SYNTHESIS OF ENVIRONMENTAL SOUNDS IN INTERACTIVE MULTIMODAL SYSTEMS

Federico Avanzini

University of Padova
Dept. of Information Engineering
Padova, Italy
avanzini@dei.unipd.it

ABSTRACT

This review paper discusses the literature on perception and synthesis of environmental sounds. Relevant studies in ecological acoustics and multimodal perception are reviewed, and physically-based sound synthesis techniques for various families of environmental sounds are compared. Current research directions and open issues, including multimodal interfaces and virtal environments, automatic recognition and classification, and sound design, are discussed. The focus is especially on applications of physically-based techniques for synthesis of environmental sounds in interactive multimodal systems. The paper reports on ongoing research on bimodal (audio-haptic) rendering of virtual objects.

[Keywords: Environmental sounds, multimodal rendering, physical models]

1. INTRODUCTION

Research on environmental sounds, which has its roots in ecological acoustics, is currently receiving interest in many domains, including multimodal interfaces and virtal environments, automatic recognition and classification (with applications to context-aware and surveillance systems as well as automatic classification of sound effects and automatic synthesis of soundscapes), and sound design.

This paper provides a review of the literature on perception and synthesis of environmental sounds, and discusses some relevant current research directions. The focus is especially on physically-based techniques for synthesis of environmental sounds, and application of these techniques in interactive multimodal systems.

The paper is organized as follows. Section 2 is devoted to perception of environmental sounds, with emphasis on studies in ecological acoustics and multimodal perception; Section 3 discusses synthesis techniques for various families of environmental sounds; Section 5 reports on our current research on physically-based models and applications to joint audio-haptic rendering.

2. PERCEPTION OF ENVIRONMENTAL SOUNDS

2.1. Ecological acoustics

The "ecological" approach to perception, originated in the work of Gibson [1], differs from more established views in two main respects: first, perception is an achievement of animal-environment systems, not simply animals (or their brains); second, the main purpose of perception is to guide action. The gibsonian approach is considered controversial because of one central and strong claim: perception is *direct*, that is, there exists a 1:1 correspondence between patterns of sensory stimulation and the underlying aspects of physical reality. This assumption implies that anything that can be perceived can also be measured in the physical world.

Gibson worked on visual perception and introduced the concept of *optic flow*, which indicates the structure in changing patterns of light at a given point of observation. Perceivers exploit particular patterns – *invariants* – to guide their activities. These considerations also apply to other senses, including audition. Recent research has introduced the concept of *global array* [2], according to which individual forms of energy are subordinate components of a higher-order spatio-temporal structure. The general claim underlying this concept is that observers are not separately sensitive to structures in the optic and acoustic flows, but are directly sensitive to patterns that extend across these flows.

Two companion papers by Gaver [3, 4] have greatly contributed to the build-up of a solid framework for *ecological acoustics*, by introducing such concepts as the acoustic array and acoustic invariants that can be associated to sound events: as an example, several attributes of a vibrating solid (e.g., size, shape, density), determine the frequencies of the sounds it produces. A single physical parameters can influence simultaneously many different sound parameters: these complex patterns of change may serve as information distinguishing the physical parameters responsible.

Gaver also coined the term *everyday listening*, the experience of listening to events rather than sounds, and proposed an "ecological taxonomy" of environmental sounds (see Fig. 1). Sounds generated by *solid objects* are structured by the type of their *interaction*, the materials, and the geometry and configuration of the objects. Sounds involving *liquids* also depend on an initial deformation counter-acted by restoring forces, but in this case sounds are created by the resonant cavities (bubbles) that form and oscillate and in the surface of the liquid. *Aerodynamic sounds* are caused by atmospheric pressure differences (e.g. an exploding balloon), or situations in which changes in pressure set objects into vibration (e.g. the wind passing through a wire). Any environmental sound is originated from *basic* events in any of the above categories. Many sounds can be described as temporal *patterns* of simpler events: breaking is a complex event involving patterns of simpler impacts. *Compound* events involve more than one type of basic level event: a door slam involves the squeak of scraping hinges and the impact of the door on its frame. *Hybrid* events involve yet another level of complexity in which more than one basic type of material is involved (e.g. the sounds of water dripping on a reverberant surface).

Although relatively young, the literature on ecological acoustics has produced a number of relevant results. Most are concerned

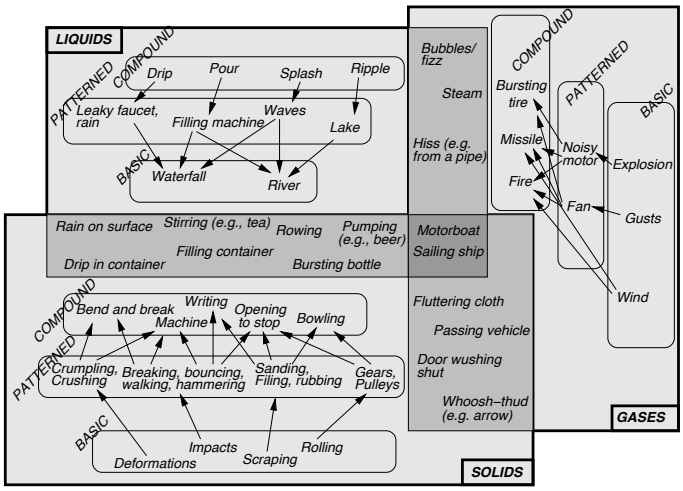


Figure 1: *An ecological taxonomy of environmental sounds. Complexity increases towards the center. Figure based on [3].*

with *basic* interactions of solids objects, while sound-producing events that involve liquids and aerodynamic interactions have been addressed less frequently. Many studies have investigated the perception of object material from impact sounds [5, 6, 7, 8]. Another relevant ecological dimension of impact sounds is the hardness of collision [9, 8]. With respect to continuous contact (e.g. scraping), a relevant ecological dimension is surface roughness: research by Lederman and coworkers has investigated the role of auditory feedback in both tactile [10] and vibratory roughness perception [11]. The auditory perception of geometric properties of interacting objects, e.g. length, has also been investigated [12].

Studies on patterned sounds include bouncing and breaking events [13], hands clapping [14], and walking sounds [15]. A recent paper by Gygi *et al*. [16] does not focus on a specific sound event and instead uses a large (70) and varied catalog of sounds that include patterned, compound, and hybrid sources. The authors investigate the role of temporal features of the sound envelope (periodicities, amount of silence, roughness) in the identification of the events.

2.2. Multimodal perception

Humans achieve robust perception through the combination and integration of information from multiple sensory modalities. Two general strategies can be identified [17]: (sensory *combination*) is used to maximize non-redundant information delivered from different sensory modalities, while sensory *integration* is used to reduce the variance in the sensory estimate and increase reliability.

In general 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 [18] is often cited when trying to explain which modality dominates under what circumstances, and states that discrepancies are always resolved in favour of the more precise or more appropriate modality. Vision dominates the integrated percept in many tasks. As an example, vision can bias the perceived location of sounds whereas sounds rarely influence visual localization. One key reason for this *visual capture* is that vision provides more accurate location information. For temporal judgments however the situation is reversed and audition, being the more appropriate modality, usually dominates over vision. Recent studies have provided evidence of *auditory capture* effects in temporal judgments, showing e.g. that the number of auditory beeps influences the perceived number of visual flashes [19] or that auditory events can even alter the perceived timing of target lights [20].

Similar capture effects can also occur between audition and tactile perception: some authors [21, 22] have extended the finding of the auditory-visual illusion established by [19] to the auditory-tactile domain. Other authors have studied auditory-tactile integration in surface texture perception. Lederman and coworkers have shown that audition has little influence on tactile texture perception [10]. However, when the contact is made via a rigid probe, with a consequent increase of touch-related sound and a degradation of tactile information, auditory and tactile cues are integrated [11]. 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 experiment on forced-choice discrimination of the roughness of abrasive surfaces [23] showed that properly processed texture sounds lead to a bias towards an increased perception of tactile smoothness or roughness.

3. SYNTHESIS OF ENVIRONMENTAL SOUNDS

Sound synthesis techniques traditionally developed for computer music applications (e.g. additive, subtractive, frequency modulation [24]) are less effective for the generation of environmental sounds. On the other hand, physically-based sound modeling approaches generate sound from computational structures that respond to physical input parameters, and therefore they automatically incorporate complex responsive acoustic behaviors. A second quality of these techniques is interactivity and ease in associating motion to sound control, so that the sound feedback responds in a natural way to user gestures and actions. Traditionally developed in the computer music community and mainly applied to the faithful simulation of existing musical instruments, physical models have now gained popularity for sound rendering in interactive applications [25].

3.1. Contact sounds

As already remarked an important class of sound events is that of *contact* sounds between solids, i.e. sounds generated when solid objects come in contact with each other (see Fig. 1). Various modeling approaches have been proposed in the literature.

Modal synthesis [26] was proposed in [27, 28] as an efficient yet accurate framework for describing the acoustic properties of objects. If a resonating object is described as a network of 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 are generally referred to as *modal displacements*, and obey a second-order linear oscillator equation. If the driving force is an impulse, the response of each mode is a damped sinusoid. Any pre-computed contact force signal can then be convolved to the impulse response and thus used to drive the modal synthesizer. The modal representation of a resonating object can be linked to many *ecological* dimensions of the corresponding sounds. As an example, in [7] the modal representation proposed by [27] has been applied to the synthesis of impact sounds with material information.

A different physically-based approach was proposed in [29, 30], which 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. However, heavy computational load prevents real-time sound generation and the use of the method in interactive applications.

3.2. Other classes of sounds

The map of everyday sounds developed by Gaver (see Fig. 1) comprises three main classes: solids, liquids, and gases. Research on sound modeling is clearly biased toward the first of these classes, while less has been done for the others.

A physically-based liquid sound synthesis methodology was developed in [31]. 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 developed a sound model for isolated single bubbles and validated it with a small user study. A stochastic model for the real-time interactive synthesis of complex liquid sounds such as produced by streams, pouring water, rivers, rain, and breaking waves is based on the synthesis of single bubble sounds. It is shown in [31] that realistic complex high dimensional sound spaces can be synthesized in this manner.

A method for creating aerodynamic sounds was presented in [32]. 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 proposed a method for creating sound textures for aerodynamic sound by making use of computational fluid dynamics. Next, they have developed a method using the sound textures for real-time rendering of aerodynamic sound according to the motion of objects or wind velocity.

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 of Fig. 1). These are topics that need more research to be carried out in the future.

4. CURRENT RESEARCH DIRECTIONS

4.1. Multimodal rendering and display

Multisensory information is essential for designing immersive virtual worlds: being able to hear sounds of objects in a virtual environment, while touching and manipulating them, provides a sense of immersion in the environment not obtainable otherwise [33]. Properly designed and synchronized haptic and auditory displays are likely to provide much greater immersion in a virtual environment than a high-fidelity visual display alone. Moreover, by skewing the relationship between the haptic and visual and/or auditory displays, the range of object properties that can be effectively conveyed to the user can be significantly enhanced.

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 [34, 35] 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 synchronization 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.

A particularly interesting problem is simultaneous audio-haptic rendering. In order to be perceived as realistic, auditory and haptic cues have to be properly synchronized and perceptually similar. Syncronizing 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.

Properly designed auditory 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. Ranges for haptic stiffnesses are usually estimated by requiring the system to be passive [36], thus guaranteeing stability of the interaction, while higher stiffness values can cause the system to become unstable, i.e., to oscillate uncontrollably.

The influence of auditory information on the perception of object stiffness through a haptic interface was studied in [37]. Prerecorded sounds of contact between several pairs of objects were played through headphones during tapping of virtual objects through a haptic interface, and were shown to modulate the perception of object stiffness. These results suggest that the range of object stiffnesses that can be displayed by a haptic interface with a limited force-bandwidth can be perceptually extended by the addition of properly designed impact sounds.

While the auditory display adopted by [37] was rather poor (the authors used recorded sounds), a more sophisticated approach amounts to synthesize both auditory and haptic feedback using physically-based models. In [38] the modal synthesis techniques described in [27] 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 audio rate (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 synchronization between auditory and haptic contact events.

4.2. Recognition and classification of environmental sounds

Automatic recognition of environmental sounds is a currently active research direction, and the MPEG-7 multimedia standard provides a framework for sound recognition [39]. The biggest challenges in this context are perhaps sound soure separation and completeness/generality: it is very difficult to design descriptors that identify every sound source in a given environment.

Many studies in this field are based on supervised training with

preselected training material. In [40] Hidden Markov models were applied to the recognition of specific classes of sounds (a wooden door opened and shut, a metal tool dropped in a container, and water poured in a container). A study on the recognition of "familiar" environmental sounds (i.e. sounds on which the recognition system was previously trained) was reported in [41], which included smoke alarm, barking dogs, bouncing balls, water running in bathtubs, vacuum cleaner motor, and so on.

Many application scenarios are currently being investigated. One is context-aware computing: research in this field has mostly focused on applications that are aware of absolute spatial and temporal location, while other aspects of context have been relatively neglected. Recent works have tried to exploit automatic recognition of environmental sounds as a contextual cue for context-aware applications, e.g. by trying to classify the "noise context" in typical everyday environments (office, car, city street) [42].

Another area of application is in automatic surveillance systems: with the increasing use of audio sensors in surveillance and monitoring applications, event detection from audio streams has emerged as an important research problem. Methods for a scenario where a system is installed in an unknown environment (specifically an office room) were presented in [43]: it was shown that a system combining both supervised and unsupervised training methods could have potential for practical applications.

An application area that is closer to the interests of the ICAD community is classification of sound effects. The technology behind sound effect libraries is still text-search: sounds are tagged with descriptive keywords, with several consequent limitations (the annotation work is error-prone and time-consuming, natural language is imprecise and ambiguous, and so on). Automatic annotation methods are not mature enough for labeling with great detail any possible sound. A general sound recognition tool requires a taxonomy that represents common sense knowledge of the world, and thousands of specialized classifiers. Recent studies [44] have shown that a general sound annotator can be constructed using a taxonomy built on top on a semantic network such as WordNet and an all-purpose sound recognition system based on nearest-neighbor classification rule.

Systems for the automatic recognition and classification of environmental sounds will make it possible to approach synthesis of soundscapes from a high-level perspective, in which sound objects can be identified and recombined in a flexible way. Creating soundscapes from libraries of individual environmental sounds is a convenient alternative to recorded "ambiances", which are not flexible (e.g. it is hard to add/remove individual sounds or change panning) and are available in a limited number. In [45] semi-automatic generation of simple soundscapes (like those of a pub or a farm) was proposed using a semantic enabled search engine: the system creates ambiances on demand given text queries by fetching relevant sounds from a large sound effect database and importing them into a sequencer multi track project.

4.3. Sound design

Product design is going to be profoundly affected by new technologies that can change the appearance of objects (e.g., electronic ink, dynamic actuators, etc.). Recent research projects [46] argue that, as microprocessors and loudspeakers can be already embedded into objects, the "sonic appearance" of objects is already easily changeable. It can therefore be expected that research on "product sound design" will became a solid and established discipline in the near future.

Many studies have addressed sound quality measurement in past years. Evaluation based on psychophysical methods has been applied to sounds of domestic objects (light switches, vacuum cleaners, etc.), equipment (car motors, air conditioners, etc.) and so on, with the aim of characterizing acoustic annoyance or preference. However environmental sounds also have emotional connotations, which precede their cognitive interpretation and influence the way a listener perceives a given sound [47]. This is clearly a fundamental aspect for product designers, since users decide to approach positive and avoid negative objects largely on that basis of an emotional response. Even more, emotional response can affect the cognitive level of interaction [48].

Sound also has functional qualities. In many cases designers associate sounds to desired meanings on the basis of empirical criteria. Research on everyday sounds could help to extract auditory attributes and patterns in order to create unambiguous sounds to fulfill specific functions. Recommendations for the designers have to be adjusted by perceptual results. An experiment with sounds currently used in automotive interfaces [49] showed that these sounds do not fulfill their intended function. The authors then proposed a methodology which draws on acoustics and semiotics and applied it to the specific sound design problem under investigation.

However, knowledge about environmental sounds is still insufficient with regard to the relations between physical characteristics, perceptual descriptions, and functional and aesthetic qualities. Techniques to relate the functional and aesthetic qualities of sound to emotional and cognitive responses may come by linking pattern analysis techniques with results of psychophysical experimentation, in such a way that mathematical models, generalizations, and classifications are conducted on functionally selected sound databases and on parameter sets of synthetic sound models [46].

5. CONTACT SOUNDS IN MULTIMODAL RENDERING ARCHITECTURES

5.1. Toward a taxonomy of contact sound models

We have proposed a modal representation of resonating objects analogous to the one adopted in [27, 28]. The main difference with the above mentioned works lies in the approach to contact force modeling. Instead of a feed-forward scheme in which the interacting resonators are set into oscillation with driving forces that are externally computed or recorded, the models proposed in [50] 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. Finally, physical parameters of the contact force models provide control over other ecological dimensions of the sound events.

The impact model used in [50], and originally proposed in [51], describe the non-linear impact force f as

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

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. In [52] a stick-slip friction model was proposed, which is derived from [53]. Microscopic irregularities of contacting surfaces can be interpreted as a large number of elastic "bristles", that will randomly deflect like damped springs when a tangential force is applied to each bristle, and start to slip when the strain exceeds a certain level. The friction force f is described by the equations

$$\begin{aligned} \dot{z}(v,z) &= v\left[1-\alpha(v,z)\frac{z}{z_{ss}(v)}\right], \\ f(z,\dot{z},v,w) &= \sigma_0 z + \sigma_1 \dot{z} + \sigma_2 v + \sigma_3 w, \end{aligned} \tag{2}$$

where z is the average bristle deflection and v is the relative velocity between the two surfaces The coefficient σ_0 is the bristle stiffness, σ_1 is the bristle damping, and the term $\sigma_2 v$ accounts for linear viscous friction. The function $z_{ss}(v)$ is the steady-state friction characteristic: steady state conditions in the sliding regime (i.e., $\dot{z}=0$, with $v \neq 0, \alpha = 1$) are met if and only if $z = z_{ss}$. The function $\alpha(v,z)$ is an adhesion map that controls the rate of change of z. Parametrizations of α must guarantee that $\alpha \equiv 0$ when z is smaller than a given breakaway displacement (purely elastic presliding regime, $\dot{z} = v$), and $\alpha \equiv 1$ for large values of z (transition to the plastic regime). The component $\sigma_3 w(t)$ is not part of the original formulation in [53]. The term w is related to surface roughness and is needed in order to simulate scraping and sliding effects, whereas the original elasto-plastic formulation only accounts for stick-slip phenomena. The w component can be modeled as fractal noise [28].

It has been shown that these *low-level* contact sound models can be taken as the basic building blocks from where sound events of increasing complexity can be simulated: this kind of approach allows for a translation of the map of Fig. 1 into a hierarchical structure in which "patterned" and "compound" sounds models are built upon impact and friction events. Models for bouncing, breaking, rolling, crumpling sounds are described in [54, 55].

5.2. Linking physical parameters and ecological dimensions

The studies in ecological acoustics mentioned in Sec. 2.1 identify features of environmental sounds that convey information about generating events, and thus provide mappings between sound signal parameters and ecological dimensions. In order to link physical parameters of the sound models to ecological dimensions, a second level of mapping is needed, which relates physical parameters to relevant sound signal features. In certain cases such mappings can be derived straightforwardly: as an example, perception of *material* is determined to a large extent by the decay characteristics of an impact sound, which are in turn directly linked to the damping parameters of a modal resonator.

A less simple case of mapping between physical parameters and ecological dimensions concerns the perception of stiffness in impact sounds. Freed [9] showed that the useful information for hardness rating is contained in the attack transients of the sounds: in his study loudness and descriptors related to the spectral centroid (average value and temporal variability in the first 300 ms) were found to account for 75% of the variance of the hardness ratings. Giordano [8] argues that the duration τ of the contact between the two objects during the stroke has an influence on hardness perception, and that τ variations are likely to explain at least in part data from [9]: an increase in τ determines a decrease in the loudness of the radiated signal, and in the amount of energy at high frequencies (and thus in the spectral centroid), since vibrational modes with a period shorter than τ are minimally excited.

Based on similar considerations, in [56] we have investigated the dependence of contact time τ and the attack spectral centroid on the parameters of the impact force model (1). The following equation was derived for τ:

$$\tau = \left(\frac{m}{k}\right)^{\frac{1}{\alpha+1}} \cdot \left(\frac{\mu^2}{\alpha+1}\right)^{\frac{\alpha}{\alpha+1}} \cdot \int_{v_{out}}^{v_{in}} g(v, v_{in}, \mu) dv, \tag{3}$$

where v_{in}, v_{out} are the normal velocities before/after collision, respectively, and $\mu = \lambda/k$ is a mathematically convenient term. Equation (3) states in particular a power-law dependence of τ on the force stiffness: $\tau(k) \sim k^{-1/\alpha+1}$. A study in [57] on synthetic impact sounds obtained from model (1) provided quantitative results that confirm the correlation between spectral centroid of the attack transients and τ. The dissipative component of the contact force also has a slight effect on the centroid: as λ is lowered, the amount of energy transferred to the higher partials is increased, and the centroid increases accordingly, even though τ remains approximately constant. Similarly, the centroid increases significantly as α decreases, even though the contact time varies slowly: high values of α can produce multiple bounces of the vibrating surface on the striking object, with a consequent increase of the centroid.

5.3. Multimodal interaction

In [58] the contact sound models described above were integrated into a multimodal rendering architecture, schematically depicted in Fig. 2, which extends typical haptic-visual architectures. 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 rendering scheme allows tight synchronization of the modalities, as well as a high degree of interactivity and responsiveness of the sound models to gestures and actions of a user.

This architecture was implemented as two processes which communicate by means of a shared memory area. The first process is responsible for graphic and haptic rendering. An event catching engine driven by a function callback model is adopted to monitor contact events. When such an event occurs, haptic data necessary for sound synthesis are written into the shared memory area. The second process reads data from the shared memory area and renders contact sounds according to the current physical parameters. Low communication latency is critical in order to ensure unitary perception rather than perception of two distinct auditory and haptic events: it has been shown that the latency in this implementation is well below typical experimental estimates for temporal windows of auditory-tactile integration (see e.g. [22, 21]).

The setup was used in [59] to run an experiment on the relative contributions of haptic and auditory information to bimodal judgments of contact stiffness. Rendering a virtual surface, i.e. simulating the interaction forces that arises when touching a stiff object, is the prototypical haptic task. We have investigated to what extent the addition of auditory feedback can affect the haptic perception of stiffness and possibly compensate for physical limitations of the haptic device so to enhance the range of perceived stiffness conveyed to the user (as an example, the nominal maximum closed-loop control stiffness for a Phantom® Omni™ device is 500 N/m, which is far from typical values for stiff surfaces).

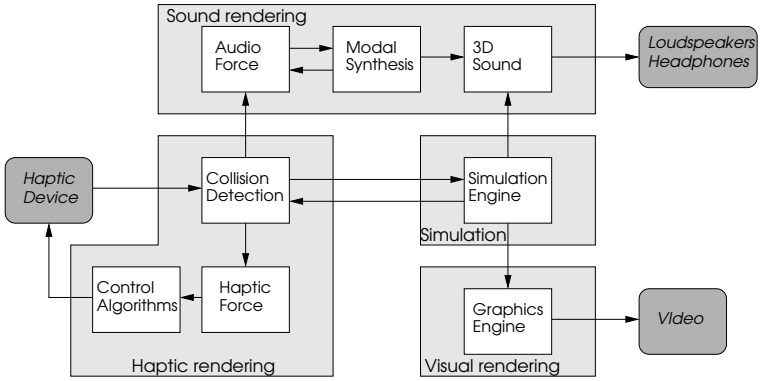


Figure 2: *An architecture for multimodal rendering of contact interactions.*

Figure 3(a) shows the visual display that was presented to the subjects: this did not change between conditions, and was intentionally composed of stylized objects, in order to limit as much as possible the amount of visual information delivered to subjects. Perceived stiffness was determined through an absolute magnitude-estimation procedure on a scale ranging from "extremely soft" (1) to "extremely stiff" (8), and the results reported in Fig. 3(b) support the effectiveness of auditory feedback in modulating haptic perception of stiffness. Moreover, about 40% perceived the haptic feedback changing together with audio and based their rating also on haptic feedback (although the haptic stiffness had the same value in all conditions), suggesting that properly designed and synchronized contact sounds can elicit an auditory-haptic illusion and modulate the haptic perception of stiffness.

5.4. Discussion

The experiment reported above relies on verbal descriptions and judgements, which are in many ways arbitrary. For instance, what should the subjects judge: "stiffness", or "hardness", or "force", or . . . ? Would that be the stiffness of the hammer, or that of the hit bar, or both?

We are currently working on a different experimental set-up, in which the focus is on performance rather than on quantitative judgments of verbally described qualities. Specifically we are working on a set in which subjects are required to hit a bar similar to the one given in Fig. 3(a), with the task of repeating a given "target performance", in terms e.g. of rythm, amplitude, impact velocity. The goal is to investigate how the task is influenced by changes in the visual, auditory, and haptic feedback and consequently assess what modality dominates in this task.

This kind of approach has two main interesting aspects: first, it avoids completely any verbal decription of the task to be performed; second, it focuses the experimental design directly on the loop between action and perception, in a way that has rarely been investigated in the literature [60].

6. ACKNOWLEDGEMENTS

I am grateful to all the people with whom I have been working on this topic so far: Davide Rocchesso, Bruno Giordano (who is the main person responsible for the ideas reported in Sec. 5.4), Stefania Serafin, Matthias Rath, and Paolo Crosato. This research was supported by the EU Sixth Framework Programme – IST Information Society Technologies (Network of Excellence "Enactive Interfaces" IST-1-002114, http://www.enactivenetwork.org).

7. REFERENCES

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

[2] T. A. Stoffregen and B. G. Bardy, "On specification and the senses," *Behavioral and Brain Sciences*, vol. 24, no. 2, pp. 195–213, Apr. 2001.

[3] W. W. Gaver, "What in the world do we hear? an ecological approach to auditory event perception," *Ecological Psychology*, vol. 5, no. 1, pp. 1–29, 1993.

[4] ——, "How do we hear in the world? explorations of ecological acoustics," *Ecological Psychology*, vol. 5, no. 4, pp. 285–313, 1993.

[5] R. P. Wildes and W. A. Richards, "Recovering material properties from sound," in *Natural Computation*, W. A. Richards, Ed. Cambridge, Mass.: MIT Press, 1988, pp. 357–363.

[6] R. A. Lutfi and E. L. Oh, "Auditory discrimination of material changes in a struck-clamped bar," *J. Acoust. Soc. Am.*, vol. 102, no. 6, pp. 3647–3656, Dec. 1997.

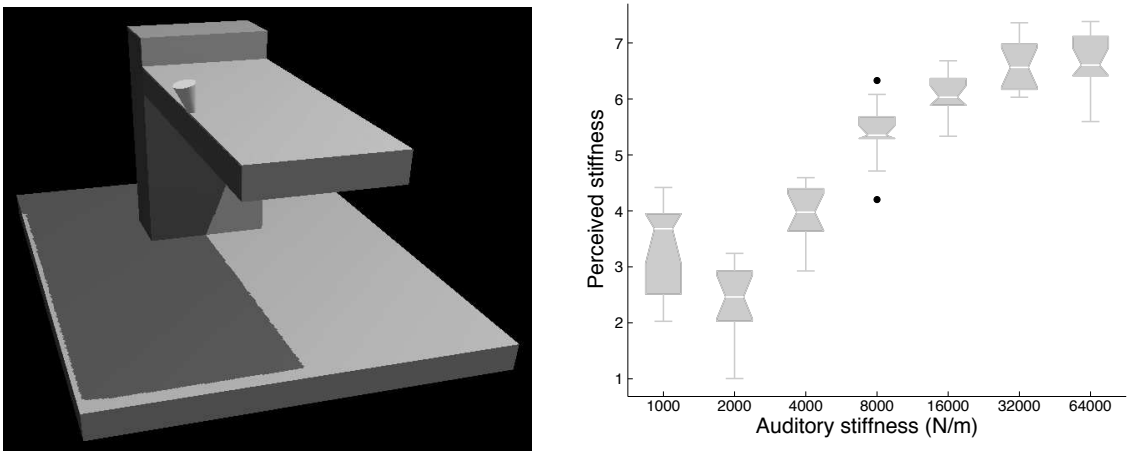
[7] R. L. Klatzky, D. K. Pai, and E. P. Krotkov, "Perception of material from contact sounds," *Presence: Teleoperators and Virtual Environment*, vol. 9, no. 4, pp. 399–410, Aug. 2000.

[8] B. Giordano, "Sound source perception in impact sounds," Ph.D. dissertation, Department of General Psychology, University of Padova, Italy, 2006.

[9] D. J. Freed, "Auditory correlates of perceived mallet hardness for a set of recorded percussive events," *J. Acoust. Soc. Am.*, vol. 87, no. 1, pp. 311–322, Jan. 1990.

[10] S. J. Lederman, "Auditory texture perception," *Perception*, vol. 8, no. 1, pp. 93–103, Jan. 1979.

[11] 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)*, Orlando, FL, 2002, pp. 97–104.



(a) (b)

Figure 3: *Experiment on bimodal stiffness judgements: (a) visual display that presented to the subjects, and (b) experimental results.*

[12] C. Carello, K. L. Anderson, and A. Kunkler-Peck, "Perception of object length by sound," *Psychological Science*, vol. 9, no. 3, pp. 211–214, May 1998.

[13] W. H. Warren and R. R. Verbrugge, "Auditory perception of breaking and bouncing events: Psychophysics," in *Natural Computation*, W. A. Richards, Ed. Cambridge, Mass.: MIT Press, 1988, pp. 364–375.

[14] B. H. Repp, "The sound of two hands clapping: an exploratory study," *J. Acoust. Soc. Am.*, vol. 81, no. 4, pp. 1100–1109, Apr. 1987.

[15] X. Li, R. J. Logan, and R. E. Pastore, "Perception of acoustic source characteristics: Walking sounds." *J. Acoust. Soc. Am.*, vol. 90, no. 6, pp. 3036–3049, Dec. 1991.

[16] B. Gygi, G. R. Kidd, and C. S. Walson, "Spectral-temporal factors in the identification of environmental sounds," *J. Acoust. Soc. Am.*, vol. 115, no. 3, pp. 1252–1265, Mar. 2004.

[17] M. O. Ernst and H. H. Bülthoff, "Merging the senses into a robust percept," *TRENDS in Cognitive Sciences*, vol. 8, no. 4, pp. 162–169, Apr. 2004.

[18] R. B. Welch and D. H. Warren, "Intersensory interactions," in *Handbook of Perception and Human Performance – Volume 1: Sensory processes and perception*, K. R. Boff, L. Kaufman, and J. P. Thomas, Eds. New York: John Wiley & Sons, 1986, pp. 1–36.

[19] L. Shams, Y. Kamitani, and S. Shimojo, "Visual illusion induced by sound," *Cognitive Brain Research*, vol. 14, no. 1, pp. 147–152, June 2002.

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

[21] K. Hötting and B. Röder, "Hearing Cheats Touch, but Less in Congenitally Blind Than in Sighted Individuals," *Psychological Science*, vol. 15, no. 1, p. 60, Jan. 2004.

[22] 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*, vol. In press, 2005.

[23] S. Guest, C. Catmur, D. Lloyd, and C. Spence, "Audiotactile interactions in roughness perception," *Exp. Brain Research*, vol. 146, no. 2, pp. 161–171, Sep. 2002.

[24] U. Zölzer, Ed., *DAFX – Digital Audio Effects*. John Wiley & Sons, 2002.

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

[26] J.-M. Adrien, "The missing link: Modal synthesis," in *Representations of Musical Signals*, G. De Poli, A. Piccialli, and C. Roads, Eds. Cambridge, MA: MIT Press, 1991, pp. 269–297.

[27] K. van den Doel and D. K. Pai, "The sounds of physical shapes," *Presence: Teleoperators and Virtual Environment*, vol. 7, no. 4, pp. 382–395, Aug. 1998.

[28] 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*, Los Angeles, CA, Aug. 2001, pp. 537–544.

[29] J. F. O'Brien, P. R. Cook, and G. Essl, "Synthesizing sounds from physically based motion," in *Proc. ACM SIGGRAPH 2001*, Los Angeles, CA, Aug. 2001, pp. 529–536.

[30] J. F. O'Brien, C. Shen, and C. M. Gatchalian, "Synthesizing sounds from rigid-body simulations," in *Proc. ACM SIGGRAPH 2002*, San Antonio, TX, July 2002, pp. 175–181.

[31] K. van den Doel, "Physically based models for liquid sounds," *ACM Trans. Appl. Percept.*, vol. 2, no. 4, pp. 534–546, Oct. 2005.

[32] 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*, San Diego, CA, July 2003, pp. 732–740.

[33] J. K. Hahn, H. Fouad, L. Gritz, and J. W. Lee, "Integrating sounds in virtual environments," *Presence: Teleoperators and Virtual Environment*, vol. 7, no. 1, pp. 67–77, Feb. 1998.

[34] N. Technologies, "The interchange of haptic information," in *Proc. Seventh Phantom Users Group Workshop (PUG02)*, Santa Fe, Oct. 2002.

[35] Sensegraphics, "Website," http://www.sensegraphics.se.

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

[37] D. E. DiFranco, G. L. Beauregard, and M. A. Srinivasan, "The effect of auditory cues on the haptic perception of stiffness in virtual environments," Proceedings of the ASME Dynamic Systems and Control Division, (DSC-Vol.61), 1997.

[38] 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.

[39] M. Casey, "Mpeg-7 sound recognition tools," *IEEE Trans. Circ. Systems for Video Tech.*, vol. 11, no. 6, pp. 737–747, June 2001.

[40] J. P. Woodard, "Modeling and classification of natural sounds by product code hidden markov models," *IEEE Trans. Sig. Process.*, vol. 40, no. 7, pp. 1833–1835, July 1992.

[41] R. S. Goldhor, "Recognition of environmental sounds," in *Proc. IEEE Int. Conf. Acoust. Speech and Signal Process.*, vol. 1, Minneapolis, Apr. 1993, pp. 149–152.

[42] D. Smith, L. Ma, and N. Ryan, "Acoustic environment as an indicator of social and physical context," *Personal and Ubiquitous Computing*, vol. 10, no. 4, pp. 241–254, May 2006.

[43] A. Härmä, M. F. McKinney, and J. Skowronekh, "Automatic surveillance of the acoustic activity in our living environment," in *Proc. IEEE Int. Conf. on Multimedia and Expo*, Amsterdam, July 2005.

[44] P. Cano, M. Koppenberger, S. L. Groux, J. Ricard, N. Wack, and P. Herrera:, "Nearest-neighbor automatic sound classification with a wordnet taxonomy," *J. Intellingent Information Systems*, vol. 24, no. 2, pp. 99–111, May 2005.

[45] P. Cano, L. Fabig, F. Gouyon, M. Koppenberger, A. Loscos, and A. Barbosa, "Semi-automatic ambiance generation," in *Proc. COST-G6 Conf. Digital Audio Effects*, Naples, Oct. 2004, pp. 319–323.

[46] P. Susini, N. Misdariis, G. Lemaitre, D. Rocchesso, P. Polotti, and K. Franinovic, "Closing the loop of sound evaluation and design," in *Proc. 2nd ISCA/DEGA Tutorial & Research Workshop on Perceptual Quality of Systems*, Berlin, Sep. 2006.

[47] D. Västfjäll and M. Kleiner, "Emotion in product sound design," in *Proc. Journes Design Sonore*, Paris, Mar. 2002.

[48] N. Tractinsky, A. Shoval-Katz, and D. Ikar, "What is beautiful is usable," *Interacting with Computers*, vol. 13, pp. 127–145, 2000.

[49] C. Suied, P. Susini, N. Misdariis, S. Langlois, B. Smith, and S. M. Adams, "Toward a sound design methodology: Application to electronic automotive sounds," in *Proc. Int. Conf. Auditory Display*, Limerick, July 2005, pp. 146–153.

[50] F. Avanzini, M. Rath, D. Rocchesso, and L. Ottaviani, "Low-level sound models: resonators, interactions, surface textures," in *The Sounding Object*, D. Rocchesso and F. Fontana, Eds. Firenze: Mondo Estremo, 2003, pp. 137–172.

[51] K. H. Hunt and F. R. E. Crossley, "Coefficient of restitution interpreted as damping in vibroimpact," *ASME J. Applied Mech.*, vol. 42, pp. 440–445, June 1975.

[52] F. Avanzini, S. Serafin, and D. Rocchesso, "Interactive simulation of rigid body interaction with friction-induced sound generation," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 6, pp. 1073–1081, Nov. 2005.

[53] P. Dupont, V. Hayward, B. Armstrong, and F. Altpeter, "Single state elasto-plastic friction models," *IEEE Trans. Automat. Control*, vol. 47, no. 5, pp. 787–792, May 2002.

[54] M. Rath and F. Fontana, "High-level models: bouncing, breaking, rolling, crumpling, pouring," in *The Sounding Object*, D. Rocchesso and F. Fontana, Eds. Firenze: Mondo Estremo, 2003, pp. 173–204.

[55] M. Rath and D. Rocchesso, "Continuous sonic feedback from a rolling ball," *IEEE Multimedia*, vol. 12, no. 2, pp. 60–69, Apr. 2005.

[56] D. Rocchesso, L. Ottaviani, F. Fontana, and F. Avanzini, "Size, shape, and material properties of sound models," in *The Sounding Object*, D. Rocchesso and F. Fontana, Eds. Firenze: Mondo Estremo, 2003, pp. 95–110.

[57] F. Avanzini and D. Rocchesso, "Physical modeling of impacts: theory and experiments on contact time and spectral centroid," in *Proc. Int. Conf. Sound and Music Computing (SMC06)*, Paris, Oct. 2004, pp. 287–293.

[58] F. Avanzini and P. Crosato, "Integrating physically-based sound models in a multimodal rendering architecture," *Comp. Anim. Virtual Worlds*, vol. 17, no. 3-4, pp. 411–419, July 2006.

[59] ——, "Haptic-auditory rendering and perception of contact stiffness," in *Proc. Workshop on haptic and audio interaction design (HAID06)*, D. McGookin and S. Brewster, Eds. Glasgow: Lecture Notes in Computer Science 4129, Springer Verlag, Aug. 2006, pp. 24–35.

[60] B. Mantel, B. Bardy, and T. A. Stoffregen, "Intermodal specification of egocentric distance in a target reaching task," in *Studies in Perception and Action*, H. Heft and K. L. Marsh, Eds. Mahwah, NJ: Lawrence Erlbaum Associates, 2005, pp. 173–176.