown.edu/~warren/`.

Robert B. Welch and David H. Warren. Intersensory interactions. In Kenneth R. Boff, Lloyd Kaufman, and James P. Thomas, editors, *Handbook of Perception and Human Performance – Volume 1: Sensory processes and perception*, pages 1–36. John Wiley & Sons, New York, 1986. URL `http://pbagroup.arc.nasa.gov/welch.htm`.

Richard P. Wildes and Whitman A. Richards. Recovering material properties from sound. In Whitman A. Richards, editor, *Natural Computation*, pages 357–363. MIT Press, Cambridge, Mass., 1988. URL `http://www.cs.yorku.ca/~wildes/`.

Chapter 8

Perception and Cognition: from Cochlea to Cortex

8.1 Introduction

This chapter is devoted to auditory perception and cognition. Two aspects are addressed: on one hand the mainstream body of established results in psychoacoustics, physiology, and theory of hearing. On the other the "cutting edge" of recent progress in this field. This reflects the outlook of the S2S^{2}initiative, and its ambitious goal of bridging the gap between *sound* as it reaches our ears, and *sense* as reflects action, understanding, or pleasure.

The first section reviews established fields of knowledge in auditory perception and cognition, with emphasis on aspects of their relevance to other fields of the S2S^{2}initiative ("what we know"). The second section reviews the various approaches and methodologies ("how we know it"). The third section is oriented towards challenging aspects that the S2S^{2}initiative has vocation to address ("what we still need to find out").

The path from Sound to Sense is via perception and cognition. This direct path also participates in the "control loop" of the reverse path from Sense to Sound. Perception and cognition are thus central to the argument of this book. From a practical point of view, knowledge of auditory perception is useful in technological fields that involve sound analysis (what dimensions of the signal are relevant?) and sound synthesis (what dimensions require control?). The established body of psychoacoustic knowledge, including detection and discrimination limens, masking

properties, characterization of elementary perceptual dimensions, etc. is of use for this purpose. However, applications increasingly involve the semantic and aesthetic dimensions evoked by sound. To cater for them, we must learn to predict these higher order dimensions from the sound signal. Before this is possible, we must bridge important gaps in our understanding of how higher-level perceptual and cognitive mechanisms work.

A number of "gaps" in our understanding may be identified. A first gap separates low-level processing at the levels of the cochlea, auditory nerve and brainstem (as revealed by anatomy, physiology and psychoacoustics), and high-level processing at the level of the thalamus and cortex (as revealed by brain imaging and neuropsychological investigation techniques). A second gap is between the simple skills studied by classic psychoacoustics (loudness, pitch or timbre), and higher level cognitive skills as revealed by studies involving speech or music. A third gap is between the scientific *knowledge* we elaborate to embody our understanding of these processes, and the real-world *actions* we can derive from this understanding to perform real tasks. This is the traditional gap between fundamental research and applications.

These gaps are at the "cutting edge" of research in hearing, and it is these gaps that we set out to bridge. However this effort can be effective only if it is firmly backed by the database of knowledge that has accumulated over decades (or even centuries) of research in psychoacoustics, physiology, and model-building. This review accordingly devotes attention to all of these aspects.

8.2 Skills and functions

This section is devoted to our knowledge of auditory skills and functions. It reviews textbooks, influential papers, and useful resources. Since the invention of psychophysics by Fechner in the 19th century Boring [1942], and its rapid development the 20th century, much effort has been devoted to the humble task of charting out the properties and performance limits of the auditory system. Early efforts were based on introspection, but later efforts use more reliable experimental methods based on objective tasks.

8.2.1 Sound qualities

Traditionally, psychoacoustics has focused on the perception of sound "dimensions". Introspective studies initially led to a relatively large number of dimensions, some (such as "volume" or

"density", e.g. [Boring, 1926]) that have since faded because their perceptual reality (outside the mind of the introspector) failed to be established reliably, or because they were not really independent from other existent dimensions. Today one speaks mainly of "pitch", "loudness" and "timbre", the latter enclosing possibly multiple perceptual dimensions. In every case, it is important to distinguish *perceptual* dimensions from the *physical* dimensions that principally determine them. The essence of psychophysics is to relate these terms. Good sources are Moore [2003], Yost et al. [1993] or Moore (1995) Moore [1995] or, for signal-related topics, Hartmann [1997].

Psychoacoustics entails two difficulties: the accurate measurement of a pschological dimension, and the appropriate definition of the corresponding "physical dimension". The former requires specialized methods (see Sect. 8.3.1), while the latter involves defining some operation to perform on the stimulus in order to obtain a "predictor" of the perceptual dimension. While a simple quantity may sometimes be adequate (such as frequency of a pure tone to predict its pitch), other dimensions such as loudness or timbre may require the definition of a more complex predictor (often called *descriptor* in the context of content-based indexing). Many different "descriptors" may be found to have predictive power: the art of descriptor design is to choose the definition that leads to the simplest psychoacoustical relation.

Loudness

The concept of loudness evolved from the need to formalize the notion of perceptual "strength" of a sound, correlate of the intensity of the sound-producing phenomenon (or sound itself) that evokes it. Although our common experience is that loudness increases with sound pressure, the precise relation between signal characteristics and subjective loudness is complex. The basic psychoacoustics of loudness are well described by e.g. [Moore, 2003] or Plack and Carlyon [1995]. Studies to characterize the dependency of loudness on the spectral and temporal characteristics of sounds have led to the development of loudness predictors (also called "loudness models"). The principles behind the design of such models are well explained by Hartmann (1997). The classic model of loudness is Zwicker and Scharf [1965]. More recent models are Moore and Glasberg [1996] or Moore et al. [1997]. These models predict the overall loudness of a relatively short isolated sound, but ongoing sounds also appear to fluctuate in loudness, an aspect that is less well characterized Susini et al. [2002].

Pitch

The concept of pitch has roots that extend far back in time [de Cheveigné, 2005]. The basic pschophysics of pitch are well described by Moore [2003], Houtsma [1995], or Hartmann [1997] for signal-related aspects. A good review of the current status of knowledge on pitch is Oxenham and Plack [2005].

Since antiquity, pitch perception has focused much interest, as exemplary of auditory perception itself. [de Cheveigné, 2005] reviews pitch perception theory from a historical perspective. Current models of pitch have roots that extend back in time to the greek philosophers, via Galileo, Mersenne, du Verney, Fourier, Helmholtz and Licklider. This chapter opens with a "crash course on pitch models" that presents the operating principles of the two major approaches to pitch: pattern-matching and autocorrelation. It then reviews the roots of the idea of resonance that led first to Helmholtz's "place" theory and from there to pattern matching, and the idea of "counting" that led to Rutherford's "telephone" theory and from there to autocorrelation. A unifying concept is the vibrating string, that underlies the monocord of Pythagoras, the measurement of frequencies of musical tones by Mersenne, and Helmholtz's metaphor of the ear as a piano. The physics of the vibrating string embody ideas of both pattern-matching and autocorrelation. The chapter then reviews several hot topics in pitch (the "cutting edge" of pitch theory), and ends with a discussion of how pitch theory might evolve from now. de Cheveigné [1998] presents a model of pitch perception based on the notion of *harmonic cancellation*. The model is formally similar to the autocorrelation model of citeLicklider51 or Meddis and Hewitt [1991], multiplication being replaced by subtraction, and the search for a maximum by that of a minimum. An advantage of the new model is that it relates pitch perception to pitch-based segregation [de Cheveigné, 1993, 1997], and this allows it to be extended to account for perception of *multiple pitches* as evoked by several instruments playing together [de Cheveigné and Kawahara, 1999]. Several lines of evidence suggest that the auditory system uses the cancellation principle as embodied by these models. For example in de Cheveigné [1999], a cancellation-based model allows prediction of the subtle shifts in the pitch of a partial within a harmonic complex, as the partial is mistuned.

Much progress has been made recently in the understanding of pitch. Using a task in which listeners were required to detect changes in a simple melody, Pressnitzer et al. [2001c] determined that the lower limit of melodic pitch is approximately 30Hz. Similar limits have been found in other studies (e.g. Krumbholz et al. [2000]), but in each case there was doubt as to the musical nature of the pitch percept involved in the task. Intonation in speech or

vibrato in music use frequency modulation as an expressive means. A series of studies by L. Demany and colleagues has exhibited a strong asymetry in the perception of FM shapes. A rising-then-decreasing pitch glide, or FM peak, is much more salient than a decreasing-then-rising glide, or FM trough. A computational modeling attempt has been made to account for the asymetry [de Cheveigné, 2000b], but its neural bases are still unknown. Pressnitzer et al. [2002] recorded ensemble cortical responses to FM-peaks and troughs using functional brain imaging (magnetoencephalography). The dipole model of the data exhibited a correlate of the perceptual data in the amplitude of late evoked responses. Although speculative, a possible interpretation of the data is consistent with the temporal integration mechanisms observed in the case of the continuity illusion. Periodic sounds such as voiced speech or musical notes produce a strong sensation of pitch. Pressnitzer et al. [2001a] investigated the limits of the periodicity detection mechanisms by introducing various types of regularities in artificial click trains. By doing so, they found a paradoxical pitch shift effect that could not be accounted for by any existing pitch model. Single unit recordings showed that the pitch shift was correlated with statistics of the time-interval distributions of spike trains at the level of the ventral cochlear nucleus Pressnitzer et al. [2001b].

Timbre

Whereas qualities such as loudness or pitch may be understood as unidimensional in first approximation, the study of timbre is complicated by its inherently *multidimensional* nature. Timbre is defined as the quality that allows to judge that are "two sounds similarly presented and having the same loudness and pitch are dissimilar" ANSI [1960], a definition that potentially includes the full richness of sound (other than pitch and loudness). One might thererfore expect timbre to be, not only multidimensionsal, but of such high dimensionality as to escape any attempt at psychoacoustic characterization. An important result of the pioneering studies on timbre using multidimensional scaling (MDS) techniques is that listeners repeatably base their judgements of timbre on a small number of psychological dimensions. This is not to say that these few dimensions exhaust the richness of sound, but rather that they carry a relatively strong weight in listener's jugments. Perceptual dimensions derived from MDS can be related to signal-based measures (timbre descriptors) in what can be seen as a psychoacoustics of timbre. Progress in this field is in part the result of progress in MDS techniques.

The psychoacoustics of timbre are described e.g. in Handel [1995]. Recent progress

are described in McAdams et al. [2002, 1999], McAdams [1999, 1994], Krimphoff et al. [1994], McAdams [1992], McAdams and Cunibile [1992], Marozeau et al. [2003]. Marozeau et al. [2003] address the dependency of the timbre of a musical instrument on the note being played. Most previous multidimensional scaling studies of timbre were performed using sets of sounds from instruments all playing the same note, because of methodological difficulties related to the perceptual salience of pitch differences. Methodological difficulties were overcome in this study, that showed (a) that timbre comparisons can be made across F_0s, (b) that significant timbre changes occur with F_0, but they are neither systematic nor large, at least for the instruments studied, (c) that adjustments must be made to descriptors of timbre (such as spectral centroid) in order to allow them to predict timbre over a full range of F_0s. This latter result is important for indexing and retrieval applications.

The perception of timbre is closely related to that of the *identity* of sources such as musical instruments, or the nature of *physical processes* that give rise to sound McAdams et al. [2004]. [develop and add pointers to other chapters].

8.2.2 Scene analysis

Traditionally, psychoacoustics has considered the relation of the sound of an *isolated* source to the quality or percept it evokes. However most acoustic scenes are cluttered, and this limits the applicability of such a simple relation. Recent years have seen renewed interest in the problems raised by the perceptual organization of complex acoustic environments, particularly under the impulse of Bregman (1990). The problem of auditory scene analysis is separating the acoustic waveform that arrives at the listener's ears into the components that correspond to the different sources in the scene, assigning perceptual identity to those sources, and taking the appropriate action.

Brain mechanisms for scene analysis and object formation are among the major problems in cognitive neuroscience. The study of scene analysis has a long history in the visual field where it has been shown that different brain areas are responsible for extracting and coding different features (such as color, shape, movement) which are grouped together by pre-attentive mechanisms into objects, that serve as inputs for higher order processing. The idea that similar processes may be operating in the auditory domain is of interest. In recent years, attention has expanded from studying the encoding of simple auditory features to more complicated, high level processing that is related to how these features group together to create the listeners

impression of sound Nelken et al. [2005]. This has led to increasing use of the term "Auditory Object" Kubovy and Van Valkenburg [2001], Griffiths and Warren [2004]. A priori, it is not clear that there should be an analogy between the ways in which visual and auditory information are represented in the brain. The nature of the incoming signals is different, in particular the crucial role that time plays in the coding of auditory information, compared to visual information. However are several reasons to draw a parallel between visual and auditory scene analysis: the long known gestalt principles of organization of a visual scene (i.e similarity, good continuation, etc..) have been shown to have counterparts in the auditory domain Bregman [1990]. This might be related to the fact that, in natural conditions, visual objects are often sources of auditory events, and so are subject to the same physical organization rules. In that respect it is interesting that Cusack et al. [2000] report deficits in allocating attention to auditory objects in patients that are diagnosed with a deficit in allocating attention between visual objects. One possible implication of these findings is that there exists a representation of 'objectness' independent of modality. Such a possibilty is further strengthened by evidence of cross-modal binding, where visual events are grouped with concurrent auditory events (e.g. [Driver and Spence, 2000]).

Neurophysiological evidence for feature extraction in the auditory domain (reviewed in [Cusack and Carlyon, 2003]) has led to the speculation that, as in vision, auditory features such as pitch, timbre and loudness are extracted by separate mechanisms. The task of the system is then to combine them together to form a representation of the external object that generated them Nelken [2004]. However it is not clear whether the problem facing the auditory system should be stated in terms of a "binding problem" or in terms of a "separation problem" de Cheveigné [2000a, 2004].

A common paradigm in the study of auditory scene analysis is stream segregation. In a variety of tasks, streams have been shown to behave similarly to visual 'objects' in the sense that tones that have similar features such as pitch, timbre or amplitude modulation tend to group together and be perceived as a single object Bregman [1990]. The study of the conditions under which simple tones group together is a step towards understanding how listeners treat more complicated acoustic streams such as speech or music (how one is able to listen to a piano and a violin and listen to each instrument separately and the tune as a whole). Of particular interest is the ambiguity point, where streams are sometimes perceived as segregated and sometimes as integrated. Studying these conditions allows the research of brain mechanisms that underlie perception (since the acoustic stimulus is exactly identical in all conditions, but the percept is different).

Are streams 'auditory objects'? The term 'auditory' object has recently generated significant debate, partly because the term is used with a wide range of meanings, that are often quite poorly defined. Zatorre and Belin [2005 (to appear] make the useful distinction between "auditory sources" and "auditory patterns", the former being the carrier of the latter. This distinction allows for the separation of the *source properties*, such as a person's voice or an instruments pitch or timbre, from the temporal properties of the acoustic signals. These two properties might be processed by different systems since computing the source requires less binding across time than computing the temporal pattern, and it might be that the properties that influence the streaming (temporal aspect) are influenced by computation of the source.

Although physiological evidence for these processes is still at it's infancy, this is a path along which research in the auditory system is progressing. There has been an attempt to characterize a two-pathway "what/where" organization of the auditory system, in the same way that was described for the visual system. However so far evidence, mostly obtained with fMRI/PET is conflicting Zatorre and Belin [2005 (to appear]. Whereas brain imaging techniques, such as fMRI and PET can mostly address questions relating to which brain areas are involved in computation, non-invasive electrophysiological techniques like EEG and MEG are superior at investigating the time course of activation. Although much more work needs to be done in this area, we are beginning to understand the temporal properties of auditory stream colleagues Alain et al. [2002], Dyson and Alain [2004], in EEG studies of concurrent sound segregation, reported that the perception of a mistuned harmonic as a separate sound is associated with a negative wave, referred to as object related negativity, peaking at about 150ms after sound onset.

de Cheveigné [2000a] argues that the most difficult and important task that confronts the auditory system is that of parsing complex acoustic scenes, and that it has been shaped by evolution so as to perform that task. Ecologically relevant tasks, such as detection and localization of auditory objects in noisy environments, involve comparison of acoustic signals across ears. Interaural correlation (IAC) - the degree of similarity of the waveforms at the two ears (defined as the cross-correlation coefficient of the signals), is a basic cue for binaural processing Therefore the investigation of the neural mechanisms that are sensitive to interaural correlation is particularly informative in the study of how listeners analyze the auditory scene and react to changes in the order of the environment. An MEG study from our lab Chait et al. [2005 (submitted] has recently investigated the neural mechanisms of interaural correlation processing in cortex and compared with behavior. We demonstrate differences in location and time course of neural processing: transitions from correlated noise are processed by a distinct neural population, and with greater

speed, than transitions from uncorrelated noise. This asymmetry is reflected in behavior: listeners are faster and better at detecting transitions from correlated noise than same-sized transitions from uncorrelated noise. An 80 ms difference between the first response to transitions from correlated noise, versus from uncorrelated noise, may be the neural correlate of behavioral response time differences. These findings have interesting implications to the processes by which the central nervous system analyzes changes in the structure of the environment.

Sequential

Correlates of perceptual illusions offer an opportunity to probe the nature of neural representations. In the auditory continuity illusions, listeners report hearing a sound as continous even though it is repeatedly interrupted by bursts of noise. Establishing perceptual continuity in such circumstances is important, for instance to follow a conversation in a noisy environment. We launched a collaborative project, that has received support of the CNRS in the form of an ACI funding, to study the neural mechanisms of the illusion. Magnetoencephalography recordings combined with psychophysical measurements indicated that the illusion is not a case of filling-in: the neural activity associated with the interrupted sound was not restored in the periods of illusory perception. However, correlates of the illusion were found in a suppression of the middle-latency and late responses (evoked gamma-band response, M100). We put forward the hypothesis that these responses might be markers of the start and end of integration into a single auditory event. The auditory illusion would then act on the temporal boundaries of integration rather than on the encoding of features of the sound event Pressnitzer et al. [2004a,b].

Simultaneous

The classical view of signal detection in noise is based on the decomposition of the incoming sounds into separate frequency bands at the level of the cochlea. When noise is added to the bands that code a signal, masking occur, and more noise implies more masking. This spectral model is adequate to predict masking with stationnary signals. However, most real-world auditory scenes comprise time-varying sounds. In this case, it has been found that the temporal structure of the noise and signal can have a large effect on signal detection, which cannot be accounted for by the classical view. For instance, if some speech and a background noise have different temporal structures, a sizeable masking release can occur. Using single unit recordings in the ventral cochlear nucleus of the anaesthetized guinea-pig, we showed that an across-frequency pooling

of information by inhibitory neurons could provide a quantitative correlate of the behavioral masking release. A computational model was also proposed to reproduce the neural recordings and masking thresholds [Pressnitzer et al., 2001b, Meddis et al., 2002, Verhey et al., 2003].

8.2.3 Sound-based cognition

[TBD]

Scene analysis is a first step in "making sense" of acoustic information, but many others have been studied particularly in the case of speech, music, and more recently environmental sounds. [pointers to other chapters]

Speech

Music

Environment

In Fourier analysis, signals start at the beginning of time and go on forever. The information about how precisely they start and end is somewhat hidden in the phase of the spectral components, an information that some auditory theories discard altogether. We are nevertheless exquisitely sensitive to the shape of amplitude transients, as the noticeable difference between synthesized and real instruments indicates. Also, the transients information is preserved in hearing-impaired patients with cochlear implants (Lorenzi, Gallégo and Patterson, 1997). We found that neurons in the early auditory nucleii enhance the shape of the transients, both in the cochlear nucleus Pressnitzer et al. [2003] and in the inferior colliculus Neuert et al. [2001], to an extent that parallels behavioral performance. Cortical investigations are under way using magnetoencephalography Kult et al. [2003].

8.2.4 Ecology of sound perception

[TBD]

Perception is ultimately at the service of *action* to ensure the survival of the organism or the species. The ecological perspective is useful to gain understanding the forces that structured

the perceptual systems of our ancestors.

de Cheveigné [2005 in pressb] reviews the auditory perception of space from the perspective of its role for the survival of our ancestors. de Cheveigné [2005 in pressa] discusses the relations between the auditory perception of space and action. de Cheveigné [2000a] argues that the most difficult and important task that confronts the auditory system is that of parsing complex acoustic scenes, and that it has been shaped by evolution so as to perform that task.

8.3 Approaches and methodology

[Tools of the trade. TBD]

8.3.1 Psychoacoustics

8.3.2 Physiology

8.3.3 Brain imaging

8.3.4 Modeling

8.4 Bridging the gaps

[TBD]

Much has been learned, but there are considerable "blind spots" in our knowledge of auditory perception and cognition. These concern our knowledge of the phenomena and how they depend on stimuli, and also our understanding of the processes at work within the auditory system. Shortcomings are most obvious when we try to design systems to do tasks that seem natural to us: we then realize that we do not know how the job is done. The $S2S^2$project is an appropriate environment to map out these areas in which our understanding is seriously lacking, and to focus new efforts on bridging the gaps.

8.4.1 From sensation to cognition

Psychoacoustics deals well with elementary sensation and perception, but there is a severe upper bound on the richness of the phenomena that can be studied. Introspective and descriptive approaches such as musicology or philosophy have no such upper limit, but they lack the power to make strong statements about the generality of their results.

8.4.2 From cochlea to cortex

Our knowledge of processing within the auditory system is fed from two sources: anatomical and physiological studies of animal models, and brain imaging in humans. Much is known about response properties of the cochlea and brainstem, but beyond the level of the inferior colliculus responses are complex, diverse, and labile. Conversely, brain imaging techniques open a window on cortical responses, but are blind to subcortical events. Lacking are satisfactory *models* of how low- and high-level responses relate to one another, and especially, how they contribute to performing the tasks required by survival.

8.4.3 From model to method

Arguably the best test of knowledge is if one can *do* something with it. If a method derived from a model is succesful, we know that the processing principles that they share are effective. The translation from model to method strips away details irrelevant for the task, and this may lead to a more abstract and deeper understanding. Furthermore, once the method has been expressed in engineering terms it may be improved, and the improved methods may serve to build new models. Intercourse between model and method is a driving force for knowledge.

Bibliography

C. Alain, B.M. Schuler, and K.L. McDonald. Neural activity associated with distinguishing concurrent auditory objects. *J. Acoust. Soc. Am.*, 111:990–995, 2002.

ANSI. *American National Standards Institute (H960). USA Standard Acoustical Terminology (Including Mechanical Shock and Vibration) Sl.1-1960 (R1976).* American National Standards Institute., New York, 1960.

E.G. Boring. Auditory theory with special reference to intensity, volume and localization. *Am. J. Psych.*, 37:157–188, 1926.

E.G. Boring. *Sensation and perception in the history of experimental psychology*. Appleton-Century, New York, 1942.

A. S. Bregman. *Auditory scene analysis*. MIT Press, Harvard, MA, 1990.

M. Chait, D. Poeppel, A. de Cheveigné, and J.Z. Simon. Human auditory cortical processing of changes in interaural correlation. *J. Neurosc.*, 2005 (submitted).

R. Cusack and R.P. Carlyon. [perceptual asymmetries, p&p]. 2003.

R. Cusack, R.P. Carlyon, and I.H. Roberston. Neglect between but not within auditory objects. *Jounal of Cognitive Neuroscience*, 12:1056–1065, 2000.

A. de Cheveigné. Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing. *Journal of the Acoustical Society of America*, 93:3271–3290, 1993. URL `$<$ahref="http://www.ircam.fr/pcm/cheveign/abstracts/deCh93.html"$>$abstract$<$/a$>$`.

A. de Cheveigné. Concurrent vowel identification iii: A neural model of harmonic interference cancellation. *Journal of the Acoustical Society of America*, 101:2857–2865, 1997. URL $<$ahref="http://www.ircam.fr/pcm/cheveign/abstracts/97_concurrent_vowel_III.html"$>$abstract

A. de Cheveigné. Cancellation model of pitch perception. *Journal of the Acoustical Society of America*, 103:1261–1271, 1998. URL $<$ahref="http://www.ircam.fr/pcm/cheveign/abstracts/98_cancellation_pitch.html"$>$abstract$<

A. de Cheveigné. Pitch shifts of mistuned partials: a time-domain model. *Journal of the Acoustical Society of America*, 106:887–897, 1999. URL $<$ahref="http://www.ircam.fr/pcm/cheveign/abstracts/98_shift_model.html"$>$abstract$<$/a$>$.

A. de Cheveigné. The auditory system as a separation machine. In A. J. M. Houtsma, A. Kohlrausch, V.F. Prijs, and R. Schoonhoven, editors, *Physiological and Psychophysical Bases of Auditory Function*, pages 453–460. Shaker Publishing BV, Maastricht, The Netherlands, 2000a.

A. de Cheveigné. A model of the perceptual asymmetry between peaks and troughs of frequency modulation. *Journal of the Acoustical Society of America*, 107:2645–2656, 2000b. URL $<$ahref="http://www.ircam.fr/pcm/cheveign/abstracts/98_demany.html"$>$abstract$<$/a$>$.

A. de Cheveigné. The cancellation principle in acoustic scene analysis. In P. Divenyi, editor, *Perspectives on Speech Separation*. Kluwer, New York, 2004. URL $<$AHREF="http://www.ircam.fr/pcm/cheveign/pss/2003_montreal.pdf"$>$PDF$<$/A$>$.

A. de Cheveigné. Pitch perception models. In C. Plack and A. Oxenham, editors, *Pitch*. Springer Verlag, New York, 2005. URL $<$AHREF="http://www.ircam.fr/pcm/cheveign/pss/2004_pitch_SHAR.pdf"$>$PDF$<$/A$>$.

A. de Cheveigné. Audition, action, espace. In Thinus-Blanc and C. Bullier, J., editors, *Agir dans l'espace*. Paris, 2005 in pressa. URL $<$AHREF="http://www.ircam.fr/pcm/cheveign/pss/2004_AuditionAction"$>$PDF$<$/A$>$.

A. de Cheveigné. Espace et son. In A. Berthoz, editor, *Les espaces de l'homme*. Odile Jacob, Paris, 2005 in pressb. URL $<$AHREF="http://www.ircam.fr/pcm/cheveign/pss/2004_Espace_Son.pdf"$>$PDF$<$/A$>$.

A. de Cheveigné and H. Kawahara. Multiple period estimation and pitch perception model. *Speech Communication*, 27:175–185, 1999. URL $<$ahref="http://www.ircam.fr/pcm/cheveign/abstracts/98_multipitch.html"$>$abstract$<$/a$>$an

J. Driver and C. Spence. Multisensory perception: beyond modularity and convergence. *Current Biology*, 10:R731–R735, 2000.

B.J. Dyson and C. Alain. Representation of concurrent acoustic objects in primary auditory cortex. *J. Acoust. Soc. Am.*, 115:280–288, 2004.

T.D. Griffiths and J.D. Warren. What is an auditory object? *Nature Reviews in Neuroscience*, 5: 887–892, 2004.

S. Handel. Timbre perception and auditory object identification. In B.C.J. Moore, editor, *Hearing*, pages 425–461. Academic Press, San Diego, 1995.

W.M. Hartmann. *Signals, sound and sensation*. AIP, Woodbury, N.Y., 1997.

A.J.M. Houtsma. Pitch perception. In B.C.J. Moore, editor, *Hearing*, pages 267–295. Academic Press, London, 1995.

J. Krimphoff, S. McAdams, and S. Winsberg. Caractérisation du timbre des sons complexes. ii: Analyses acoustiques et quantification psychophysique. *Journal de Physique*, 4(C5):625–628, 1994.

K. Krumbholz, R. D. Patterson, and D. Pressnitzer. The lower limit of pitch as revealed by rate discrimination thresholds. *J. Acoust. Soc. Am.*, 108:1170–1180, 2000.

M. Kubovy and D. Van Valkenburg. Auditory and visual objects. *Cognition*, 80:97–126, 2001.

A. Kult, A. Rupp, D. Pressnitzer, M. Scherg, and S. Supek. Meg study on temporal asymmetry processing in the human auditory cortex. In *Human Brain Mapping*, page in press. 2003.

J. Marozeau, A. de Cheveigné, S. McAdams, and S. Winsberg. The dependency of timbre on fundamental frequency. *Journal of the Acoustical Society of America*, 114:2946–2957, 2003.

S. McAdams. Perception and memory of musical timbre. *International Journal of Psychology*, 27 (3-4):146(A), 1992.

S. McAdams. Big sister pitch's little brother timbre comes of age. In I. Deliège, editor, *3rd International Conference on Music Perception and Cognition*, pages 41–45, Liège, 1994. ESCOM.

S. McAdams. Perspectives on the contribution of timbre to musical structure. *Computer Music Journal*, 23(2):96–113, 1999.

S. McAdams, J. Beauchamp, and S. Meneguzzi. Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters. *Journal of the Acoustical Society of America*, 105:882–897, 1999.

S. McAdams, A. Caclin, and B. Smith. A confirmatory analysis of four acoustic correlates of timbre space. *Journal of the Acoustical Society of America*, 112:2239 (A), 2002.

S. McAdams, A. Chaigne, and V. Roussarie. The psychomechanics of simple sound sources: Material properties of impacted bars. *Journal of the Acoustical Society of America*, 115:1306–1320, 2004.

S. McAdams and J.C. Cunibile. Perception of timbral analogies. *Philosophical Transactions of the Royal Society, London, series B*, 336:383–389, 1992.

R. Meddis, R. Delahaye, L. O'Mard, C. Sumner, D. A. Fantini, I. M. Winter, and D. Pressnitzer. A model of signal processing in the cochlear nucleus: Comodulation masking release. *Acta acustica united with Acustica*, 88:387–398, 2002.

R. Meddis and M.J. Hewitt. Virtual pitch and phase sensitivity of a computer model of the auditory periphery. i: Pitch identification. *J. Acoust. Soc. Am.*, 89:2866–2882, 1991.

B.C.G. Moore and B.R. Glasberg. A revision of zwicker's loudness model. *Acustica*, 82:335–345, 1996.

B.C.J. Moore. *Hearing*. Academic Press, San Diego, 1995.

B.C.J. Moore. *An introduction to the psychology of hearing*. Academic Press, London, 2003.

B.C.J. Moore, B.R Glasberg, and T. Baer. A model for the prediction of thresholds, loudness and partial loudness. *J. Audio Eng. Soc.*, 45:224–240, 1997.

I. Nelken. Processing of complex stimuli and natural scenes in the auditory cortex. *Curr. Opin. Neurobiol.*, 14:474–480, 2004.

I. Nelken, N. Ulanovsky, L. Las, O. Bar-Yosef, M. Anderson, G. Chechik, N. Tishby, and E. Young. Transformation of stimulus representations in the ascending auditory system. In D. Pressnitzer, A. de Cheveigné, S. McAdams, and L. Collet, editors, *Auditory signal processing: physiology, psychoacoustics, and models*, pages 265–274. Springer, New York, 2005.

V. Neuert, D. Pressnitzer, R. D. Patterson, and I. M. Winter. The response of single units in the inferior colliculus of the guinea pig to damped and ramped sinusoids. *Hearing Research*, 159:36–52, 2001. URL $<$AHREF="http://www.ircam.fr/pcm/pdf/neuert-2001-damped_ramped_IC.pdf"$>$PDF$<$/A$>$.

A. Oxenham and C.J. Plack. *Pitch - Neural coding and perception*. Springer, New York, 2005.

C.J. Plack and R.P. Carlyon. Loudness perception and intensity coding. In B.C.J. Moore, editor, *Hearing*, pages 123–160. Academic press, San Diego, 1995.

D. Pressnitzer, A. de Cheveigné, and I. M. Winter. Perceptual pitch shift for sounds with similar waveform autocorrelation. *Acoustical Research Letters Online*, 3:1–6, 2001a. URL $<$AHREF="http://www.ircam.fr/pcm/pdf/pressnitzer-2002-pitch_shift_kxx.pdf"$>$PDF$<$/A$>$.

D. Pressnitzer, A. de Cheveigné, and I. M. Winter. Physiological correlates of the perceptual pitch shift of sounds with similar waveform autocorrelation. *Acoustic Research Letters Online*, 5:1–6, 2003. URL $<$AHREF="http://www.ircam.fr/pcm/pdf/pressnitzer-2003-physiology_kxx.pdf"$>$PDF$<$/A$>$.

D. Pressnitzer, L. Demany, and A. Rupp. The perception of frequency peaks and troughs: psychophysical data and functional brain imaging data. *Acta Acustica*, 2002.

D. Pressnitzer, R. Meddis, R. Delahaye, and I. M. Winter. Physiological correlates of comodulation masking release in the mammalian vcn. *J. Neuroscience*, 21:6377–6386, 2001b. URL $<$AHREF="http://www.ircam.fr/pcm/pdf/pressnitzer-2001-CMR_VCN.pdf"$>$PDF$<$/A$>$.

D. Pressnitzer, R. D. Patterson, and K. Krumbholz. The lower limit of melodic pitch. *J. Acoust. Soc. Am.*, 109:2074–2084, 2001c. URL $<$AHREF="http://www.ircam.fr/pcm/pdf/pressnitzer-2001-lower_limit_melodic_pitch.pdf"$>$PDF$<

D. Pressnitzer, R. Ragot, A. Ducorps, D. Scwartz, and S. Baillet. Is the auditory continuity illusion based on a change-detection mechanism? a meg study. *Acta Acustica*, page in press (A), 2004a.

D. Pressnitzer, J. Tardieu, R. Ragot, and S. Baillet. Mechanisms underlying the auditory continuity illusion. *Journal of the Acoustical Society of America*, page in press (A), 2004b.

P. Susini, S. McAdams, and B. Smith. Global and continuous loudness estimation of time-varying levels. *ACUSTICA - acta acustica*, 88:536–548, 2002.

J. Verhey, D. Pressnitzer, and I. M Winter. The psychophysics and physiology of comodulation masking release. *Experimental Brain Research*, 153:405–415, 2003.

W. Yost, A. N. Popper, and R.R Fay. *Human Psychophysics*. Springer, New York, 1993.

R.J. Zatorre and P. Belin. Auditory cortex processing streams: where are they and what do they do? In *[Prague proceedings]*, pages 241–254, 2005 (to appear).

E. Zwicker and B. Scharf. A model of loudness summation. *Psychological review*, 72:3–26, 1965.

Chapter 9

Sound Design and Auditory Displays

Amalia de Goetzen, Pietro Polotti, Davide Rocchesso

Università degli studi di Verona; Video, Image Processing and Sound Group

Abstract

The goal of this chapter is to define the state of the art of research in Sound Design and Auditory Display. The aim is to provide a wide overview of the extremely different fields, where these relatively new disciplines find application. These fields range from warning design and computer auditory display to Architecture and Media.

9.1 Introduction

Sounds in human-computer interfaces have always played a minor role as compared to visual and textual components. Research efforts in this segment of human-computer interaction have also been relatively little, as testified by the relatively new inclusion of Sound and Music Computing (H.5.5) as a sub-discipline of Information Interfaces and Presentation (H.5). The words sound or audio do not appear in any other specification of level-one or level-two items of the hierarchy. On the other hand, for instance, computer graphics is a level-two item on its own (I.3), and Image

Processing and Computer Vision is another level-two item (I.4).

So, the fact the scarcity of the literature, especially the lack of surveys of the field, do not come as a surprise. Indeed, a survey was published in 1994 by Hereford **?**, where a deep investigation of the state of the art of sound usage in Human-Computer Interaction was presented. The main important topics of this overview are: Earcons (symbolic and iconic), and sound in data sonification and in virtual reality environments. The literature study follows some important applications, pointing out successes and problems of the interfaces, always pushing the reader to think about lack of knowledge and need of further explorations. The paper ends with useful guidelines for the interface designer who uses sound, trying to stress the need to improve the knowledge about how people interpret auditory messages and about how sound can be used in human-computer interface to convey information about data. The knowledge about sound perception is not enough to perform good interactions, as the nature of the interface affects the creation of users' mental models of the device.

The rest of the chapter intends to go beyond Hereford's survey, in several ways. First, we consider a selection of major works that appeared in the field in the last couple of decades. This works have been either very influential for the following researches, or have appeared in respected journals thus being likely to affect wide audiences. Second, the division of the chapter into sections and subsections gives a sort of taxonomical organization of the field of Sound Design and Auditory Display.

9.2 Warnings, Alerts and Audio Feedback

Auditory warnings are perhaps the only kind of auditory displays that have be thoroughly studied and for whom solid guidelines and best design practices have been formulated. A milestone publication summarizing the multifaceted contributions to this sub-discipline is the book edited by Neville A. Stanton and Judy Edworthy **?**. This book opening chapter summarizes well the state of the art in human factors for auditory warnings as it was in the late nineties. Often warnings and alerts are designed after anecdotical evidence, and this is also the first step taken by the authors as they mention problems arising in pilot cockpits or central control rooms. Then, auditory displays are confronted against visual displays, to see how and when to use one sensory channel instead of the other. A good observation is that hearing tends to act as a natural warning

sense. It is the ears-lead-the-eyes pattern[1] that should be exploited. The authors identify four areas of applications for auditory warnings: personal devices, transport, military, and control rooms. Perhaps a fifth important area is geographic-scale alerts, as found in **?**.

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 such as those of Patterson **?**. He stated, for instance, that alarms should be set between 15 and 25 dB above the masked threshold of environment. Patterson faced also the issue of design for understanding, by suggesting a sound coding system that would allow mapping different levels of urgency.

The possibility that using naturalistic sounds may be better for retention is discussed, especially with reference to the works of Blattner and Gaver **??**. The problem of the legacy with traditional warnings is also discussed (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, 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. In the final part of the introductory chapter **?**, the authors focus on their own work on the classification of alarm-related behaviors, especially Alarm-Initiated Activities (AIA) in routine events (where ready-made responses are adequate) and critical events (where deductive reasoning is needed). In the end, designing good warnings means balancing between attention-getting quality of sound and impact on routine performance of operators.

In the same book, the chapter **?** is one of the early systematic investigations 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. Here, a methodology for designing "ecological" auditory warnings is given, and it unrolls through the phases of highlighting a reference function, finding or generating appropriate sounds, ranking the sounds for appropriateness, evaluating properties in terms of learning and confusion, mapping urgency onto sounds. A study aimed at testing the theory of auditory affordances is conducted by means of nomic (heartbeat for ECG monitor), symbolic (nursery chime for infant warmer), or metaphoric (bubbles for syringe pump) sound associations. Some results are that:

[1] http://c2.com/cgi/wiki?SonificationDesignPatterns

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.

However, for affordances that are learnt through long-time practice, performance may still be poor if an abstract sound is chosen. As a final remark for further research, the authors recommend to get the end users involved when designing new alarms. This is a call for more participatory design practices that should apply to auditory interface components in general, and not only to warnings.

If one considers a few decades of research in human-machine interfaces, the cockpit is one of the most extensively studied environments, even from an acoustic viewpoint. It is populated of alarms, speech communications, and it is reached by "natural" sounds, here intended as produced by system processes or events, such as mechanical failures. In the framework of the functional sounds of Auditory Warning Affordances, in **?** Ballas proposes five linguistic functions used to analyze the references to noise in accidents briefs: exclamation, deixis (directing attention), simile (interpretation of an unseen process), metaphor (referring to another type of sound-producing event), and onomatopoeia. To see how certain acoustic properties of the sounds affect the identification of brief sound phenomena, an acoustic analysis was performed on a set of 41 everyday sounds. A factor related to perceptual performance turned out to be the union of (i) harmonics in continuous sounds or (ii) similar spectral patterns in bursts of non-continuous sounds. This union is termed H_{st} and it describes a form of spectral/temporal entropy. The author notices that the warning design principles prescribe similar spectral patterns in repeated bursts, a property similar to H_{st}. An innovative point of this paper is that counting the pulses may give a hint for identification performance. Experimental results give some evidence that the repetition of a component improves identification, whereas the aggregation of different components impairs identification. In the last section, the chapter describes the work of F. Guyot, who investigated the relationship between cognition and perception in categorization of everyday sounds. She suggested three levels for the categorization (abstraction) process:

1. type of excitation,
2. movement producing the acoustic pattern,
3. event identification.

Her work is related with the work of Schafer **?**, Gaver **?**, and Ballas' **?** own investigations on the set of 41 sounds. In particular, Ballas' perceptual and cognitive clustering resulted in the categories:

water-related,

signalling and danger-related,

doors and modulated noises,

two or more transient components.

Finally, Ballas' chapter provides a connection with the soundscape studies of ecological acousticians.

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. In **?** the authors face the problem of a system alert for the town of Venice, periodically flooded by the so-called "acqua alta", 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 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 fundamental frequency was set between 400 and 500 Hz, which maximize the audibility at large distances. 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.

9.3 Earcons

In **?** Blattner introduced the concept of *earcons*, defining them as "non-verbal audio messages that are used in the computer/user interface 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 (Gaver 1989), where natural/everyday sounds are used in order to build auditory interfaces.

In **?**, Brewster 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. In particular, this work concerns environments like 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 from each other. The three experiments described in the abstract and presented in the paper explore different aspects of the earcons. The first one is more "abstract" and it aims at defining easily recognizable and distinguishable earcons. The second one addresses the very concrete problem of lo-fi situations, where mono signals and a limited bandwidth (a typical telephone-based scenario) is a strong limitation. In the same experiment the fundamental aspect of "musical memory" is considered: the navigation test was carried out first after the training and repeated after one week. In this latter aspect very good results were achieved: there was no significant difference between the results of the test right after the training and after one week. On the contrary, in some cases the listeners were even more skilled in recognizing the earcons one week later than immediately after the training. An interesting feedback coming from the experiments was that the listeners developed mnemonic strategies based on the identification of the earcons with something external as geometric shapes (triangles and so on). This could be a good cue for earcon sound design. The third experiment is a bit less intriguing: the idea is to identify a sound (timbre+ register) with numbers and to represent hierarchies in a book-like style (chapter, sections, subsections) by means of "sounding numbers". In general, these experiments show how problematic the design of earcons is, where many hierarchical levels are involved or where many items are present: one needs to think about complex-composed or even polyphonic earcons challenging the listening skills of the user. In any case, situations which do not present very complex navigation requirements (as in the case of TBI applications), can build upon earcons a robust and extensible method of representing hierarchies.

Another work done by Brewster[**?**] is one of the solid papers about earcons. Namely, it faces the problem of concurrent earcon presentation. Before treating such problem it gives a very good three-page survey about auditory display, sonification, auditory icons, earcons, etc. Then it gives a few ideas about auditory scene analysis and its principles, because they could be used to design more robust earcons. Two experiments are presented, which are also exemplary for their use of statistical analysis and workload measures. In the first experiment, the goal is to see how recognition of earcons and their parameters gets worse as the number of concurrent earcons is increased. In the second experiment, new design solutions are tested in their ability to increase the earcon robustness against concurrent presentation. It turns out that using multiple timbres or staggering the onsets will improve attribute identification. As a practical final result of the

experiments, four guidelines for designing robust concurrent earcons are given.

9.4 Auditory Icons

Another concept has been introduced in the nineties by Bill Gaver as an earcon counterpart: auditory icons. The basic idea is to use natural and everyday sounds to represent actions and sounds within an interface. The two papers in **??** can be considered as a foundation for later works on everyday listening: Gaver presents a fundamental aspect of our way of perceiving the surrounding environment by means of our auditory system. Trying to reply the question "what do we hear in the world?" **?** a first and most apparent result is that while a lot of research efforts were and are devoted to the study of musical perception, 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. Gaver writes the truth: 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. Traditional research on audition was and is concerned mainly with a Fourier approach, whose parameters are frequency, amplitude phase and duration. On the contrary, new research on everyday sounds focuses on the study of different features and dimensions, i.e. those concerning the sound source.
The new approach to perception is "ecological". New perceptual dimensions like size and force are introduced by Gaver. More generally, the fundamental idea is that complex perceptions are related to complex stimuli (Gaver writes about "perceptual information" too) and not to the integration of elementary sensations: "For instance, instead of specifying a particular waveform modified by some amplitude envelope, one can request the sound of an 8-inch bar of metal struck by a soft mallet". The map of everyday sounds compiled by Gaver is based on the knowledge of

how a sound source first and the environment afterwards determine the structure of an acoustical signal. "Sound provides information about an interaction of materials at a location in an environment".
Here, Gaver makes a fundamental distinction among three categories: solid, liquid and aerodynamic sounds. First, he considers sounds produced by vibrating solids. Then he analyzes the behavior of sounds produced by changes in the surface of a liquid. Finally, he takes into consideration sounds produced by aerodynamic causes. Each of these classes is divided according to the type of interaction between materials. For example, sounds generated by vibrating solids are divided in rolling, scraping, impact and deformation sounds. These classes are denoted as "basic level sound-producing events". Each of them makes the properties of different sound sources evident. At a higher level three types of complex events should be considered: those defined by a "temporal patterning" of basic events (e.g., bouncing is given by a specific temporal pattern of impacts); "compound", resulting from the overlap of different basic level events; "hybrid events", given by the interaction between different types of basic materials (i.e., solids, liquids and gasses). Each of these complex events should possess, potentially, the same sound source properties made available by the component basic events plus other properties (e.g., bouncing events may provide us information concerning the symmetry of the bouncing object). In more general terms, we can hear something that is not the size or the shape or the density of an object, but the effect of the combination of these attributes. Finally, Gaver tries to define maps based on a hierarchical organization of everyday sounds. Another interesting remark appears in the paper: what is the result of a simple question such as: "what do you ear?" If the source of a sound is identified, people tend to answer in terms of an object and a space-time context, i.e. an event and, possibly, a place in the environment. The answer concerns the perceptual attributes of the sound only if the source is not identified.

The complementary part of the paper discussed above is the one in which Gaver introduces the question "how do we hear in the world?" **?**.

While in the previous paper the relevant perceptual dimensions of the sound generation events were investigated, here the focus is on the acoustical information through which we gather information about the events. The starting point is once again the difference between the experience of sounds themselves (e.g. musical listening) and "the perception of the sound-producing events" (e.g. everyday listening). Taking into account the framework developed in the companion article, Gaver proposes a variety of algorithms that allow everyday sounds to be synthesized and controlled along some dimensions of their sources. He proposes to use the analysis and

synthesis approach to study everyday sounds: both sounds and events can be analyzed in order to reduce the data, re-synthesized and then compared to the originals. While in synthesizing sounds of traditional musical instruments the goal is to achieve a perceptually identical version of the original, the main difference here is that we just need to convey the same information about a given aspect of the event. In order to suggest relevant source attributes, the acoustical analysis of a sound event must then be supported by a physical analysis of the event itself. Gaver gives some examples of algorithms, starting from the three basic sound events, which are impact, scraping, and dripping, and concludes with three examples of temporally-complex events, which are breaking, bouncing and spilling. An example of informal physical analysis in describing complex machine sounds is given too. All these examples provide a methodology to explore acoustic information using casual listening to guide the development of the algorithms. The final discussion of the paper concerns the methodological issues that are connected to the validation of synthesis models and suggests their application to the creation of auditory icons.

This investigation is developed in another paper written in 1993 **?** which can be considered a fundamental one in the history of sound design and auditory icon definition: Gaver defines in a very clear way the goals of auditory icons as vectors of "useful information about computer events". Being a paper from the beginning of the 90s, it is not surprising that it is still very concerned about lack-of-capability and computational inefficiency of digital sampling and synthesis techniques. Some of these concerns, related to the parametrization of sounding objects, are of course still open problems, while some other issue seems to belong to the past (fortunately). According to what are the main computer events and interaction with a computer desktop, the author follows the classification described in the previous paper and analyzes different kinds of interaction-sounds as: a) Impact sounds, b) Breaking/Bouncing and Spilling sounds and c) Scraping sounds. A final section is devoted to machine sounds.

This subdivision of interaction sounds is extremely clear, ranging from a simple impact to groups of impact sounds involving temporal patterns and organization, and concluding with continuous interaction (scraping). All these classes of sounds are considered in terms of their psychoacoustic attributes, in the perspective of the definition of some physical model or spectral considerations aiming at the synthesis and parametric control of auditory icons. The fundamental statement of Gaver is that the nonlinearity of the relationship between physical parameters and perceptual results should be bypassed through a simplification of the model. The result is what he calls cartoon sounds.

9.5 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.

The chapter **?** has been important for describing the role of mediating structures between the data and the listener or, in other words, about 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, as explained in the introductory chapter of that book, 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. Audiation is the third term introduced here. As compared to audification and sonification, it had less fortune among the researchers. It is used to indicate all those cases where recall of the sonic experience (or auditory imagery) is necessary. Kramer gives a nice description of "parameter nesting", a method to codify many data dimensions (multivariate information) into sound signals. He distinguishes between loudness, pitch, and brightness nesting. Nesting is resemblant of the procedure proposed by Patterson to design auditory warnings **?**. The chapter continues discussing the advantages and drawbacks of realistic vs. abstract sounds. Then, the important issue of parameter overlap and orthogonality is discussed. When the same audible variable is used on different time scales, it is likely that a loss of clarity results. More generally, changing one sound parameter may affect another parameter. This may be advantageous for mapping related variables, otherwise it may be a problem. It is argued that orthogonality between parameters, although desirable in principle, is very difficult if not impossible to achieve. The design of a balanced display can be achieved through a combination of scaling, multiple mappings, and experiments with map sequencing and interpolation. In designing a sonification, it is important to use beacons, which are points of orientation for data analysis. The concept of beacon is also used in navigation of virtual environments, even with an acoustic sense. In this chapter, conversely, the orientation provided by beacons is not necessarily spatial. Beacons are considered as the cornerstones to build mappings, or routes from data to auditory parameters.

Examples are given in process monitoring and data analysis, where the role of emerging gestalts from multivariate auditory streams is recognized. Data are often naturally grouped in families and it is useful to preserve the relationships in the auditory display. A way to do that is by using streams, as defined and researched by Bregman **?**. Data may be mapped to streams through per stream, inter-stream, and global variables. These concepts are exemplified well by an example with different plant species. A problem with streams is that it is not possible to follow more than one at a given time, even though ensembles of streams may be perceived as gestalts or, as some other people like to say, as textures. The chapter is concluded by discussing problems related to memory, cognition, and affection. A major problem is how to recall the mappings (see also **?**). 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.

9.5.1 Direct (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, as it is described in two papers of the sixties **??**. After those exploratory works, however, no seismic audio activities had been recorded in the literature until the presentations made at early ICAD conferences and until the paper **?**. Here it is argued that audification (direct transposition of data into sound with a minimal amount of processing) makes sense in a few cases, but seismic data offer one such case because they are produced by physical phenomena (elastic waves) that are similar for propagation of earthquakes in rocks and for propagation of acoustic waves in air. So, 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. The authors gives a clear and brief introduction to the problems of seismology, distinguishing between exploration seismology and planetary seismology, and highlighting the compelling explosion discrimination problem. Then, an extensive list of possible applications of auditory display in seismology is given, including education for operators, quality control of measurements, and event recognition. 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. All these techniques are described fairly accurately in the text, with special emphasis on frequency doubling: for this effect it is recommended to derive the analytic signal via Hilbert transform and from this calculate the double-frequency signal via straightforward trigonometric formulas. The technique works well for sine waves but it is not free of artifacts for real signals. The most valuable part of the article is in the audification examples, which are given with reference to the soundfiles enclosed in the companion CD. First, synthetic data from an earth model are sonified, then field data of different types are analyzed. Listening to these examples is the best way to have an idea of the possibilities of audification of seismic data. A remark is given for annotation tones, which are necessary to help orienting the listener. These are similar to beacons in other applications of auditory display, but Hayward recommends to make them "similar to the signal or the listener will perceive two unrelated streams and it will be difficult to relate the timing between the two". This problem of the accurate temporal localization of diverse auditory events is a relevant phenomenon that should be always considered when designing auditory displays. To conclude, the contribution by Hayward has been very important to launch further studies and experimentations in seismic audification. As the author wrote, the best use of audification will be obtained when these techniques will be integrated with visual displays and given to operators for their routine work.

9.5.2 Naturalistic

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.

The chapter **?** is probably the first rigorous study that tries to compare the auditory and visual sensory channels in a complex monitoring task, where actions have to be taken in response to a

variety of configurations of system variables. The system to be monitored is the human body, and the visual display is chosen from standard practice in anesthesiology. The auditory display is designed from scratch as a hybrid of realistic and abstract sounds. The work of Gaver **?** is explicitly cited as a source of design inspiration and guidelines for realistic (everyday) auditory displays. Moreover, layers of parameters are superimposed on the auditory streams by the principles of parameter nesting **?**. No use of spatial cues in sound is made. This choice is theoretically well founded as it is supported by the evidence that, while space is the principal dimension of vision, time is the principal dimension of audition. This echoes the theory of indispensable attributes by Kubovy **?**. The task under analysis and the data here considered have a temporal structure and are inherently concurrent. The authors make sure that there is informational equivalence between the visual and the auditory display, and this is assessed in the early experimental stages by measuring the accuracy in recognizing different configurations. The experimental results show that, for this situation, users react faster with the auditory display than with the visual display. Moreover, a mixed audio-visual display does not give any advantage over the pure visual display, thus indicating a visual bias in presence of dual stimuli. The authors emphasize the emergence of gestalts from complex auditory information. In other words, users are capable to process an ensemble of audio streams as a whole and to readily identify salient configurations.

9.5.3 Abstract

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. Some researchers looked at the massive data generated by stock market exchanges to see if sonification could help enhancing the predictive capabilities of operators. The paper **?** documents a large-scale effort aimed at providing a multimodal display for the exploration of stock market data, where 3D immersive graphics is combined with 2D manipulated voice sounds. For visuals, the proposed mapping supports focus and context. For sounds, at the schema level a Bid-Ask landscape metaphor is used, where the audio echoes the changing tension between buyers and sellers. At the perceptual level, data are mapped to timbre (bids vs. asks), loudness, and pitch. With a within-subjects test, the prediction capabilities under auditory, visual, or multisensory feedback are tested. Data show that the auditory and multisensory

feedbacks perform similarly, thus indicating sensory redundancy for this specific application. Analyzing comments, it emerged that such redundancy turned into increased sense of presence and decreased workload.

9.5.4 Musical

Music has its own laws and organizing principles, but sometimes these can be bent to follow flows of data. The paper **?** is an example of the use of music for auditory display of complex time-varying information. The idea is simple: since many people use low-volume FM radio to mask the background noise of their offices, why not using a continuous musical stream that has the additional property of varying according to important system information? In this case the information comes from accesses to web servers, as these are of interest for webmasters and system administrators. Not much space is dedicated to how the mapping from data to music is actually done, even though it is reported that the authors used "a musical structure that's neutral with respect to the usual and conventional musical themes". Instead, the paper focuses on architectural aspects of the auditory display system. Three software layers collaborate in a pipeline that goes from the HTTP requests to the sound rendering. In the middle, a Collector processes the events provided by the web server and sends requests to the WebPlayer component. The information is of three different kinds: server workload, errors, details on normal behavior. Within an Apache module, information is processed according to a set of rules and directives describing which data are relevant and how they must be mapped into sounds. Support for visual peripheral display is also provided. Unfortunately, the project web address does not work anymore and no further details are given about musical mapping strategies. However, details on a user study and the musical material used are given in the previous paper **?**.

9.6 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. Carla Scaletti **?** 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."In the following sections we analyze two aspects of sonification, i.e. the definition of a

methodology for representing information by means of sound (Section 9.6.1) and sonification in an interactive context (Section 9.6.2).

9.6.1 Information Sound Spaces (ISS)

In his thesis work **?**, Stephen Barrass 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

Also, the auditory display specifications should be device-independent.

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

Syntactic and grammar-based (es. 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.

About the concept of sign in general, Barrass writes: "The concept that the sign stands for is called the "denotation" and additional signifieds are called "connotations". Cultural associations generate connotations by metonym and metaphor. A metonym invokes an idea or object by some detail or part of the whole - a picture of a horseshoe may be a metonym for a horse. A metaphor expresses the unfamiliar in terms of the familiar - a picture of a tree may be a metaphor for a genealogy."This point will be very important in the definition of Barrass' methodology for auditory display design described in the following chapters.

¿From a practical point of view, another kind of problems arises in the auditory display design: Identifiability and learnability, i.e. how much intuitive is the auditory display. Also, potential problems have to be considered, concerning masking, discriminability and conflicting mappings.

Once defined the state of the art and the opened problems, Barrass proposes different approaches for auditory display design. Among the others calls a Pragmatic Approach and a Task-oriented Approach. The Pragmatic Approach concerns design principles of warnings and alarms. A set of rules can be asserted as:

Use two stages signals 1) attention demanding 2) designation signal

Use interrupted or variable signals

Use modulated signals

Do not provoke startling

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. In this context, three main tasks can be assigned to auditory displays:

Analysis/diagnosing

Monitoring

Controlling (Create, delete, enter, edit etc.).

Here a task is denoted as a "small closed actions."From an operative point of view, a set of purposes of a task could be to confirm, to identify, to judge and to compare.

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. The ultimate goal is, thus, the user. User-centered design has become a prominent topic in Human-Computer Interaction. 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.

At this point, it becomes important to define what do we mean by means of the term information. Information is characterized by a type, a range and an organization typology. A "reading "attribute is also defined with two values: conventional (to be learned) vs. direct (intuitively understandable). The type is of four kinds: boolean, nominal, ordinal and ratio.

The organization can be in terms of category/set, time, space, mnemonic (e.g. alphabet) or continuum. It is, in other words, a case-based design from stories about listening. The stories in many professional design journals and magazines are case-studies, which convey information about a good design in a way that can be easily understood and assimilated. In HCI an intelligent borrowing of good ideas as well as reuse of a previously successful design is one of the main strategies. Barrass calls the case studies with the term "Earbenders ", from a colloquial expression

denoting short stories coming out from everyday life and worthy to listen to (to bend one's ear to).

It is then a problem of collecting and selecting good data. In order to collect "stories", Barrass proposes techniques like asking questions as: "During the next few weeks if you notice yourself using your hearing to help you, say, find a lost pin, or search for hollow spaces in a wall, or notice something wrong with your car, or tell whether the dog is hungry, anything at all, then please email me with a brief story about the occasion."

In order to define e methodology, Barrass characterizes application scenarios by means of a story and some precise keys:

A "critical" question

An answer

A characterizing subject

A characterizing sound

Possible approaches for Sound Design are: Metonymic (e.g. Gaver's Auditory Icons), Metaphoric (e.g. "stock prices and internet traffic don't normally make sounds, so are likely candidates for a metaphorical design. But what metaphor to choose?") and Pattern-based. A pattern is a regularity in the mapping between the problem and solution domains. The pattern method begins by identifying features shared by different solutions to the same problem that may capture patterns in the mapping between domains. The pattern analysis is realized by appending a description of auditory characteristics to each EarBenders case. Finally, the characteristics of sound are: Nature (everyday, musical, vocal), Level (analytic vs. holistic), Streams (in the sense of Bregmann) and others.

Examples of EarBender design by means of database of Earbenders and keywords-based searching is presented. In the first Appendix of Barrass' thesis a rich EarBenders stories database is available.

Another important part of Barrass' work 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 a usual representation of a color choosing tool by means of a circle with the hues corresponding

to different sectors as a function of the angle, the saturation levels mapped along the rays and, as a third parameter, the brightness controlled by means of a separated slider. In order to build a similar tool, representing a "sound space", Barrass analyzes many different possibilities of mapping from one of the dimensions of the color chooser to different sound parameters. In Barrass' ISS (Information Sound –Space) representation, the third dimension (the slider) becomes the height of a cylinder built on the circular pedestal of the color chooser. Barrass obtains a dimension with a categorical (not ordered) organization of the information (the sectors of the circle) a dimension with a perceptual metric (ordered) along the radial spokes and a vertical axle also with a perceptual metric as well. Then, he considers different possibilities of attribution of the parameters. Pitch, formants, static and dynamic timbers are alternatively mapped to the circle and the different mappings are tested with listening experiments. Finally a "sound chooser" is designed, where the three dimensions of the ISS are related to Timbre, Brightness and Pitch (TBP), the brightness corresponding to the radial dimension and the pitch to the height dimension. As a raw material for sound design, Barrass uses traditional musical instrument samples from the McGill University Master Samples (MUMS) collection.

Barrass designs and implement a tool for computer-aided design for auditory display that integrates the TaDa approach to auditory display design with an interface for sound manipulation. A TaDa panel allows the user to define the characteristics of the information to be represented. Consequently, a set of rules determines the kind of tool for sound manipulation (or "tailoring") to be used. For example, a spiral or the line tools allow to draw and then to move along a spiral or a straight line in the 3D representation of the ISS.

In the last chapter of his thesis, Barrass describes the design of auditory displays for different information processing scenarios. The first one is the RiverAndRain scenario about the organization of "a new sewerage treatment works to minimize the environmental impact on a river system." The second one, called PopRock concerns the assessment of the risk in digging a mineshaft. The third is the cOcktail scenario, a modelling of climate change according to the measurements of oxygen isotopes in sea-bed drill-core sites. The last one is the LostIn-Space scenario, concerning a visualization of 3D irregular structures, where the problem is to be able to navigate back to some place. The author claims that "Experiences with the multimedia interfaces that were implemented shows that the sounds can provide information that is difficult to obtain visually, and can improve the usefulness of the display."

In a following paper **?** Barrass and Kramer discuss about sonification scenarios envision-

ing applications made by nano-guitars and garage-band-bacteria revealing themselves to the stethoscope of the doctor. From this science fiction-like application, a long series of considerations about sonification starts and an overview of the subject is drawn. The work analyzes a set of already existing applications ranging from auditory displays for visually impaired people to auditory feedback for people working together. The main points raised by the authors in terms of advantages of sonification are the possibility of perceiving cycles or temporal patterns in general as well as very short events and the possibility of perceiving multidimensional data sets (the ear is polyphonic). Problems of learnability, synthesis/design skills, unpleasantness and incomprehensibility are discussed with respect to different approaches to auditory display design such as earcons, auditory icons and parameter mappings.

A work similar to Barrass' for the representation of timbre space was done by **?**. They present a tool for accessing sounds or collections of sounds using sound spatialization and context-overview visualization techniques. Audio file are mapped with symbols and colors and displayed in a 2D environment. This view is not unique: there are multiple view: starfield view, TreeMap view, HyperTree view, TouchGraph view and a file information window is always available for detailed information. In addition to the visualization components there are several interaction devices. This devices allow the filtering on the sounds. One mechanism uses sliders to control various arbitrary user classifications. These classifications are related to specific sound classifications as well as the shape and color properties of an object. Another mechanism is a simple text based filtering mechanism that uses arbitrary user classifications of the objects. The system seem to improve performance in browsing audio files, but sometimes lead to ambiguity as users' classifications of sounds may differ. Hyperbolic layout (HyperTree) for browsing makes it both easy and enjoyable. The TreeMap view presented poor results and visualization of the data.

9.6.2 Interactive Sonification

In **?** the authors propose a new approach to data sonification, starting with a deep investigation on the link between sound and meaning. The idea is to find a way to use data sonification without using musical listening, e.g. without the need of a training time. The Model Based Sonification that the authors propose provides a natural mean for interacting with a sonification system and allows the development of auditory displays for arbitrary data sets. Both results are achieved using a virtual object in the interaction and a parameterized sound model as auditory display.

The authors argue that the MBS has many advantages: fewer parameters to be tuned, a natural connection between sound and data, a softer learning slope, an intuitive interface, a continuous natural control etc. The MBS has been illustrated using an example of particle trajectories with a prototype of a tangible physical representation-interface. There is no real evaluation of the sonification method and of the interaction.
- Continuous sonic feedback from a rolling ball
The paper written by Rocchesso and Rath **?** explains how continuous sonic feedback made by physical models can be used in Human-Computer Interaction. 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 nature behavior: trigger sounds are quite unusual, while we are always using continuous sounds to be informed 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 to perform the test that has been analyzed later in the paper: this sound is particularly informative, conveying information about direction, velocity, shape and surface textures of the contacting objects.
The rolling model is realized in a hybrid architecture with higher level structures: the lowest level is the physics-based impact model, while the higher one is the perception-oriented structure through a connecting signal-processing algorithm. A rolling filter is used to reduce the dimensions of the impact model simply to the perpendicular to the global direction of the surface.
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".
Considering the higher level modelling, the authors pointed out the importance of macroscopic characteristic renderings, such as the periodic patterns of timbre and intensity which are featured by a rolling sound: the rolling frequency is very important for the perception of size and speed.
The main idea here is to take into account rolling objects that do not show perfect circular symmetry: the height of the center of mass will vary during the movement performing macroscopic asymmetries that lead to periodic modulations of the effective gravity force. The use of this approach has many advantages that can be summarized as follows:

1. the synthesized sound is always new and repetition–free
2. there is no need to store large amounts of sound samples

3. all the ecological attributes can be varied on the fly allowing a continuous real time interaction with the model

The last part of the paper describes the test which has been done to verify the naturalness and effectiveness of interaction. The balancer metaphor has been used: the subjects are asked to balance a virtual ball on a tiltable track, with and without a video feedback (the target areas were realized in four different sizes). The test results are very interesting:

1. the sound of the virtual rolling ball is easier to recognize than the sound of a real rolling ball; users can describe better the physical characteristics (e.g. the size) of the virtual ball than the real one
2. subjects intuitively understood the modelled metaphor without a learning phase
3. the cartoonification approach appears to be effective in such a control metaphor
4. the performance measurements show that there is an improvement from 9% for bigger displays to 60% for smaller displays
5. all subjects solved the task using auditory feedback only

This investigation suggests to the authors that continuous feedback can be used for sensory substitution of haptic or visual feedback and equilibrium tasks can be seen as possible exploitation areas, beside the more obvious one: video games and virtual environments.

- Interactive simulation of rigid body interaction with friction-induced sound generation

Another example of interactive sonification is done in **?**: the paper starts with the description of a complete physical model of the complex mechanics of friction, taking into account numerical methods to make the model running real time on low cost platforms. The main idea of the friction model is based on a "bristle-based" interpretation of friction contact which is made by a number of asperities (e.g. microscopic irregularities) of two facing surfaces. The LuGre friction model and its development made by Dupon are analyzed in order to be improved: a parameter has been added in the definition of the friction force in order to simulate scraping and sliding effects other than the stick-slip phenomena. The model is divided in two main parts: the excitation and the resonators of the vibrating system. The resonating objects are modelled according to the modal

synthesis approach as lumped mechanical systems. In order to use the model in interactive settings (e.g. real time) a numerical implementation is discussed. A decomposition of the system into a linear differential system coupled to a memory-less non-linear map is done in order to apply efficient numerical methods to the model (e.g. K method). The final part of the article is the discussion of a number of applications of the model to the simulation of many everyday friction phenomena: all these examples explain how physical models can be used in a multimodal context, such as sound modelling for computer animation. The model has been implemented as a plugin to pd (Pure Data) and then used in several examples of acoustic systems with induced frictional vibrations. The animations presented in the paper are the following: braking effects, wineglass rubbing and door squeaks. All these examples show a high degree of interactivity of the model: the user can control one of the virtual objects in the animation through a simple pointing device (e.g. a mouse), controlling at the same time some of the physical parameters involved in the friction (e.g. the force acting on the exciter). The positions and velocities which are returned by the synthesis engine can be used to drive both the graphic rendering and the audio feedback.

9.7 Sound Design

9.7.1 Sound Objects

In **?**, Michel Chion tries to define an omni-comprehensive but synthetic review of the thinking of Pierre Schaeffer. The book is a considerable effort aiming at making the intuitions and concepts developed by Schaeffer systematic. The main thesis of the book are the definition of Acousmatic music, the definition of Reduced Listening, the definition of Concrete (vs. Abstract) Music, the definition of Sound Object and the definition of a new Solfège for the development of a new music. All these concepts provide a summa of the theoretical work of Schaeffer **?**. Acousmatic comes from the ancient Greek and means a sound that we hear, without seeing the source. Acousmatic is here meant as opposed to Direct Listening. The acousmatic situation corresponds to an inversion of the normal way of listening: it is not any more a question of studying how a subjective listening deforms the reality but the listening itself becomes the phenomenon to study. Two listening experiences due to the use of the tape recorder are mentioned as fundamental for the evolution of the concept of acousmatic music: the looped tape and the cut of a sound of a

bell (a sound without the attack). These kind of experiments allow us to become aware of our perceptive activity. This awareness is also called Epoché and is directly related to the second main concept defined by Schaeffer: the Reduced Listening. This new way of listening is thought as opposed to what he calls trivial listening (a listening that goes directly to the causality of sound events), to the pragmatic listening (a gallop can have the meaning of a danger that is possibly coming or be just a rhythmical event) and cultural listening (that looks for a meaning). In other words, the Reduced Listening places out of our consideration anything related in a more or less direct way to a sound (sources, meanings, etc.) and considers only the sound itself. The sound itself becomes an object on its own. In order to define the concept of the Sound Object, Schaeffer adopts a negative approach. A sound object is nor the body of the sound source (sounding object), neither a physical signal. The sound object is nor a recorded sound neither a symbol on a score. Also, the sound object is not a state of our spirit. These negative specifications delimit what in positive could be defined as "the sound itself", a definition that could be vague. By means of the manipulation of sound objects it becomes possible to build a new kind of music: the Concrete Music. Classical music starts from an abstract notation and the musical performance come afterwards. Conversely, the new music starts from the concrete phenomenon of sound and tries to extract musical values from it. In other words,recalling the difference between phonetics and phonology, the whole work of Schaeffer can be considered as a path from Acoustics to what he defines as "Acoulogy", i.e. from a gross sound, conceived as an acoustic object, to a sound that is analyzed and considered in a musical sense. The development of a complete method for the analysis of sounds and the synthesis of new sounds as well as the production of a new music is the ultimate goal of the work of Schaeffer. The plan for this new methodology forms the "new Solfège" and is articulated into 5 steps: typology, morphology, characterology, analysis and synthesis. Schaeffer developed only the first two steps, while the other three were only planned. More into detail, Typology performs a first approximate sorting (a kind of elementary morphology). Morphology describes and qualifies sounds and defines classes. Characterology realizes a sort of taxonomy, a kind of new Lutherie. Analysis defines musical structures by means of the perceptual fields. Finally, Synthesis forms the innovative Lutherie, i.e. a way for the creation of new sound objects according to the results of the analysis. A noteworthy remark is that in a preliminary definition of what he calls Analysis, Schaeffer defines the analytical tools as Natural Perceptual Field, i.e. pitch, duration and intensity. These criteria are in a sense "natural", even if, for what concerns pitch, it seems to be quite a traditional choice.

The typology-morphology task is developed quite into detail. The whole process is defined

in terms of identification and qualification. In other words the main tasks are those of isolating an object from the context and then describing its properties. The crucial point in this activity is the retrieval of the character, i.e. the structural elements (e.g. the timbre of a sound) with respect to the values (or version) of a particular sound, i.e. the variable elements that do not contradict the character (e.g. the three perceptual fields). The three tasks of the typo-morphology are: identifying (typology), classifying (typology) and describing (morphology) within the perspective of a reduced listening, i.e. independently from any reference to causes/origins of the sounds or to what they could evoke).

In the context of a new Solfège, the equivalent of a musical dictation becomes the task of recognizing and defining the version of a sound object and the art of improving the listening skills.

One of the main points in order to achieve an identification, is the definition of some segmentation criteria, able to isolate the single sound objects. This is not evident, since one does not want to use either musical criteria or natural systems of identification (source detection). The chosen units correspond to syllables, i.e. units that are negligible in a linguistic context. The distinction between articulation and prolongation, i.e. the identification of the breaking of the sonic-continuum in subsequent and distinct elements (consonant) and sounds with a structure that maintains its characteristics over time (vowel) is the way pointed out by the author.
Finally, the author defines a set of descriptors that have some generality, even if they are possibly not completely exhaustive as the author claims they are.
The typology classification is based on the distinction between:

- Impulses,
- Iterations (sequences of impulses),
- Tonics (voiced sounds),
- Complex sounds (fixed mass but no pitch),
- Complex and variable sounds.

The classification of sounds is based upon the morphology principles, which subdivide sounds according to:

- Matter criteria (Mass, Harmonic timbre, Grain)
- Form criteria (Allure - a kind of generalized vibrato- and Dynamics)
- Variation criteria (Melodic profile and Mass profile)

Some recapitulation summary tables are reported at the end of the book, providing an analytical grid for sound analysis
Murray Schafer **?** as well talks about reduced listening by introducing the concepts of schizophonia and of sound looping, both related to the studies of Pierre Schaeffer. Schizophonia points out the new listening scenario introduced by the recoding supports and the reproduction by means of loudspeakers, where the sound sources disappear from our visual feedback and are thus separated (Schizo) from the sounds (Phonia). We do not see any source any more and the sounds become objects on their own. This is what he calls the effect on audio of the electric revolution.

9.7.2 Sounding Objects

In **?** three partners of the European Project SOb (Sounding Object: www.soundobject.org) draw their conclusions about the results of a three year-long research project from a high level perspective. Their approach is somehow complementary with respect to the concept of Sound Objects. Instead of analyzing and medelling sound by itself,the idea is to model the source in terms of its physical behavior. Perception analysis, cartoonification/simplification and control of physically meaningful parameters were the three main guidelines of the work. The term cartoonification refers to the cartoons and to the technique of reducing the complexity of the audio and visual information to its essential elements. These reduced elements are then emphasized and clarified as much as possible in order to provide a cartoonified version of the reality. The great advantage of a cartoonified version of reality is its augmented intelligibility. This in terms of display purposes is extremely important.

After pointing out the lack of a comprehensive study on everyday sounds and the importance of these sounds in the context of auditory display design, the paper presents some examples in terms both of sound design techniques and psychology of perception. First a psychology experiment was conducted in order to explore the perception of sounds of filling/emptying bottles. The principle of cartoonification was illustrated by the example of a cartoon mouse drinking

with a straw. The sound is designed according to a precise decomposition of the physical events occurring in the action. Each one of these events is then treated separately for what concerns sound modelization
Another important aspect outlined by the authors is the temporal organization of the sounds as in the case o a bouncing ball, where many similar events take place in a certain recognizable sequence. If, from one side, one principle is the simplification of the physical model aiming also at achieving a higher intelligibility, on the other side the possibility of using the complexity of human gesture in order to control the parameters of the model is crucial in order to obtain a natural sound object. As a result of their experience, the authors propose a set of principles for sound design. According to these principles, the main identity of sounds can be defined and reproduced by means of a set of basic physical interaction models reproducing the sound sources. The quality of these sounds can be then refined and enhanced by means of signal processing techniques. Finally a spatial and temporal organization of sound objects is of extreme importance in order to enlarge the vocabulary and information contents that one wants to convey by means of sound.

9.7.3 Cartoon Sounds

The potentialities of cartoon sounds were deeply addressed by William Gaver . In his work Gaver investigates a fundamental aspect of our way of perceiving the surrounding environment by means of our auditory system **??**. A lot of research efforts were and are devoted to the study of musical perception. Nevertheless 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 says: "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 town 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 filed is scarce. It is true, as Gaver says, that we do not really know how we are able to gather so much information from a situation as the one of the approaching car described before.

Traditional research on audition was and is concerned mainly with a Fourier approach, whose parameters are frequency, amplitude phase and duration. On the contrary, new research on everyday sounds focuses on the study of different features and dimensions, i.e. those concerning the sound source. The new approach to perception is "ecological". New perceptual dimensions as size and force are introduced by Gaver. More generally the fundamental idea is that complex perceptions are related to complex stimuli (Gaver talks also about "perceptual information") and not on the integration of elemental sensations: "For instance, instead of specifying a particular waveform modified by some amplitude envelope, one can request the sound of an 8-inch bar of metal struck by a soft mallet". The map of everyday sounds compiled by Gaver is based on both the knowledge about how a sound source first and the environment afterwards determine the structure of an acoustical signal. "Sound provides information about an interaction of materials at a location in an environment".

At this point, Gaver makes a fundamental distinction between three categories: solid, liquid and aerodynamic sounds. First he considers sounds produced by vibrating solids. Then he analyzes the behavior of sounds produced by changes in the surface of a liquid. Finally he takes into consideration sounds produced by aerodynamic causes. Each of these classes is divided according to the type of interaction between materials. For example, sounds generated by vibrating solids are divided in rolling, scraping, impact and deformation sounds. These classes are denoted as "basic level sound-producing events". Each of them makes the properties of different sound sources evident.

At a higher level one has to consider three types of complex events: those defined by a "temporal patterning" of basic events (e.g., bouncing is given by a specific temporal pattern of impacts); "compound", given by the overlap of different basic level events; "hybrid events", given by the interaction between different types of basic materials (i.e., solids, liquids and gasses). Each of these complex events should potentially yield the same sound source properties, made available by the component basic events but also other properties (e.g., bouncing events may provide us information concerning the material but also the symmetry of the bouncing object). More in general, we can hear something that is not the size or the shape or the density of an object, but the effect of the combination of these attributes.

Finally Gaver tries to define map based on a hierarchical organization of everyday sounds In conclusion, an interesting remark about everyday listening is: what is the result of a simple question as: "what do you ear?" If the source of a sound is identified, people answer in terms of an object and a space-time context, i.e. an event and, possibly, a place in some environment. Only if the source is not identified, then the answer concerns the perceptual attributes of the sound.

9.7.4 Soundscape

The word soundscape was born as a counterpart of landscape, denoting the discipline that studies sound in its environmental context, both naturalistic and urbanistic. This discipline grew up first in Canada, then in other countries as the Scandinavian ones, Australia and others. A milestone publication on this subject was written by Murray Schafer **?**. Entitled Soundscape, Schafer's book is a long trip through a novel conception of sound. One of the main goal is the recovering of a clear hearing (claireaudience), and of a hi-fi (high fidelity) soundscape as opposed to the lo-fi (low fidelity) soundscape of our nowadays world. A subjective and an objective perspectives that are intimately related. One of the main thesis of the book is, in fact, that the soundscape is not an accidental by-product of a society, but, on the contrary, it is a construction, a more or less unconscious "composition" that can be of high or low quality. Evaluation criteria are widely investigated by Schafer, leading to a basic classification of soundscapes into two categories: The already mentioned hi-fi and lo-fi scenarios that will be described more in detail later on. This forms the objective platform. On the other side, our subjective effort should be to pursuit a more and more refined ear cleaning process, in order to become able to hear, evaluate and interact-with sounds of the surrounding environments. Hearing is an intimate sense similar to touch: The acoustic waves touch our hearing apparatus. Also, the ears do not have lids. It is thus a delicate and extremely important task to take care of the sounds that forms the soundscape of our daily life. Schafer even says that a lo-fi, confused and chaotic soundscape is an indicator of decadency of a society.

The book is divided into three parts. In the first one the author considers the natural elements (air, water, earth), the animals, the different geographical landscapes and, in a historical perspective, sounds through the centuries with a particular attention to the effect of the industrial and the electric revolutions. In the second part he analyzes the sounds in its sonic and semantic contents. In the third part Schafer moves towards the definition of an Aesthetics of Acoustic Design.

In the beginning of the book the definition of the main elements of a soundscape are given: keynotes, signals and sound prints. Keynotes are sounds related to geography, belonging

somehow to our unconscious background and that are in the background. On the contrary a signal is anything that we listen to and that conveys some information about the surrounding environment. Finally a sound print is again something that belongs to the background, but, in this case, it is a product of the human life and society. Sound prints are the main concern of Schafer's investigation.

In the historical sections, the analysis is always related to sociological aspects. The rural soundscape, the industrial "sound revolution" and the consequences of the electric revolution are analyzed in depth. In the pre-industrial society the SPL in a rural village was never above 40 dB, except when the bells or the organ played or a mill was working. On the other side reports from the past tell us that the sound in the big towns of the pre-industrial era were unbearable. Nevertheless these sounds were variegated and their dynamics was spike-like and always changing. The main sound sources were people screaming (especially hawkers, street musicians and beggars), hand-worker activities, horses and other animals. As opposed to this kind of sounscape, the industrial revolution introduces continuous, not-evolving and repeating sounds. This is one of the main characteristic of a lo-fi soundscape. Schafer says that a spike-like and varying amplitude envelope was substituted by a continuous, linear amplitude envelope, which fixes a stable persistent, and unnatural (psychologically disturbing) sound dynamic evolution.

Besides the industrial revolution, the electric revolution plays a particularly relevant role in the first part of the book: The electricity allows recording, reproducing and amplifying the sound. How this influenced the soundscape is also matter of discussion in the book. In particular the concepts of schizophonia and of looping a sound are investigated. Both of them are related to the studies of Pierre Schaeffer and the definition of Concrete Music. Schizophonia points out the new listening scenario introduced by the recording supports and the reproduction by means of loudspeakers, where the sound sources disappear from our visual scene and are thus separated (Schizo) from the sounds (Phonia). We do not see any performance any more and the sounds become objects on their own.

Another crucial point of Schafer's analysis is the diffusion and prevailing in modern (pop) music of bass sounds with respect to mid-range frequency. All of these aspects deteriorate the quality of the "soundscape": the sound-to-noise ratio increases and we pass from a hi-fi soudscape to a lo-fi soundscape. The physical-symbolic meaning of low frequencies is quite clear:

Basses propagate farther and longer in time than high frequencies. Due to diffraction phenomena they pass obstacles. Also, it is difficult to localize a low frequency. The global effect is that of a sense of immersiveness that cannot be achieved by other musical traditions. As an interesting exception, Schafer observes that a very similar immersive effect was characteristic of a completely different scenario as that of the ancient romanic and gothic churches, when a choir was singing. In this case, it was the reverberation that created the effect of prolongation, diffraction and delocalization typical of bass sounds.

In the second part of his book, Schafer illustrates a proposal of notation for sounds. First, he criticizes the ordinary representation, typically based on the spectrogram. He points out how this kind of representation misleads the attention from the auditory channel to the visual one. He suggests not to consider seriously an evaluation of a sound based on some diagram: "if you don't hear it, don't trust it". Then, he defines a sort of taxonomy of sounds, a set of parameters relevant for the characterization of a sound timbre, and a symbolic notation representing these parameters. This part ends with an extensive inquiry about how people in different cultural and geographical contexts consider more or less annoying different categories of sound. An extremely interesting collection of data emerges from such an analysis. In general, both for the production of sounds as for the reaction that they generate, a sociological approach is always considered.

In the last part of the book, Schafer moves towards the definition of an Acoustic Design practice. According to the author, the tasks of a sound designer should be: the preservation of sound prints, especially those that are going to disappear, and the definition and development of strategies for improving a soundscape. The principles that a sound designer should follow are:

- To respect for the ear and voice, i.e. the SPL of the soundscape has to be such that human voices are clearly audible.
- To be aware of the symbolic contents of sounds.
- To know the rhythms and tempi of the natural soundscape.
- To understand the balancing mechanism by which an eccentric soundscape may be turned back to a balanced condition.

An interesting example of sound design for a urban context is given by the work of Karmen Franinovic and Yon Visell (http://www.zero-th.org/) with the sound installation Recycled

Soundscape- Sonic Diversion in the City. The authors aim at stimulating people attention and awareness about the soundscape of their town. The installation is formed by a mobile recording cube (approximately 1 meter high), supporting a microphone placed in front of a parabolic surface so that it created a sharp directional microphone for recording sounds from the environment. The installation was completed by two mobile cubes, which played back the recorded sound after a significant "recycling" processing. The idea is to get people involved in playful interactions in the urban setting and to make them sensitive about the cacophony that surrounds us in the city soundscape and let them try to put some order in the city. Also, "..these recordings of the context, made by previous and current users of the system.. are woven into the remixed soundscape", i.e. the other idea of the installation Recycled Soundscape. Their work demonstrates how the human behavior associated to a place, and our perception of it, may be disrupted by public diversions that offer the possibility for the extension of perception and for collaborative creation.

9.7.5 Space and Architecture

We could start this section with a point outlined in **?**, i.e. the lack of awareness about sound in the architectural context. The main concern of architects is usually how to eliminate sound. No attention is devoted to active intervention in architectonic projects.
In this section we will discuss about sound and space from two different point of view: the relation between sound and architecture, meant as sonification of the environment, from one side, and the technical aspects related to the acoustic space rendering.

A very first and well-known example of an architectural project that gave extreme importance to sound was the Philips Pavilion at the Expo in Bruxelles in 1958. In that case Le Corbusier and Xenakis built a structure, where the sound played a fundamental role: the work was a organic project where space and sound were conceived together in order to be experienced together as an unicum. The Philips Pavilion was more than a building at the fair, it was a multimedia experience displaying the technological prowess of the Philips company by combining light, sound, and color. The music was composed by Edgar Varèse and entitled "Poème électronique." The radical concept behind this first experience and any other project involving sound and architecture, is that sound modifies the perception of space. We could say that space is in sound or that (a different) space comes out, when sound rings. In this sense a new discipline as Electroacoustic Soundscape Design, i.e. the electroacoustic sonification of the environments

and buildings, takes a relevant place in the frame of sound design.

For what concerns the technical aspects, **?** provides a good and concise overview of the Real Time spatial processing techniques available up to 1999 for room simulation with application to multimedia and interactive HCI. He goes through the models of a) directional encoding and rendering over loudspeakers, including conventional recording and ambisonic B format b) binaural processing and c) artificial reverberation, with an extension to the dynamic case for acoustic-source-distance rendering (Chowning's model) and to Moore's ray-tracing method. In this overview the advantages and weak points of each approach, for instance the limit of the "sweet spot" for the methods of point a). He states the criteria for establishing perceptually-based spatial sound processing: Tunability, Configurability and, last but not least Computational efficiency and scalability. In particular the first criterion includes the definition of source direction (azimuth and elevation) and descriptor of the room. Configurability implies the possibility of changing output format (headphones vs. different loudspeaker configurations). Finally the author presents SPAT, the spatialization tool realized by IRCAM. The great novelty at the time for such a tool was the high-level interface. SPAT does not present to the user physical parameters but only perceptual parameters classified as: Source perception (source presence, brilliance and warmth) b) Source/room interaction (room presence). c) Room perception (heaviness and liveliness). These parameters are chosen according to the studies of psychoacoustic done at IRCAM, specifically for the perceptual characterization of room acoustic quality. A series of application scenarios is then analyzed, ranging from VR and multimedia to live performance and architectural acoustics.

9.7.6 Media

In the book entitled "L'audio-vision. Son et image au cinéma", **?** two main thesis appears from the very beginning of the book: a) Audiovision is a further dimension: different from bare vision and different from bare audio b) Between images and sound (music in particular) there is no necessary relationships. As an example of these thesis, he starts with the analysis of the movie Persona by Ingmar Bergman and of "Les Vacances de Monsieur Hulot" by Jacques Tati. Chion demonstrates that the bare vision of a sequence of mute images is something completely different from the audiovision of the same sequence (as an example, the prologue of Persona is played first without and then with the soundtrack). Also, a contrasting audio and video situation are

not only possible but sometimes extremely effective from many point of views (e.g. the scene on the beach in "Les Vacances de Monsieur Hulot", where a group of annoyed people are "superimposed" on a joyful hilarious soundscape of children playing and screaming. The whole scene is of a great comic effect). A second major point in his book is the evidence that our listening is first of all vococentric, i.e. our principal source of information is the voice and the words of other human beings. The text is, thus, a predominant element in the cinema. On the other side, there are other kinds of listening beside the semantic (human voice) one: A causal listening and a reduced listening. The first one is related to the class of sounds that involves a question of the kind: "What is the source of this sound?". The reduced listening is somehow the opposite: It occurs, when we listen to the qualities of a sound by themselves independently from its cause and meaning. This is related to the idea of acousmatic listening that will be discussed more in detail later on (see also 9.7.1).
The author considers then the different roles of sound in a movie: Sound has the function of a temporal (overlapping effect) of isolated events. Also, it functions as a spatial connection: the acoustical unity of the environment of the scene (reverberation). Concerning the "ground" level of sound, i.e. silence, Chion quotes Bresson: It was the synchronism between sound and image that introduced silence. Silence as pauses in between sound events. The mute cinema, on the contrary, is a continuous suggestion of sound. On the other side, there are sounds in the cinema used as a metaphor of silence: Animals that cry in the far, clocks in the neighbor apartment, any very soft but present (near) noise.
An interesting, possibly central matter of debate is given by the off-field sounds. An articulated classification follows the analysis of all of the possible situations. Altogether with the "off sound" (the sound that does not belong to the time and the space of the scene) and the "in-sounds" (the sounds, whose sources appears in the scene), they form the so-called tri-circle. The off-field sounds can take different attributes: acousmatic, objective/subjective, past/present/ future, giving raise to a wide scope of expressive possibilities. Also, off-field sounds can be of different kind, trash (e.g. explosions, catastrophes noises), active (acousmatic sound that provokes questions as "what is it?" or "where is it?") and passive (they create an environment that involves the image and give a feeling of stability).
Another dimension of a sound is its extension: null extension corresponds to an internal voice, while a wide extension: corresponds, for example, to a situation, where the traffic sounds from the near street are audible and amplified in an unrealistic way. Expressive effects can be obtained by playing with variations of the extension within a scene (e.g. alternatively from outdoor to indoor and to the internal dimension).

A further element is the definition of the listening point. One possibility is to refer to a spatial perspective: from which point in the space we listen? But a subjective point of view is possible too: which character is listening? A particular case is given by weak sounds, which give the impression to be heard only by a character, as if the sounds were near to the ears. In all the previous examples, the unrealistic element is essential. In general, there is no reason for the audiovision relationships (between images and sounds) should be the same as in real-life. A stylization strategy of representation is possible and can open wide and various horizons for expressivity. This is also related to the discussion of Section 9.7.3.

In the followings of his book, Chion points out the difference between definition and fidelity: definition is a technical term denoting the range of reproduction/rendering possibilities of the system. On the contrary, fidelity is a dangerous term: fidelity evokes realistic reproduction that is a very debatable concept: cinema is also a metaphoric representation, involving augmentation, unrealistic points of view, distorsion of time and space, of soundscape and landscape. Sound should be veridical, not realistic. The goal of a veridical sound is to render the associated sensations, not to reproduce the sound realistically. A realistic sound, if detached from the image is often not comprehensible, a deception. Sound technicians are skeptic about the possibility of recognizing sources from sounds. Chion quotes a couple of examples: The same sound can be used in relationship with a smashed head in a war movie or a squeezed water-melon in a comic movie. The same gargling sound can be used for a tortured Russian prince (Andrej Rublov by Tarkowskj) and for the gurgling of Peter Sellers in a comic movie.

Another interesting aspect is the dialectics between realistic and unrealistic. Reverberation contributes to a realistic rendering of the spatial dimension of sound. On the other side, unrealistic reverberation can give an impression of dematerialization and symbolism. An interesting case of unrealistic sounds is given by the sounds and noises that children use to evoke the life of their puppets, dolls and little cars (especially their movement): Where do they come from? In general, in movie sound tracks, noises were always neglected with respect to dialogues and music. Only with the Dolby and multichannel systems, noise has gained a relevant place. Another important aspect of sound in audiovision is the dynamic changes of the acousmatic nature of sounds. For example, a deacousmatization wisely prepared, is always extremely effective and important in the dramatic evolution. On the contrary, the acousmatization of the final part of an event, e.g. a violent scene is also a very common technique adopted in cinema. A significantly different case is given by the television. The main point between cinema and television is the difference of position occupied by the sound. Television is rather an illustrated radio. The sound of words has always a principal role, in a sense it is never off-field. Even in the news the images are rather

a "decoration" of the verbal information. A typical TV effect is given by different voices that speak together, provoking a short circuit w.r.t the visual element. The radio-like attributes of the television increases when the TV is constantly switched on, for example, in public places. In this case the image is not any more the structural element, but only the exception, the "surprise". An interesting case deserving some considerations is given by tennis matches. Tennis is the most "sonified" sport: the different impact sounds of the ball, the cries of the players, the audience exclamations. It is the only sport, where the speaker can stop talking even for 30 seconds and more.

Finally, it is necessary to spend a few words about video-clips. The structure of video-clips and their stroboscopic effect make them a different case. It is not any more a dramatic time, but rather the turning of the faces of a prism. The success of a video-clips relies mainly on a simple punctual synchronism between sound and images. Also, in the video-clip sound looses its linearity character.

Chapter 10

Content processing of musical audio signals

Fabien Gouyon, Xavier Amatriain, Jordi Bonada, Pedro Cano, Emilia Gomez, Perfecto Herrera, Alex Loscos
Universitat Pompeu Fabra, Institut Universitari de l'Audiovisual, Music Technology Group

About this chapter

In this chapter, we provide an overview of state-of-the-art algorithms for the automatic description of musical audio signals, both from a low-level perspective (focusing on signal characteristics) and a more musical perspective (focusing on musically-meaningful dimensions). We also provide examples of applications based on this description, such as music identification, music browsing and musical signal transformations. A special focus is put on promising research directions.

10.1 Introduction

Music Information Retrieval (MIR) is a young and very active research area. This is clearly shown in the constantly growing number and subjects of articles published in the Proceedings of

the annual International Conference on Music Information Retrieval (ISMIR, the first established international scientific forum for researchers involved in MIR) and also in related conferences and scientific journals such as ACM Multimedia, IEEE International Conference on Multimedia and Expo or Wedelmusic, to name a few. In MIR, different long-tradition disciplines such as musicology, signal processing, psychoacoustics, information science, computer science, or statistics, converge by means of a multidisciplinary approach in order to address the wealth of scenarios for interacting with music posed by the digital technologies in the last decades (the standardization of world-wide low-latency networks, the extensive use of efficient search engines in everyday life, the continuously growing amount of multimedia information on the web, in broadcast data streams or in personal and professional databases and the rapid development of on-line music stores as e.g. Apples iTunes, Walmart or MusicMatch). Applications are manifold, consider for instance automated music analysis, personalized music recommendation, on-line music access, query-based retrieval (e.g. "by-humming," "by-example") and automatic play-list generation.

Among the vast number of disciplines and approaches to MIR (an overview of which can be found in Downie [2003a], content processing of audio signals plays in important role. Music comes in many forms but content-based audio processing is only concerned with one of them: audio signals.[1] This chapter does not deal with the analysis of symbolic music representations as e.g. digitized scores or structured representation of music events as MIDI. The relatively new direction of research concerning the automated analysis of social, cultural and marketing dimensions of music networks is addressed in XXXREF TO WIDMER CHAPTERXXX, see also Cano et al. [2005a].

This section defines the notion of music content at diverse levels of abstraction and what we understand by *processing* music content: both its *description* and its *exploitation*. We also shortly mention representation issues in music content processing. Section 10.2 provides an overview of audio content description according to low-level features and diverse musically-meaningful dimensions as pitch, melody and harmony (see page 396), rhythm (see page 405), and musical genre (see page 411). The organization follows increasing levels of abstraction. In section 10.3, we address content exploitation and present different applications to content-based audio description. Finally, promising avenues for future work in the field are summarized in section 10.4.

[1] Hence the undifferentiated use in this chapter of the terms "music content processing" and "audio content processing."

10.1.1 Music content: A functional view

A look at a dictionary reveals, at least, three senses for the word "content":

> everything that is included in a collection;
>
> what a communication that is about something is about;
>
> the sum or range of what has been perceived, discovered, or learned.

The disciplines of information science and linguistics offer interesting perspectives on the meaning of this term. However, we will rather focus on a more pragmatic view. The Society of Motion Picture and Television Engineers (SMPTE) and the European Broadcasting Union (EBU) have defined content as the combination of two entities termed *metadata* and *essence*. Essence is the raw program material itself, the data that directly encodes pictures, sounds, text, video, etc. Essence can also be referred to as media (although the former does not entail the physical carrier). In other words, essence is the encoded information that directly represents the actual message, and it is normally presented in a sequential, time-dependent manner. On the other hand, metadata (literally, "data about the data") is used to *describe* the essence and its different manifestations. Metadata can be classified, according to SMPTE/EBU, into several categories:

> Essential (meta-information that is necessary to reproduce the essence, like the number of audio channels, the Unique Material Identifier, the video format, etc.)
>
> Access (to provide control and access to the essence, i.e. Copyright information)
>
> Parametric (to define parameters of the essence capture methods like camera set-up, microphones set-up, perspective, etc.)
>
> Relational (to achieve synchronization between different content components, e.g. timecode)
>
> Descriptive (giving a description of the actual content or subject matter in order to facilitate the cataloging, search, retrieval and administration of content; i.e. title, cast, keywords, classifications of the images, sounds and texts, etc.),

In a quite similar way the National Information Standards Organization considers three main types of metadata:

Descriptive metadata, which describe a resource for purposes such as discovery and identification; they can include elements such as title, abstract, author, and keywords.

Structural metadata, which indicate how compound objects are put together, for example, how visual or audio takes are ordered to form a seamless audiovisual excerpt.

Administrative metadata, which provide information to help manage a resource, such as "when" and "how" it was created, file type and other technical information, and who can access it. There are several subsets of administrative data; two that sometimes are listed as separate metadata types are:

- Rights management metadata, which deals with intellectual property rights, and
- Preservation metadata, which contains information needed to archive and preserve a resource.

In accordance with these rather general definitions of the term "metadata," we propose to consider as content *all* that can be predicated from a media essence.

Any piece of information related to a musical piece that can be annotated, extracted, and that is in any way meaningful (that carries semantic information) to some user, can be technically denoted as metadata. Along this rationale, the MPEG-7 standard defines a content descriptor as "a distinctive characteristic of the data which signifies something to somebody" Manjunath et al. [2002]. This rather permissive view on the nature of music contents has a drawback: as they represent many different aspects of a musical piece, metadata are not ensured to be understandable by *any* user. This is part of the "user-modeling problem," whose lack of precision participates to the so-called *semantic gap*, that is, "the lack of coincidence between the information that one can extract from the (sensory) data and the interpretation that the same data has for a user in a given situation" Smeulders et al. [2000]. That has been signaled by several authors Smeulders et al. [2000], Lew et al. [2002], Jermyn et al. [2003] as one of the recurrent open issues in systems dealing with audiovisual content.

It is therefore important to consider metadata together with their *functional values* and address the question of which content means what to which users, and in which application Gouyon and Meudic [2003]. A way to address this issue is to consider content hierarchies with *different levels of abstraction*, any of them potentially useful for *some* users. In that sense, think of how different would a content description of a musical piece be if the targeted user was a naive listener or an expert musicologist. Even a low-level descriptor such as the spectral envelope

of a signal can be thought of as a particular level of content description targeted for the signal processing engineer. All these specifically targeted descriptions can be thought of as different instantiations of the same, general, content description scheme.

Following Amatriain and Herrera [2001] and Lesaffre et al. [2003], let us here propose the following distinction between descriptors of low, mid and high levels of abstraction (the latter being also sometimes referred to as "semantic" descriptors):

> A low-level descriptor can be computed from the essence data in a direct or derived way (i.e. after signal transformations like Fourier or Wavelet transforms, after statistical processing like averaging, after value quantization like assignment of a discrete note name for a given series of pitch values, etc.). Most of low-level descriptors make little sense to the majority of users but, on the other hand, their exploitation by computing systems are usually easy. They can be also referred to as "signal-centered descriptors" (see on page 387).
>
> Mid-level descriptors require an induction operation that goes from available data towards an inferred generalization about them. This kind of descriptors usually pave the way for labeling contents, as for example a neural network model that makes decisions about musical genre or about tonality, or a Hidden Markov Model that makes possible to segment a song according to timbre similarities. Machine learning and statistical modeling make mid-level descriptors possible, but in order to take advantage of those techniques and grant the validity of the models, we need to gather large sets of observations. Mid-level descriptors are also sometimes referred to as "object-centered descriptors."
>
> The jump from low- or mid-level descriptors to high-level descriptors requires "bridging the semantic gap." Semantic descriptors require an induction that has to be carried by means of a user-model (in order to yield the "interpretation" of the description), and not only a data-model as it was in the case of mid-level descriptors. As an example, let us imagine a simplistic "mood" descriptor consisting on labels "happy" and "sad." In order to compute such labels, one may[2] compute the tonality of the songs (i.e. "major" and "minor") and the tempo by means of knowledge-based analyzes of spectral and amplitude data. Using these mid-level descriptors, a model for computing the labels "happy" and "sad" would be elaborated by getting users' ratings of songs in terms of "happy" and "sad" and studying the relationships between these user-generated labels and values for tonality and tempo. High-level descriptors can also be referred to as "user-centered descriptors."

[2] and it is only a speculation here

Standards In order to be properly exploited, music content (either low-, mid- or high-level content) has to be organized into knowledge structures such as taxonomies, description schemes, or ontologies. The Dublin Core and MPEG-7 are currently the most relevant standards for representing music content. The Dublin Core (DC) was specified by the Dublin Core Metadata Initiative, an institution that gathers organizations such as the Library of Congress, the National Science Foundation, or the Deutsche Bibliotek, to promote the widespread adoption of interoperable metadata standards. DC specifies a set of sixteen metadata elements, a core set of descriptive semantic definitions, which is deemed appropriate for the description of content in several industries, disciplines, and organizations. The elements are Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights, and Audience. Description, for example, can be an abstract, a table of contents, a graphical representation or free text. DC also specifies a list of qualifiers that refine the meaning and use of the metadata elements, which open the door to refined descriptions and controlled-term descriptions. DC descriptions can be represented using different syntaxes, such as HTML or RDF/XML.

On the other hand, MPEG-7 is a standardization initiative of the ISO/IEC Moving Picture Expert Group that, contrasting with other MPEG standards, does not address the *encoding* of audiovisual essence. MPEG-7 aims at specifying an interface for the description of multimedia contents. MPEG-7 defines a series of elements that can be used to describe content, but it does not specify the algorithms required to compute values for those descriptions (in some cases, the algorithms are still to be found!). The building blocks of MPEG-7 description are descriptors, description schemes (complex structures made of aggregations of descriptors and description schemes), and the Description Definition Language (DDL), which defines the syntax that an MPEG-7 compliant description has to follow. The DDL makes hence possible the creation of non-standard, but compatible, additional descriptors and description schemes. This is an important feature because different needs will call for different kinds of structures, and for different instantiations of them. Depending on the theoretical and/or practical requirements of our problem, the required descriptors and description schemes will vary but, thanks to the DDL, we may build the proper structures to tailor our specific approach and required functionality. MPEG-7 descriptions are written in XML but a binary format has been defined to support their compression and streaming. The MPEG-7 standard definition covers eight different parts: Systems, DDL, Visual, Audio, Multimedia Description Schemes, Reference, Conformance and Extraction. In the audio section, we find music-specific descriptors for melody, rhythm or timbre, and in the Multimedia Description Schemes we find structures suitable to define classification

schemes and a wealth of semantic information. As argued in Gomez et al. [2003a,b] by some authors of this chapter, the status of the original standard (see Manjunath et al. [2002] for an overview), as to representing music contents, is nevertheless a bit deceiving and will require going beyond the yet-to-be-approved version 3 (which will probably include score descriptions) in order to be seriously adopted by the digital music community.

10.1.2 Processing music content: Description and exploitation

"Processing," beyond its straight meaning of "putting through a prescribed procedure," usually denotes a functional or computational approach to a wide range of scientific problems. "Signal processing" is the main term of reference here, but we could also mention "speech processing," "language processing," "visual processing" or "knowledge processing." A processing discipline focuses on the algorithmic level as defined by Marr [1982]. The algorithmic level describes a system in terms of the steps that have to be carried out to solve a given problem. This type of description is, in principle, independent of the implementation level (as the algorithm can be effectively implemented in different ways). However, it is important to contrast the meaning of content processing with that of signal processing. The object of signal processing are the raw data captured by sensors, whereas content processing deals with an object that is within the signal, embedded in it like a second-order code, and to which we refer to using the word metadata. The processes of extraction and modeling these metadata require the synergy of, at least, four disciplines (signal processing, artificial intelligence, information retrieval, and cognitive science) in order to make use, among other elements, of:

Powerful signal analysis techniques that make possible to address complex real-world problems, and to exploit context- and content-specific constraints in order to maximize their efficacy.

Reliable automatic learning techniques, that aid to build models about classes of objects that share specific properties, about processes that show e.g. temporal trends.

Availability of large databases of describable objects, and the technologies required to manage (index, query, retrieve, visualize) them.

Usable models of the human information processing that is involved in the processes of extracting and exploiting metadata (i.e. how humans perceive, associate, categorize,

remember, recall, and integrate into their behavior plans the information that might be available to them by means of other content processing systems).

Looking for the origins of music content processing, we can spot different forerunners depending on the contributing discipline that we consider.

When focusing on the discipline of *Information Retrieval*, the acknowledged pioneers seem to be Kassler [1966], Lincoln [1967]. The former defines music information retrieval as "the task of extracting, from a large quantity of musical data, the portions of that data with respect to which some particular musicological statement is true" (p. 66) and presents a computer language for addressing those issues. The latter discusses three criteria that should be met for automatic indexing of musical materials: eliminating the transcription by hand, effective input language for music, and an economic means for printing the music. This thread was later followed by Byrd [1984], Downie [1994], McNab et al. [1996], Blackburn [2000] with works dealing with score processing, representation and matching of melodies as strings of symbols, or query by humming.

Another batch of forerunners can be found when focusing on *digital databases* concepts and problems. Even though the oldest one dates back to 1988 Eaglestone [1988], the trend towards databases for "content processing" emerges more clearly in the early nineties de Koning and Oates [1991], Eaglestone and Verschoor [1991], Feiten et al. [1991], Keislar et al. [1995]. These authors address the problems related to extracting and managing the acoustic information derived from a large amount of sound files. In this group of papers, we find questions about computing descriptors at different levels of abstraction, ways to query a content-based database using voice, text, and even external devices, and exploiting knowledge domain to enhance the functionalities of the retrieval system.

To conclude with the antecedents for music content processing, we must also mention the efforts made since the last 30 years in the field of *music transcription*, whose goal is the automatic recovering of symbolic scores from acoustic signals. See Klapuri [2004a] for an exhaustive overview of music transcription research and Scheirer [2000] for a critical perspective on music transcription. Central to music transcription is the segregation of the different musical streams that coexist in a complex music rendition. Blind Source Separation (BSS) and Computational Auditory Scene Analysis (CASA) are two paradigms that address musical stream segregation. An important conceptual difference between them is that, unlike the latter, the former intends to actually separate apart the different streams that summed together make up the multi-instrumental music signal. BSS is the agnostic approach to segregate musical streams, as it usually does not

assume any knowledge about the signals that have been mixed together. The strength of BSS models (but at the same time its main problem in music applications) is that only mutual statistical independence between the source signals is assumed, and no a priori information about the characteristics of the source signals Casey and Westner [2000], Smaragdis [2001]. CASA, on the other hand, is partially guided by the groundbreaking work of Bregman —and originally coined Auditory Scene Analysis (ASA)— on the perceptual mechanisms that enables a human listener to fuse or fission concurrent auditory events Bregman [1990], see also XXXREF TO CHEVEIGNE CHAPTERXXX. CASA addresses the computational counterparts of ASA. Computer systems embedding ASA theories assume, and implement, specific heuristics that are hypothesized to play a role in the way humans perceive the music, as e.g. Gestalt principles. Worth to mention here are the works by Mellinger [1991], Brown [1992], Ellis [1996], Kashino and Murase [1997].

A comprehensive characterization of the field of music content processing was offered recently by Leman in Leman [2003]: "the science of musical content processing aims at explaining and modeling the mechanisms that transform information streams into meaningful musical units (both cognitive and emotional)." Musical content processing, for Leman, is the object of study of his particular view of Musicology, much akin to the so-called systematic musicology than to historic musicology. He additionally provides a definition of music content processing by extending it along three dimensions:

> The intuitive-speculative dimension, which includes semiotics of music, musicology, sociology, and philosophy of music. These disciplines provide a series of concepts and questions from a culture-centric point of view; music content is, following this dimension, a culture-dependent phenomenon.
>
> The empirical-experimental dimension, which includes research in physiology, psychoacoustics, music psychology, and neuro-musicology. These disciplines provide most of the empirical data needed to test, develop or ground some elements from the intuitive dimension; music content is, following this dimension, a percept in our auditory system.
>
> The computation-modeling dimension, which includes sound analysis and also computational modeling and simulation of perception, cognition and action. Music content is, following this dimension, a series of processes implemented in a computer, intended to emulate a human percept.

However, these three dimensions address mainly the *descriptive* part of music content processing, and according to Aigrain, "content processing is meant as a general term covering feature

extraction and modeling techniques for enabling basic retrieval, interaction and creation functionality" Aigrain [1999]. He also argue that music content processing technologies will provide "new aspects of listening, interacting with music, finding and comparing music, performing it, editing it, exchanging music with others or selling it, teaching it, analyzing it and criticizing it." We see here that music content processing can be characterized by two different tasks: *describing* and *exploiting* content. Furthermore, as mentioned above, the very meaning of "music content" cannot be entirely grasped without considering its *functional* aspects and including specific applications, targeted to specific users. Hence, in addition to describing music content (as reviewed in section 10.2), music content processing is also concerned with the design of computer systems that open the way to a more pragmatic content *exploitation* according to constraints posed by Leman's intuitive, empirical and computational dimensions (this is the object of section 10.3).

10.2 Audio content description

10.2.1 Low-level audio features

Many different low-level features can be computed from audio signals. literature in signal processing and speech processing provide us with an dramatic amount of techniques for signal representation and signal modeling over which features can be computed. Parametric methods (as e.g. AR modeling, Prony modeling) provide directly such features, while additional post-processing is necessary to derive features from non-parametric methods (as e.g. peaks can be extracted from a spectral or cepstral representations). A comprehensive overview of signal representation and modeling techniques and their associated features is clearly out of the scope of this chapter, we will only mention the features most commonly used in musical audio signal description, with a special focus on the recent work published in the music transcription and MIR literature.

Commonly, the audio signal is first digitized (if necessary) and converted to a general format, e.g. mono PCM (16 bits) with a fixed sampling rate (ranging from 5 to 44.1 KHz). A key assumption is that the signal can be regarded as stationary over intervals of a few milliseconds. Therefore, the signal is divided into frames (short chunks of signal) of e.g. 10 ms. The number of frames computed per second is called *frame rate*. A tapered window function (e.g. a Gaussian or Hanning window) is applied to each frame to minimize the discontinuities at the beginning and end. Consecutive frames are usually considered with some *overlap* for smoother analyzes. The

analysis step, the *hop size*, equals the frame rate minus the overlap.

Temporal features Many audio features can be computed directly from the temporal representation of these frames, for instance, the *mean* (but also the *maximum* and the *range* of) amplitude of the samples in a frame, the *energy*, the *zero-crossing rate*, the *temporal centroid* Gomez et al. [2005] and *auto-correlation coefficients* Peeters [2004].

Some low-level features have also shown to correlate with perceptual attributes, for instance, amplitude is loosely correlated with the *loudness*, see XXXREF TO CHEVEIGNE CHAPTERXXX. Equal-loudness curves (called isophones) show that there is a logarithmic relation between the physical measure of amplitude and loudness, with an additional frequency dependent trend. Although this curves have proved only valid for the stable part of pure sinusoids (more than 500 ms long), they have been used as a quite robust approximation for measuring loudness of complex mixtures Pfeiffer [1999].

Spectral features It is also very common to compute features on a different representation of the audio, as for instance the spectral representation. Hence, a spectrum is obtained from each signal frame by applying a Discrete Fourier Transform (DFT), usually with the help of the Fast Fourier Transform (FFT), this procedure is called Short-Time Fourier Transform (STFT). Sometimes, the time-frequency representation is further processed by taking into account perceptual processing that take place in human auditory system as for instance the filtering performed by the middle-ear, loudness perception, temporal integration or frequency masking Moore [1995], see also XXXREF TO CHEVEIGNE CHAPTERXXX. Many features can be computed on the obtained representation, as e.g. the spectrum *energy*, energy values in several frequency sub-bands (as e.g. the perceptually-motivated *Bark bands* Moore [1995]), the *mean*, *geometric mean*, *spread*, *centroid*, *flatness*, *kurtosis*, *skewness*, *spectral slope*, *high-frequency content* and *roll-off* of the spectrum frequency distribution or the *kurtosis* and *skewness* of the spectrum magnitude distribution, see Peeters [2004] and Gomez et al. [2005] for more details on these numerous features.

Further modeling of the spectral representation can be achieved through sinusoidal modeling McAulay and Quatieri [1986] or sinusoidal plus residual modeling Serra [1989], Amatriain et al. [2002]. Other features can be computed on the series of *spectral peaks* corresponding to each frame and on the spectrum of the *residual* component. Let us mention, for instance, the mean (and the accumulated) *amplitude* of sinusoidal and residual components, the *noisiness*, the *harmonic distortion*, the *harmonic spectral centroid*, the *harmonic spectral tilt* and different ratios of peak amplitudes as

the first, second and third *tristimulus* or the *odd-to-even ratio* Serra and Bonada [1998], Gomez et al. [2005].

Bear in mind that other transforms can be applied instead of the DFT as e.g. the Wavelet Kronland-Martinet et al. [1987] or the Wigner-Ville transforms Cohen [1989].

Cepstral features *Mel-Frequency Cepstrum Coefficients* (MFCCs) are widespread descriptors in speech research. The cepstral representation has been shown to be of prime importance in this field, partly because of its ability to nicely separate the representation of voice excitation (the higher coefficients) from the subsequent filtering performed by the vocal tract (the lower coefficients). Roughly, lower coefficients represent spectral envelope (i.e. the formants) while higher ones represent finer details of the spectrum, among which the pitch Oppenheim and Schafer [2004]. One way of computing the Mel-Frequency Cepstrum from a magnitude spectrum is as follows:

1. Projection of the frequency axis from linear scale to the Mel scale, of lower dimensionality (i.e. 20, by summing magnitudes in each of the 20 frequency bands of a Mel critical-band filter-bank)
2. Magnitude logarithm computation
3. Discrete Cosine Transform (DCT)

The number of output coefficients of the DCT is variable. It is often set to 13, as in the standard implementation of the MFCCs detailed in the widely-used speech processing software Hidden Markov Model Toolkit (HTK).[3]

Temporal evolution of frame features Apart from the instantaneous, or frame, feature values, many authors focus on the temporal evolution of features. This can be computed for instance as the derivative of feature values, which can be estimated by a first-order differentiator. The degree of change can also be measured as the feature differential normalized with its magnitude, e.g. Klapuri et al. [2005]. This is supposed to provide a better emulation of human audition, indeed, according to Weber's law, for humans, the just-noticeable-difference in the increment of a physical attribute depends linearly on its magnitude before incrementing. That is, $\Delta x/x$ (where x is a specific feature and Δx is the smallest perceptual increment) would be constant.

[3] `http://htk.eng.cam.ac.uk/`

10.2.2 Segmentation and region features

Frame features represent a significant reduction of dimensionality with respect to the audio signal itself, however, it is possible to further reduce the dimensionality by focusing on features computed on groups of consecutive frames, often called *regions*. The main issue here is the determination of relevant region boundaries, or the *segmentation* process. Once a given sound has been segmented into regions, it is possible to compute features as statistics of all frame features over the whole region (for example, the mean and variance of the amplitude of sinusoidal and residual components Serra and Bonada [1998]).

Segmentation Segmentation comes in different flavors, for McAdams and Bigand [1993], it "refers to the process of dividing an event sequence into distinct groups of sounds. The factors that play a role in segmentation are similar to the grouping principles addressed by Gestalt psychology." This definition implies that the segmentation process represents a step forward in the level of abstraction of data description. However, it may not necessarily be the case. Indeed, consider an adaptation of a classic definition coming from the visual segmentation area Pal and Pal [1993]: "(sound) segmentation is a process of partitioning the sound file/stream into non-intersecting regions such that each region is homogeneous and the union of no two adjacent regions is homogeneous." The notion of homogeneity in this definition implies a property of signal, or feature stationarity that may equate to a perceptual grouping process, but not necessarily.

In what is sometimes referred to as *model-free* segmentation, the main idea is using the amount of change of a feature vector, as a boundary detector: when this amount is higher than a given threshold, a boundary change decision is taken. Threshold adjustment requires a certain amount of trial-and-error, or fine-tuned adjustments regarding different segmentation classes. Usually, a smoothing window is considered in order to weight contributions from closer observations [Vidal and Marzal, 1990, p. 45].

It is also possible to generalize the previous segmentation process to multidimensional feature vectors. There, the distance between consecutive frames can be computed with the help of different measures as e.g. the Mahalanobis distance Tzanetakis and Cook [1999]. In the same vein, Foote [2000] uses MFCCs and the cosine distance measure between pairs of frames (not only consecutive frames), which yields a dissimilarity matrix that is further correlated with a specific kernel. Different kernels can be used for different types of segmentations (from short- to

long-scale).

The level of abstraction that can be attributed to the resulting regions may depend on the features used in the first place. For instance, if a set of low-level features is known to correlate strongly with a human percept (as the fundamental frequency correlates with the pitch and the energy in Bark bands correlates with the loudness) then the obtained regions may have some relevance as features of mid-level of abstraction (e.g. musical notes in this case).

Model-based segmentation on the other hand is more directly linked to the detection of mid-level feature boundaries. It corresponds to a focus on mid-level features that are thought, *a priori*, to make up the signal. A classical example can be found in speech processing where dynamical models of phonemes, or words, are built from observations of labeled data. The most popular models are Hidden Markov Models (HMM) Rabiner [1989]. Applications of HMMs to segmentation of music comprise e.g. segmentation of fundamental frequency envelopes in musical notes Raphael [1999] and segmentation of MFCC-based temporal series in regions of globally-homogeneous timbres Batlle and Cano [2000]. Other examples of model-based segmentation can be found in Rossignol [2000] who reports on the performance of different induction algorithms, Gaussian Mixture Models (GMM), k-Nearest Neigbours (k-NN) and Artificial Neural Networks (ANN), in the tasks of speech/music segmentation and intra-note segmentation. In the more general context of signal segmentation (not just music signals), Basseville et al. [1993] propose many segmentation techniques, some of which entailing the use of signal models as for instance, a time-domain based technique in which two windows are used, one with fixed size and a growing one, and some distance estimation is computed between two AR models (derived from the cross entropy between the conditional distributions of the two models), one built on each window. Here also, a threshold is used to determine whether the distance should be considered representative of a boundary or not. Application of this technique to musical signals can be found in Jehan [1997], Thornburg and Gouyon [2000].

We refer to Herrera and Gomez [2001] for a complete review of musical signal segmentation techniques.

Note onset detection The detection of *note onsets* in musical signals has attracted many computer music researchers since the early eighties Gordon [1984]. Several methods have been designed, making use of diverse low-level features. The simplest focus on the temporal variation of a single feature, as for instance the energy or the pitch. However, the combined used of multiple features (as energy *and* pitch) seems to provide better estimates, state-of-the-art algorithms

often making use of band-wise energy processing Klapuri [1999], Bello [2003]. Model-based note onset segmentation has also been an active research field in the last years Thornburg and Gouyon [2000]

The literature in onset detection is extremely furnished and a review is out of the scope of this chapter, for an exhaustive review, see Bello [2003].

Intra-note segmentation In addition to note onset detection, some research has also been dedicated to the segmentation of musical signals in terms of Attack, Sustain and Release regions. This is especially relevant, from a feasibility point of view, when dealing with isolated instrument samples or musical phrases played by a monophonic instrument Jenssen [1999], Maestre and Gomez [2005].

Given start and end boundaries of these regions, it is possible to compute a number of features that relate to their durations as e.g. the *log-attack time* Peeters [2004]. Some authors also focus on the variations of low-level frame features in these regions, such as the energy Maestre and Gomez [2005] or the fundamental frequency in sustain regions, characterizing therefore the *vibrato* Herrera and Bonada [1998].

Speech/Music segmentation A large body of work in automatic segmentation of audio signal also concerns the determination of boundaries of speech regions and musical regions. This is usually achieved by model-based segmentation of multiple low-level features Scheirer and Slaney [1997], Harb and Chen [2003], Pinquier et al. [2003].

10.2.3 Audio fingerprints

Audio fingerprints have attracted a lot of attention for their usefulness in audio identification applications. Compact content-based signatures summarizing audio recordings (the audio fingerprints) can be extracted from a musical audio piece and stored in a database. Fingerprint of unlabeled pieces of audio can be calculated and matched against those stored in the database, providing a link to corresponding metadata (e.g. artist and song name). Section 10.3.1 provides more details on the main requirements of fingerprinting systems and application scenarios.

For a general functional framework of audio fingerprinting systems and an overview of

current technologies, see Cano et al. [2002a].[4] This section provides a short overview of audio features commonly used in the design of audio fingerprints.

Fingerprint extraction The fingerprint extraction derives a set of features from a recording in a concise and robust form. Fingerprint requirements include:

Discrimination power over huge numbers of other fingerprints,

Invariance to distortions,

Compactness,

Computational simplicity.

The simplest approach one may think of —using directly the digitized waveform— is neither efficient nor effective. A more efficient implementation of this approach could use a hash method, such as MD5 (Message Digest 5) or CRC (Cyclic Redundancy Checking), to obtain a compact representation of the binary file. However, hash values are fragile, a single bit flip is sufficient for the hash to completely change. They are also not robust to compression or distortions.

Most fingerprint extraction systems consist of a front-end and a fingerprint modeling block (see Figure 10.1). The front-end computes low-level features from the signal and the fingerprint model defines the final fingerprint representation, we now briefly describe them in turn.

Front-End Several driving forces co-exist in the design of the front-end: dimensionality reduction, perceptually meaningful parameters (similar to those used by the human auditory system), invariance/robustness (to channel distortions, background noise, etc.) and temporal correlation (systems that capture spectral dynamics).

After the first step of audio digitization, the audio is sometimes preprocessed to simulate the channel, e.g: band-pass filtered in a telephone identification task. Other types of processing are a GSM coder/decoder in a mobile phone identification system, pre-emphasis, amplitude normalization (bounding the dynamic range to $[-1, 1]$).

[4]Note that fingerprinting should not be mistaken for watermarking, differences are explained in Gomes et al. [2003].

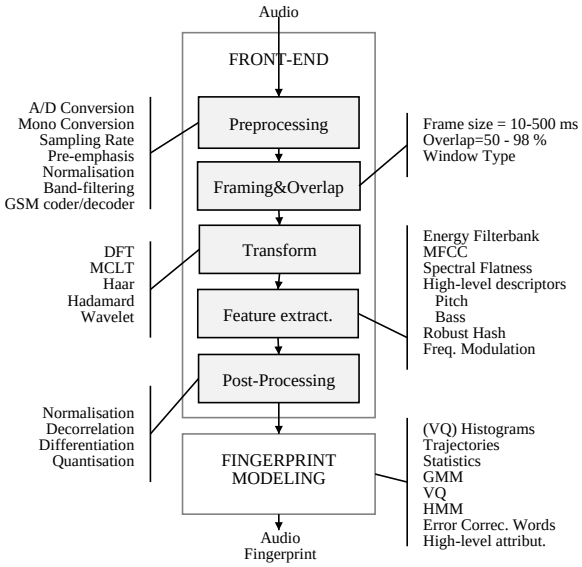

Figure 10.1: Fingerprint Extraction Framework: Front-end (top) and Fingerprint modeling (bottom).

After framing the signal is small windows, overlap must be applied to assure robustness to shifting (i.e. when the input data is not perfectly aligned to the recording that was used for generating the fingerprint). There is a trade-off between the robustness to shifting and the computational complexity of the system: the higher the frame rate, the more robust to shifting the system is but at a cost of a higher computational load.

Then, linear transforms are usually applied (see Figure 10.1). If the transform is suitably chosen, the redundancy is significantly reduced. There are optimal transforms in the sense of information packing and de-correlation properties, like Karhunen-Loeve (KL) or Singular Value Decomposition (SVD). These transforms, however, are computationally complex. For that reason, lower complexity transforms using fixed basis vectors are common (as e.g. the DFT).

Additional transformations are then applied in order to generate the final acoustic vectors. In this step, we find a great diversity of algorithms. The objective is again to reduce the dimensionality and, at the same time, to increase the invariance to distortions. It is very common to include knowledge of the transduction stages of the human auditory system to extract more perceptually meaningful parameters. Therefore, many systems extract several features performing a critical-band analysis of the spectrum. Resulting features are e.g. MFCCs, energies in Bark-scaled bands, geometric mean of the modulation frequency estimation of the energy in

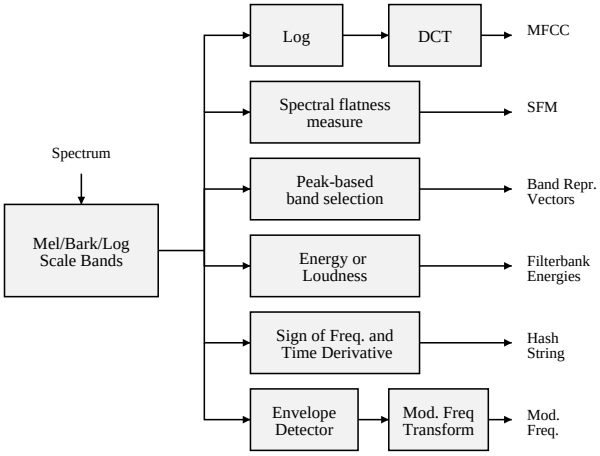

Figure 10.2: Feature Extraction Examples

bark-spaced band-filters, etc., or many of the features presented in section 10.2.1. Some examples are given in Figure 10.2.

Most of the features described so far are absolute measurements. In order to better characterize temporal variations in the signal, higher order time derivatives are added to the signal model. Some systems compact the feature vector representation using transforms as Principal Component Analysis (PCA). It is also quite common to apply a very low resolution quantization (ternary or binary) to the features, the purpose of which is to gain robustness against distortions and reduce the memory requirements.

Fingerprint Models The sequence of features calculated on a frame by frame basis is then further reduced to a fingerprint model that usually implies statistics of frame values (mean and variance) and redundancies in frame vicinity. A compact representation can also be generated by clustering the feature vectors. The sequence of vectors is thus approximated by a much lower number of representative code vectors, a codebook. The temporal evolution of audio is lost with this approximation, but can be kept by collecting short-time statistics over regions of time or by HMM modeling Batlle et al. [2002].

At that point, some systems also derive musically-meaningful attributes from low-level features, as the *beats* Kirovski and Attias [2002] (see on page 405) or the *predominant pitch* Blum et al. [1999] (see on page 400).

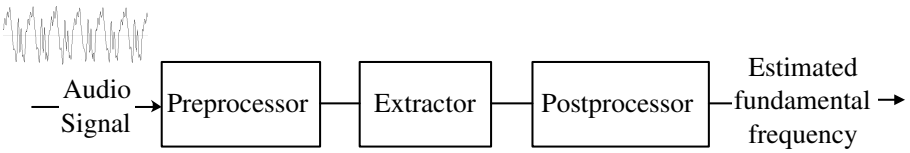

Figure 10.3: Steps of the fundamental frequency detection process

10.2.4 Tonal descriptors: from pitch to key

This section first reviews computational models of *pitch* description and then progressively addresses tonal aspects of higher levels of abstraction that imply different combinations of pitches: *melody* (sequence of single pitches combined over time), *pitch classes* and *chords* (simultaneous combinations of pitches), and *chord progressions*, *harmony* and *key* (temporal combinations of chords).

Pitch

Fundamental frequency is the main low-level descriptor to consider when describing melody and harmony. Due to the significance of pitch detection for speech and music analysis, a lot of research has been made on this field. We present here a brief review of the different approaches for pitch detection: fundamental frequency estimation for monophonic sounds, multi-pitch estimation and predominant pitch estimation. We refer to Gomez et al. [2003c] for an exhaustive review.

Fundamental frequency estimation for monophonic sounds As illustrated on Figure 10.3, the fundamental frequency detection process can be subdivided into three successive steps: the preprocessor, the basic extractor, and the post-processor Hess [1983]. The basic extractor converts the input signal into a series of fundamental frequency estimates, one per analysis frame. Pitched/unpitched measures are often additionally computed to decide whether estimates are valid or should be discarded Cano [1998]. The main task of the pre-processor is to facilitate the fundamental frequency extraction. Finally, the post-processor performs more diverse tasks, such as error detection and correction, or smoothing of an obtained contour. We now describe these three processing blocks in turn.

Concerning the main extractor processing block, the first solution was to adapt the tech-

niques proposed for speech Hess [1983]. Later, other methods have been specifically designed for dealing with music signals. These methods can be classified according to their processing domain: *time-domain* algorithms vs *frequency-domain* algorithms. This distinction is not always so clear, as some of the algorithms can be expressed in both (time and frequency) domains, as the autocorrelation function (ACF) method. Another way of classifying the different methods, more adapted to the frequency domain, is to distinguish between *spectral place* algorithms and *spectral interval* algorithms Klapuri [2000]. The spectral place algorithms weight spectral components according to their spectral location. Other systems use the information corresponding to spectral intervals between components. Then, the spectrum can be arbitrarily shifted without affecting the output value. These algorithms work relatively well for sounds that exhibit inharmonicity, because intervals between harmonics remain more stable than the places for the partials.

Time-domain algorithms The simplest time-domain technique is based on counting the number of times the signal crosses the 0-level reference, the *zero-crossing rate* (ZCR). This method is not very accurate when dealing with noisy signals or harmonic signals where the partials are stronger than the fundamental.

Algorithms based on the *time-domain autocorrelation* function (ACF) have been among the most frequently used fundamental frequency estimators. ACF-based fundamental frequency detectors have been reported to be relatively noise immune but sensitive to formants and spectral peculiarities of the analyzed sound Klapuri [2000].

Envelope periodicity algorithms find their roots in the observation that signals with more than one frequency component exhibit periodic fluctuations in their time domain amplitude envelope. The rate of these fluctuations depends on the frequency difference of each two frequency components. In the case of a harmonic sound, the fundamental frequency is clearly visible in the amplitude envelope of the signal. The most recent models of human pitch perception calculate envelope periodicity separately at distinct frequency bands and then combine the results across channels **?**Terhardt et al. [1981]. These methods attempt to estimate the perceived pitch, not pure physical periodicity, in acoustic signals of various kinds. (See XXXREF TO CHEVEIGNE CHAPTERXXX for more references on the perception of pitch.)

The *parallel processing approach* of Gold and Rabiner [1969], Rabiner and Schafer [1978], designed to deal with speech signals, has been successfully used in a wide variety of applications. Instead of designing one very complex algorithm, the basic idea is to tackle the same problem with several, more simple processes in parallel and later combine their outputs. As Bregman

points out in Bregman [1998], human perception appears to be redundant at many levels, several different processing principles seem to serve the same purpose, and when one of them fails, another is likely to succeed.

Frequency-domain algorithms Noll Noll [1967] introduced this idea of *Cepstrum analysis* for pitch determination of speech signals. The Cepstrum computation (see on page 389) nicely separates the transfer function (spectral envelope) from the source, hence the pitch. Cepstrum fundamental frequency detection is closely similarity to autocorrelation systems Klapuri [2000].

Spectrum autocorrelation methods were inspired by the observation that a periodic but non-sinusoidal signal has a periodic magnitude spectrum, the period of which is the fundamental frequency. This period can be estimated by ACF Klapuri [2000].

Harmonic matching methods extract a period from a set of spectral peaks of the magnitude spectrum of the signal. Once these peaks in the spectrum are identified, they are compared to the predicted harmonics for each of the possible candidate note frequencies, and a measure to fit can be developed. A particular fitness measure is described in Maher and Beauchamp [1993] as a "Two Way Mismatch" procedure. This method is used in the context of Spectral Modeling Synthesis (SMS), with some improvements, as pitch-dependent analysis window, enhanced peak selection, and optimization of the search Cano [1998].

The idea behind *Wavelet based algorithms* is to filter the signal using a wavelet with derivative properties. The output of this filter will have maxima where glottal-closure instants or zero crossings happen in the input signal. After detection of these maxima, the fundamental frequency can be estimated as the distance between consecutive maxima.

Klapuri Klapuri [2000] proposes a *band-wise processing algorithm* that calculates independent fundamental frequencies estimates in separate frequency bands. Then, these values are combined to yield a global estimate. This method presents several advantages: it solves the "inharmonicity" problem, it is robust to heavy signal distortions, where only a fragment of the frequency range is reliable.

Preprocessing methods The main task of a preprocessor is to suppress noise prior to fundamental frequency estimation. Some preprocessing methods used in speech processing are detailed in Hess [1983]. Methods specifically defined for musical signals are detailed in Klapuri [2004b]

Post-processing methods The estimated series of pitches may be noisy and may present isolated errors, different methods have been proposed for correcting these. The first is low-pass filtering (linear smoothing) of the series. This may remove much of the local jitter and noise, but does not remove local gross measurements errors, and, in addition, it smears the intended discontinuities at the voiced-unvoiced transitions Hess [1983]. Non-linear smoothing has been proposed to address these problems Rabiner et al. [1975]. Another procedure consists in storing several possible values for the fundamental frequency for each analysis frame Laroche [1995], assigning them a score (e.g. the value of the normalized autocorrelation). Several tracks are then considered and ranked (according to some continuity evaluation function) by e.g. dynamic programming. This approach minimizes the abrupt fundamental frequency changes (e.g. octave errors) and gives good results in general. Its main disadvantage is its estimation delay and non-causal behavior. It is also usually useful to complement the forward estimation by a backward estimation Cano [1998].

Multi-pitch estimation Multi-pitch estimation is the simultaneous estimation of the pitches making up a polyphonic sound (e.g. a polyphonic instrument or several instruments playing together).

Some algorithms used for monophonic pitch detection can be adapted to the simplest polyphonic situations Maher and Beauchamp [1993]. However, they are usually not directly applicable to general cases, they require, among other differences, significantly longer time frames (around 100 ms) Klapuri [2000].

Relatively successful algorithms implement principles of the perceptual mechanisms that enables a human listener to fuse or fission concurrent auditory streams (see references to "Auditory Scene Analysis" on page 386 and XXXREF TO CHEVEIGNE CHAPTERXXX). For instance, Kashino et al. implement such principles in a Bayesian probability network, where bottom-up signal analysis can be integrated with temporal and musical predictions Kashino et al. [1995]. A recent example following the same principles is detailed in Walmsley et al. [1999], where a comparable network estimates the parameters of a harmonic model jointly for a number of frames. Godsmark and Brown have developed a model that is able to resolve melodic lines from polyphonic music through the integration of diverse knowledge Godsmark and Brown [1999]. Other methods are listed in Klapuri [2000].

The state-of-the-art multi-pitch estimators operate reasonably well for clean signals, frame-level error rates increasing progressively with the number of concurrent voices. However, the

number of concurrent voices is often underestimated and the performance usually decreases significantly in the presence of noise Klapuri [2000].

Predominant pitch estimation Predominant pitch estimation also aims at estimating pitches in polyphonic mixtures, however, contrarily to multi-pitch estimation, it assumes that a specific instrument is predominant and defines the melody. For instance, the system proposed in Goto [2000] detects melody and bass lines in polyphonic recordings using a multi-agent architecture by assuming that they occupy different frequency regions.

Other relevant methods are reviewed in Gomez et al. [2003c] and Klapuri [2004b].

Melody

Extracting melody from note sequences We have presented above several algorithms whose outputs are time sequences of pitches (or simultaneous combinations thereof). Now, we present some approaches that, building upon those, aim at identifying the notes that are likely to correspond to the main melody. We refer to Gomez et al. [2003c] for an exhaustive review of the sate-of-the-art in melodic description and transformation from audio recordings.

Melody extraction can be considered not only for polyphonic sounds, but also for monophonic sounds as they may contain notes that do not belong to the melody (as for example grace notes, passing notes or the case of several interleaved voices in a monophonic stream).

As argued in Nettheim [1992] and [Selfridge-Field, 1998, Section 1.1.3.], the derivation of a melody from a sequence of pitches faces the following issues:

> A single line played by a single instrument or voice may be formed by movement between two or more melodic or accompaniment strands.
>
> Two or more contrapuntal lines may have equal claim as "the melody."
>
> The melodic line may move from one voice to another, possibly with overlap.
>
> There may be passages of figuration not properly considered as melody.

Some approaches try to detect note groupings. Experiments have been done on the way listeners achieve melodic grouping, see [Scheirer, 2000, p.131] and McAdams [1994]. These provide heuristics that can be taken as hypothesis in computational models.

Other approaches make assumptions on the type of music to be analyzed. For instance, methods can be different according to the complexity of the music (monophonic or polyphonic music), the genre (classical with melodic ornamentations, jazz with singing voice, etc) or the representation of the music (audio, MIDI, etc.).

In Uitdenbogerd and Zobel [1998], Uitdenbogerd evaluates algorithms that extract melody from MIDI files containing channel information. Considering the highest notes as the one carrying the melody gives the best results. According to Uitdenbogerd and Zobel [1998], it appears that very few researches focus on this area of research in comparison to the interest that is granted to other tasks such as melody matching and pattern induction.

Melodic Segmentation The goal of melodic segmentation is to establish a temporal structure on a sequence of notes. It may involve different levels of hierarchy, as those defined by Lerdahl and Jackendoff Lerdahl and Jackendoff [1983], and may include overlapping, as well as unclassified, segments.

One relevant method proposed by Cambouropoulos Cambouropoulos [2001] is the Local Boundary Detection Model (LBDM). This model computes the transition strength of each interval of a melodic surface according to local discontinuities. In Cambouropoulos [2001], not only pitch is used, but also temporal (inter-onset intervals, IOIs) and rest intervals. He compares this algorithm with the punctuation rules defined by Friberg et al. from KTH,[5] getting coherent results. The LBDM has been used by Melucci and Orio Melucci and Orio [1999] for content-based retrieval of melodies.

Another approach can be found in the *Grouper*[6] module of the *Melisma* music analyzer, implemented by Temperley and Sleator. This module uses three criteria to select the note boundaries. The first one considers the *gap score* for each pair of notes, that is, the sum of the IOIs and the offset-to-onset interval (OOI). Phrases receive a weight proportional to the gap score between the notes at the boundary. The second one considers an optimal phrase length in number of notes. The third one is related to the metrical position of the phrase beginning, relative to the metrical position of the previous phrase beginning.

Spevak et al. Spevak et al. [2002] have compared several algorithms for melodic segmentation: LBDM, the Melisma Grouper, and a memory-based approach, the Data-Oriented Parsing

[5] see `http://www.speech.kth.se/music/performance_rules.html`
[6] see `http://www.link.cs.cmu.edu/music-analysis/grouper.html`

(DOP) from Bod Bod [2001]. They also describe other approaches to melodic segmentation. To explore this issue, they have compared manual segmentation of different melodic excerpts. However, according to them, "it is typically not possible to determine one 'correct' segmentation, because the process is influenced by a rich and varied set of context."

Miscellaneous melodic descriptors Other descriptors can be derived from a numerical analysis of the pitches of a melody and be used in diverse applications as comparative analysis Toiviainen and Eerola [2001], melody retrieval Kostek [1998], Tzanetakis [2002], and algorithmic composition Towsey et al. [2001]. Some of these descriptors are computed using features related to structural, musical or perceptual aspects of sound. Some others are computed from note descriptors (therefore they require algorithms for note segmentation, see on page 391). Yet other descriptors can be computed as statistics of frame or sample features. One example are the pitch histogram features proposed in Tzanetakis [2002].

Pitch class distribution

Many efforts have been devoted to the analysis of *chord sequences* and *key* in MIDI representations of classical music, but little work has dealt directly with audio signals and with other musical genres. Adapting MIDI-oriented methods would require a previous step of automatic transcription of polyphonic audio, which, as argued in Scheirer [2000], Klapuri [2004b], is far from being solved.

Some approaches extract information related to the pitch class distribution of music without performing automatic transcription. The pitch-class distribution is directly related to the chords and the tonality of a piece. Chords can be recognized from the pitch class distribution without requiring the detection of individual notes. Tonality can be also estimated from the pitch class distribution without a previous procedure of chord estimation.

Fujishima Fujishima [1999] proposes a system for chord recognition based on the pitch-class profile (PCP), a 12-dimensional low-level vector representing the intensities of the twelve semitone pitch classes. His chord recognition system compares this vector with a set of chord-type templates to estimate the played chord. In Sheh [2003], chords are estimated from an audio recordings by modeling sequences of PCPs with an HMM.

In the context of a key estimation system, Gomez Gomez [In print] proposes the Harmonic

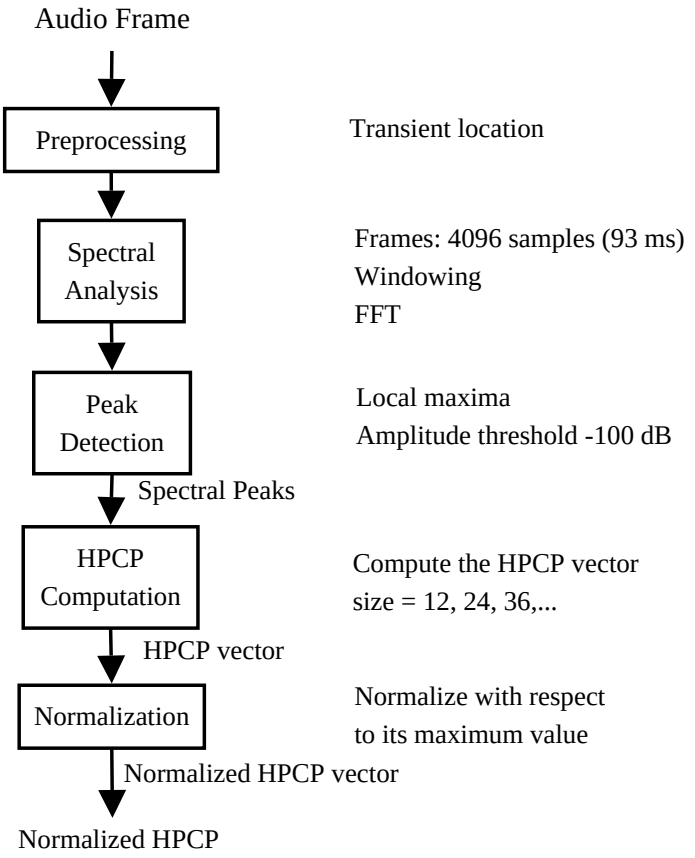

Figure 10.4: Block Diagram for HPCP Computation

PCPs (HPCPs) as extension of the PCPs: only the spectral peaks in a certain frequency band are used ([100, 5000] Hz), a weight is introduced into the feature computation and a higher resolution is used in the HPCP bins (decreasing the quantization level to less than a semitone). The procedure for HPCP computation is illustrated in Figure 10.4. A transient detection algorithm Bonada [2000] is used as preprocessing step in order to discard regions where the harmonic structure is noisy, the areas located 50 ms before and after the transients are not analyzed. As a post-processing step, HPCPs are normalized with respect to maximum values for each analysis frame, in order to store the *relative relevance* of each of the HPCP bins.

In the context of beat estimation of drum-less audio signals, Goto and Muraoka Goto [1999] also introduced the computation of a histogram of frequency components, used to detect chord changes (this method does not identify chord names).

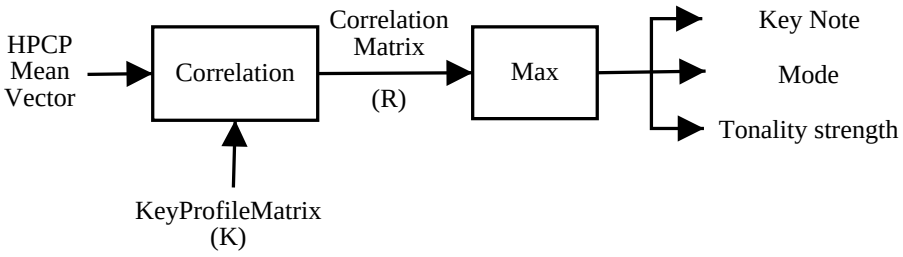

Figure 10.5: Block Diagram for Key Computation using HPCP

Constant Q profiles have also been used to characterize the tonal content of audio Purwins et al. [2000]. Constant Q profiles are twelve-dimensional vectors, each component referring to a pitch class, which are computed with the constant Q filter bank Brown J. C. [1992]. Purwins et al. Purwins et al. [2003] present examples where constant Q profiles are used to track tonal centers. In recent works, they use these features to analyze the interdependence of pitch classes and key as well as key and composer.

Tzanetakis Tzanetakis [2002] proposes a set of features related to audio harmonic content in the context of musical genre classification. These features are derived from a pitch histogram that can be computed from MIDI or audio data: the most common pitch class used in the piece, the frequency of occurrence of the main pitch class and the pitch range of a song.

Tonality: from chord to key

Pitch class distributions can be compared (correlated) with a tonal model to estimate the *chords* (when considering small time scales) or the *key* of a piece (when considering a larger time scale). This is the approach followed by Gomez in Gomez [In print] to estimate the tonality of audio pieces at different temporal scales, as shown on Figure 10.5.

To construct the key-profile matrix shown on Figure 10.5, Gomez Gomez [In print] follows a model proposed by Krumhansl for key estimation of MIDI file Krumhansl [1990]. This model considers that tonal hierarchies may be acquired through internalization of the relative frequencies and durations of tones. The algorithm estimates the key from a set of note duration values, measuring how long each of the 12 pitch classes of an octave (C, C♯, etc.) have been played in a melodic line. In order to estimate the key of the melodic line, the vector of note durations is correlated to a set of key profiles or probe-tone profiles. These profiles represent the tonal

hierarchies of the 24 major and minor keys, and each of them contains 12 values, which are the ratings of the degree to which each of the 12 chromatic scale tones fit a particular key. They were obtained by analyzing human judgments with regard to the relationship between pitch classes and keys [Krumhansl, 1990, pp. 78-81]. Gomez adapts this model to deal with HPCPs (instead of note durations) and polyphonies (instead of melodic lines), details of evaluations can be found in Gomez [In print], together with an exhaustive review of computational models of tonality.

10.2.5 Rhythm

Representing rhythm

Imagine the following musical scene. Somebody (or some machine) is making music: musical events are generated at given instants (onset times) and make up a temporal sequence. One way to represent the rhythm of this sequence could be to specify an exhaustive and accurate list of onset times, maybe together with some other musical feature characterizing those events as e.g. durations, pitches or intensities (as is done in MIDI). However, the problem to this representation is the *lack of abstraction*. There is more to rhythm than the absolute timings of successive musical events, namely *tempo*, *meter* and *timing* Honing [2001].

Tempo Cooper et al. Cooper and B. [1960] define a pulse as "[...] one of a series of regularly recurring, precisely equivalent stimuli. [...] Pulses mark off equal units in the temporal continuum." Commonly, 'pulse' and 'beat' are often used indistinctly and refer *both* to one element in such a series and to the whole series itself.

The *tempo* is defined as the number of beats in a time unit (usually the minute). There is usually a *preferred* pulse, which corresponds to the rate at which most people would tap or clap in time with the music. However, the perception of tempo exhibits a degree of variability. It is not always correct to assume that the pulse indicated in a score (Maelzel Metronome) corresponds to the "foot-tapping" rate, nor to the actual "physical tempo" that would be an inherent property of audio flows Drake et al. [1999]. Differences in human perception of tempo depend on age, musical training, musical preferences and general listening context Lapidaki [1996]. They are nevertheless far from random and most often correspond to a focus on a different metrical level and are quantifiable as simple ratios (e.g. 2, 3, $\frac{1}{2}$ or $\frac{1}{3}$).

Meter The metrical structure (or meter) of a musical piece is based on the coexistence of several pulses (or "metrical levels"), from low levels (small time divisions) to high levels (longer time divisions). The segmentation of time by a given low-level pulse provides the basic time span to measure musical event accentuation whose periodic recurrences define other, higher, metrical levels. The duration-less points in time, the *beats*, that define this discrete time grid obey a specific set of rules, formalized in the Generative Theory of Tonal Music (GTTM, Lerdahl and Jackendoff [1983]). Beats must be equally spaced. A beat at a high level must also be a beat at each lower level. At any metrical level, a beat which is also a beat at the next higher level is called a downbeat, and other beats are called upbeats.

The notions of time signature, measure and bar lines reflect a focus on solely two (or occasionally three) metrical levels. Bar lines define the slower of the two levels (the measure) and the time signature defines the number of faster pulses that make up one measure. For instance, a $\frac{6}{8}$ time signature indicates that the basic temporal unit is an eighth-note and that between two bar lines there is room for six units. Two categories of meter are generally distinguished: duple and triple. This notion is contained in the numerator of the time signature: if the numerator is a multiple of two, then the meter is duple, if not a multiple of two but of three, the meter is triple.

The GTTM specifies that there must be a beat of the metrical structure for every note in a musical sequence. Accordingly, given a list of note onsets, the quantization (or rhythm-parsing) task aims at making it fit into Western music notation. Viable time points (metrical points) are those defined by the different coexisting metrical levels. Quantized durations are then rational numbers (e.g. 1, $\frac{1}{4}$, $\frac{1}{6}$) relative to a chosen time interval: the time signature denominator.

Timing A major weakness of the GTTM is that it does not deal with the departures from strict metrical timing which are apparent in almost all styles of music. Thus it is only really suitable for representing the timing structures of musical scores, where the expressive timing is not represented. There are conceptually two types of non-metrical timing: long-term tempo deviations and short-term timing deviations (as e.g. Swing).

One of the greatest difficulties in analyzing performance data is that the two dimensions of expressive timing are projected onto the single dimension of time. Mathematically, it is possible to represent any tempo change as a series of timing changes and vice-versa, but these descriptions are somewhat counterintuitive Honing [2001].

Challenges in automatic rhythm description

Contrarily to what it may seem, automatically describing musical rhythm is not obvious. First of all because it seems to entail two dichotomic processes: a bottom-up process enabling very rapidly the percept of pulses from scratch, and a top-down process (a persistent mental framework) that lets this induced percept guide the organization of incoming events Desain and Honing [1999]. Embodying in a computer program both reactivity to the environment and persistence of internal representations is a challenge.

Rhythm description does not solely call for the handling of timing features (onsets and offsets of musical tones). The definition, and understanding of the relevance, of other musical features such as intensities or pitches are still open research topics.

Rhythm involves two dichotomic aspects that are readily perceived by humans: there is both a strong and complex structuring of phenomena occurring at different time scales and widespread departures from exact metrical timing. Indeed, inexact timings always occur because of expressive performances, sloppy performances and inaccurate collection of timing data (e.g. onset detection may have poor time precision and suffer false-alarms).

Furthermore, recent research indicates that even if perceived beats are strongly correlated to onsets of musical tones, they do not necessarily line up exactly with them, our perception rather favoring smooth tempo curves Dixon [2005].

Functional framework

The objective of automatic rhythm description is the parsing of acoustic events that occur in time into the more abstract notions of tempo, timing and meter. Computer programs described in the literature differ in their goals. Among others, some derive the beats and the tempo of a single metrical level, others aim at deriving complete rhythmic transcriptions (i.e. scores) from musical performances, others aim at determining some timing features from musical performances (such as tempo changes, event shifts or swing factors), others aim at classifying musical signals by their overall rhythmic similarities and still others aim at the determination of rhythm patterns. Nevertheless, these computer programs share some functional aspects that we illustrate as functional blocks of a general diagram on Figure 10.6 and briefly explain in turn in the following paragraphs; we refer to Gouyon and Dixon [2005] for a more complete survey.

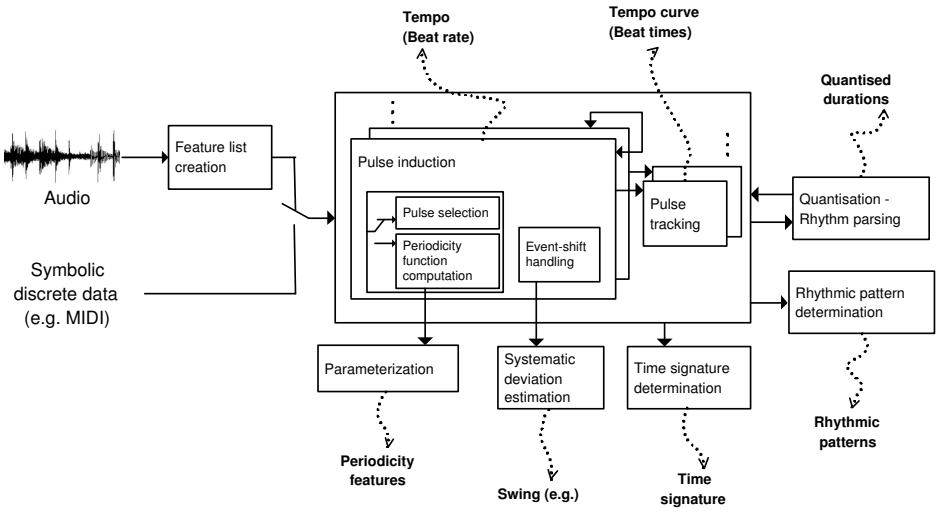

Figure 10.6: Functional units of rhythm description systems.

Feature list creation Either starting from MIDI, other symbolic formats (e.g. in the format of files containing solely onset times and durations Brown [1993]) or audio data, the first analysis step is the creation of a feature list, i.e. the parsing, or "filtering," of the data at hand into a sequence assumed to convey the predominant information relevant to a rhythmic analysis.

These feature lists are defined here broadly, to include frame-based feature vectors as well as lists of symbolic events. The latter include onset times, durations Brown [1993], relative amplitude Dixon [2001], pitch Dixon and Cambouropoulos [2000], chords Goto [2001] and percussive instrument classes Goto [2001]. Some systems refer to a data granularity of a lower level of abstraction: frames. Section 10.2.1 describes usual low-level features. In rhythm analysis, common frame features are energy values and energy values for frequency sub-bands. Some systems also measure energy variations between consecutive frames Scheirer [2000], Klapuri et al. [2005]. Low-level features other than energy (e.g. spectral flatness, temporal centroid) have also been recently advocated Gouyon and Herrera [2003].

Pulse induction A metrical level (a pulse) is defined by the periodic recurrence of some musical event. Therefore, computer programs generally seek periodic behaviors in feature lists in order to select one (or some) pulse period(s) and also sometimes phase(s). This is the process of *pulse induction*. For pulse induction, computer programs either proceed by *pulse selection*, i.e. evaluating the salience of a *restricted number* of possible periodicities Parncutt [1994], or by

computing a periodicity function, i.e. generating a continuous function plotting pulse salience versus pulse period (or frequency) with the help of e.g. the Fourier transform, Wavelet transforms, the autocorrelation function, bank of comb filters, etc.

In pulse induction, a fundamental assumption is made: The pulse period (and phase) is *stable* over the data used for its computation. That is, there is no speed variation in that part of the musical performance used for inducing a pulse. In that part of the data, remaining timing deviations (if any) are assumed to be short-time ones (considered as either errors or expressiveness features). They are either "smoothed out", by considering tolerance intervals or smoothing windows, or cautiously handled in order to derive patterns of systematic short-time timing deviations as e.g. the swing (see on page 410).

Another step is needed to output a discrete pulse period (and optionally its phase) rather than a continuous periodicity function, this is usually achieved by a peak-picking algorithm.

Pulse tracking Pulse tracking and pulse induction often occur as complementary processes. Pulse induction models consider short term timing deviations as noise, assuming a relatively stable tempo, whereas a pulse tracker handles the short term timing deviations and attempts to determine changes in the pulse period and phase, without assuming that the tempo remains constant. Another difference is that induction models work bottom-up, whereas tracking models tend to follow top-down approaches, for example, driven by the pulse period and phase computed by the pulse induction module, tracking is often a process of reconciliation between predictions (driven by previous period and phase computations) and the observed data. Diverse formalisms and techniques have been used in the design of pulse trackers: rule-based Desain and Honing [1999], problem-solving Allen and Dannenberg [1990], agents Dixon [2001], adaptive oscillators Large and Kolen [1994], dynamical systems Cemgil et al. [2001], Bayesian statistics Raphael [2002] and particle filtering Hainsworth and Macleod [2004]. A complete review can be found in Gouyon and Dixon [2005].

Some systems rather address pulse tracking by repeated induction e.g. Scheirer [2000], Laroche [2003], Klapuri et al. [2005]. A pulse is induced on a short analysis window (e.g. around 5 s of data), then the window is shifted in time and another induction takes place. Determining the tempo evolution then amounts to connecting the observations at each step. In addition to computational overload, one problem that arises with this approach to tracking is the lack of continuity between successive observations and the difficulty to model sharp tempo changes.

Quantization and time signature determination Few algorithms for time signature determination exist. The simplest approach is based on parsing the peaks of the periodicity function to find two significant peaks, which correspond respectively to a fast pulse, the time signature denominator, and a slower pulse, the numerator Brown [1993]. The ratio between the pulse periods defines the time signature. Another approach is to consider all pairs of peaks as possible beat/measure combinations, and compute the fit of all periodicity peaks to each hypothesis Dixon et al. [2003]. Another strategy is to break the problem into several stages: determining the time signature denominator (e.g. by tempo induction and tracking), segmenting the musical data with respect to this pulse and compute features at this temporal scope and finally detecting periodicities in the created feature lists Gouyon and Herrera [2003].

Quantization (or "rhythm parsing") can be seen as a by-product of the induction of several metrical levels, which together define a metrical grid. The rhythm of a given onset sequence can be parsed by assigning each onset (independently of its neighbors) to the closest element in this hierarchy. The weaknesses of such an approach are that it fails to account for musical context (e.g. a triplet note is usually followed by 2 more) and distortions of the metrical structure. Some models as Desain and Honing [1989] do account for musical context and possible distortions of the metrical structure. However such distortions would in turn be easier to determine if the quantized durations were known Allen and Dannenberg [1990]. Therefore, rhythm parsing is often considered as a process *simultaneous* with tempo tracking, rather than subsequent to it (hence the bi-directional arrow between these two modules in Figure 10.6 on page 408), see e.g. Raphael [2002] and Cemgil and Kappen [2003].

Systematic deviation characterization In the pulse induction process, short-term timing deviations can be "smoothed out" or cautiously handled so as to derive patterns of short-term timing deviations such as *swing*: a long-short timing pattern of consecutive eight-notes. For instance, Laroche [2001] proposes to estimate the swing jointly with tempo and beats at the half-note level, assuming constant tempo: all pulse periods, phases and eight-note long-short patterns are enumerated and a search procedure determines which best match the onsets.

Rhythmic pattern determination Systematic short-term timing deviations are important musical features. In addition, repetitive rhythmic patterns covering a *longer* temporal scope can also be characteristic of some music styles. For instance, many electronic synthesizers feature templates of prototypical patterns such as Waltz, Cha Cha and the like. The length of such patterns

is typically one bar, or a couple or them. Few algorithms have been proposed for the automatic extraction of rhythmic patterns; they usually require the knowledge (or previous extraction) of part of the metrical structure, typically the beats and measure Dixon et al. [2004].

Periodicity features Other rhythmic features, with a musical meaning less explicit than e.g. the tempo or the swing, have recently been advocated, in particular in the context of designing rhythm similarity distances. Most of the time, these features are derived from a parameterization of a periodicity function, as e.g. the salience of several prominent peaks Gouyon et al. [2004], their positions Tzanetakis and Cook [2002], Dixon et al. [2003], selected statistics (high-order moments, flatness, etc.) of the periodicity function considered as a probability density function Gouyon et al. [2004] or simply the whole periodicity function itself Foote et al. [2002].

Future research directions

Current research in rhythm description addresses all of these aspects, with varying degrees of success. For instance, determining the tempo of music with minor speed variations is feasible for almost all musical styles, if we do not insist that the system finds a specific metrical level Gouyon et al. [in press]. Recent pulse tracking systems also reach high levels of accuracy. On the other hand, accurate quantization, score transcription, determination of time signature and characterization of intentional timing deviations are still open questions. Particularly, it remains to be seen how well recently proposed models generalize to different musical styles. New research directions include the determination of highly abstract rhythmic features required for music content processing and music information retrieval applications, the definition of the best rhythmic features and the most appropriate periodicity detection method.

10.2.6 Genre

Most music can be described in terms of dimensions such as melody, harmony, rhythm, etc. These high-level features characterize music and at least partially determine its genre, but, as mentioned in previous sections, they are difficult to compute automatically from audio. As a result, most audio-related music information retrieval research has focused on low-level features and induction algorithms to perform genre classification tasks. This approach has met with some success XXXREF TO WIDMER CHAPTERXXX, but it is limited by the fact that the low level of

representation may conceal many of the truly relevant aspects of the music. See XXXREF TO WIDMER CHAPTERXXX for a review and more information on promising directions in genre classification.

10.3 Audio content exploitation

We consider in this section a number of applications of content-based descriptions of audio signals. Although audio retrieval (see page 412) is the one that has been addressed most often, others deserve a mention, as content-based transformations (see page 418).

10.3.1 Content-based search and retrieval

Searching a repository of musical pieces can be greatly facilitated by automatic description of audio and musical content (as e.g. fingerprints, melodic features, tempo, etc.).

A content-based music retrieval system is a search engine at the interface of a repository, or organized database, of musical pieces. Typically,

1. it receives a query, defined by means of musical strategies (e.g. humming, tapping, providing an audio excerpt or some measures of a score) or textual strategies (e.g. using "words" and/or "numbers" that describe some musical feature like tempo, mood, etc.) referring to audio or musical descriptors,
2. it has access to the set of musical features extracted from the musical files in the repository,
3. it returns a list of ranked files or excerpts that
   (a) are all relevant to the query (i.e. with high precision) or
   (b) constitute the set of all relevant files in the database (i.e. high recall),
4. and, optionally, it processes some user-feedback information in order to improve its performance in the future,

Identification

With fingerprinting systems it is possible to identify an unlabeled piece of audio and therefore provide a link to corresponding metadata (e.g. artist and song name). Depending on the application, different importance may be given to the following requirements:

Accuracy: The number of correct identifications, missed identifications, and wrong identifications (false positives).

Reliability: This is of major importance for copyright enforcement organizations.

Robustness: Ability to accurately identify an item, regardless of the level of compression and distortion or interference in the transmission channel. Other sources of degradation are pitching, equalization, background noise, D/A-A/D conversion, audio coders (such as GSM and MP3), etc.

Granularity: Ability to identify whole titles from excerpts a few seconds long. It requires to deal with shifting, that is lack of synchronization between the extracted fingerprint and those stored in the database and it adds complexity to the search (it needs to compare audio in all possible alignments).

Security: Vulnerability of the solution to cracking or tampering. In contrast with the robustness requirement, the manipulations to deal with are designed to fool the fingerprint identification algorithm.

Versatility: Ability to identify audio regardless of the audio format. Ability to use the same database for different applications.

Scalability: Performance with very large databases of titles or a large number of concurrent identifications. This affects the accuracy and the complexity of the system.

Complexity: It refers to the computational costs of the fingerprint extraction, the size of the fingerprint, the complexity of the search, the complexity of the fingerprint comparison, the cost of adding new items to the database, etc.

Fragility: Some applications, such as content-integrity verification systems, may require the detection of changes in the content. This is contrary to the robustness requirement, as the fingerprint should be robust to content-preserving transformations but not to other distortions.

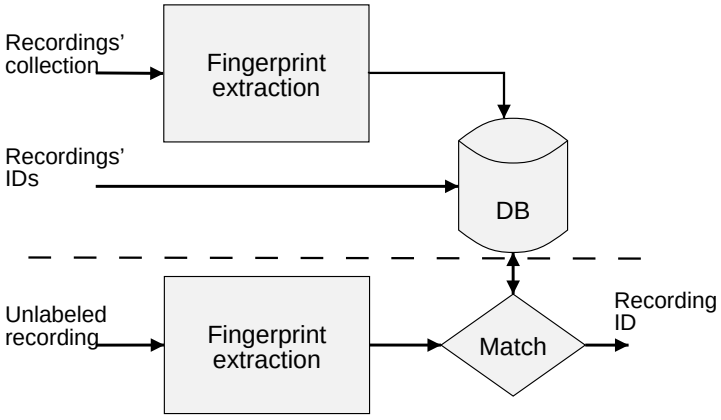

Figure 10.7: Content-based audio identification framework

The requirements of a complete fingerprinting system should be considered together with the fingerprint requirements listed in section 10.2.3. Bear in mind that improving a certain requirement often implies losing performance in some other.

The overall identification process mimics the way humans perform the task. As seen in Figure 10.7, a memory of the recordings to be recognized is created off-line (top); in the identification mode (bottom), unlabeled audio is presented to the system to look for a match.

Audio Content Monitoring and Tracking One of the commercial usages of audio identification is that of remotely controlling the times a piece of music has been broadcasted, in order to claim the broadcaster for doing the proper clearance of the involved rights Cano et al. [2002b].

Monitoring at the distributor end Content distributors may need to know whether they have the rights to broadcast certain content to consumers. Fingerprinting helps identify unlabeled audio in TV and radio channels repositories. It can also identify unidentified audio content recovered from CD plants and distributors in anti-piracy investigations (e.g. screening of master recordings at CD manufacturing plants).

Monitoring at the transmission channel In many countries, radio stations must pay royalties for the music they air. Rights holders are eager to monitor radio transmissions in order to verify whether royalties are being properly paid. Even in countries where radio stations can freely air music, rights holders are interested in monitoring radio transmissions for statistical

purposes. Advertisers are also willing to monitor radio and TV transmissions to verify whether commercials are being broadcast as agreed. The same is true for web broadcasts. Other uses include chart compilations for statistical analysis of program material or enforcement of cultural laws (e.g. in France a certain percentage of the aired recordings must be in French). Fingerprinting-based monitoring systems can be used for this purpose. The system "listens" to the radio and continuously updates a play list of songs or commercials broadcast by each station. Of course, a database containing fingerprints of all songs and commercials to be identified must be available to the system, and this database must be updated as new songs come out. Examples of commercial providers of such services are: `http://www.bdson-line.com`, `http://www.musicreporter.net`, `http://www.audiblemagic.com` and `http://www.yacast.fr`.

Additionally, audio content can be found in Web pages and Web-based Peer-to-Peer networks. Audio fingerprinting combined with a web crawler can identify their content and report it to the corresponding right owners (e.g. `http://www.baytsp.com`).

Monitoring at the consumer end In usage-policy monitoring applications, the goal is to avoid misuse of audio signals by the consumer. We can conceive a system where a piece of music is identified by means of a fingerprint and a database is contacted to retrieve information about the rights. This information dictates the behavior of compliant devices (e.g. CD and DVD players and recorders, MP3 players or even computers) in accordance with the usage policy. Compliant devices are required to be connected to a network in order to access the database.

Added-value services Some systems store metadata related to audio files in databases accessible through the internet. Such metadata can be relevant to a user for a given application and covers diverse types of information related to an audio file: e.g. how it was composed and how it was recorded, the composer, year of composition, the album cover image, album price, artist biography, information on the next concerts, etc. Fingerprinting can then be used to identify a recording and retrieve the corresponding metadata. For example, MusicBrainz (`http://www.musicbrainz.org`), Id3man (`http://www.id3man.com`) or Moodlogic (`http://www.moodlogic.com`) automatically label collections of audio files; the user can download a compatible player that extracts fingerprints and submits them to a central server from which metadata associated to the recordings is downloaded. Gracenote (`http://www.gracenote.com`) recently enhanced their technology based on CDs' table of contents with audio fingerprinting.

Another application consists in finding or buying a song while it is being broadcast, by

means of mobile-phone transmitting its GPS-quality received sound (e.g. `http://www.shazam.com`).

Summarization

Summarization, or thumbnailing, is essential for providing fast-browsing functionalities to content processing systems. An audiovisual summary that can be played, skipped upon, replayed or zoomed can save time to the user and help him/her to get a glimpse of "what the music is about," especially when using personal media devices. Music summarization consists in determining the key elements of a musical sound file and rendering them in the most efficient way. There are two tasks here: first extracting structure, and then creating aural and visual representations of this structure. Extracting a good summary from a sound file needs a comprehensive description of its content, plus some perceptual and cognitive constraints to be derived from users. An additional difficulty here is that different types of summaries can coexist, and that different users will probably require different summaries. Because of this amount of difficulty, the area of music summarization is still very underdeveloped, a review of recent promising approaches can be found in Ong and Herrera [2004].

Play-list generation

This area concerns the design of lists of music pieces that satisfy some ordering criteria, with respect to content descriptors previously computed, indicated (explicitly or implicitly) by the listener. Play-list generation is usually constrained by time-evolving conditions (i.e. "start with slow-tempo pieces, then progressively increase tempo") Pachet et al. [2000]. Besides the play-list construction problem, we can also mention related problems such as achieving seamless transitions (in user-defined terms such as tempo, tonality, loudness) between the played pieces.

Music browsing and recommendation

Music browsing and recommendation are very demanded functionalities, especially among youngsters. Recommendation consists in suggesting, providing guidance, or advising a potential consumer, in order he/she may find an interesting musical file in e.g. an on-line music store.

Nowadays, this is possible by querying artist or song names (or other types of editorial data such as genre), or by browsing recommendations generated by collaborative filtering, i.e.

using recommender systems that exploit information of the type "users that bought this album also bought this album." An obvious drawback of the first approach is that consumers need to know the name of the song or the artist beforehand. The second approach is only suitable when a number of consumers has heard and rated the music. This situation makes it difficult for users to access and discover the vast amount of music composed and performed by unknown artists which is available in an increasing number of sites (e.g. `http://www.magnatune.com`) and which nobody yet rated nor described.

Content-based methods represent an alternative to these approaches. User musical preferences can be estimated for recommendation purposes by automatically analyzing (e.g. via background processes) the songs they store on hardware, or e.g. those they listen more often. See Cano et al. [2005b] for the description of a recent large-scale music browsing and recommendation system based on automatic description of music content.[7]

It is reasonable to assume that these different approaches will merge in the near future and result in improved music browsing and recommendation systems.

Content visualization

The last decade has witnessed a great progress in the field of data visualization. Massive amounts of data can be represented in multidimensional graphs in order to facilitate comparisons, grasp the patterns and relationships between data, and improve our understanding of them. Four purposes of information visualization can be distingushed Hearst [1999]:

Exploration, where visual interfaces can also be used as navigation and browsing interfaces.

Computation, where images are used as tools for supporting the analysis and reasoning about information. Data insight is usually facilitated by good data visualizations.

Communication, where images are used to summarize what otherwise would need many words and complex concepts to be understood. Music visualization tools can be used to present concise information about relationships extracted from many interacting variables.

Decoration, where content data are used to create attractive pictures whose primary objective is not the presentation of information but the aesthetic amusement.

[7] See also `http://musicsurfer.iua.upf.edu`

It is likely that in the near future we witness an increasing exploitation of data visualization techniques in order to enhance song retrieval, collection navigation and music discovery.

10.3.2 Content-based audio transformations

Transformations of audio signals have a long tradition Zoelzer [2002]. A recent trend in this area of research is the editing and transformation of musical audio signals triggered by explicit musically-meaningful representational elements, in contrast to low-level signal descriptors. These recent techniques have been coined content-based audio transformations Amatriain et al. [2003], or adaptive digital audio effects Verfaille et al. [in print], and are based on the type of description of audio signals detailed above in section 10.2. In this section, we give examples of such techniques, following increasing levels of abstraction in the corresponding content description.

Loudness modifications

The most commonly known effects related to loudness are the ones that modify the sound intensity level: volume change, tremolo, compressor, expander, noise gate and limiter Verfaille et al. [in print].

However, when combined with other low-level features, loudness is correlated to higher-level descriptions of sounds, such as the timbre or the musical intentions of a performer. It can therefore be used as a means to control musically-meaningful aspects of sounds.

In the case of a piano, for example, it is possible to obtain the whole range of possible musical loudness values from the analysis of a single note Sola [1997]. In Sola [1997], the best implementation is based on taking the highest possible dynamic as a starting point, the remaining lower loudness values being obtained by subtraction of higher-range spectral information.

In the case of the singing voice, some studies have been carried out and are summarized by Sundberg in Sundberg [1987]. Using Sundberg's nomenclature, it is possible, under certain conditions, to infer the source spectrum modifications from uttering the same vowel at different loudness of phonation. Building upon this assumption, Fabig and Janer [2004] propose a method for modifying the loudness of the singing voice by detecting automatically the excitation slope.

Time scaling

In a musical context, time scaling can be understood changing the pace of a musical signal, its tempo. If a musical performance is time-scaled to a different tempo, we should expect to listen to the same notes starting at a scaled time pattern, but with durations modified linearly according to the tempo change. The pitch of the notes should however remain unchanged, as well as the perceived expression. Thus, for example, vibratos should not change their depth, tremolo or rate characteristics. And of course, the audio quality should be preserved in such a way that if we had never listened to that musical piece, we would not be able to know if we were listening to the original recording or to a transformed one.

Time-scale modifications can be implemented in different ways. Generally, algorithms are grouped in three different categories: time domain techniques, phase-vocoder and variants, and signal models. In the remainder of this section we explain the basics of these approaches in turn.

Time domain techniques Time domain techniques are the simplest methods for performing time-scale modification. The simplest (and historically first) technique is the variable speed replay of analog audio tape recorders McNally [1984] A drawback of this technique is that during faster playback, the pitch of the sound is raised while the duration is shortened. On the other hand, during slower playback, the pitch of the sound is lowered while the duration is lengthened. Many papers show good results without scaling frequency by segmenting the input signal into several windowed sections and then placing these sections in new time locations and overlapping them to get the time scaled version of the input signal. This set of algorithms is referred to as Overlap-Add (OLA). To avoid phase discontinuities between segments, the synchronized OLA algorithm (SOLA) uses a cross-correlation approach to determine where to place the segment boundaries Wayman et al. [1989]. In TD-PSOLA Moulines et al. [1989], the overlapping operation is performed pitch-synchronously to achieve high quality time-scale modification. This works well with signals having a prominent basic frequency and can be used with all kinds of signals consisting of a single signal source. When it comes to a mixture of signals, this method will produce satisfactory results only if the size of the overlapping segments is increased to include a multiple of cycles thus averaging the phase error over a longer segment making it less audible. More recently, WSOLA Verhelst and Roelands [1993] uses the concept of waveform similarity to ensure signal continuity at segment joints, providing high quality output with high algorithmic and computational efficiency and robustness. All the aforementioned

techniques consider equally transient and steady state parts of the input signal, thus time-scale them both in the same way. To get better results, it is preferable to detect the transients regions and not time-scale them, just translate them into a new time position, while time-scaling the non-transient segments. The earliest mention of this technique can be found in the *Lexicon 2400* time compressor/expander from 1986. This model detected transients, and only time-scales the remaining audio using TD-PSOLA style algorithm. In Lee et al. [1997] it is shown that using time-scale modification on only non-transient parts of speech improves the intelligibility and quality of the resulting time-scaled speech.

Phase vocoder and variants The phase-vocoder is a relative old technique that dates from the 70's Portnoff [1976]. It is a frequency domain algorithm computationally quite more expensive than time domain algorithms. However it can achieve high-quality results even with high time-scale factors. Basically, the input signal is split into many frequency channels, uniformly spaced, usually using FFT. Each frequency band (bin) is decomposed into magnitude and phase parameters, which are modified and re-synthesized by the IFFT or a bank of oscillators. With no transformations, the system allows a perfect reconstruction of the original signal. In the case of time-scale modification, the synthesis hop size is changed according to the desired time-scale factor. Magnitudes are linearly interpolated and phases are modified in such a way that phase consistency are maintained across the new frame boundaries. The phase-vocoder introduces signal smearing for impulsive signals due to the loss of phase alignment of the partials.

A typical drawback of the phase vocoder is the loss of vertical phase coherence that produces reverberation or loss of presence in the output. This effect is also referred to as phasiness. Recently, the synthesis quality has been improved applying phase-locking techniques Laroche and Dolson [1999] among bins around spectral peaks. Note that adding peak tracking to the spectral peaks, the phase-vocoder resembles the sinusoidal modeling algorithms, which is introduced in the next paragraph.

Another traditional drawback of the phase vocoder is the bin resolution dilemma: the phase estimates are incorrect if more than one sinusoidal peak reside within a single spectral bin. Increasing the window may solve the phase estimation problem, but it implies a poor time resolution and smooths the fast frequency changes. And the situation gets worse in the case of polyphonic music sources because then the probability is higher that sinusoidal peaks from different sources will reside in the same spectrum bin. A recent technology allows different temporal resolutions at different frequencies by a convolution of the spectrum with a variable

kernel function Hoek [1999]. Thus, long windows are used to calculate low frequencies, while short windows are used to calculate high frequencies. Other approaches approximate a constant-Q phase-vocoder based on wavelet transforms or nonuniform sampling.

Techniques based on signal models Signal models have the ability to split the input signal into different components which can be parameterized and processed independently giving a lot of flexibility for transformations. Typically these components are sinusoids, transients and noise. In sinusoidal modeling McAulay and Quatieri [1986] the input signal is represented as a sum of sinusoids with time-varying amplitude, phase and frequency. Parameter estimation can be improved by using interpolation methods, signal derivatives and special windows. Time-scale using sinusoidal modeling achieves good results with harmonic signals, especially when keeping the vertical phase coherence. However it fails to successfully represent and transform noise and transient signals. Attacks are smoothed and noise sounds artificial. The idea of subtracting the estimated sinusoids from the original sound to obtain a residual signal is proposed in **?**, this residual can then be modeled as a stochastic signal. This method allows to split e.g. a flute sound into the air flow and the harmonics components, and to transform both parts independently. This technique successfully improves the quality of time-scale transformations but fails to handle transients, which are explicitly handled in Verma et al. [1997]. Then, all three components (sinusoidal, noise and transient) can be modified independently and re-synthesized. When time-scaling an input signal, transients can successfully be translated to new onset location, preserving their perceptual characteristics.

Timbre modifications

Timbre is defined as all those characteristics that distinguish two sounds of the same pitch, duration and loudness. As a matter of fact, timbre perception depends on many characteristics of the signal such as its instantaneous spectral shape and its evolution, the relation of its harmonics, and some other features related to the attack, release and temporal structure.

Timbre instrument modification can be achieved by many different techniques. One of them is to modify the input spectral shape by *timbre mapping*. Timbre mapping is a general transformation performed by warping the spectral shape of a sound by means of a mapping function $g(f)$, that maps frequencies of the transformed spectrum (f_y) to frequencies of the initial spectrum (f_x) via a simple equation $f_y = g(f_x)$.

Linear *scaling* (compressing or expanding) is a particular case of timbre mapping in which the mapping function pertains to the family $f_y = k * f_x$, where k is the scale factor, usually between 0.5 and 2. The timbre scaling effect resembles modifications of the size and shape of the instrument.

The *shifting* transformation is another particular case of the timbre mapping as well in which $g(f)$ can be expressed as $f_y = f_x + c$, where c is an offset factor.

Morphing Another way of accomplishing timbre transformations is to modify the input spectral shape by means of a secondary spectral shape. This is usually referred to as *morphing* or *cross-synthesis*. In fact, morphing is a technique with which, out of two or more elements, we can generate new ones with hybrid properties. In the context of video processing, morphing has been widely developed and enjoys great popularity in commercials, video clips and films where faces of different people change one into another or chairs mutate into e.g. elephants. Analogously, in the context of audio processing, the goal of most of the developed morphing methods has been the smooth transformation from one sound to another. Along this transformation, the properties of both sounds combine and merge into a resulting hybrid sound.

With different names, and using different signal processing techniques, the idea of audio morphing is well known in the computer music community Serra [1994], Slaney et al. [1996]. In most algorithms, morphing is based on the interpolation of sound parameterizations resulting from analysis/synthesis techniques, such as the short-time Fourier transform, linear predictive coding or sinusoidal models.

Voice timbre Whenever the morphing is performed by means of modifying a reference voice signal in matching its individuality parameters to another, we can refer to it as voice conversion. Some applications for the singing voice exist in the context of karaoke entertainment Cano et al. [2000], see also Amatriain et al. [2002] and Bonada [2005] on the related topics of gender change and unison choir generation, respectively.

Still for the particular case of voice, other finer-grained transformations exist to modify the timbre character (without resorting to a morphing between two spectral shapes): e.g. *rough*, *growl*, *breath* and *whisper* transformations.

Roughness in voice can come from different pathologies such as biphonia, or diplophonia, and can combine with many other voice tags such as “hoarse” or “creaky.” However here we

will refer to a rough voice as the one due to cycle to cycle variations of the fundamental frequency (jitter), and the period amplitude (shimmer). The most common techniques used to synthesize rough voices work with a source/filter model and reproduce the jitter and shimmer aperiodicities in time domain Childers [1990]. These aperiodicities can be applied to the voiced pulse-train excitation by taking real patterns that have been extracted from rough voice recordings or by using statistical models Schoentgen [2001]. Spectral domain techniques have also proved to be valid to emulate roughness Loscos and Bonada [2004].

Growl phonation is often used when singing jazz, blues, pop and other music styles as an expressive accent. Perceptually, growl voices are close to other dysphonic voices such as "hoarse" or "creaky," however, unlike these others, growl is always a vocal effect and not a permanent vocal disorder. According to Sakakibara et al. [2004], growl comes from simultaneous vibrations of the vocal folds and supra-glottal structures of the larynx. The vocal folds vibrate half periodically to the aryepiglottic fold vibration generating sub-harmonics. Growl effect can be achieved by adding these sub-harmonics in frequency domain to the original input voice spectrum Loscos and Bonada [2004]. These sub-harmonics follow certain magnitude and phase patterns that can be modeled from spectral analyzes of real growl voice recordings.

Breath can be achieved by different techniques. One is to increase the amount of the noisy residual component in those sound models in which there is a sinusoidal-noise decomposition. For sound models based on the phase-locked vocoder (see on page 419) a more breathy timbre can be achieved by filtering and distorting the harmonic peaks.

The *whisper* effect can be obtained by equalizing a previously recorded and analyzed template of a whisper utterance. The time behavior of the template is preserved by adding to the equalization the difference between the spectral shape of the frame of the template currently being used and an average spectral shape of the template. An "anti-proximity" filter may be applied to achieve a more natural and smoother effect Fabig and Janer [2004].

Rhythm transformations

In addition to tempo changes (see section 10.3.2), existing music editing softwares provide several rhythm transformation functionalities. For instance, any sequencer provides the means to adjust MIDI note timings to a metrical grid ("quantization") or a predefined rhythmic template. By doing a appropriate mapping between MIDI notes and audio samples, it is therefore possible to apply similar timing changes to audio mixes. But when dealing with general polyphonic

musical excerpts, without corresponding MIDI scores, these techniques cannot be applied. A few commercial applications implement techniques to transform the rhythm of general polyphonic musical excerpts, they are restricted to swing transformations, a review can be found in Gouyon et al. [2003].

A technique for swing transformation has also been proposed in Gouyon et al. [2003] which consists in a description module and a transformation module. The description module does onset detection and rhythmic analysis. Swing is relative to the length of consecutive eighth-notes, it is therefore necessary to determine the beat indexes of eighth-notes. But this is not sufficient, one must also describe the excerpt at a the next higher (slower) metrical level, the quarter-note, and determine the eighth-note "phase", that is, determine in a group of two eighth-notes which is the first one. Indeed, it is not at all the same to perform a long-short pattern as a short-long pattern. The existing ratio between consecutive eighth-notes is also be estimated. This ratio can be changed, in the transformation module, by shortening or lengthening the first eighth-notes of each quarter-note, and lengthening or shortening accordingly the second eighth-notes. This is done with time scaling techniques, see 10.3.2 for a review of such techniques. In Gouyon et al. [2003], time-scaling is done in real-time and can be controlled by a "User Swing Ratio," while playing back the audio file (in a loop), the user can continuously adjust the swing ratio.

Having found evidence for the fact that deviations occurring within the scope of the smallest metrical pulse are very important for musical expressiveness, Bilmes [1993] proposes additional rhythmic transformations based on a high-level description of the rhythmic content of audio signals.

Melodic transformations

Melodic transformations such as *pitch discretization to temperate scale* and *intonation* apply direct modifications to the fundamental frequency envelope. Arguably, these transformations may be considered low level transformations, however, they do change the way a high-level descriptor, namely the melody, is perceived by the listener.

Intonation transformations are achieved by stretching or compressing the difference between the analysis pitch envelope and a low pass filtered version of it. The goal of the transformation is to increase or decrease the sharpness of the note attack, it is illustrated on Figure 10.8.

Pitch discretization to temperate scale can be accomplished by forcing the pitch to take the

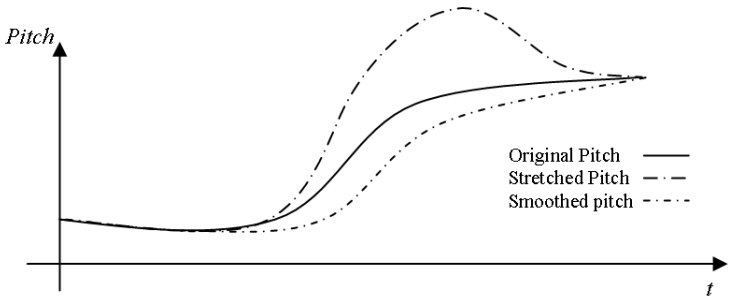

Figure 10.8: Intonation transformation.

nearest frequency value of the equal temperate scale. It is indeed a very particular case of pitch transposition where the pitch is quantified to one of the 12 semitones of an octave Amatriain et al. [2003].[8]

Other melodic transformations can be found in the software *Melodyne*[9] and in Gomez et al. [2003d] as *transposition* (global change of pitch), *horizontal symmetry*, in which the user can choose a pitch value (arbitrary or some global descriptor related to pitch distribution as minimum, maximum or mean pitch value of the melody) and perform a symmetric transformation of the note pitches with respect to this value on a horizontal axis, *contour direction changes* in which the user can change the interval direction without changing the interval depth (e.g. converting an ascending octave to a descending one), etc. Although these transformations are conceptually simple, they correspond to usual music composition procedures and can create dramatic changes that may enhance the original material (if used in the right creative context).

Finally, melodies of monophonic instruments can also be transformed by applying changes on other high-level descriptors in addition to the pitch, such as tempo curves Grachten et al. [2004] and note timing and loudness Ramirez et al. [2004], see XXXREF GOEBL CHAPTERXXX for more information on analysis and generation of expressive musical performances.

Harmony transformations

According to Amatriain et al. [2002], Verfaille et al. [in print], *harmonizing* a sound can be defined as mixing a sound with several pitch-shifted versions of it. This requires two parameters: the

[8] See also Antares' *Autotune*, `http://www.antarestech.com/`.
[9] `http://www.celemony.com/melodyne`

number of harmonies and the pitch for each of these. Pitches of the voices to generate are typically specified by the key and chord of harmonization. In case the key and chord is estimated from the analysis of the input pitch and the melodic context Pachet and Roy [1998], some refer to "intelligent harmonizing."[10]

An application of harmonizing in real-time monophonic solo voices is detailed in Amatriain et al. [2002].

10.4 Perspectives

All areas of high level description of musical audio signals will, without doubt, witness rapid progresses in the near future. We believe however that a critical element to foster these progresses lies in the systematic use of large-scale evaluations Cano et al. [2004].

Evaluations Developing technologies related to content processing of musical audio signals requires data. For instance, implementing algorithms for automatic instrument classification requires annotated samples of different instruments. Implementing a voice synthesis and transformation software calls for repositories of voice excerpts sung by professional singers. Testing a robust beat-tracking algorithm requires songs of different styles, instrumentation and tempi. Building models of musical content with a machine learning rationale calls for large amounts of data. Besides, running an algorithm on big amounts of diverse data is a requirement to ensure its quality and reliability.

In other scientific disciplines long-term improvements have shown to be bounded to systematic evaluation of models. For instance, text retrieval techniques significantly improved over the year thanks to the TREC initiative (see `http://trec.nist.gov`). TREC evaluations proceed by giving access to research teams to a standardized, large-scale test collection of text, a standardized set of test queries, and requesting a standardized way of generating and presenting the results. Different TREC tracks have been created over the years (text with moving images, web retrieval, speech retrieval, etc.) and each track has developed its own special test collections, queries and evaluation requirements. The standardization of databases and evaluation metrics greatly facilitated progress in the fields of Speech Recognition Przybocki and Martin

[10] see TC-Helicon's *Voice Pro*, `http://www.tc-helicon.com/VoicePro`.

[1989], Pearce and Hirsch [2000], Machine Learning Guyon et al. [2004] or Video Retrieval (see `http://www-nlpir.nist.gov/projects/trecvid/`).

In 1992, the visionary Marvin Minsky declared Minsky and Laske [1992]: "the most critical thing, in both music research and general AI research, is to learn how to build a common music database." More than 10 years later, this is still an open issue. Since a few years, the music content processing community has recognized the necessity to conduct rigorous and comprehensive evaluations Downie [2002, 2003b]. However, we are still far from having set a clear path to be followed for evaluating research progresses. Downie Downie [2003b] has listed the following urgent methodological problems to be addressed to by the research community:

1. there is no standard collection of music against which to test content description or exploitation techniques;
2. there are no standardized sets of performance tasks;
3. there are no standardized evaluation metrics.

As a first step, an audio description contest took place during the fifth edition of the ISMIR, in Barcelona, Spain, in October 2004. The goal of this contest was to compare state-of-the-art audio algorithms and systems relevant for some tasks of music content description, namely genre recognition, artist identification, tempo extraction and melody extraction Cano et al.. It is the first published large-scale evaluation of audio description algorithms, and the first initiative to make data and legacy metadata publicly available (see `http://ismir2004.ismir.net` for more details). However, it addresses a small part of the bulk of research going on in music content processing. Future editions of the ISMIR are likely to continue this effort and will certainly widen its scope, from evaluation of content description algorithms to evaluations of complete MIR systems.

Acknowledgments

This work has been partially funded by the European IST project 507142 SIMAC (Semantic Interaction with Music Audio Contents),[11] and the HARMOS E-Content project.[12] The authors

[11] `http://www.semanticaudio.org`
[12] `http://www.harmosproject.com`

wish to thank their colleagues in the Music Technology Group for their help. Thanks also to Simon Dixon (from ÖFAI) for participating in previous versions of part of the material of section 10.2.5.

Bibliography

P. Aigrain. New applications of content processing of music. *Journal of New Music Research*, 28 (4):271–280, 1999.

P. Allen and R: Dannenberg. Tracking musical beats in real time. In Allen P. and Dannenberg R., editors, *Proc. International Computer Music Conference*, pages 140–143, 1990.

X. Amatriain, J. Bonada, A. Loscos, J. Arcos, and V. Verfaille. Content-based transformations. *Journal of New Music Research*, 32(1):95–114, 2003.

X. Amatriain, J. Bonada, A. Loscos, and X. Serra. Spectral processing. In U. Zoelzer, editor, *DAFX Digital Audio Effects*, pages 373–439. Wiley & Sons, 2002.

X. Amatriain and P. Herrera. Transmitting Audio Content as Sound Objects. In *Proceedings of the AES 22nd Conference on Virtual, Synthetic, and Entertainment Audio*, Helsinki, 2001. Audio Engineering Society.

M. Basseville, , and I. V. Nikiforov. *Detection of abrupt changes: Theory and application*. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1993.

E. Batlle and P. Cano. Automatic segmentation for music classification using competitive hidden markov models. In *Proceedings of International Symposium on Music Information Retrieval*, 2000.

E. Batlle, J. Masip, and E. Guaus. Automatic song identification in noisy broadcast audio. In *Proceedings of the International Conference on Signal and Image Processing*, 2002.

J. P. Bello. *Towards the Automated Analysis of Simple Polyphonic Music: A Knowledge-Based Approach*. PhD thesis, Department of Electronic Engineering, Queen Mary University of London, 2003.

J. Bilmes. Techniques to foster drum machine expressivity. In *Proc. International Computer Music Conference*, 1993.

S. Blackburn. *Content based retrieval and navigation of music using melodic pitch contour*. PhD thesis, University of Southampton, 2000.

T. Blum, D. Keislar, J. Wheaton, and E. Wold. Method and article of manufacture for content-based analysis, storage, retrieval and segmentation of audio information, U.S. patent 5,918,223, June 1999.

R. Bod. Memory-based models of melodic analysis: challenging the gestalt principles. *Journal of New Music Research*, 30(3), 2001.

J. Bonada. Automatic technique in frequency domain for near-lossless time-scale modification of audio. In *Proceedings of International Computer Music Conference 2000*, Berlin, Germany, 2000.

J. Bonada. Voice solo to unison choir transformation. In *Proceedings of 118th Audio Engineering Society Convention*, 2005.

A. S. Bregman. *Auditory scene analysis*. MIT Press, Harvard, MA, 1990.

A. S. Bregman. Psychological data and computational auditory scene analysis. In D Rosenthal and H. G. Okuno, editors, *Computational auditory scene analysis*. Lawrence Erlbaum Associates, Inc., 1998.

G. J. Brown. *Computational auditory scene analysis: A representational approach*. PhD thesis, University of Sheffield, Department of Computer Science, 1992.

J. Brown. Determination of the meter of musical scores by autocorrelation. *Journal of the Acoustical Society of America*, 94(4):1953–1957, 1993.

M. S. Puckette Brown J. C. An efficient algorithm for the calculation of a constant q transform. *JASA*, pages 2698–2701, 1992.

D. Byrd. *Music notation by computer*. PhD thesis, Indiana University, 1984.

E. Cambouropoulos. The local boundary detection model and its application in the study of expressive timing. In *Proc. International Computer Music Conference*, 2001.

P. Cano. Fundamental frequency estimation in the SMS analysis. In *COSTG6 Conference on Digital Audio Effects (DAFX)*, 1998.

P. Cano, E. Batlle, T. Kalker, and J. Haitsma. A review of algorithms for audio fingerprinting. In *International Workshop on Multimedia Signal Processing*, 2002a.

P. Cano, E. Batlle, H. Mayer, and H. Neuschmied. Robust sound modeling for song detection in broadcast audio. In *Proc. AES 112th Int. Conv.*, 2002b.

P. Cano, O. Celma, M. Koppenberger, and J. Martin-Buldu. The topology of music artists' graphs. In *Proceedings XII Congreso de Fisica Estadistica*, 2005a.

P. Cano, E. Gomez, F. Gouyon, P. Herrera, M. Koppenberger, B. Ong, X. Serra, S. Streich, and N. Wack. ISMIR 2004 audio description contest. *submitted*.

P. Cano, M. Koppenberger, S. Ferradans, A. Martinez, F. Gouyon, V. Sandvold, V. Tarasov, and N. Wack. MTG-DB: A repository for music audio processing. In *Proceedings of 4th International Conference on Web Delivering of Music*, 2004.

P. Cano, M. Koppenberger, N. Wack, J. Garcia, J. Masip, O. Celma, D. Garcia, E. Gomez, F. Gouyon, E. Guaus, P. Herrera, J. Massaguer, B. Ong, M. Ramirez, S. Streich, and X. Serra. An industrial-strength content-based music recommendation system. In *Proceedings of 28th Annual International ACM SIGIR Conference*, 2005b.

P. Cano, A. Loscos, J. Bonada, M. de Boer, and X. Serra. Voice morphing system for impersonating in karaoke applications. In *Proceedings of International Computer Music Conference*, 2000.

M. A. Casey and A. Westner. Separation of mixed audio sources by independent subspace analysis. In *Proceedings of the International Computer Music Conference*, 2000.

A. Cemgil and B. Kappen. Monte carlo methods for tempo tracking and rhythm quantization. *Journal of Artificial Intelligence Research*, 18:45– 81, 2003.

A. Cemgil, B. Kappen, P. Desain, and H. Honing. On tempo tracking: Tempogram representation and Kalman filtering. *Journal of New Music Research*, 28(4):259–273, 2001.

D.G. Childers. Speech processing and synthesis for assessing vocal disorders. *IEEE Magazine on Engineering in Medicine and Biology*, 9:69–71, 1990.

L. Cohen. Time-frequency distributions - A review. *Processings of the IEEE*, 77(7), 1989.

G. W. Cooper and Meyer L. B. *The rhythmic structure of music*. University of Chicago Press, 1960.

K. de Koning and S. Oates. Sound base: Phonetic searching in sound archives. In *Proceedings of the International Computer Music Conference*, pages 433–466, 1991.

P. Desain and H. Honing. The quantization of musical time: A connectionist approach. *Computer Music Journal*, 13(3):55–66, 1989.

P. Desain and H. Honing. Computational models of beat induction: The rule-based approach. *Journal of New Music Research*, 28(1):29–42, 1999.

S. Dixon. Automatic extraction of tempo and beat from expressive performances. *Journal of New Music Research*, 30(1):39–58, 2001.

S. Dixon. Perceptual smoothness of tempo in expressively performed music. *Music Perception*, 23, 2005.

S. Dixon and E. Cambouropoulos. Beat tracking with musical knowledge. In *Proc. European Conference on Artificial Intelligence*, pages 626–630, 2000.

S. Dixon, F. Gouyon, and G. Widmer. Towards characterisation of music via rhythmic patterns. In *Proc. International Conference on Music Information Retrieval*, pages 509–516, 2004.

S. Dixon, E. Pampalk, and G. Widmer. Classification of dance music by periodicity patterns. In *Proc. International Conference on Music Information Retrieval*, 2003.

J. Downie. *The MIR/MDL evaluation project white paper collection*. Proceedings of the International Conference on Music Information Retrieval, 2002.

J. S. Downie. The musifind musical information retrieval project, phase II: User assessment survey. In *Proceedings of the 22nd Annual Conference of the Canadian Association for Information Science*, 1994.

J. S. Downie. Music information retrieval. *Annual Review of Information Science and Technology*, 37, 2003a.

J. S. Downie. The scientific evaluation of music information retrieval systems: foundations and the future. *Computer Music Journal*, 28(2), 2003b.

C. Drake, L. Gros, and A. Penel. How fast is that music? the relation between physical and perceived tempo. In S. Yi, editor, *Music, Mind and Science*. Seoul National University Press, 1999.

B. M. Eaglestone. A database environment for musician-machine interaction experimentation. In *Proceedings of the International Computer Music Conference*, pages 20–27, 1988.

B. M. Eaglestone and A. P. C. Verschoor. An intelligent music repository. In *Proceedings of the International Computer Music Conference*, pages 437–440, 1991.

D. P. W. Ellis. *Prediction-driven computational auditory scene analysis*. PhD thesis, Massachussetts Institute of Technology, 1996.

L. Fabig and J. Janer. Transforming singing voice expression - the sweetness effect. In *Proceedings of 7th International Conference on Digital Audio Effects*, 2004.

B. Feiten, R. Frank, and T. Ungvary. Organizing sounds with neural nets. In *Proceedings of the International Computer Music Conference*, pages 441–444, 1991.

J. Foote. Automatic audio segmentation using a measure of audio novelty. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 452–455, 2000.

J. Foote, M. Cooper, and U. Nam. Audio retrieval by rhythmic similarity. In *International Conference on Music Information Retrieval*, 2002.

T. Fujishima. Realtime chord recognition of musical sound: a system using common lisp music. In *International Computer Music Conference*, pages 464–467, 1999.

D. Godsmark and G. J. Brown. A blackboard architecture for computational auditory scene analysis. *Speech Communication*, 27:351–366, 1999.

B. Gold and L. Rabiner. Parallel processing techniques for estimating pitch periods of speech in the time domain. *Journal of the Acoustic Society of America*, 46:442–448, 1969.

L. Gomes, P. Cano, E. Gomez, M. Bonnet, and E. Batlle. Audio watermarking and fingerprinting: For which applications? *Journal of New Music Research*, 32(1):65–82, 2003.

E. Gomez. Tonal description of polyphonic audio for music content processing. *INFORMS Journal on Computing, Special Issue on Computation in Music*, In print.

E. Gomez, J. P. Bello, M. Davies, D. Garcia, F. Gouyon, C. Harte, P. Herrera, C. Landone, K. Noland, B. Ong, V. Sandvold, S. Streich, and B. Wang. Front-end signal processing and low-level descriptors computation module. Technical Report D2.1.1, SIMAC IST Project, 2005.

E. Gomez, F. Gouyon, P. Herrera, and X. Amatriain. Mpeg-7 for content-based music processing. In *Proceedings of the 4th WIAMIS-Special session on Audio Segmentation and Digital Music*, 2003a.

E. Gomez, F. Gouyon, P. Herrera, and X. Amatriain. Using and enhancing the current mpeg-7 standard for a music content processing tool. In *Proceedings of the 114th AES Convention*, 2003b.

E. Gomez, A. Klapuri, and B. Meudic. Melody description and extraction in the context of music content processing. *Journal of New Music Research*, 32(1):23–40, 2003c.

E. Gomez, G. Peterschmitt, X. Amatriain, and P. Herrera. Content-based melodic transformations of audio for a music processing application. In *Proceedings of 6th International Conference on Digital Audio Effects*, 2003d.

J. W. Gordon. *Perception of attack transients in musical tones*. PhD thesis, CCRMA, Stanford University, 1984.

M. Goto. A robust predominant-f0 estimation method for real-time detection of melody and bass lines in cd recordings. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages 757–760, 2000.

M. Goto. An audio-based real-time beat tracking system for music with or without drums. *Journal of New Music Research*, 30(2):159–171, 2001.

Y. Muraoka Goto, M. Real-time beat tracking for drumless audio signals: chord change detection for musical decisions. *Speech Communication*, (27):311–335, 1999.

F. Gouyon and S. Dixon. A review of automatic rhythm description systems. *Computer Music Journal*, 29(1):34–54, 2005.

F. Gouyon, S. Dixon, E. Pampalk, and G. Widmer. Evaluating rhythmic descriptors for musical genre classification. In *Proceedings of 25th International AES Conference*, 2004.

F. Gouyon, L. Fabig, and J. Bonada. Rhythmic expressiveness transformations of audio recordings: swing modifications. In *Proceedings of 6th International Conference on Digital Audio Effects*, 2003.

F. Gouyon and P. Herrera. Determination of the meter of musical audio signals: Seeking recurrences in beat segment descriptors. In *Proceedings of Audio Engineering Society, 114th Convention*, 2003.

F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano. An experimental comparison of audio tempo induction algorithms. *IEEE Transactions on Speech and Audio Processing*, in press.

F. Gouyon and B. Meudic. Towards rhythmic content processing of musical signals: Fostering complementary approaches. *Journal of New Music Research*, 32(1):41–64, 2003.

M. Grachten, J. Ll. Arcos, and R. Lopez de MÃ¡ntaras. Tempoexpress, a cbr approach to musical tempo transformations. In *Proceedings of the 7th ECCBR*, 2004.

I. Guyon, S. Gunn, A. Ben Hur, and G. Dror. Result analysis of the nips 2003 feature selection challenge. In *Proceedings of the Neural Information Processing Systems Conference*, 2004.

S. Hainsworth and M. Macleod. Particle filtering applied to musical tempo tracking. *EURASIP Journal on Applied Signal Processing*, 15:2385–2395, 2004.

H. Harb and L. Chen. Robust speech music discrimination using spectrum's first order statistics and neural networks. In *Proceedings of the 7th International Symposium on Signal Processing and Its Applications*, pages 125 – 128, 2003.

M. Hearst. User interfaces and visualization. In R. Baeza-Yates and B. Ribeiro-Neto, editors, *Modern information retrieval*. Harlow, Essex: ACM Press, 1999.

P. Herrera and J. Bonada. Vibrato extraction and parameterization in the spectral modeling synthesis framework. In *Proceedings of the Digital Audio Effects Conference*, 1998.

P. Herrera and E. Gomez. Study report on audio segmentation. Technical report, CUIDADO internal report, Music Technology Group, Pompeu Fabra University, 2001.

W. Hess. *Pitch Determination of Speech Signals. Algorithms and Devices.* Springer Series in Information Sciences. Springer-Verlag, Berlin, New York, Tokyo, springer-verlag edition, 1983.

S. M. J. Hoek. Method and apparatus for signal processing for time-scale and/or pitch modification of audio signals, U.S. patent 6266003, 1999.

H. Honing. From time to time: The representation of timing and tempo. *Computer Music Journal*, 25(3):50–61, 2001.

T. Jehan. *Musical signal parameter estimation*. MSc thesis, Institut de Formation Superieure en Informatique et Communication, Universite de Rennes I, France, 1997.

K. Jenssen. Envelope model of isolated musical sounds. In *Proceedings of the Digital Audio Effects Conference*, 1999.

I. Jermyn, C. Shaffrey, and N. Kingsbury. The methodology and practice of the evaluation of image retrieval systems and segmentation methods. Technical Report 4761, Institut National de la Recherche en Informatique et en Automatique, 2003.

K. Kashino, T. Kinoshita, and H. Tanaka. Organization of hierarchical perceptual sounds: music scene analysis with autonomous processing modules and a quantitative information integration mechanism. In *International Joint Conference On Artificial Intelligence*, Montreal, 1995.

K. Kashino and H. Murase. Sound source identification for ensemble music based on the music stream extraction. In *Proceedings of the International Joint Conference on Artificial Intelligence Workshop of Computational Auditory Scene Analysis*, pages 127–134, 1997.

M. Kassler. Toward musical information retrieval. *Perspectives of New Music*, 4, 1966.

D. Keislar, T. Blum, J. Wheaton, and E. Wold. Audio analysis for content-based retrieval. In *Proceedings of the International Computer Music Conference*, pages 199–202, 1995.

D. Kirovski and H. Attias. Beat-ID: Identifying music via beat analysis. In *5th IEEE Int. Workshop on Multimedia Signal Processing: special session on Media Recognition*, 2002.

A. Klapuri. Sound onset detection by applying psychoacoustic knowledge. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1999.

A. Klapuri. Qualitative and quantitative aspects in the design of periodicity estimation algorithms. In *European Signal Processing Conference*, 2000.

A. Klapuri. Automatic music transcription as we know it today. *Journal of New Music Research*, 33(3):269–282, 2004a.

A. Klapuri. *Signal Processing Methods for the Automatic Transcription of Music*. PhD thesis, Tampere University of Technology, Tampere, Finland, 2004b.

A. Klapuri, A. Eronen, and J. Astola. Analysis of the meter of acoustic musical signals. *IEEE Trans. Speech and Audio Processing*, 2005.

B. Kostek. Computer-based recognition of musica phrases using the rough-set approach. *Information Sciences*, 104:15–30, 1998.

R. Kronland-Martinet, J. Morlet, and Grossman. Analysis of sound patterns through wavelet transforms. *International Journal on Pattern Recognition and Artificial Intelligence*, 1(2), 1987.

C. L. Krumhansl. *Cognitive Foundations of Musical Pitch*. New York, 1990.

E. Lapidaki. *Consistency of tempo judgments as a measure of time experience in music listening*. PhD Thesis, Northwestern University, Evanston, IL, 1996.

E. Large and E. Kolen. Resonance and the perception of musical meter. *Connection Science*, 6: 177–208, 1994.

J. Laroche. Traitement des signaux audio-frequences. Technical report, Ecole National Supeieure de Telecommunications, 1995.

J. Laroche. Estimating tempo, swing and beat locations in audio recordings. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 135–138, 2001.

J. Laroche. Efficient tempo and beat tracking in audio recordings. *Journal of the Acoustical Society of America*, 51(4):226–233, 2003.

J. Laroche and M. Dolson. Improved phase-vocoder. time-scale modification of audio. *IEEE Transactions on Speech and Audio Processing*, 7:323–332, 1999.

S. Lee, H. D. Kin, and H. S. Kim. Variable time-scale modification of speech using transient information. In *Proceedings of the International Conference of Acoustics, Speech, and Signal Processing*, 1997.

M. Leman. Foundations of musicology as content processing science. *Journal of Music and Meaning*, 1(3), 2003. URL `http://www.musicandmeaning.net/index.php`.

F. Lerdahl and R. Jackendoff. *A generative theory of tonal music*. MIT Press, Cambridge, Massachusetts, 1983.

M. Lesaffre, M. Leman, K. Tanghe, B. De Baets, H. De Meyer, and J. P. Martens. User-dependent taxonomy of musical features as a conceptual framework for musical audio-mining technology. In *Proceedings of the Stockholm Music Acoustics Conference*, 2003.

M. S. Lew, N. Sebe, and J. P. Eakins. Challenges of image and video retrieval. In *Proceedings of the International Conference on Image and Video Retrieval*, 2002.

H. B. Lincoln. *Electronishe datenverarbeitung in der Musikwissenschaft*, chapter Some criteria and techniques for developing computerized thematic indices. Regensburg: Gustave Bosse Verlag, 1967.

A. Loscos and J. Bonada. Emulating rough and growl voice in spectral domain. In *Proceedings of 7th International Conference on Digital Audio Effects*, 2004.

E. Maestre and E. Gomez. Automatic characterization of dynamics and articulation of expressive monophonic recordings. In *Proceedings of 118th Audio Engineering Society Convention*, 2005.

R. C. Maher and J. W. Beauchamp. Fundamental frequency estimation of musical signals using a two-way mismatch procedure. *Journal of the Acoustic Society of America*, 95:2254–2263, 1993.

B. S. Manjunath, P. Salembier, and T. Sikora. *Introduction to MPEG-7: Multimedia Content Description Language*. Wiley and Sons, New York, 2002.

D. Marr. *Vision*. W.H. Freeman and Co., San Fransisco, 1982.

S. McAdams. Audition: physiologie, perception et cognition. In M. Robert J. Requin and M. Richelle, editors, *Traite de psychologie experimentale*, pages 283–344. Presses Universitaires de France, 1994.

S. McAdams and E. Bigand. *Thinking in Sound: The Cognitive Psychology of Human Audition*. Clarendon, Oxford, 1993.

R. J. McAulay and T. F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(4), 1986.

R. McNab, L. A. Smith, and I. H. Witten. Signal processing for melody transcription. In *Proceedings of the 19th Australasian Computer Science Conference*, 1996.

G. W. McNally. Variable speed replay of digital audio with constant output sampling rate. In *Proceedings 76th AES Convention*, 1984.

D. K. Mellinger. *Event formation and separation in musical sound*. PhD thesis, Stanford University Department of Computer Science, 1991.

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

M. Minsky and O. Laske. A conversation with Marvin Minsky. *AI Magazine*, 13(3):31–45, 1992.

B. Moore. *Hearing - Handbook of perception and cognition*. Academic Press, Inc., London, 2nd edition, 1995.

E. Moulines, F. Charpentier, and C. Hamon. A diphone synthesis system based on time-domain prosodic modifications of speech. In *Proceedings of the International Conference of Acoustics, Speech, and Signal Processing*, pages 238–241, 1989.

N. Nettheim. On the spectral analysis of melody. *Journal of New Music Research*, 21:135–148, 1992.

A. M. Noll. Cepstrum pitch determination. *Journal of the Acoustic Society of America*, 41:293–309, 1967.

B. S. Ong and P. Herrera. Computing structural descriptions of music through the identification of representative excerpts from audio files. In *Proceedings of 25th International AES Conference*, 2004.

A. Oppenheim and R. Schafer. From frequency to quefrency: A history of the cepstrum. *IEEE Signal Processing Magazine*, 21(5):95–106, 2004.

F. Pachet and P. Roy. Reifying chords in automatic harmonization. In *Proceedings of the ECAI Workshop on Constraints for Artistic Applications*, 1998.

F. Pachet, P. Roy, and D. Cazaly. A combinatorial approach to content-based music selection. *IEEE Multimedia*, 7:44–51, 2000.

N. R. Pal and S. K. Pal. A review of image segmentation techniques. *Pattern Recognition*, 26: 1277–1294, 1993.

R. Parncutt. A perceptual model of pulse salience and metrical accent in musical rhythms. *Music Perception*, 11(4):409–464, 1994.

D. Pearce and H. Hirsch. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proceedings of the International Conferences on Spoken Language Processing*, 2000.

G. Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. Technical report, CUIDADO IST Project, 2004.

S. Pfeiffer. The importance of perceptive adaptation of sound features in audio content processing. In *Proceedings of SPIE Storage and Retrieval for Image and Video Databases*, pages 328–337, 1999.

J. Pinquier, J.-L. Rouas, and R. Andre-Obrecht. A fusion study in speech/music classification. In *Proceedings of the International Conference on Multimedia and Expo*, pages 409–412, 2003.

M. R. Portnoff. Implementation of the digital phase vocoder using the fast fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24:243–248, 1976.

M. Przybocki and A. Martin. NIST speaker recognition evaluations. In *Proceedings of the International Conference on Language Resources and Evaluations*, pages 331–335, 1989.

H. Purwins, B. Blankertz, and K. Obermayer. A new method for tracking modulations in tonal music in audio data format. *Proceeding International Joint Conference on Neural Network*, pages 270–275, 2000.

H. Purwins, T. Graepel, B. Blankertz, and K. Obermayer. Correspondence analysis for visualizing interplay of pitch class, key, and composer. In E. Puebla, G. Mazzola, and T. Noll, editors, *Perspectives in Mathematical Music Theory*. Verlag, 2003.

L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–285, 1989.

L. R. Rabiner, M. R. Sambur, and C. E. Schmidt. Applications of a nonlinear smoothing algorithm to speech processing. *IEEE Trans. ASSP*, 23(6), 1975.

L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

R. Ramirez, A. Hazan, E. Gomez, and E. Maestre. A machine learning approach to expressive performance in jazz standards. In *Proceedings of 10th International Conference on Knowledge Discovery and Data Mining*, 2004.

C. Raphael. Automatic segmentation of acoustic musical signals using hidden markov models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(4):360–370, 1999.

C. Raphael. A hybrid graphical model for rhythmic parsing. *Artificial Intelligence*, 137(1-2): 217–238, 2002.

S. Rossignol. *Separation, segmentation et identification d'objets sonores. Application a la representation, a la manipulation des signaux sonores, et au codage dans les applications multimedias*. PhD thesis, IRCAM, Paris, France, 2000.

K. I. Sakakibara, L. Fuks, H. Imagawa, and N. Tayama. Growl voice in ethnic and pop styles. In *Proc. International Symposium on Musical Acoustics*, 2004.

E. D. Scheirer. *Music listening systems*. PhD thesis, Program in Arts and Sciences, MIT, 2000.

E. D. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Proceedings of the IEEE International Conference on Audio, Speech and Signal Processing*, pages 1331–1334, 1997.

J. Schoentgen. Stochastic models of jitter. *Journal of the Acoustical Society of America*, 109:1631–1650, 2001.

E. Selfridge-Field. *Conceptual and Representational Issues in Melodic Comparison*. Melodic Similarity: Concepts, Procedures, and Applications. MIT Press, Cambridge, Massachusetts, 1998.

X. Serra. *A System for Sound Analysis/Transformation/Synthesis based on a Deterministic plus Stochastic Decomposition*. PhD thesis, Stanford University, 1989.

X. Serra. Sound hybridization techniques based on a deterministic plus stochastic decomposition model. In *Proceedings of the ICMC*, 1994.

X. Serra and J. Bonada. Sound transformations based on the SMS high level attributes. In *Proceedings of COST G6 Conference on Digital Audio Effects 1998*, Barcelona, 1998.

D. Ellis Sheh, A. Chord segmentation and recognition using em-trained hidden markov models. In *Proceedings of ISMIR 2003 - 4rd International Conference on Music Information Retrieval*, 2003.

M. Slaney, M. Covell, and B. Lassiter. Automatic audio morphing. In *Proc. IEEE ICASSP*, pages 1001–1004, 1996.

P. Smaragdis. *Redundancy Reduction for Computational Audition, a Unifying Approach*. PhD thesis, Massachusetts Institute of Technology, 2001.

A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 2000.

J. Sola. *Disseny i implementacio d'un sintetitzador de piano*. MSc. Thesis, Universitat Politecnica de Catalunya, Barcelona, 1997.

C. Spevak, B. Thom, and K. Hothker. Evaluating melodic segmentation. In *Proc. International Conference on Music and Artificial Intelligence*, 2002.

J. Sundberg. *The Science of the Singing Voice*. Northern Illinois University Press, 1987.

E. Terhardt, G. Stoll, and M. Seewann. Algorithm for extraction of pitch and pitch salience from complex tonal signals. *Journal of the Acoustic Society of America*, 71:679–688, 1981.

H. Thornburg and F. Gouyon. A flexible analysis/synthesis method for transients. In *Proceedings of the International Computer Music Conference*, 2000.

P. Toiviainen and T. Eerola. A method for comparative analysis of folk music based on musical feature extraction and neural networks. In *Proc. VII International Symposium on Systematic and Comparative Musicology and III International Conference on Cognitive Musicology*, 2001.

M. Towsey, A. Brown, S. Wright, and J. Diederich. Towards melodic extension using genetic algorithms. *Educational Technology & Society*, 4(2), 2001.

G. Tzanetakis. *Manipulation, analysis and retrieval systems for audio signals*. PhD thesis, Computer Science Department, Princeton University, June 2002.

G. Tzanetakis and P. Cook. Multifeature audio segmentation for browsing and annotation. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 1999.

G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5), 2002.

A. L. Uitdenbogerd and J. Zobel. Manipulation of music for melody matching. In *ACM Multimedia*, 1998.

V. Verfaille, U. Zoelzer, and D. Arfib. Adaptive digital audio effects (A-DAFx): A new class of sound transformations. *IEEE Transactions on Speech and Audio Processing*, in print.

W. Verhelst and M. Roelands. An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech. In *Proceedings of the International Conference of Acoustics, Speech, and Signal Processing*, 1993.

T. S. Verma, S. N. Levine, and T. Meng. Transient modeling synthesis: a flexible analysis/synthesis tool for transient signals. In *Proceedings of the Internationtal Computer Music Conference*, 1997.

E. Vidal and A. Marzal. A review and new approaches for automatic segmentation of speech signals. In L. Torres, E. Masgrau, and Lagunas M. A., editors, *Signal Processing V: Theories and Applications*. 1990.

P. J. Walmsley, S. J. Godsill, and P. J. W. Rayner. Bayesian graphical models for polyphonic pitch tracking. In *Proc. Diderot Forum*, 1999.

J. L. Wayman, R. E. Reinke, and D. L. Wilson. High quality speech expansion, compression, and noise filtering using the SOLA method of time scale modification. In *Proceedings of 23d Asilomar Conference on Signals, Systems and Computers*, volume 2, pages 714–717, 1989.

U. Zoelzer, editor. *DAFX Digital Audio Effects*. Wiley & Sons, 2002.