Chapter 9

Interactive Sound

Federico Avanzini

Department of Information Engineering, University of Padua

About this chapter

This chapter tries to trace a route that, starting from studies in ecological perception and action-perception loop theories, goes down to sound modelling and design techniques for interactive computer animation and virtual reality applications.

We do not intend to provide an in-depth discussion about different theories of perception. We rather review a number of studies from experimental psychology that we consider to be relevant for research in multimodal virtual environments and interfaces, and we argue that such research needs to become more aware of studies in ecological perception and multimodal perception.

The chapter starts with an analysis of relevant literature in perception, while sound modelling techniques and applications to multimodal interfaces and VR are addressed in the last part of the chapter. The technically inclined reader may turn the chapter upside-down and start reading the last sections, referring to the initial material when needed. Where necessary, we will make use of the notions about physics-based sound synthesis techniques reviewed in Chapter 8.

9.1 Introduction

Most of Virtual Reality (VR) applications make use of visual displays, haptic devices, and spatialised sound displays. Multisensory information is essential for designing immersive virtual worlds, as an individual's perceptual experience is influenced by interactions among sensory modalities. As an example, in real environments visual information can alter the haptic perception of object size, orientation, and shape (Welch and Warren, 1986). Similarly, being able to hear sounds of objects in an environment, while touching and manipulating them, provides a sense of immersion in the environment not obtainable otherwise (Hahn et al., 1998). Properly designed and synchronised haptic and auditory displays are likely to provide much greater immersion in a virtual environment than a high-fidelity visual display alone. Moreover, by skewing the relationship between the haptic and visual and/or auditory displays, the range of object properties that can be effectively conveyed to the user can be significantly enhanced.

The importance of multimodal feedback in computer graphics and interaction has been recognised for a long time (Hahn et al., 1998) and is motivated by our daily interaction with the world. Streams of information coming from different channels complement and integrate each other, with some modality possibly dominating over the remaining ones, depending on the task (Welch and Warren, 1986; Ernst and Bülthoff, 2004). Research in ecological acoustics (Gaver, 1993a,b) demonstrates that auditory feedback in particular can effectively convey information about a number of attributes of vibrating objects, such as material, shape, size, and so on (see also Chapter 10).

Recent literature has shown that sound synthesis techniques based on physical models of sound generation mechanisms allow for high quality synthesis and interactivity, since the physical parameters of the sound models can be naturally controlled by user gestures and actions. Sounds generated by solid objects in contact are especially interesting since auditory feedback is known in this case to provide relevant information about the scene (e.g. object material, shape, size). Sound models for impulsive and continuous contact have been proposed for example in the papers by van den Doel and Pai (1998) and by Avanzini et al. (2003). Physically-based sound models of contact have been applied by DiFilippo and Pai (2000) to the development of an audio-haptic interface for contact interactions.

A particularly interesting research direction is concerned with bimodal (auditory and haptic) perception in contact interaction. Starting from a classic work by Lederman (1979), many studies have focused on continuous contact (i.e. scraping or sliding) and have investigated the relative contributions of tactile and auditory information to judgments of roughness of both real surfaces (Lederman, 1979; Lederman et al., 2002; Guest et al., 2002) and synthetic haptic and auditory textures (McGee et al., 2002). Impulsive contact interactions (i.e. impact) are apparently less investigated. A few studies have investigated the effect of auditory feedback on haptic stiffness perception (DiFranco et al., 1997; Avanzini and Crosato, 2006). Again, results from ecological acoustics (Freed, 1990; Giordano, 2006) provide useful indications about which auditory cues are relevant to stiffness/hardness perception, and can be exploited in the design of synthetic sound feedback.

The chapter is organised as follows. In Section 9.2 we provide a concise overview of the ecological approach to perception, and we focus on the literature on ecological acoustics. Section 9.3 addresses the topic of multisensory perception and interaction, and introduces some powerful concepts like sensory combination/integration, embodiment and enaction, sensory substitution. Finally, Section 9.4 discusses recent literature on interactive computer animation and virtual reality applications with a focus on multimodal feedback and especially auditory feedback. We will emphasise the relevance of studies in ecological acoustics and multimodal perception in aiding the design of multimodal interfaces and virtual environments.

9.2 Ecological acoustics

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 provided by Gibson (1986) and Michaels and Carello (1981). 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 more established views. First, perception is an achievement of animal-environment systems, not simply animals (or their brains). What makes up the environment of a particular animal is part of this theory of perception. Second, the main purpose of perception is to guide action, so a theory of perception cannot ignore what animals do. The kinds of activities that a particular animal does, e.g. how it eats and moves, are part of this theory of perception.

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

A theory of direct perception, in contrast, argues that sensory stimulation is determined in such a way that there exists a 1:1 correspondence between patterns of sensory stimulation and the underlying aspects of physical reality (Gibson, 1986). This is a very strong assumption, since it basically says that reality is fully specified in the available sensory stimulation. Gibson (1986) provides the following example in the domain of visual perception, which supports, in his opinion, the direct perception theory. If one assumes that objects are isolated points in otherwise empty space, then their distances on a line projecting to the eye cannot be discriminated, as they stimulate the same retinal location. Under this assumption it is correct to state that distance is not perceivable by eye alone. However Gibson argues that this formulation is inappropriate for describing how we see. Instead he emphasises that the presence of a continuous background surface provides rich visual structure.

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

Energy flows and invariants

Consider the following problem in visual perception: how can a perceiver distinguish the motion of an object from his/her own motion? 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. A first example of additional input is the information on whether any muscle commands had been issued to move the eyes or the head or the legs. If no counter-acting motor command is detected, then object motion can be concluded; on the contrary, if such motor commands are present then this will allow the alternative conclusion of self-motion. When the observer is moved passively (e.g. in a train), other input must be taken into account: an overall (global) change in the pattern of light indicates self-motion, while a local change against a stationary background indicates object motion.

This argument opened a new field of research devoted to the study of the structure in changing patterns of light at a given point of observation: the optic flow. The goal of this research is to discover particular patterns, called invariants, which are relevant to perception and hence to action of an animal immersed in an environment. Perceivers exploit invariants in the optic flow, in order to effectively guide their activities. Carello and Turvey (2002) provide the following instructive example: a waiter, who rushes towards the swinging door of the restaurant kitchen, adjusts his motion in order to control the collision with the door: he maintains enough speed to push through the door, and at the same time he is slow enough not to hurt himself. In order for his motion to be effective he must know when a collision will happen and how hard the collision will be. One can identify structures in the optic flow that are relevant to these facts: these are examples of quantitative invariants.

The above considerations apply not only to visual perception but also to other senses, including audition (see Section 9.2.2). 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 many dimensions of energy. The general claim underlying this concept is that observers are not separately sensitive to structures in the optic and acoustic flows but, rather, observers are directly sensitive to patterns that extend across these flows, that is, to patterns in the global array.

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. Studies of this effect use audio-visual recordings in which the visual portion shows a speaker saying one syllable, while the audio track contains a different syllable. Observers are instructed to report the syllable that they hear, 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 wide interest in the McGurk effect arises in part from the need to explain why and how the final percept differs from the patterns in both the optic and acoustic arrays. In particular, Stoffregen and Bardy (2001) claim that the McGurk effect is consistent with the general idea 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 is unnatural (or at least uncommon), for observers who can both see and hear the speaker, to report only what they hear. 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 (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 behaviours are afforded. A stair with a certain proportion of a person's leg length affords climbing (is climbable); a surface which is rigid relative to the weight of an animal affords stance and traversal (is traversable); a ball which is falling with a certain velocity, relative to the speed that a person can generate in running toward it, affords catching (is catchable), and so on. Therefore, affordances are the possibilities for action of a particular animal-environment setting; they are usually described as "-ables", as in the examples above. What is important is that affordances are not determined by absolute properties of objects and environment, but depend on how these relate to the characteristics of a particular animal, e.g. size, agility, style of locomotion, and so on (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 behaviour. Instead, affordances are defined relative to the animal and refer to behaviour (i.e. they are animal-environment relations that afford some behaviour). The concept of affordance thus emphasises the relevance of activity to defining the environment to be perceived.

9.2.2 Everyday sounds and the acoustic array

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

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

Two companion papers by Gaver (1993a,b) have greatly contributed to the build-up of a solid framework for ecological acoustics. Specifically, Gaver (1993a) 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

Gaver (1993a) introduces the concept of everyday listening, as opposed to musical listening. When a listener hears a sound, she/he might concentrate on attributes like pitch, loudness, and timbre, and their variations over time. Or she/he might notice its masking effect on other sounds. Gaver refers to these as examples of musical listening, meaning that the considered perceptual dimensions and attributes have to do with the sound itself, and are those used in the creation of music.

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

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

Everyday listening needs a different theoretical framework, in order to understand listening and manipulate sounds along source-related dimensions instead of sound-related dimensions. Such a framework must answer two fundamental questions. First, it has to develop an account of ecologically relevant perceptual attributes, i.e. the features of events that are conveyed through listening. Thus the first question asked by Gaver (1993a) is: "What do we hear?". Second, it has to develop an ecological acoustics, that describes which acoustic properties of sounds are related to information about the sound sources. Thus the second question asked by Gaver (1993b) is: "How do we hear it?"

Acoustic flow and acoustic invariants

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

Sound also conveys information about the environment in which the event have occurred. In everyday conditions, a listener's ear is reached not only by the direct sound but also by the reflections of sound over various other objects in the environment, resulting in a coloration of the spectrum.

In addition, the transmitting medium also has an influence on sound signals: dissipation of energy, especially at high-frequency, increases with the path travelled by the sound waves and thus carries information about the distance of the source. Another example is the Doppler effect, which is produced when sound sources and listeners are in relative motion, and results in a shift of the frequencies. Changes in loudness caused by changes in distance from a moving sound source may provide information about time-to-contact in a fashion analogous to changes in visual texture. The result is an acoustic array, analogous to the optical array described previously.

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

Maps of everyday sounds

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

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

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

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

Although all sound-producing events involve any of the above categories (vibrating solids, aerodynamic, or liquid interactions), many also depend on complex patterns of simpler events. As an example, footsteps are temporal patterns of impact sounds. The perception of these patterned sounds is also related to the timing of successive events, (e.g. successive footstep sounds must occur within a range of rates and regularities in order to be perceived as walking). A slightly more complex example is a door slam, which involves the squeak of scraping hinges and the impact of the door on its frame. This kind of compound sounds involve mutual constraints on the objects that participate in related events: concatenating the creak of a heavy door closing slowly with the slap of a slammed light door would probably not sound natural.

Starting from these considerations, Gaver derived a tentative map of everyday sounds, which is shown in figure 9.1 and discussed in the following.

- Basic Level Sources: consider, for example, the region describing sounds made by vibrating solids. Four different sources of vibration in solids are indicated as basic level events: deformation, impacts, scraping and rolling.

- Patterned Sources involve temporal patterning of basic events. For instance, walking, as described above, but also breaking, spilling, and so on,

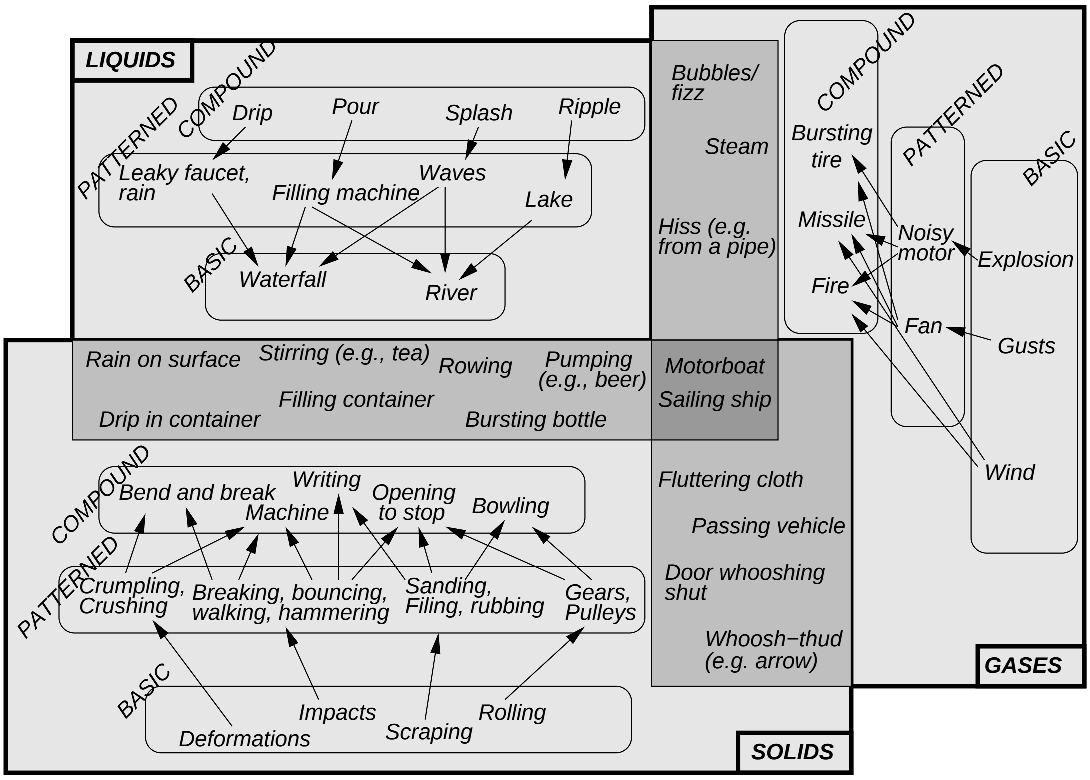

Figure 9.1: A map of everyday sounds. Complexity increases towards the center. Figure based on Gaver (1993a).

are all complex events involving patterns of simpler impacts. Similarly, crumpling or crushing are examples of patterned deformation sounds. In addition, other sorts of information are made available by their temporal complexity. For example, the regularity of a bouncing sound provides information about the symmetry of the bouncing object.

- Compound events involve more than one type of basic level event. An example is the slamming door discussed above. Other examples are the sounds made by writing, which involve a complex series of impacts and scrapes over time, while those made by bowling involve rolling followed by impact sounds.
- Hybrid events involve yet another level of complexity in which more

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

9.2.3 Relevant studies

Although still quite "young", the literature on ecological acoustics has produced a number of relevant results in the last 20 years. In the following we briefly review some of the most influential studies, classified according to the categorisation by Gaver discussed above: basic, patterned, compound, and hybrid sources. It has to be noted that most of these studies are concerned with sound events produced from interactions of solids objects, while sound-producing events that involve liquids and aerodynamic interactions have been addressed less frequently. A reason for this is probably that sounds from solids are especially interesting when talking about interaction: auditory feedback is frequently generated 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

Many studies have investigated the perception of object material from impact sounds. Wildes and Richards (1988) tried to find an acoustical parameter that could characterise material type independently from variations in other features (e.g. size or shape). Materials can be characterised using a coefficient of internal friction, which measures anelasticity (in ascending order of anelasticity we have steel, glass, wood and rubber). This coefficient is measurable using both the quality factor Q and the decay time t_e of vibration, the latter being the time required for amplitude to decrease to $1/e$ of its initial value. Decreasing anelasticity results in increasing Q and t_e.

Lutfi and Oh (1997) performed a study on material discrimination in synthetic struck clamped bar sounds. Stimuli were synthesised by varying elasticity and density of the bars, with values taken in the ranges of various metals, glass, and crystal. Perturbations on parameter values were applied either to all the frequency components together (lawful covariation) or independently to each of them (independent perturbation). Listeners were presented with a pair of stimuli, were given a target material (either iron or glass), and had to tell which of two presented stimuli was produced by the target materials. 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 et al. (2000) also investigated material discrimination in stimuli with variable frequency and internal friction. In a first experimental setup subjects were presented with pairs of stimuli and had to judge on a continuous scale the perceived difference in the materials. In another experiment they were presented with one stimulus and had to categorise the material using four response alternatives: rubber, wood, glass and steel. Results indicated that judgments of material difference were significantly influenced by both the friction coefficient and the fundamental frequency. An effect of both these variables was found in a categorisation 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.

Besides material, another relevant ecological dimension of impact sounds is the hardness of collision. Freed (1990) tried to relate hardness to some attack-related timbral dimensions. His stimuli were generated by percussing four cooking pans, with variable diameter, by means of 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 analysis of results showed that the useful information for mallet hardness rating was contained in the first 300 ms of the signals. Four acoustical indices were measured in this sound attack portion: 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 (time weighted average). These acoustical indices were used as predictors in a multiple regression analysis and were found to account for 75% of the variance of the ratings.

When we consider continuous contact (e.g. scraping) instead of impulsive contact, a relevant ecological dimension is surface roughness. In a classic study, Lederman (1979) compared the effectiveness of tactile and auditory information in judging the roughness of real surfaces. Roughness of aluminum plates was manipulated by varying the distance between adjacent grooves of fixed width, or by varying the width of the grooves. Subjects were given the task to rate numerically the roughness. In one condition participants only listened to the sounds generated by the experimenter by moving his fingertips on the plate. In a second condition subjects were asked to move their fingertips onto the plate while wearing cotton plugs and earphones. In a third condition they were able to hear the sounds they generated when touching the plate. Results showed that when both tactile and auditory information were present, the tactile one dominated in determining experimental performance. Roughness estimates were shown to increase as both the distance between grooves and the width of the grooves decreased. More recent research by Lederman and coworkers has focused on roughness perception when the surface is explored using a rigid probe rather than with the bare skin: as the probe provides a rigid link between the skin and the surface, vibratory roughness perception occurs in this case. Lederman et al. (2002) investigated relative contributions of haptic and auditory information to roughness judgments. Stimuli were obtained by asking subjects to explore with a probe a set of plates with periodic textures of varying inter-element spacings. Three conditions were used: touch-only, audition-only, and touch+audition. Results showed that, although dominance of haptic information was still found, sound played a more relevant role than in the case of direct contact with fingers. The authors argue that this may be due not only to the different interaction conditions, but also to the fact that the amplitude of the produced sounds is considerably greater for probe-based exploration than for bare skin contact.

The auditory perception of geometric properties of interacting objects has also been investigated. Carello et al. (1998) studied the recognition of the length of wood rods dropped on the floor. In their experiments subjects judged the perceived length by adjusting the distance of a visible surface in front of them. Subjects were found to be able to scale length of the rods consistently and physical length was found to correlate strongly with estimated length. 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 predicted length estimates better than actual length. Length estimates were then 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.

Patterned and compound sources

According to figure 9.1, patterned sound sources include bouncing, breaking, walking, and so on. Many of these everyday sounds have been investigated in the literature. Warren and Verbrugge (1988) studied acoustic invariants in bouncing and breaking events, and distinguished between two classes of invariants: structural invariants that specify the properties of the objects, and transformational invariants that specify their interactions and changes. Warren and Verbrugge investigated the nature of the transformational invariants that allow identification of breaking and bouncing events. On the basis of a physical analysis the authors hypothesised that the nature of these invariants was essentially temporal, static spectral properties having little or no role. 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 synthesised by superimposing four trains of damped quasi-periodic pulses, each one generated from a recorded frame of a bouncing glass sound, all with the same damping. The breaking event was synthesised by superimposing the same sequences, but using different damping coefficients for each of them. Identification performance was extremely accurate in all cases, despite the strong simplifications of the spectral and temporal profile of the acoustical signal. The transformational invariants for bouncing was then 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 (1987) reports a study on auditory perception of another patterned sound composed of impact events: hands clapping. In particular he hypothesised that subjects are able to recognise size and configuration of clapping hands from the auditory information. Recognition of hands size was also related to recognition of the gender of the clapper, given that males have in general bigger hands than females. In a first experiment, clapper gender and hand size recognition from recorded clapping sounds were investigated. 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 a second experiment subjects had to recognise the configuration of clapping hands. In this case performance was quite good: 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.

A study on gender recognition in walking sounds is reported by Li et al. (1991). Subjects were asked to categorise the gender of the walker on the basis of four recorded walking sequences. Results show that recognition levels are well above chance. Several anthropometric measures were collected on the walkers (height, weight and shoe size). Duration analysis on the recorded walking excerpts indicated that female and male walkers differed with respect to the relative duration of stance and swing phases, but not with respect to walking speed. Nonetheless judged maleness was significantly correlated with this latter variable, and 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 were then derived from principal components analysis, and were then used as predictors for both physical and judged gender. Overall male walkers were characterised by lower spectral centroid, mode and high frequency energy, and by higher skewness, kurtosis and low-frequency slope. These results were then tested in a further experiment. Stimuli were generated by manipulating the spectral mode of the two most ambiguous walking excerpts. 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 the previous studies, the work of Gygi et al. (2004) did not focus on a specific event or feature. Instead the authors use for their experiments a large (70) and varied catalogue of sounds, which covers "nonverbal human sounds, animal vocalisations, 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) are included, e.g. beer can opening, bowling, bubbling, toilet flushing, etc. The experiments applied to non-verbal sound an approach adopted in speech perception studies, namely the use of low-, high-, and bandpass filtered speech to assess the importance of various frequency regions for speech identification. The third experiment is perhaps the most interesting one. The authors seem to follow an approach already suggested by Gaver (1993b): "[. . .] 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.". Results from this experiment show that identifiability is heavily affected by experience and has a strong variability between sounds. The authors tried to quantify the relevance of temporal structures through a selection of time- and frequency-domain parameters, including statistics of the envelope (a measure of the envelope "roughness"), autocorrelation statistics (to reveal periodicities in the waveform), and moments of the long term spectrum (to see if some spectral characteristics were preserved when the spectral information was drastically reduced). Correlation of these parameters with the identification results showed that three variables were mainly used by listeners: number of autocorrelation 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.

9.3 Multimodal perception and interaction

9.3.1 Combining and integrating auditory information

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

Sensory combination and integration

Looking at how multisensory information is combined, two general strategies can be identified (Ernst and Bülthoff, 2004): the first is to maximise information delivered from the different sensory modalities (sensory combination). The second strategy is to reduce the variance in the sensory estimate to increase its reliability (sensory integration).

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

By contrast, sensory integration describes interactions between redundant signals. Ernst and Bülthoff (2004) illustrate this concept with an example:

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

There are a number of studies that show that vision dominates the integrated percept in many tasks, while other modalities (in particular audition and touch) have a less marked influence. This phenomenon of visual dominance is often termed visual capture. As an example, it is known that in the spatial domain vision can bias the perceived location of sounds whereas sounds rarely influence visual localisation. 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 (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. As an example, the visual modality usually dominates in spatial tasks, 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 (2004) note that the terminology modality precision and modality appropriateness can be misleading because it is not the modality itself or the stimulus that dominates: 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 in the context of this chapter since, as already noted, temporal aspects are especially related to audition.

Auditory capture and illusions

Psychology has a long history of studying intermodal conflict and illusions in order to understand mechanisms of multisensory integration. Much of the literature on multisensory perception has focused on spatial interactions: an example is the ventriloquist effect, in which the perceived location of a sound shifts towards a visual stimulus presented at a different position. Identity interactions are also studied: an example is the already mentioned McGurk effect (McGurk and MacDonald, 1976), in which what is being heard is influenced by what is being seen (for example, when hearing /ba/ but seeing the speaker say /ga/ the final perception may be /da/).

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

There is a variety of cross-modal effects that demonstrate that, outside the spatial domain, audition can bias vision. In a recent study, 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 by 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 by the authors 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.

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). In particular some authors have investigated whether audition can influence tactile perception similarly to what Shams et al. (2002) have done for vision and audition. 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. (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. (2005) 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 (2004) merely reflected an original but very specific illusion. In a first experiment, the auditory and tactile sequences were always presented simultaneously. Results showed 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. In a second experiment, the authors investigate how sensory integration is affected by manipulation of the timing between auditory and tactile sequences. Results showed 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. (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.

The studies discussed above 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. (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 authors have studied auditory-tactile integration in surface texture perception. Lederman and coworkers, already mentioned in Section 9.2.3, 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, with a consequent 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

A related study was conducted by Guest et al. (2002). In their experimental setup, participants had 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 particularly interesting from a sound-design perspective, since it investigates 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).

9.3.2 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 view of the relation between perception and action makes then two assumptions. First, the causal flow between perception and action is primarily one-way: perception is input from world to mind, action is output from mind to world, and thought (cognition) is the mediating process. Second, perception and action are merely instrumentally related to each other, so that each is a means to the other. If this kind of "input-output" picture is right, then it must be possible, at least in principle, to disassociate capacities for perception, action, and thought.

Although everyone agrees that perception depends on processes taking place in the brain, and that internal representations are very likely produced in the brain, more recent theories have questioned such a modular decomposition in which cognition interfaces between perception and action. The ecological approach discussed in Section 9.2 rejects the one-way assumption, but not the instrumental aspect of the traditional view, so that perception and action are seen as instrumentally interdependent. Others argue that a better alternative is to reject both assumptions: the main claim of these theories is that it is not possible to disassociate perception and action schematically, and that every kind of perception is intrinsically active and thoughtful: perception is not a process in the brain, but a kind of skillful activity on the part of the animal as a whole. As stated by Noë (2005), only a creature with certain kinds of bodily skills (e.g. a basic familiarity with the sensory effects of eye or hand movements, etc.) could be a perceiver.

One of the most influential contributions in this direction is due to Varela et al. (1991) (see O'Regan and Noë, 2001, for a detailed review of other works based on similar ideas). They presented an "enactive conception" of experience, which is not regarded as something that occurs inside the animal, but rather as something that the animal enacts as it explores the environment in

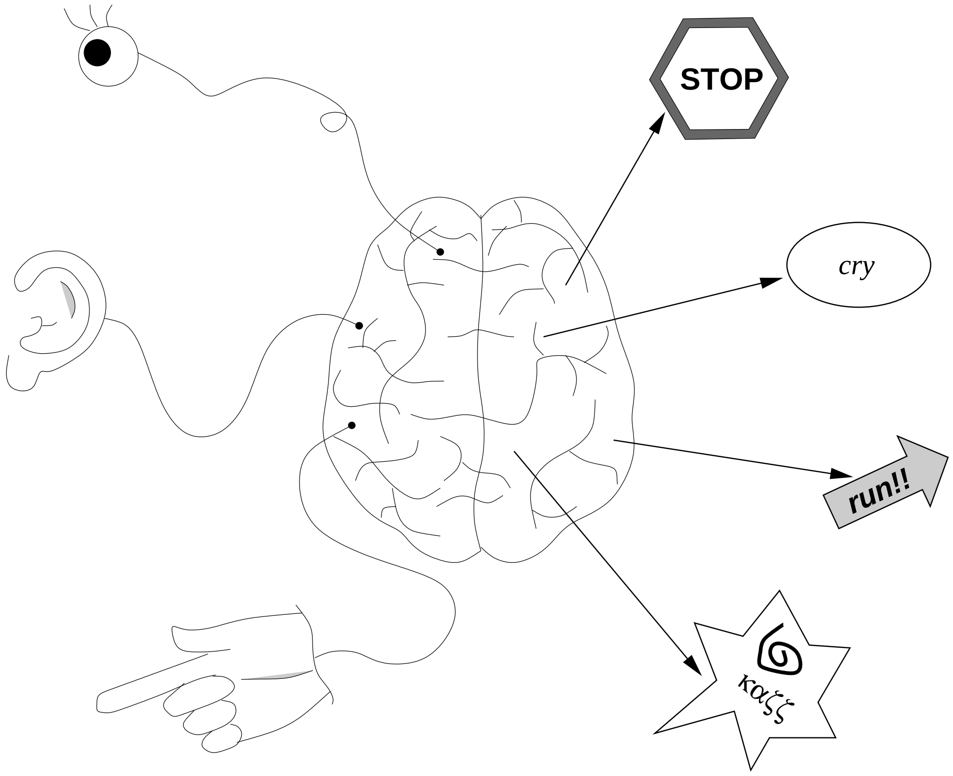

Figure 9.2: A cartoon representation of traditional views of the perception-action functions as a causal one-way flow.

which it is situated. In this view, the subject of mental states is the embodied, environmentally-situated animal. The animal and the environment form a pair in which the two parts are coupled and reciprocally determining. Perception is thought of in terms of activity on the part of the animal. The term “embodied” is used by the authors as a mean to highlight two points: first, cognition depends upon the kinds of experience that are generated from specific sensorimotor capacities. Second, these individual sensorimotor capacities are themselves embedded in a biological, psychological, and cultural context. Sensory and motor processes, perception and action, are fundamentally inseparable in cognition.

O’Regan and Noë (2001) have introduced closely related concepts, 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 characterised by a variety of such possible patterns of interaction with the sponge. The authors 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 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 is the extent to which that input can cause automatic orienting behaviours that capture cognitive processing resources. According to these definitions, vision, touch, hearing, and smell have not only high corporality but also high alerting capacity. 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. An immediate counter-objection is 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 Noë (2001) refer 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. In the "change blindness" experiments, however, conditions were arranged in such a way that these transients were hidden by superimposing a brief global flicker over the whole visual field at the moment of the change. It was shown that in this condition 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. The authors regard 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.

A related example, again in the domain of visual perception, is discussed by Noë (2005) who introduces 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 (caused by e.g. cataracts, retinal injury, and so on), 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. The author 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 loose their receptive power. A number of studies are reported by Noë (2005), which show that images stabilised 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" lies in the fact that, among other things, one is seeing if there is a large change in sensory input when blinking; on the other hand, one 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 is provided by 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). 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 supports 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.

A certain amount of studies investigate sensory substitution phenomena that involve audition. One research stream deals with 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 studied as prostheses for the blind. 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 down-converted 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 localised sound images corresponding to the direction and the size of the obstacles. Results showed that the auditory feedback was successfully used for the recognition of small obstacles, and also for discriminating between several obstacles at the same time without any virtual image.

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 people 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 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 (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 chosen 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. (1991), since it provides a representation of the entire scene rather than simply detecting obstacles and isolated objects. The approach followed by Mejer resembles closely the work by Bach-y-Rita, except that audition instead of tactile stimulation is used as 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 non-intuitive. Accordingly, Meijer (1992) remarks 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.

[1] The experience of a visually impaired user, who explicitly described herself as seeing with the visual-to-auditory substitution device, is reported at `http://www.seeingwithsound.com/tucson2002f.ram`

9.4 Sound modelling for multimodal interfaces

In this final section we discuss recent literature on interactive computer animation and virtual reality applications. All of these applications involve direct interaction of an operator with virtual objects and environments and require multimodal feedback in order to enhance the effectiveness of the interaction. We will especially focus on the role of auditory feedback, and will emphasise the relevance of studies in ecological acoustics and multimodal perception, which we have previously discussed, in aiding the design of multimodal interfaces and virtual environments.

The general topic of the use of sound in interfaces is also addressed in Chapter 10.

9.4.1 Interactive computer animation and VR applications

The need for multisensory feedback

Typical current applications of interactive computer animation and VR applications (Srinivasan and Basdogan, 1997) include medicine (surgical simulators for medical training, manipulation of micro and macro robots for minimally invasive surgery, remote diagnosis for telemedicine, aids for the disabled such as haptic interfaces for non-sighted people), entertainment (video games and simulators that enable the user to feel and manipulate virtual solids, fluids, tools, and avatars), education (e.g. interfaces giving students the feel of phenomena at nano, macro, or astronomical scales, "what if" scenarios for non-terrestrial physics, display of complex data sets), industry (e.g. CAD systems in which a designer can manipulate the mechanical components of an assembly in an immersive environment), and arts (virtual art exhibits, concert rooms, museums in which the user can log in remotely, for example to play musical instruments or to touch and feel haptic attributes of the displays, and so on). Most of the virtual environments (VEs) built to date contain complex visual displays, primitive haptic devices such as trackers or gloves to monitor hand position, and spatialised sound displays. However it is being more and more acknowledged that accurate auditory and haptic displays are essential in order to realise the full promise of VEs.

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 synchronizing even simple haptic and auditory displays with the visual one, than by increasing the complexity of the visual display alone. Moreover, by skewing the relationship between the haptic and visual and/or auditory displays, the range of object properties that can be effectively conveyed to the user can be significantly enhanced. Based on these considerations, many authors (see for example Hahn et al., 1998 and Srinivasan and Basdogan, 1997) emphasise the need to make a more concerted effort to bring the three modalities together in VEs.

The problem of generating effective sounds in VEs has been addressed in particular by Hahn et al. (1998), who identify three sub-problems: sound modelling, sound synchronisation, and sound rendering. The first problem has long been studied in the field of computer music (see also Chapter 8). However, the primary consideration in VE applications is the effective parametrisation of sound models so that the parameters associated with motion (changes of geometry in a scene, user's gestures) can be mapped to the sound control parameters, resulting in an effective synchronisation between the visual and auditory displays. Finally, sound rendering refers to the process of generating sound signals from models of objects and their movements 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 perceptual processing of the sound signal may be needed in order to take into account listener effects (e.g. filtering with Head Related Transfer Functions). The whole process of rendering sounds can be seen as a rendering pipeline analogous to the image rendering pipeline.

Until recently the primary focus for sound generation in VEs has been in spatial localisation of sounds. On the contrary, research about models for sound sources and mappings between object motion/interaction and sound control is far less developed. In Section 9.4.2 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 and Bardy (2001) is a very powerful one: the global array provides information that can optimise 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 is already 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 these parameters 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 (1993b) provides a framework for the design of environmental sounds. Gaver emphasises that, since it is often difficult to identify the acoustic information of 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.

The studies on multimodal perception reviewed in Section 9.3 also provide a number of useful guidelines and even quantitative data. We have seen that streams of information coming from different channels complement and integrate each other, with some modality possibly dominating over the remaining ones depending on the features to be evaluated or the tasks to be accomplished (the modality precision or modality appropriateness hypothesis by Welch and Warren, 1986). In particular, when senses interact in time, a window of synchrony between the feedback of different modalities (e.g. auditory and visual, or auditory and haptic feedbacks) is crucial for multisensory integration. Many of the studies previously discussed (e.g., Shams et al., 2002; Guest et al., 2002; Bresciani et al., 2005) report quantitative results about "integration windows" between modalities. These estimates can be used as constraints for the synchronisation of rendering pipelines in a multimodal architectures.

9.4.2 Sound models

Physics-based approaches

Sound synthesis techniques traditionally developed for computer music applications (e.g. additive, subtractive, frequency modulation, Zölzer, 2002) provide abstract descriptions of sound signals. Although well suited for the representation of musical sounds, these techniques are in general not effective for the generation of non-musical interaction sounds. We have seen in Section 9.2 that research in ecological acoustics points out that the nature of everyday listening is rather different and that auditory perception delivers information which goes beyond attributes of musical listening.

On the other hand, physically-based sound modelling approaches (see Chapter 8) generate sound from computational structures that respond to physical input parameters, and therefore they automatically incorporate complex responsive acoustic behaviours. Moreover, the physical control parameters do not require in principle manual tuning in order to achieve realistic output. Again, results from research in ecological acoustics aid in determining what sound features are perceptually relevant, and can be used to guide the tuning process.

A second advantage of physically-based approaches is interactivity and ease in associating motion to sound control. As an example, the parameters needed to characterise collision sounds, e.g. relative velocity at collision, are computed in the VR physical simulation engine and can be directly mapped into control parameters of a physically-based sound model. The sound feedback consequently responds in a natural way to user gestures and actions. This is not the case with traditional approaches to sound synthesis, where the problem of finding a motion-correlated parametrisation is not a trivial one. Think about the problem of parameterizing real recorded sounds by their attributes such as amplitude and pitch: this corresponds to a sort of "reverse engineering" problem where one tries to determine how the sounds were generated starting from the sounds themselves. Designing effective mappings between user gestures and sound control parameters is important especially in the light of the studies in action-perception loop, that we have addressed in Section 9.3.2.

Finally, physically-based sound models can in principle allow the creation of dynamic virtual environments in which sound rendering attributes are incorporated into data structures that provide multimodal encoding of object properties: shape, material, elasticity, texture, mass, and so on. In this way a unified description of the physical properties of an object can be used to control the visual, haptic, and sound rendering, without requiring the design of separate properties for each thread. This problem has already been studied in the context of joint haptic-visual rendering, and recent haptic-graphic APIs (Technologies, 2002; Sensegraphics, 2006) adopt a unified scene graph that takes care of both haptics and graphics rendering of objects from a single scene description, with obvious advantages in terms of synchronisation and avoidance of data duplication. Physically-based sound models may allow the development of a similar unified scene, that includes description of audio attributes as well.

For all these reasons, it would be desirable to have at disposal sound modelling techniques that incorporate complex responsive acoustic behaviours and can reproduce complex invariants of primitive features: physically-based models offer a viable way to synthesise naturally behaving sounds from computational structures that respond to physical input parameters. Although traditionally developed in the computer music community and mainly applied to the faithful simulation of existing musical instruments, physical models are now gaining popularity for sound rendering in interactive applications (Cook, 2002).

Contact sounds

As already remarked in Section 9.2, 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 9.1). Various modelling approaches have been proposed in the literature.

Van den Doel and coworkers (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. Modal synthesis techniques have been already presented in Chapter 8. Here, we recall that if a resonating object is modelled as a network of N masses connected with springs and dampers, then a geometrical transformation can be found that turns the system into a set of decoupled equations. The transformed variables $\{q_n\}_{n=1}^{N}$ are generally referred to as modal displacements, and obey a second-order linear oscillator equation:

$$\ddot{q}_n(t) + g_n\dot{q}_n(t) + \omega_n^2 q_n(t) = \frac{1}{m_n} f(t), \tag{9.1}$$

where q_n is the oscillator displacement and f represents any driving force, while ω_n is the oscillator center frequency. The parameter $1/m_n$ controls the "inertial" properties of the oscillator (m_n has the dimension of a mass), and g_n is the oscillator damping coefficient and relates to the decay properties of the system. Modal displacements q_n are related to physical displacement through an $N \times K$ matrix $\boldsymbol{A}$, whose elements a_{nk} weigh the contribution of the nth mode at a location k. If the force f is an impulse, the response q_n of each mode is a damped sinusoid and the physical displacement at location k is given by

$$x_k(t) = \sum_{n=1}^{N} a_{nk} q_n(t) = \sum_{n=1}^{N} a_{nk} e^{-g_n t/2} \sin(\omega_n t). \tag{9.2}$$

Any pre-computed contact force signal can then be convolved to the impulse response and thus used to drive the modal synthesiser.

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 the paper by Klatzky et al. (2000), already discussed in Section 9.2. In this work, the modal representation proposed by van den Doel and Pai (1998) has been applied to the synthesis of impact sounds with material information.

An analogous modal representation of resonating objects was also adopted by Avanzini et al. (2003). The main difference with the above mentioned works lies in the approach to contact force modelling. 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 by Avanzini et al. (2003) embed direct computation of non-linear contact forces. Despite the complications that arise in the synthesis algorithms, this approach provides some advantages. Better quality is achieved due to accurate audio-rate computation of contact forces: this is especially true for impulsive contact, where contact times are in the order of few ms. Interactivity and responsiveness of sound to user actions is also improved. This is especially true for continuous contact, such as stick-slip friction (Avanzini et al., 2005). Finally, physical parameters of the contact force models provide control over other ecological dimensions of the sound events. As an example, the impact model used by Avanzini et al. (2003), and originally proposed by Hunt and Crossley (1975), describe the non-linear contact force as

$$f(x(t), v(t)) = \begin{cases} kx(t)^{\alpha} + \lambda x(t)^{\alpha} \cdot v(t)\,(1 + \mu v(t)) & x > 0, \\ 0 & x \leq 0, \end{cases} \tag{9.3}$$

where x is the interpenetration of the two colliding objects and $v = \dot{x}$. Then force parameters, such as the force stiffness k, can be related to ecological dimensions of the produced sound, such as perceived stiffness of the impact. Similar considerations apply to continuous contact models (Avanzini et al., 2005).

It has been shown that this approach allows for a 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 9.1). Models for bouncing, breaking, rolling, crumpling sounds are described in the works by Rath and Fontana (2003) and Rath and Rocchesso (2005). See also Chapter 10 for a description of "sounding objects" synthesised with this approach.

A different physically-based approach has been proposed by 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.

Other classes of sounds

The map of everyday sounds developed by Gaver (see figure 9.1) comprises three main classes: solids, liquids, and gases. Research on sound modelling is clearly biased toward the first of these classes, while less has been done for the others.

A physically-based liquid sound synthesis methodology has been developed by van den Doel (2005). 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 those produced by streams, pouring water, rivers, rain, and breaking waves is based on the synthesis of single bubble sounds. It is shown by van den Doel (2005) how realistic complex high dimensional sound spaces can be synthesised in this manner.

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 sounds 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.

This brief overview shows that little has been done in the literature about models of everyday sounds in the “liquids” and “gases” categories (we are sticking to the terminology used by Gaver (1993a), and reported in figure 9.1). These are topics that need more research to be carried out in the future.

9.4.3 Applications to multimodal interfaces

Multimodal rendering

An important consequence of using physically-based sound models is that synchronisation with other modalities is in principle straightforward, since the parameters that are needed to characterise the sounds resulting from mechanical contact 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.

A particularly interesting problem is 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. In order to be perceived as realistic, auditory and haptic cues have to be synchronised so that they appear simultaneous. They must also be perceptually similar – a rough surface has to both sound and feel rough. Synchronizing the two modalities 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.

Rendering a virtual surface, i.e. simulating the interaction forces that arises when touching a stiff object, is the prototypical haptic task. Properly designed visual (Wu et al., 1999) and/or auditory (DiFranco et al., 1997) feedback can be combined with haptics in order to improve perception of stiffness, or even compensate for physical limitations of haptic devices and enhance the range of perceived stiffness that can be effectively conveyed to the user. Physical limitations (low sampling rates, poor spatial resolution of haptic devices) constrain the values for haptic stiffness rendering to ranges that are often far from typical values for stiff surfaces (Kuchenbecker et al., 2006). Ranges for haptic stiffness are usually estimated by requiring the system to be passive (Colgate and Brown, 1994), thus guaranteeing stability of the interaction, while higher stiffness values can cause the system to become unstable, i.e. to oscillate in an uncontrolled way.

Perceptual experiments on a platform that integrates haptic and sound displays were reported by 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 stiffness 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 by DiFranco et al. (1997) was rather poor (the authors used recorded sounds), a more sophisticated approach amounts to synthesise both auditory and haptic feedback using physically-based models. This approach was taken in the work of DiFilippo and Pai (2000). In this work the modal synthesis techniques described by 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 low latency between haptic and audio rendering (the latency is 1ms if the rate of the haptic rendering routine is 1kHz), which is below the perceptual tolerance for detecting synchronisation between auditory and haptic contact events.

A related study was recently conducted by Avanzini and Crosato (2006). In this paper the sound models proposed by Avanzini et al. (2003, 2005) were integrated into a multimodal rendering architecture, schematically depicted in Fig. 9.3, which extends typical haptic-visual architectures (Salisbury et al., 2004). The sound rendering thread runs at audio rate (e.g. 44.1kHz) in parallel with other threads. Computation of audio contact forces is triggered by collision detection from the haptic rendering thread. Computation of 3D sound can be cascaded to the sound synthesis block. It was shown that the proposed

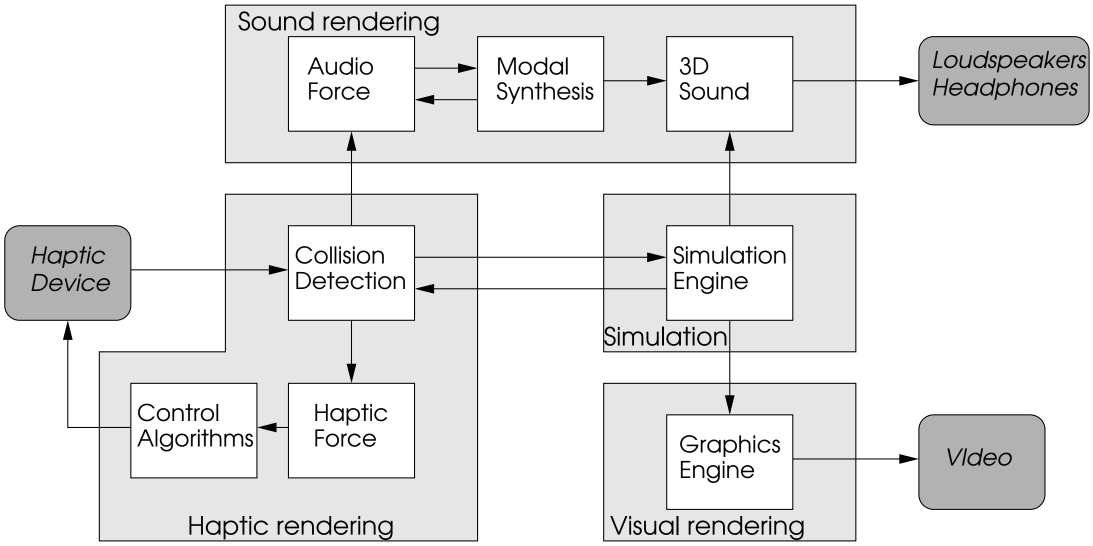

Figure 9.3: An architecture for multimodal rendering of contact interactions. Adapted from Fig. 3 in Salisbury et al. (2004).

rendering scheme allows tight synchronisation of the two modalities, as well as a high degree of interactivity and responsiveness of the sound models to gestures and actions of a user. The setup was used to run an experiment on the relative contributions of haptic and auditory information to bimodal judgments of contact stiffness: experimental results support the effectiveness of auditory feedback in modulating haptic perception of stiffness.

Substituting modalities

In Section 9.3.2 we have already reviewed some studies that address the topic of sensory substitution with applications to the design of interfaces. The focus of such studies (Ifukube et al., 1991; Kaczmarek et al., 1991; Meijer, 1992) is especially substitution systems for visually-impaired users. The very same idea of sensory substitution can be exploited in a different direction: having an interface which is not able to provide feedback of a given modality (e.g. a passive device such as a standard mouse is not able to provide haptic feedback), that modality can effectively substituted with feedback of other modalities, provided that it uses the same sensory-motor skills. We will try to clarify this concept in the remainder of this section. The studies briefly reviewed in the following, although not specifically related with audio but rather with visual and haptic feedback, contain interesting ideas that may be applied to auditory rendering.

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. The authors exemplify the general idea as follows: assume that a user manipulates a cube in a VE using a passive device like a mouse, and has to insert it inside a narrow duct. As the cube is inserted in the duct, it "visually resists" motion by reducing its speed, and consequently the user increases the pressure on the mouse which results in an increased feedback force by the device. The combined effects of the visual slowing down of the cube and the increased feedback force from the device provides the user with an "illusion" of force feedback, as if a friction force between the cube and the duct was rendered haptically. Lecuyer and coworkers have applied this idea to various interactive tasks, and have shown that properly designed visual feedback can to a certain extent provide a user with "pseudo-haptic" feedback.

Similar ideas have driven the work of 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 changes due to 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.

The same approach 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 behaviour of objects in motion, thus producing audible effects at different time scales.

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. MIT Press, Cambridge, MA, 1991.

F. Avanzini and P. Crosato. Integrating physically-based sound models in a multimodal rendering architecture. *Comp. Anim. Virtual Worlds*, 17(3-4): 411–419, July 2006.

F. Avanzini, M. Rath, D. Rocchesso, and L. Ottaviani. Low-level sound models: resonators, interactions, surface textures. In D. Rocchesso and F. Fontana, editors, *The Sounding Object*, pages 137–172. Mondo Estremo, Firenze, 2003.

F. Avanzini, S. Serafin, and D. Rocchesso. Interactive simulation of rigid body interaction with friction-induced sound generation. *IEEE Trans. Speech Audio Process.*, 13(6), Nov. 2005.

J. P. Bresciani, M. O. Ernst, K. Drewing, G. Bouyer, V. Maury, and A. Kheddar. Feeling what you hear: auditory signals can modulate tactile tap perception. *Exp. Brain Research*, In press, 2005.

C. Carello and M. T. Turvey. The ecological approach to perception. In Encyclopedia of cognitive science. London, Nature Publishing Group, 2002.

C. Carello, K. L. Anderson, and A. Kunkler-Peck. Perception of object length by sound. *Psychological Science*, 9(3):211–214, May 1998.

J. E. Colgate and J. M. Brown. Factors Affecting the Z-Width of a Haptic Display. In *Proc. IEEE Int. Conf. on Robotics & Automation*, pages 3205–3210, San Diego, May 1994.

P. R. Cook. *Real sound synthesis for interactive applications*. A. K. Peters, Natick, MA, USA, 2002.

D. DiFilippo and D. 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.

D. E. DiFranco, G. L. Beauregard, and M. A. Srinivasan. The effect of auditory cues on the haptic perception of stiffness in virtual environments. In *Proceedings of the ASME Dynamic Systems and Control Division, Vol.61*, 1997.

Y. Dobashi, T. Yamamoto, and T. 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.

M. O. Ernst and H. H. Bülthoff. Merging the senses into a robust percept. *TRENDS in Cognitive Sciences*, 8(4):162–169, Apr. 2004.

D. 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.

W. W. Gaver. What in the world do we hear? an ecological approach to auditory event perception. *Ecological Psychology*, 5(1):1–29, 1993a.

W. W. Gaver. How do we hear in the world? explorations of ecological acoustics. *Ecological Psychology*, 5(4):285–313, 1993b.

J. J. Gibson. *The ecological approach to visual perception*. Lawrence Erlbaum Associates, Mahwah, NJ, 1986.

B. Giordano. *Sound source perception in impact sounds*. PhD thesis, Department of General Psychology, University of Padova, Italy, 2006. URL `http://www.music.mcgill.ca/~bruno/`.

S. Guest, C. Catmur, D. Lloyd, and C. Spence. Audiotactile interactions in roughness perception. *Exp. Brain Research*, 146(2):161–171, Sep. 2002.

B. Gygi, G. R. Kidd, and C. S. Walson. Spectral-temporal factors in the identification of environmental sounds. *J. Acoust. Soc. Am.*, 115(3):1252–1265, Mar. 2004.

J. K. Hahn, H. Fouad, L. Gritz, and J. W. Lee. Integrating sounds in virtual environments. *Presence: Teleoperators and Virtual Environment*, 7(1):67–77, Feb. 1998.

K. Hötting and B. Röder. Hearing Cheats Touch, but Less in Congenitally Blind Than in Sighted Individuals. *Psychological Science*, 15(1):60, Jan. 2004.

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.

T. Ifukube, T. Sasaki, and C. Peng. A blind mobility aid modeled after echolocation of bats. *IEEE Trans. Biomedical Engineering*, 38(5):461–465, May 1991.

K. A. Kaczmarek, J. G. Webster, P. Bach-y-Rita, and W. J. Tompkins. Electrotactile and vibrotactile displays for sensory substitution systems. *IEEE Trans. Biomedical Engineering*, 38(1):1–16, Jan. 1991.

R. L. Klatzky, D. K. Pai, and E. P. Krotkov. Perception of material from contact sounds. *Presence: Teleoperators and Virtual Environment*, 9(4):399–410, Aug. 2000.

K. J. Kuchenbecker, J. Fiene, and G. Niemeyer. Improving Contact Realism through Event-Based Haptic Feedback. *IEEE Trans. on Visualization and Comp. Graphics*, 13(2):219–230, Mar. 2006.

A. Lécuyer, S. Coquillart, and A. Kheddar. Pseudo-haptic feedback: Can isometric input devices simulate force feedback? In *IEEE Int. Conf. on Virtual Reality*, pages 83–90, New Brunswick, 2000.

S. J. Lederman. Auditory texture perception. *Perception*, 8(1):93–103, Jan. 1979.

S. J. Lederman, R. L. Klatzki, T. Morgan, and C. 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.

X. Li, R. J. Logan, and R. E. Pastore. Perception of acoustic source characteristics: Walking sounds. *J. Acoust. Soc. Am.*, 90(6):3036–3049, Dec. 1991.

R. A. Lutfi and E. L. Oh. Auditory discrimination of material changes in a struck-clamped bar. *J. Acoust. Soc. Am.*, 102(6):3647–3656, Dec. 1997.

M. R. McGee, P. Gray, and S. Brewster. Mixed feelings: Multimodal perception of virtual roughness. In *Proc. Int. Conf EuroHaptics*, pages 47–52, Edinburgh, July 2002.

H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264 (5588):746–748, Dec. 1976.

P. B. L. Meijer. An experimental system for auditory image representations. *IEEE Trans. Biomedical Engineering*, 39(2):112–121, Feb. 1992.

C. F. Michaels and C. Carello. *Direct Perception*. Prentice-Hall, Englewood Cliffs, NJ, 1981.

S. Morein-Zamir, S. Soto-Faraco, and A. Kingstone. Auditory capture of vision: examining temporal ventriloquism. *Cognitive Brain Research*, 17:154–163, 2003.

A. Noë. *Action in perception*. MIT press, Cambridge, Mass., 2005.

J. F. O'Brien, P. R. Cook, and G. Essl. Synthesizing sounds from physically based motion. In *Proc. ACM SIGGRAPH 2001*, pages 529–536, Los Angeles, CA, Aug. 2001.

J. F. O'Brien, C. Shen, and C. M. Gatchalian. Synthesizing sounds from rigid-body simulations. In *Proc. ACM SIGGRAPH 2002*, pages 175–181, San Antonio, TX, July 2002.

J. K. O'Regan and A. Noë. A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24(5):883–917, 2001.

M. Rath and F. 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.

M. Rath and D. Rocchesso. Continuous sonic feedback from a rolling ball. *IEEE Multimedia*, 12(2):60–69, Apr. 2005.

B. H. Repp. The sound of two hands clapping: an exploratory study. *J. Acoust. Soc. Am.*, 81(4):1100–1109, Apr. 1987.

K. Salisbury, F. Conti, and F. Barbagli. Haptic Rendering: introductory concepts. *IEEE Computer Graphics and Applications*, 24(2):24–32, Mar. 2004.

Sensegraphics. Website, 2006. http://www.sensegraphics.se.

L. Shams, Y. Kamitani, and S. Shimojo. Visual illusion induced by sound. *Cognitive Brain Research*, 14(1):147–152, June 2002.

M. A. Srinivasan and C. Basdogan. Haptics in virtual environments: taxonomy, research status, and challenges. *Comput. & Graphics*, 21(4):393–404, July 1997.

T. A. Stoffregen. Affordances and events. *Ecological Psychology*, 12(1):1–28, Winter 2000.

T. A. Stoffregen and B. G. Bardy. On specification and the senses. *Behavioral and Brain Sciences*, 24(2):195–213, Apr. 2001.

Novint Technologies. The interchange of haptic information. In *Proc. Seventh Phantom Users Group Workshop (PUG02)*, Santa Fe, Oct. 2002.

K. van den Doel. Physically based models for liquid sounds. *ACM Trans. Appl. Percept.*, 2(4):534–546, Oct. 2005.

K. van den Doel and D. K. Pai. The sounds of physical shapes. *Presence: Teleoperators and Virtual Environment*, 7(4):382–395, Aug. 1998.

K. van den Doel, P. G. Kry, and D. 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.

K. 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.

F. Varela, E. Thompson, and E. Rosch. *The Embodied Mind*. MIT Press, Cambridge, MA, 1991.

W. H. Warren and R. R. Verbrugge. Auditory perception of breaking and bouncing events: Psychophysics. In W. A. Richards, editor, *Natural Computation*, pages 364–375. MIT Press, Cambridge, Mass., 1988.

R. B. Welch and D. H. Warren. Intersensory interactions. In K. R. Boff, L. Kaufman, and J. P. Thomas, editors, *Handbook of Perception and Human Performance – Volume 1: Sensory processes and perception*, pages 1–36. John Wiley & Sons, New York, 1986.

R. P. Wildes and W. A. Richards. Recovering material properties from sound. In W. A. Richards, editor, *Natural Computation*, pages 357–363. MIT Press, Cambridge, Mass., 1988.

W. C. Wu, C. Basdogan, and M. A. Srinivasan. Visual, haptic, and bimodal perception of size and stiffness in virtual environments. In *Proceedings of the ASME Dynamic Systems and Control Division, (DSC-Vol.67)*, 1999.

U. Zölzer, editor. *DAFX – Digital Audio Effects*. John Wiley & Sons, 2002.