Chapter 7
Auditory rendering and display of interactive floor cues

S. Serafin, F. Fontana, L. Turchet, and S. Papetti

Abstract A walking task engages, among other senses, that of hearing. Humans do not only perceive their own footstep sounds during locomotion; Walking also conveys auditory cues that aid its recognition by listeners who are not performing the task. As a result, footstep sounds contribute to keep walkers in the perception and action loop, meanwhile informing external listeners about several characteristics of their actions. After reviewing the literature dealing with the auditory aspects of walking, this chapter provides an overview of some current developments in the interactive synthesis and display of footstep sounds. Several novel user evaluations conducted on auditory display prototypes applying these techniques are presented, shedding light on the potential of these techniques to find application in multimodal scenarios involving foot-floor interactions.

7.1 Introduction

During locomotion, humans are often engaged in parallel tasks distracting them from a thorough visual exploration and fine analysis of the floor that must be traversed. This is the case, for example, for a person walking in the city while browsing a newspaper, or simply looking at his or her surroundings. Symmetrically, it is not uncommon to notice walking persons who cannot continuously focus on a concurrent visual task, but who, due to their ability to exploit non-visual sensory cues, are seen to be capable of distinguishing the path ahead.

Such situations reveal a tendency by humans to keep walking tasks at the periphery of attention, a decision that is at any moment challenged by potential pitfalls that may occur during locomotion [211]. From this perspective, walking represents an interesting application domain in which to study possibilities for empowering the exchange of non visual information between humans and the external world.

The sonic modality, in particular, is at the core of many interactions involving communication at the periphery of one's attention. Physical contact with grounds

that, once pressed or scraped, produce sounds, turn out to be especially informative. They also provide perceptual cues to walking persons nearby and to listeners who are outside this loop, but in the proximity of the interaction touchpoint. Such cues can reveal surface materials [111], shoe type and walking style, and can enable listeners to make inferences about the gender or other physical, biomechanical, and affective characteristics of a person [170, 310, 46].

This chapter builds upon a general hypothesis that subsumes the above situations: By providing peripheral, yet highly informative, cues, interactive walking sounds support listeners in their everyday activities involving self-locomotion.

This hypothesis is, in principle, applicable to a broad set of contexts and, hence, its implications have potential to span diverse social groups. Specifically, it would be interesting to assess walking sounds as human factors, and consequently measure their informative value in critical areas such as labor, locomotion in hostile spaces, or as navigation aids for people with different abilities [165]. Once available, these measurements may be useful toward enabling the design of auditory scenes in which such factors are optimized depending on the specific context of application.

In this chapter, we suggest that effective sonic interactions in walking can be realized through proper *augmentations* of the perceived reality. This means that future interactive floor scenarios should increase, rather than substitute the natural possibilities of grounds to convey auditory cues at floor level. In our case, additional cues will be enabled by enriching otherwise neutral floors with "layers" that provide physically-consistent sounds recalling familiar, possibly everyday ground ecologies. Examples include the superposition of sounds capable of evoking virtual aggregate grounds such as gravel, snow or mud.

Fortunately, current sensing and actuation devices allow interaction designers to come up with interface concepts, whose realization can be carried out at affordable cost and in reasonable time. Concerning, in particular, the floor interface domain, a plethora of force sensors as well as sound reproduction devices exist for detecting, then instantaneously responding to human walking through synthetic sonic feedback generated in real time by a laptop, or even smaller computing machines. In parallel it must be remembered that technologies differ in costs and feature specifications, and not all such technologies promise to scale down in price and encumbrance in their future versions. In the case of audio reproduction, the hardest constraints are perhaps given by the power consumption and weight of the amplifiers and loudspeakers. On one hand, we can opt for small, low-power devices providing fair quality and little low frequency content (bass sound), as a consequence of their limited acoustic bandwidth. On the other hand we can choose large amplification and reproduction devices with a broadband response, but bulky, by no means wearable auditory interfaces.

Whether the desired soundscapes will be finally achieved through rich and immersive, hence necessarily complex and encumbering systems or, conversely, through (even not computer-enabled) simple and miniaturized interfaces finding room in a shoe sole, is a question that will require time and further work to be answered. In this perspective, the chapter represents just a starting point toward that answer.

7.1.1 Chapter organization

This chapter discusses aspects of the design and realization process of soundscapes centered around walking sounds. In particular:

- Synthesis techniques are reviewed that allow one to simulate walking events, especially for the creation of virtual sonic layers of aggregate material on top of acoustically neutral floors. Since they are based on the simulation of simplified physical event descriptions, such techniques can be straightforwardly employed to synthesize vibrotactile signals that naturally arise out as a result of the same descriptions.
- Taken alone, a footstep sound has little perceptual meaning unless a suitable auditory context is present around it. This scene includes design of soundscapes that simulate different indoor and outdoor spaces.
- The resulting sonic scenarios must be rendered by adopting proper combinations of interactive walking sounds and soundscape descriptions. Not only they must achieve a sufficient degree of realism when displayed using a conventional reproduction arrangement, such as for instance a headset or a couple of stereo loudspeakers: they should be also flexible enough to accommodate unusual displays. This flexibility becomes especially interesting in the case of walking sound augmentations, for which some non-conventional reproduction sets are presented and then discussed.
- A user-centered evaluation of such scenarios is important toward assessing the ability of the simulations to realistically recreate ground surfaces.

7.2 Walking sound synthesis

7.2.1 Background

The first systematic attempt to synthesize walking sounds has been proposed by Cook in 2002 [65]. In this pioneering work, he introduced elements of novelty that make his work stimulating and still largely state-of-the art. The most interesting aspect in this modeling approach was the emphasis on foot-floor interactivity: thanks to an detailed procedure, which included several analysis stages, the model stored essential features from signals which were recorded during foot interactions with diverse floors; then, a reproduction of the same features could be made online by informing a parametric synthesis filter with temporal series of force envelopes, corresponding to footstep sequences. This allowed straightforward connection of the resulting system architecture to floor interfaces like sensing mats, performing a physically-informed interactive synthesis of walking sounds.

The physically-informed approach was also exploited in other work on the synthesis of walking sounds. By making use of physically-based algorithms for the reproduction of microscopic impacts [257], as early in 2003 Fontana designed a

stochastic controller that, once parameterized in parameters of force and resistance (respectively against and belonging to the floor), generated realistic sound simulations of footsteps over crumples and similar aggregate materials. Thanks to a higher-level control layer, such sounds were grouped into a footstep sequence once they were triggered by an expressive control model exposing affective parameters and musical performance rules, proposed by Bresin. The resulting real-time software architecture was a document ("patch") for the Puredata software environment, which synthesized the sound and allowed continuous control of both physical floor parameters and gestural intentions of users [93].

An attempt to integrate some biomechanical parameters of locomotion, particularly the GRF, in a real time footstep sound synthesizer was made by Farnell in 2007 [91]. The result was a patch for Puredata that was furthermore intended to demonstrate how to create an audio engine for computer games in which walking is interactively sonified.

7.2.2 Synthesizing walking

Acoustic and vibrational signatures of locomotion are the result of more elementary physical interactions, including impacts, friction, or fracture events, between objects with certain shape and surface material properties such hardness, texture etc. The decomposition of complex everyday sound phenomena in terms of more elementary ones has been an organizing idea in auditory display research during recent decades [107].

Specifically, a footstep sound can be considered the result of multiple micro-impacts between a shoe and a floor. Either they converge to form a unique percept consisting of a single impact, in the case of *solid* materials, or conversely they result in a more or less dispersed, however coherent burst of impulsive sounds in the case of *aggregate* materials. At simulation level, it is convenient to draw a main distinction between solid and aggregate ground surfaces, the latter being assumed to possess a granular structure, such as that of gravel.

An impact involves the interaction between an active *exciter*, i.e., the impactor, and a passive *resonator*. Sonic impacts between solid surfaces have been extensively investigated, and results are available which describe relationships between physical and perceptual parameters of the objects in contact [147, 305]. Such sounds are typically short in duration, with sharp temporal onsets and relatively rapid decay.

The most simple approach to synthesizing such sounds is based on a lumped source-filter model, in which a signal $s(t)$ modelling the excitation is passed through a linear filter with impulse response $h(t)$ modeling the resonator, and resulting in an output expressed by the linear convolution of these two signals: $y(t) = s(t) * h(t)$.

By borrowing terminology from the kinematics of human locomotion, the excitation force can be identified with the GRF. In our case, GRF signals acquired using microphones or force input devices have been used to control different sound

synthesis algorithms, which reproduce solid and aggregate surfaces as listed in the following of this section.

7.2.3 Physics-based modeling

The simulation of the interaction between solid surfaces can be obtained by decomposing the physical phenomenon into its basic constituents, instead of linearizing it into a series of two or more filters. The physics-based modeling approach precisely allows to deal with different kinds of interactions, by preserving their invariant phenomenological properties through this decomposition. According to this approach, situations like a foot sliding across the floor, or conversely walking on it, can be rendered respectively by starting from a friction or impact excitation component, meanwhile preserving the invariant floor properties in the resonant component.

Impact and friction are two crucial categories that affect walking perception [107]. In the impact model, the excitation corresponding to each impact $s(t)$ is assumed to possess a short temporal extent and an unbiased frequency response. A widely adopted physically-based description of this phenomenon considers the force f between the two bodies to be a function of the compression x of the exciter and velocity of impact $\dot{x}$, depending on the parameters of elasticity of the materials, masses, and local geometry around the contact surface [21]:

$$f(x,\dot{x}) = \begin{cases} -kx^{\alpha} - \lambda x^{\alpha}\dot{x}, & x > 0 \\ 0, & x \leq 0 \end{cases} \tag{7.1}$$

where k accounts for the material stiffness, λ represents the force dissipation due to internal friction during the impact, α depends on the local geometry around the contact surface. When $x \leq 0$ the two objects are not in contact.

Friction has been already implemented in sound synthesis as well, by means of a dynamic model in which the relationship between relative velocity v of the bodies in contact and friction force f is represented as a differential problem [23]. Assuming that friction results from a large number of microscopic elastic bonds, also called bristles [82], the v-to-f relationship is expressed as:

$$f(z,\dot{z},v,w) = \sigma_0 z + \sigma_1 \dot{z} + \sigma_2 v + \sigma_3 w \tag{7.2}$$

where z is the average bristle deflection, the coefficient σ_0 is the bristle stiffness, σ_1 the bristle damping, and the term $\sigma_2 v$ accounts for linear viscous friction. The fourth component $\sigma_3 w$ relates to surface roughness, and is simulated as fractal noise.

7.2.3.1 Interaction with aggregate surfaces

The sonic properties of aggregate surfaces can be reproduced by dense temporal sequences of short impact sounds. In most cases sound designers avoid modeling

these properties at a fine-grained level of detail, since a profound inspection of the microscopic phenomenon does not bring proportional advantages to the quality of the synthesis meanwhile increasing the computational burden to levels that often become intolerable[1].

The crumpling algorithm [93, 47] implements a higher-level control, that is put on top of the physically-based impact-resonator model described above. In other words, it organizes temporal sequences of physical impacts between microscopic objects, assumed to be solid. Such sequences give rise to crumpling events, each represented by a corresponding group of micro-impacts. At the same time, the energetic evolution of a sequence instantaneously informs the parameters that are responsible for the generation of micro-impacts.

The temporal distribution of micro-impacts is governed by a Poisson stochastic process, whose inter-arrival times are given by the Poisson distribution $P(\tau) = \lambda e^{\lambda \tau}$ in which λ controls the stochastic density of the micro-impacts. In parallel, the power of each micro-impact follows a stochastic law $P(\gamma) = E^{\gamma}$, controlled by the γ parameter, which is derived from the physics of crumpling [268]. As a result, i) the dissipation of energy occurring during an impact, and ii) the temporal distribution of adjacent impacts can be constantly controlled (in stochastic sense) by the energy left to the process.

The crumpling model is characterized by the *average interval* between micro-impacts and the *average power* of every event as characteristic parameters for the control of sound. Once such controls are instantaneously mapped onto the force signals coming from a foot interface, the model allows continuous control over the generation of crumpling events that can be associated to footstep sounds [47].

Further mapping can be designed in between the interface and the above controls for setting the invariant features of an aggregate ground material, like its *resistance* or *compliance* parameters, having consequences in the perceived granularity of the ground. Together with proper settings of the modal resonator parameters defining the "color" of each micro-impact, these macroscopic controls set the acoustic signature of an aggregate material. Figure 7.1 illustrates the continuous crumpling algorithm.

7.2.3.2 Ground surfaces as resonant objects

In many cases of interest, a ground surface can be modeled as a linear resonator as opposed to the exciter. The ground properties determine the resonator parameters. Solid and homogeneous floors exhibit a narrow-band (hence longer and possessing definite color) sonic signature, conversely aggregate floors can be synthesized using bursts of short, wide-band (hence more noisy) resonant sounds simulating multiple collisions of the shoe against ensembles of small resonators.

[1] Note that this simplification cannot be used a general rule holding for all sounds resulting from multiple, small-scale processes. For instance, when substances with varying contact properties are involved such as liquids in motion, more sophisticated simulations must be realized which consider also the transitions across different macroscopic states [81].

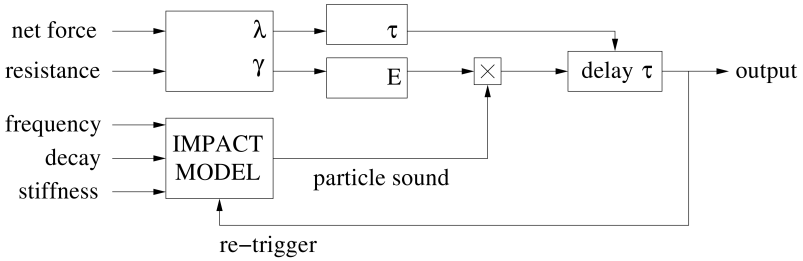

Fig. 7.1: Continuous crumpling algorithm.

Notably, the former sounds in practice are more difficult to be displayed. In fact, homogeneous floors are inherently large resonators capable of producing spatially diffused, possibly loud oscillations across a wide area (think of walking over a wooden floor or jumping on a metal grate). Conversely, aggregate materials generate localized sounds that normally can not propagate along the ground surface. The former, hence, need a powerful enough sound reproduction system: a constraint that rarely can be satisfied by a wearable interface.

As opposed to reproduction, the computational modeling of an even complex linear resonator is relatively straightforward. Modal synthesis [5] explains how to decompose a resonator that responds to an impulse with a signal $h(t)$ in terms of a number of resonant *modes*. The response, then, is modelled as a filter bank with impulse response $h(t) = \sum_i a_i e^{-b_i t} \sin(2\pi f_i t)$, where a_i represents the amplitude of the ith mode, b_i its decay rate, and f_i the associated modal frequency.

Each resonating filter is equivalent to a second-order damped linear oscillator:

$$\ddot{x}^{(r)}(t) + g^{(r)}\dot{x}^{(r)}(t) + \left[\omega^{(r)}\right]^2 x^{(r)}(t) = \frac{1}{m^{(r)}} f_{ext}(t), \tag{7.3}$$

where $x^{(r)}$ is the oscillator displacement and f_{ext} represents any external driving force, while the parameters $\omega^{(r)}$ and $g^{(r)}$ are the oscillator center frequency and damping coefficient, respectively. The parameter $1/m^{(r)}$ controls the "inertial" properties of the oscillator. Such a one-dimensional model provides a basic description of the resonator in terms of its pitch $\omega^{(r)}$ and quality factor $q^{(r)} = \omega^{(r)}/g^{(r)}$. The parameter $g^{(r)}$ relates to the decay properties of the impulse response of the system (7.3): specifically, the relation $t_e = 2/g^{(r)}$ holds, where t_e is the $1/e$ decay time of the impulse response.

7.2.4 Physics-based sound synthesis using the SDT

The Sound Design Toolkit[2] (SDT) [74] is a software product made up of a set of physically-consistent tools for designing, synthesizing and manipulating ecological sounds [106] in real time. The aim of the SDT is to provide efficient and effective instruments for interactive sonification and sonic interaction design.

The SDT consists of a collection of patches and *externals* for Puredata and Max/MSP.[3] The library is compatible with MacOSX, Windows, and Linux (Puredata only).

In the Puredata and Max/MSP terminology, an *external* is a dynamic library which provides some kind of signal processing. Depending on its communication interface (in the form of a set of *inlets* and *outlets*) an *external* can be linked to other *externals*, arithmetical operators, digital filters, sliders or other GUI elements that are natively provided by such environments. Together, all these elements find place inside patches allowing to define complete digital signal processing procedures.

In particular, each SDT's *external* represents a physically-based or -informed/-inspired algorithm for sound synthesis or control. The SDT patches make use of these *externals* to implement fully functional physically-consistent sound models. Moreover, they provide features for parametric control and routing of I/O signals.

The SDT has been used for synthesizing audio and vibrotactile feedback simulating different ground materials. Following is a brief description of the models, and how they have been used.

7.2.4.1 Realization of solid impacts

In the SDT implementation, a modal resonator can have an arbitrary number of resonant modes, each of which is represented by a linear 2nd-order oscillator in the form given by (7.3). Also, the resonating object can be endowed with a macro-dynamic behavior provided by an *inertial mode* added to the modal resonator structure. The inertial mode describes the macro-dynamics of a modal resonator as that of a pointwise mass, which is described by the Newton equation of motion:

$$\ddot{x}(t) = \frac{1}{m} f(t) \tag{7.4}$$

where x is the *displacement* of the whole object, m is its *mass*, and f is the external *force* applied to the object. When present, the inertial mode is considered as the first mode of a modal resonator. It is clear that while an inertial mode is undamped, conversely resonant modes are damped. Having described its single components, it is now possible to describe the structure of a modal resonator having N modes of index $l = 1 \ldots N$ by means of the following linear system:

[2] `http://www.soundobject.org/SDT/`

[3] Which is in a sense the advanced, yet commercial, counterpart to Puredata.

$$\begin{bmatrix} \ddot{x}_1(t) \\ \vdots \\ \ddot{x}_N(t) \end{bmatrix} + G \begin{bmatrix} \dot{x}_1(t) \\ \vdots \\ \dot{x}_N(t) \end{bmatrix} + \Omega^2 \begin{bmatrix} x_1(t) \\ \vdots \\ x_N(t) \end{bmatrix} = \bar{m} f(t) \tag{7.5}$$

where G and Ω are diagonal matrices whose elements are, respectively: $g_{l=1\ldots N}$ and $\omega_{l=1\ldots N}$, and $\bar{m} = [1/m_{l=1\ldots N}]^T$. In case the inertial mode was present, g_1 and ω_1 would be equal to 0, while x_1 would be the displacement of the entire object and m_1 its total mass.

The displacement x_j of a resonating object at a given point $j = 1\ldots N$ can be calculated as:

$$x_j(t) = \sum_{l=1}^{N} q_{jl} \cdot x_l(t) \tag{7.6}$$

where the coefficients q_{jl} are the *output weights* for each mode $l = 1\ldots N$ at the output point j. It is clear that, in case an input and an output point coincided (that is, $i = j$), their modal weights $1/m_{li}$ and output weights q_{jl} would also be the same.

The algorithms underlying solid surface sound models share a common structure which is shown in Figure 7.2: two objects interact through what is called an *interactor*, which models the actual contact interaction. The interactor contains most of

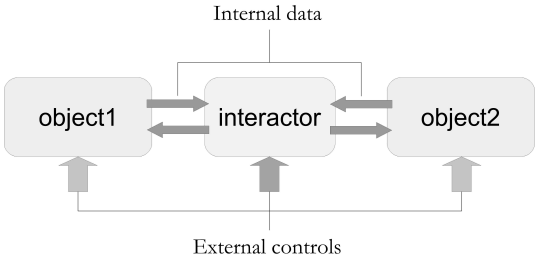

Fig. 7.2: The common structure underlying the SDT algorithms simulating solid surface contacts.

the art, that is necessary to couple two or more objects in the discrete-time domain.

In principle, in order to implement the algorithm represented in Figure 7.2 it is sufficient to couple two instances of (7.5) with (7.1) or (7.2). In practice, this elaboration must be performed while converting the same equations to the discrete-time domain. Thus, their coupling needs to solve issues about numerical stability and accuracy, as well as instantaneous propagation of the effects; for details, see [223].

In the current version of the SDT the bilinear transformation — an A-stable 2nd-order implicit method [244] — is used to translate the continuous-time equations above to the discrete-time domain. In order to discretize the system of 2nd-order differential equations of (7.5) it is useful first to rewrite a single mode (7.3) as an equivalent system of two 1st-order differential equations:

$$\begin{cases} \dot{v}(t) + gv(t) + \omega^2 x(t) = \frac{1}{m} f(t) \\ v(t) = \dot{x}(t) \end{cases} . \tag{7.7}$$

By applying the Laplace-transform to (7.7) we obtain two 1st-order equations in s. The next step is to apply the bilinear transformation, thus obtaining two equations in z, and finally apply the inverse Z-transform in order to obtain the following discrete-time system, expressed in *state-space* form:

$$\begin{bmatrix} x(n) \\ v(n) \end{bmatrix} = A \begin{bmatrix} x(n-1) \\ v(n-1) \end{bmatrix} + \begin{bmatrix} \frac{1}{4m\Delta} \\ \frac{F_s}{2m\Delta} \end{bmatrix} [f(n) + f(n-1)] \tag{7.8a}$$

where the matrix A has the following expression:

$$A = \frac{1}{\Delta} \begin{bmatrix} \Delta - \omega^2/2 & F_s \\ -F_s\omega^2 & 2F_s^2 - \Delta \end{bmatrix} \tag{7.8b}$$

with $\Delta = F_s^2 + gF_s/2 + \omega^2/4$.

As for the inertial mode, the discrete counterpart to (7.4) is easily obtained from (7.8a) and (7.8b) considering $g = 0$ and $\omega = 0$. It follows that the bilinear transformation enables to maintain a unified formulation for both the inertial and resonant modes.

Since the bilinear transformation is an implicit equation, the resulting discrete representation (7.8a) is also in implicit form: an instantaneous dependency between the state variables (displacement x and velocity v) and the input force f is present.

Finally, the discrete-time counterpart to the system made of (7.5) and (7.6), representing an entire modal object, can be written as:

$$\begin{cases} \begin{bmatrix} x_l(n) \\ v_l(n) \end{bmatrix} = A_l \begin{bmatrix} x_l(n-1) \\ v_l(n-1) \end{bmatrix} + \begin{bmatrix} \frac{1}{4m_l\Delta_l} \\ \frac{f_s}{2m_l\Delta_l} \end{bmatrix} [f(n) + f(n-1)] \\ \\ \begin{bmatrix} x_j(n) \\ v_j(n) \end{bmatrix} = \sum_{l=1}^{N} q_{jl} \begin{bmatrix} x_l(n) \\ v_l(n) \end{bmatrix} \end{cases} \tag{7.9}$$

where $l = 1 \ldots N$ and $j = 1 \ldots N$ denote respectively the mode and output point considered. The matrix A_l is as in (7.8b), but now accounts for a different $\Delta_l = F_s^2 + g_l F_s/2 + \omega_l^2/4$ for each mode l.

Due to the implicit form of the bilinear transformation, the resulting discrete-time equations are implicit as well. Hence, an instantaneous relationship is present. For instance, in the case of the impact model, while the modal resonator of (7.9) needs f_{n+1} to compute $[x_{n+1}\ v_{n+1}]^T$, the impact force f_{n+1} also has an instantaneous dependence on x_{n+1} and v_{n+1} given by the discrete-time counterpart of (7.1). Such a *delay-free loop* is not directly computable and, because of the non-linear dependence $f(x,v)$, it needs some special handling in order to be solved. In particular, the *K-method* [42] together with *Newton's method* [244] are used.

The algorithm summarized in Figure 7.2 can now be seen in more detail: at each temporal step the resonators send their internal state (namely, displacement and velocity at the interaction point) to the interactor, which in turn, after solving the delay-free loop as explained above, can send the newly computed interaction forces back

to the objects, thus putting them in condition to perform a computation for the next step. The non-linearities provide richness and improved dynamics to the resulting sounds, even when using low-order resonators.

The solid surface sound models from the SDT allow to set the number of modes of a modal resonator, and the control parameters allow manipulate their modal properties individually.

7.2.4.2 Application to footstep sounds

The SDT realization of solid impacts is a basic building block for synthesizing footstep sounds. A realization of the friction model (7.2) exist in the same library as well, whose discrete-time implementation is left out of this chapter for the peculiar numerical issues that it raises [23].

For the sake of footstep sound synthesis, SDT has been enriched with an alternative impact model implementation, called soft impact. This model allows to synthesize the sound of an impact on a soft surface, or a soft impact between two surfaces. Although avoiding an accurate simulation of the physics of contacts between spatially distributed objects, nevertheless the soft impact algorithm provides effective acoustic results by making use of a dense temporal sequence of tiny signals that excite the resonator described by (7.9). In more detail, no mutual interaction among an interactor and resonating objects is simulated. The interactor, i.e. the force f of (7.9), is instead substituted by a proper static force in the form of a noise burst that finally excites a modal resonator.

This algorithm, hence, realizes a simple feed-forward signal processing procedure. In spite of its simplicity, the idea behind can be qualitatively justified considering that smooth contacts can be reduced to dense temporal sequences of micro-impacts, in a sense modeling the surfaces of the interacting objects as multiple contact areas. Also, the use of specifically filtered noise signals can be motivated considering that such micro-impacts can exhibit a stochastic-like distribution. Besides the modal resonator parameters, the available controls include an ADSR (*attack time*, *decay time*, *sustain gain*, *sustain time*, *release time*) envelope shaper and the *cutoff frequency* parameter of two auxiliary equalization filters (respectively high- and low-pass) which process the noise burst

Finally, SDT puts available an implementation of the continuous crumpling model described in 7.2.3.1.

Practical use of the SDT has been made in the following case studies:

- while simulating impacts between solid surfaces, force data streams provided by the input sensors (see Chapter 2) were pre-conditioned and then analyzed in order to identify foot-floor contact events. Such events have been used to trigger four separate instances of the impact model, corresponding to the heel and toe of each foot: when a contact event was detected, its energy was estimated and the resulting value used to set the initial velocity of the corresponding impact model [225, 226];

- friction has been used while implementing compounds of physically-based building blocks, such as during the synthesis of creaking wood (see Section 7.3);
- concerning soft impacts, the force data stream provided by the FSR sensors (again in Chapter 2) has been treated as in the impact model. In this case however, the energy was used to control the amplitude of the noise burst directly [225, 224, 46, 224];
- as for the continuous crumpling simulations, four pre-conditioned force signals corresponding to the heel and toe of each foot were directly mapped to the respective force parameters of separate instances of the SDT's *crumpling* model. At a lower level, the micro-impacts can be synthesized either by triggering solid or soft impact events. By means of this model, realistic simulations of grounds covered with snow, brushwood or gravel have been obtained [225, 46, 224, 226].

7.2.5 Parametric modeling

In parallel with the synthesis of walking sounds obtained using physically-based models, more simple models can be realized when there is reasonable expectation to come up with realistic, low-latency sonic interactions.

Parametric synthesis has already been proposed for the synthesis of walking sounds [91]. In this section we describe a method that proved effective in the simulation of floors consisting of either homogeneous solid or aggregate surfaces. This method inherits existing analysis-and-synthesis techniques that have been proposed for the generation of walking and, later, hand clapping sounds [65, 234], furthermore it introduces simple novelties in an effort to make the synthesis process as most intuitive for the sonic interaction designer.

The idea is that of starting from the knowledge contained in a pre-recorded set of walking sounds (examples can be downloaded also for free, e.g. from the Freesound online database `www.freesound.org`). This knowledge is used to i) shape noise by means of linear filters, for instance obtained by LPC processing of the source samples or even by manual tuning of the filter parameters, and then to ii) envelope the amplitude of the filtered noise depending on the instantaneous GRF value. Figure 7.3 shows the method in more detail, as well as the design procedure behind.

Concerning the shaping of the noise source, acceptable results can be obtained by manually tuning a series of second-order bandpass recursive digital filters. Typical parameters are shown in Table 7.1, in which bandwidths are expressed in terms of quality factor (Q) parameters.

In parallel, an amplitude weighing function can be obtained from each material example by aligning and then averaging the envelopes, each obtained through straightforward nonlinear processing [65] of a corresponding source sample. This function ultimately represents a mean envelope signal, whose standard deviation from the average value is known at each temporal step.

At this point one can re-synthesize a sequence of footstep sounds over a different material: when a walking event is detected, the GRF onset (typically the initial 3 or

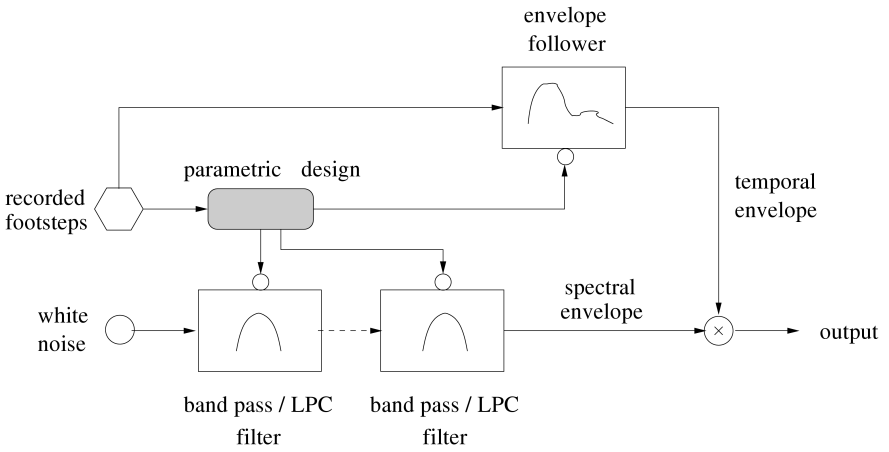

Fig. 7.3: Sketch of the method based on parametric synthesis of footstep sounds.

Material	Filters	Bandwidth (Q)	Center frequency	Gain
Snow	IIR BP	50	400 Hz	1
	IIR BP	700	660 Hz	0.4
Dead leaves	IIR BP	50	100 Hz	1
	IIR BP	500	850 Hz	0.5
	IIR BP	5000	6000 Hz	0.33
Metal (heel)	IIR BP	17	220 Hz	1
	IIR BP	500	220 Hz	1
Metal (toe)	IIR BP	12	250 Hz	1
	IIR BP	20	250 Hz	1
	IIR BP	200	400 Hz	0.24
	IIR BP	500	400 Hz	0.24
Wood (heel)	IIR BP	250	250 Hz	1
	IIR BP	250	250 Hz	1
	IIR BP	180	660 Hz	0.01
	IIR BP	550	660 Hz	0.01
Wood (toe)	IIR BP	55	130 Hz	1
	IIR BP	250	130 Hz	1
	IIR BP	200	610 Hz	0.16
	IIR BP	500	610 Hz	0.16

Table 7.1: Second-order filter parameter values for different ground materials

4 ms) is used to estimate the duration of a footstep. By considering that the energy consumed during every footstep varies to a small extent, the amplitude weighing function is shrunk/stretched along the time dimension (for instance using linear interpolation) and proportionally increased/decreased in amplitude depending on the GRF strength. Furthermore, randomness is added to the mean trajectory at every temporal step proportionally to the standard deviation of the average envelope function.

Parametric synthesis has been successfully adopted to generate accurate stimuli for the test appearing in Section 7.5.2, using linear predictive coding (LPC) to compute the shaping filter coefficients starting from envelope-normalized sound source recordings. Although in that case the procedure was performed offline, the parsimonious use of computational as well as memory resources made by the parametric synthesis procedure poses no problems toward the implementation in real time of this method.

In more detail, the online analysis on the GRF signal is not that simple for parametric synthesis purposes. More elaboration in fact would be needed on top of the onset identification, which keeps into account variations in the foot gesture occurring during the walking act: consider, for instance, a person who stops her foot before completing a step. Until this issues of analysis remain unsolved, physics-based models informed by continuous GRF data deserve more appeal compared to parametric synthesis.

7.3 Composition and parameterization of the models

The sound of a footstep potentially depends on a myriad of contextual conditions: the kind of shoes a subject wears and the type of surface a person is walking on characterize just a subspace of the entire set of possible contexts. One assumption that, for instance, can be made to reduce this space is that the shoe has a rigid sole. This assumption has notable consequences in the simulation of interactions with solid floors, whereas bringing minor effects in the case of aggregate grounds.

The algorithms proposed in the previous sections can be qualitatively informed, starting from a spectral analysis of recordings of real footsteps. From the same recordings, characteristic recurrent events can be identified and then reproduced during the re-synthesis, with an aim toward simulating their evolution across time and subsequently re-combining them into new sounds.

The sound produced while walking on dry leaves can be constructed as a combination of sonic events having relatively long duration and spectral energy at both low and high frequencies, with the addition of noisy glitches that confer “crunchiness” to the final sound. A similar example comes from walking on gravel, resulting from the contribution of physical impacts among stones of different mass, which while colliding give rise to sounds having features that depend on this mass. Such interventions must be correctly parametrized in order to obtain a good match with the components identified in the corresponding real sounds. Finally, an overall loudness must be determined to confer ecological realism to the final sound.

Using different combinations of the models described in the previous section, the following solid and aggregate surfaces have been simulated and tested: wood and creaking wood, metal, deep and low snow, gravel, beach sand, forest underbrush, dry leaves and dirt plus pebbles. [210, 302, 267]. Table 7.2 summarizes how such models have been combined to generate these sounds.

Sound	Impact	Friction	PhISM	Crumpling	GRF control	Stochastic control
Creaking wood	YES	YES	-	-	impacts	friction
Metal	YES	-	-	-	impacts	-
Fresh snow	YES	-	YES	YES	PhISM & crumpling	-
Gravel	-	-	YES	-	PhISM	-
Beach sand	-	-	YES	-	PhISM	-
Forest underbrush	-	-	YES	-	PhISM	-
Dry leaves	-	-	YES	-	PhISM	-
Dirt plus pebbles	-	-	YES	-	PhISM	-

Table 7.2: Combination of sound synthesis models for different walking sounds. Note: for some materials, up to three independent instances of PhISM are computed.

For instance, the sound of footsteps on wood was synthesized by controlling one instance of the SDT impact through GRF data, and superimposing to the impact "creaking" sounds obtained by exciting a friction model through ramp functions driving the external rubbing force. The variation and duration of such ramps was randomly set within certain ranges, before synthesizing every new footstep. The addition of randomness enhances the realism, overall introducing changes in the frequency, amplitude and duration of the sounds.

The examples proposed in this section have been validated through the experiment reported in Section 7.5.1.

7.4 Footstep sounds rendering and displaying

Sounds can be conveyed to the walker through the air and/or bone conduction, by means of speaker, headphone, or contact devices. Concerning display techniques, and depending on the device, approaches can be followed ranging from the traditional mono or stereophonic reproduction up to solutions affording more precise sound source localization, through binaural listening or physical distribution in space of one or more loudspeakers reproducing the virtual source.

Several rendering paradigms have been tested for the synthesis of auditory displays [102]. In recent years, some such paradigms have been challenged in the novel context of interactive display, in which the dynamic evolution of an auditory scene can be obtained by trading off real time constraints and computational burden.

Besides common headphone and room loudspeaker systems, more unusual solutions have been developed for the purpose of realizing ground-level auditory displays: among these, small speakers or insole shakers that are directly mounted into the shoes, and audio-tactile devices placed under the floor. In some cases, walking sounds are delivered by these devices as a by-product of broadband vibrotactile transduction, resulting also in sound as part of the overall reproduced mechanical energy.

7.4.1 Soundscapes rendering and displaying

When exploring a place by walking, at least two categories of sounds can be identified: the person's own footsteps and the surrounding soundscape. In the movie industry, footsteps sounds represent important elements. Chion writes of footstep sounds as being rich in what he refers to as *materializing sound indices*—those features that can lend concreteness and materiality to what is on-screen, or contrarily, make it seem abstracted and unreal [59]. Studies on soundscape originated with the work of R. Murray Schafer [265]. Among other ideas, Schafer proposed soundwalks as empirical methods for identifying a soundscape for a specific location. In a soundwalk people are supposed to move in a specific location, noticing all the environmental sounds heard. Schafer claimed that every place has a sound signature. The idea of experiencing a place by listening has been originally developed by Blesser [41]. By synthesizing technical, aesthetic and humanistic considerations, he describes the field of aural architecture and its importance in everyday life.

For VR purposes, three classes of soundscapes can be identified: static, dynamic, and interactive.

- *Static* soundscapes diffuse an auditory scene regardless of any specific localization effect.
- *Dynamic* soundscapes render the spatial position of one or more sound sources, even dynamically in space, regardless of any user input.
- *Interactive* soundscapes render the auditory scene also as a result of the actions and gestures of the user(s), which for instance can be tracked by a motion capture system. As an example of sound interaction, one can imagine the simulation of a forest, with sounds of fleeing animals following by the movements of a listener furthermore engaged in a walking task.

An engine has been realized in Max/MSP able to provide soundscapes belonging to any of these three classes. To include dynamic and interactive features, the ambisonic tools[4] for Max/MSP were used. Such tools, in fact, allow to move virtual sound sources along trajectories defined on a three-dimensional space [264]. At present, the engine can manage up to sixteen independent virtual sound sources, one to display the user's footsteps and the remaining fifteen handling the external sound sources populating the soundscape.

7.4.2 How to combine footsteps and soundscapes

VR studies made in the field of sound delivery methods and sound quality have recently shown that the addition of environmental cues can lead to measurable enhancement in the sense of presence [280, 61, 263]. Recently, the role of self-produced sound to enhance sense of presence in VE has been investigated. By com-

[4] Available at http://www.icst.net/research/projects/ambisonics-tools/

bining different kinds of auditory feedback consisting of interactive footstep sounds created by ego-motion with static soundscapes, it was shown how motion in VR is significantly enhanced when moving sound sources and ego-motion are rendered [208].

Specifically in our simulations, a number of soundscapes have been designed according to statistically significant indications given by subjects concerning the sonic ecologies they imagined for a specific environment, e.g. a forest. Such soundscapes were composed mainly by assembling freely available recorded material, like that existing in the Hollywood Edge sound effects library and the Freesound website.

A crucial step for the production of ecologically correct soundscapes consists of balancing the loudness of the background sounds with that of the footsteps, conversely lying in the foreground. This balance was again determined by users during magnitude-adjustment experiments, in which subjects were asked to find out the correct trade-off between the loudness of their own footsteps and the surrounding sounds.

7.5 Evaluating the engines

This section reports on experiments, that were conducted for evaluating the models and techniques described in the previous sections of this chapter.

7.5.1 Auditory recognition of simulated surfaces

The ability of subjects to identify different synthetic ground materials by listening during walking was investigated. In this experiment, subjects were asked to recognize the sounds in an active setting involving microphone acquisition at foot level and subsequent envelope extraction [65] of the subject's walking action. For this reason, the setting was acoustically isolated and subjects were asked to avoid producing sounds other than those generated by their own walking.

7.5.1.1 Methodology and protocol

Sounds were synthesized in real time using the recipes listed in Section 7.3, while subjects were walking across the isolated environment described above.

Participants were exposed to 26 trials, for a total presentation of 13 stimuli each displayed twice in randomized order. The stimuli consisted of footstep sounds on the following surfaces: beach sand, gravel, dirt plus pebbles (like in a country road), snow (in particular deep snow), high grass, forest underbrush (a forest floor composed by dirt, leaves and branches breaking), dry leaves, wood, creaking wood and

metal. To increase the ecology of the experiment, footstep sounds on wood, creaking wood and metal were enriched by including some standard room reverberation.

Fifteen participants (six male and nine female), aged between 19 and 29 (mean 22.13, std 2.47), took part in the experiment. All participants reported normal hearing conditions and were naive with respect to the experimental setup and to the purpose of the experiment. They wore sneakers, trainers, boots and other kinds of shoes with rubber sole.

Participants were asked to wear a pair of headphones and to walk in the area delimited by the microphones. They were given a list of different surfaces to be held in one hand, presented as non-forced alternate choice. The list of surfaces presented to the subjects is outlined in the first row of Table 7.3. It represents an extended list of the surfaces the subjects were exposed to.

At the end of the experiment, subjects were asked to answer some questions concerning the naturalness of the interaction with the system. Every participant took on average 24 minutes to complete the experiment.

7.5.1.2 Results

Table 7.3 shows the confusion matrix which displays the results of the experiment. The first row lists the materials that could be chosen, while the first column lists the

	BS	GL	DR	SW	HG	UB	DL	WD	CW	MT	WR	CR	MR	FS	CC	PD	WT	CP	—
BS	15	2		5			1							2				1	4
GL		21	2			1	1							4					1
DR		1	3	2		6	6		1					10					1
SW				24	1									4					1
HG	2	7	3	1	0	3	7			2									5
UB		1	3	1		19	1							3				1	1
DL	1	3	5			5	12							4					
WD			1	2				14		1						1	1		10
CW								1	28					1					
MT								1		24				1	2				2
WR									3	11	6				7				3
CR												28			1				1
MR								1					25	1	1				2

Abbreviations:

WD wood	CW creaking wood	SW snow	UB underbrush
— don't know	FS Frozen snow	BS beach sand	GL Gravel
MT metal	HG High grass	DL dry leaves	CC concrete
DR dirt	PD puddles	WT Water	CP carpet
WR wood reverb	MR metal reverb	CR creaking+ reverb	

Table 7.3: Confusion matrix: recognition of synthesized footstep sounds.

materials simulated in the stimuli subjects were exposed to. The decision of provid-

ing a wider choice of materials was taken to minimize the statistical significance of subjects guessing at random.

From this table, it is possible to notice how surfaces such as snow, creaking wood with and without reverberation, gravel and metal with reverberation were correctly recognized in a high number of trials. Recognition of surfaces such as dirt plus pebbles, high grass and wood appeared to be low. An analysis performed on the wrong answers reveals that on average subjects tended to mistakenly spread judgments over surfaces belonging to the same category (e.g., wood versus concrete, snow versus frozen snow, dry leaves versus forest underbrush) while keeping different categories distinct in their judgments (e.g., wood versus water, wood versus gravel, metal versus dry leaves).

Moreover, results show that the addition of reverberation to the sounds resulted in better recognitions for metal, and worse for wood, which was perceived most of the times as concrete. Overall, recognition rates are similar to those measured on recorded footstep sounds [210].

7.5.2 Salience of temporal and spectral cues of walking

Based on the parametric synthesis model described in Section 7.2.5, an experiment has been performed aiming to understand the salience of auditory cues of walking.

The experimental hypothesis was kept simple, by relying on a classification of such cues in spectral and temporal. Furthermore the experiment itself was made offline, in this way focusing on the auditory feedback alone while excluding vision and touch. The complete report on this activity can be found in [94].

We hypothesized that the perception of solid materials is mainly determined by *spectral* cues, conversely the perception of aggregate materials is mainly determined by *temporal* cues. In particular, we experimented using concrete (C) and wooden (W) floors, representative of solid materials, as well as with gravel (G) and dried twigs (T), representative of aggregate materials. The former, such as concrete, marble, wood, are stiff. The latter, such as gravel, dry leaves, sand, allow relative motion of their constituent units and progressively adapt to the sole profile during the interaction. Now,

- solid materials give rise to short, repeatable impacts having a definite spectral color;
- aggregate materials elicit sequences of tiny impacts having distinctive temporal density, that create a sort of “crumpling”, less resonant sound.

Figure 7.4 illustrates the hypothesis.

MATERIAL	PHYSICAL PROPERTIES	ACOUSTIC PROPERTIES
C , W	Solid	Spectral Cues
G , T	Aggregate	Temporal Cues

Fig. 7.4: Experimental hypothesis. (C: concrete, W: wood, G: gravel, T: twigs.)

7.5.2.1 Methodology and Protocol

Subjects were sitting in front of a Mac Pro PC running a Java application communicating (via the *pdj* library) with Puredata, a free software environment for real time audio synthesis also enabling simple visualizations (through the *GEM* library). They listened to the auditory stimuli through a pair of AKG K240 headphones.

Thirteen male and three female undergraduate computer science students aged aged 22 to 31 (mean = 24.62, std = 2.55) participated in the experiment. Few of them had some experience in sound processing. All of them reported to usually wear snickers.

At the end of the experiment, every subject completed a subjective questionnaire about the realism and ease of identification of the audio stimuli.

One footstep by a normally walking male wearing leather shoes was repeatedly recorded while he stepped over a tray filled with gravel and, then, dried twigs. Recordings were made inside a silent, normally reverberant room using a Zoom H2 digital hand recorder standing 0.5 m far from the tray. For either material, seven recordings were selected and randomly enqueued to create walking sequences lasting 12 s and containing 13 footsteps. In addition to the in-house recordings, high quality samples of a male walking on concrete and on wooden parquet were downloaded from the commercial database `sounddogs.com`. Using these samples, two further walking sequences were created having the same beat and average Sound Pressure Level as of those based on in-house recordings.

Temporal envelopes were extracted from every sequence, by computing the signal

$$e_M[n] = (1 - b[n])|s_M[n]| + b[n]e_M[n-1] \tag{7.10}$$

out of the corresponding sequence s_M, $M \in \mathscr{M} = \{C, W, G, T\}$. (Refer to Figure 7.4 for the meaning of the C, W, G, and T.) As in previous research on synthetic footsteps, the envelope following parameter $b[n]$ was set to 0.8 when $|s_M[n]| > e_M[n-1]$, and to 0.998 otherwise [65]. By following the input when its magnitude is greater than the envelope, and by in parallel allowing a comparably slow decay of the envelope itself when the same magnitude is smaller, this setting ensures that amplitude peaks are tracked accurately, while leaving spurious peaking components off the envelope signal e_M.

By dividing every sequence s_M by its envelope e_M, we computed signals $u_M = s_M/e_M$ in which the temporal dynamics was removed. In other words, we manipulated the footstep sequences so to have stationary amplitude along time.

What remained in u_M was a spectral color, that we extracted with a 48th-order inverse LPC filter h_M^{-1} estimated in correspondence of those parts of the signals containing footstep sounds. Using this filter order, if training the model using *one* footstep then we could not detect differences between the original sound and the correspondingly re-synthesized footstep. We emphasize that the resulting LPC filter in any case estimated one single transfer function, independently of the number of footsteps taken from the original sequence which informed the model. Since we trained the estimator with the entire sequence, the re-synthesized sound had a slightly different color compared to any other footstep belonging to the original sequence.

In the end, for every material M a highly realistic version $\tilde{s}_M$ of the original sequence s_M could be re-synthesized by convolving digital white noise w by the "coloring" filter h_M, and then multiplying its output, i.e. the synthetic version $\tilde{u}_M$ of u_M, by the envelope signal e_M:

$$\tilde{s}_M[n] = (w * h_M)[n] \cdot e_M[n] = \tilde{u}_M[n] \cdot e_M[n]. \tag{7.11}$$

This technique draws ideas from a family of physically-informed models of walking sounds [65, 303]. In the meantime it provides a simpler, more controlled re-synthesis process avoiding stochastic generation of patterns as in such models. In our case, the silent parts of the four envelopes were tailored to generate synthetic sequences having identical walking tempos. This simple manipulation ensured seamless mutual exchange of the envelopes among sequences, as explained in the following.

Sixteen stimuli were created by adding twelve *hybrid* re-syntheses to the *native* stimuli $\tilde{s}_C$, $\tilde{s}_W$, $\tilde{s}_G$, and $\tilde{s}_T$. Every hybrid stimulus $\tilde{s}_{M_t,M_f}$, $M_t, M_f \in \mathscr{M}$ was defined as to account for the spectral color of material M_f *and* the temporal envelope of material $M_t \neq M_f$:

$$\tilde{s}_{M_t,M_f}[n] = (w * h_{M_f})[n] \cdot e_{M_t}[n] = \tilde{u}_{M_f}[n] \cdot e_{M_t}[n]. \tag{7.12}$$

For each material M_f, we checked that all hybrid temporal manipulations using $M_t \neq M_f$ did not notably alter the spectral information of $\tilde{s}_{M_f}$, and thus its original color. In fact, an inspection of the spectra $E_M(\omega)$ of the various envelopes made by Fourier-transforming e_M, i.e., $E_M(\omega) = \mathscr{F}\{e_M\}(\omega)$, shows that they all have a comparable spectrum. More precisely, all spectra E_C, E_W, E_G, E_T exhibit similar magnitudes, that are shown in Figure 7.5 after removing the respective dc component for ease of inspection. This means that the spectral differences in $\tilde{S}_{M_t,M_f}(\omega)$ caused by multiplying $\tilde{u}_{M_f}$ by e_{M_t}, that is,

$$\tilde{S}_{M_t,M_f}(\omega) = \mathscr{F}\{\tilde{u}_{M_f} \cdot e_{M_t}\}(\omega) = (\tilde{U}_{M_f} * E_{M_t})(\omega), \tag{7.13}$$

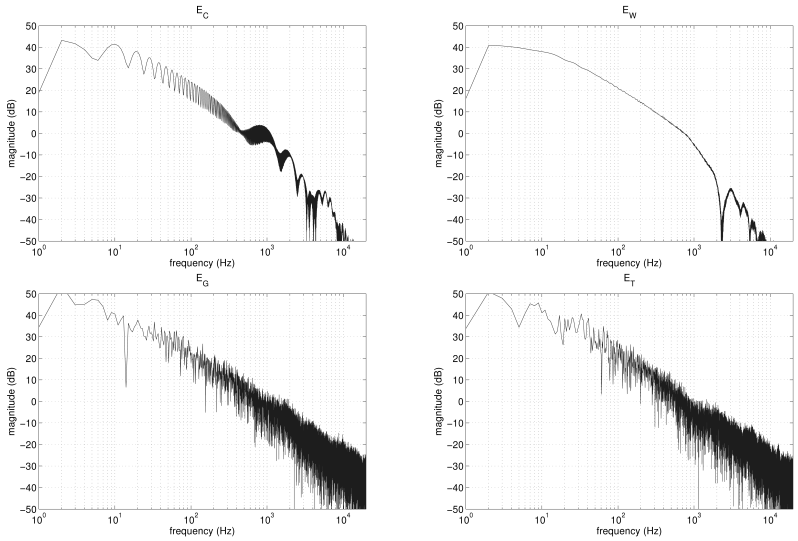

Fig. 7.5: Magnitude spectra of envelopes E_C, E_W, E_G, and E_T. Respective dc components removed for ease of inspection.

are substantially independent of the material, hence almost identical to those introduced in $\tilde{s}_{M_f}$ by its own envelope e_{M_f}.

Symmetrically, the temporal artifacts which are caused by hybridization between two different materials can be considered minor. In fact, because of the LPC design methodology, all filters h_C, h_W, h_G, h_T do transform white noise into a stationary signal independently of the material.

To become confident with the auditory stimuli, subjects trained for some minutes before starting an individual session by selecting and playing each one of the original sequences s_C, s_W, s_G, s_T for several times. Each sequence could be selected by clicking on the corresponding software button in a graphic interface.

Each individual session consisted of 192 trials, obtained by randomly playing each one of the sixteen synthetic stimuli for twelve times. At each trial the subject listened to a stimulus, and selected one material by clicking the corresponding button in the interface. When the button was released, the screen froze for two seconds and changed color to inform subjects of the conclusion of the trial. After this short pause, a new trial was performed.

The four buttons randomly switched position at each trial. Subjects could temporarily stop the experiment by clicking the pause icon '||' located in the middle of the screen, whenever they wanted to take a short break among trials. It took approximately 45 minutes for each participant to complete the session.

7.5.2.2 Results

For each participant, percentages of selection for the four materials C, W, G, T were analyzed. We considered the fraction of participants who showed significantly correct (random is 25%) material recognitions from the four synthetic stimuli $\tilde{s}_C$, $\tilde{s}_W$, $\tilde{s}_G$, $\tilde{s}_T$, across the twelve repetitions. The critical value (with $\alpha = 0.05$) of the one-tailed binomial test Bin(12,0.25) is 7 trials (i.e., 58.33%): only the participants with an auditory recognition of the original materials higher than 58.33% were included in the analysis. After this check, 16 participants were considered for the recognition of dried twigs, 15 for gravel, 16 for wood, and 10 for concrete.

The results of the analysis are presented in Figure 7.6. In these plots, a bar exhibiting a low percentage means that the correspondingly substituted information (either temporal or spatial) is important for the recognition of the original material, represented by the leftmost bar in the same plot. The difference from random percentage (25%) was tested using one-proportion (two-tailed) z tests.

By aggregating the data, we also evaluated the auditory recognition of material categories. Again, this analysis was conducted using data from participants exhibiting an auditory recognition significantly higher than random concerning the two sets of stimuli accounting for the respective categories (24 trials for each category). In this case, the critical value (this time computed by a one-proportion/one-tailed z test to account for the larger number of trials, with $\alpha = 0.05$) is 10 trials, corresponding to 41.67%. All the participants passed the check.

The results of the new analysis are presented in Figure 7.7. For the different percentages of selection, the difference relative to random (25%) was tested using one-proportion (two-tailed) z tests. Thus, for the aggregate category, the percentages of selection in native (82.29%) and frequency manipulated (52.78%) conditions were significantly different from random ($z = 25.98$, $p < 0.001$ and $z = 21.77$, $p < 0.001$, respectively). By contrast, the percentage of selection in the time manipulated condition (23.44%) was not significantly different from random ($z = -1.22$, $p = 0.22$). On the other hand, for the solid category, the percentages of selection in native (78.39%) and frequency manipulated (35.59%) conditions were significantly different from random ($z = 24.16$, $p < 0.001$ and $z = 8.30$, $p < 0.001$, respectively). By contrast, the percentage of selection in the time manipulated condition (26.65%) was not significantly different from random ($z = 1.29$, $p = 0.20$). The differences between the three audio conditions were tested with two-proportion (two-tailed) z tests.

A correction for experiment-wise error was realized by using Bonferroni-adjusted alpha level (p divided by the number of tests). Thus, in order to compare the three audio conditions (native, frequency manipulated, and time manipulated), the alpha level was adjusted to $p = 0.05/3 = 0.0167$. For the aggregate category, the analysis showed that the native condition was significantly different from the frequency manipulated ($z = 10.23$, $p < 0.05$) and time manipulated ($z = 20.56$, $p < 0.05$) conditions. The difference between frequency manipulated and time manipulated conditions was significantly different ($z = 14.50$, $p < 0.05$) as well. For the solid category, the analysis indicated that the native condition was significantly different from

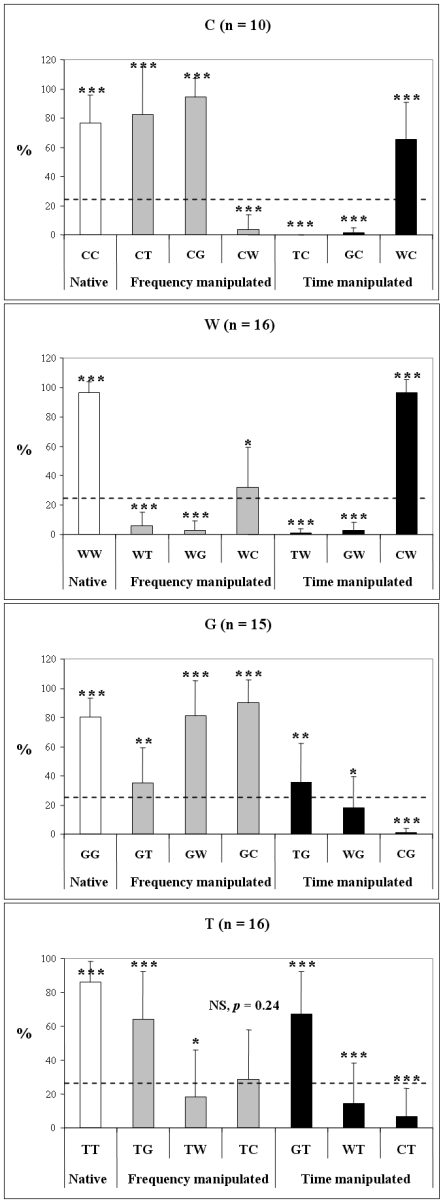

Fig. 7.6: Mean percentages of selection (lines represent std) for C,W,G,T as a function of the auditory stimulus $\tilde{s}_{M_t,M_f}$. The difference from random selection (line at 25%) was tested using one-proportion (two-tailed) z tests. Note: $*: p < 0.05$, $**: p < 0.01$, $***: p < 0.001$, NS: not significant, n: number of subjects.

the frequency manipulated ($z = 14.57$, $p < 0.05$) and time manipulated ($z = 17.96$, $p < 0.05$) conditions. The difference between frequency manipulated and time manipulated conditions was also significantly different ($z = 4.63$, $p < 0.05$).

After the experiment, a questionnaire was proposed in which each participant had to grade from 1 to 7 the four native stimuli according to two subjective crite-

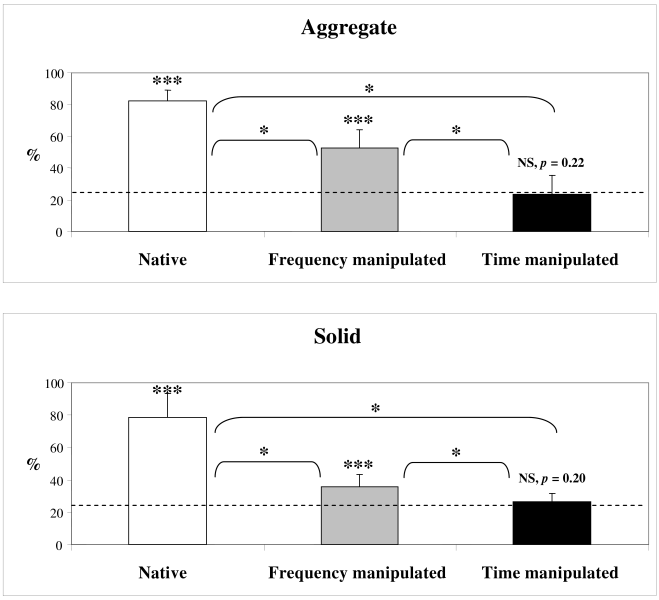

Fig. 7.7: Mean percentages of selection (lines represent std) for material categories (Aggregate and Solid) as a function of the auditory stimulus $\tilde{s}_{M_t,M_f}$. The difference from random selection (line at 25%) was tested using one-proportion (two-tailed) z tests. The differences between the three audio conditions were tested with two-proportion z tests (two-tailed and Bonferroni-adjusted alpha level with $p = 0.05/3 = 0.0167$). Note: $*: p < 0.05$, $**: p < 0.01$, $***: p < 0.001$, NS: not significant.

ria: realism, and ease of identification. Wilcoxon signed rank (two-tailed) tests with Bonferroni correction showed significant differences only for the realism of sounds: between concrete and dried twigs ($z = -3.28,\ p = 0.001$), and between concrete and gravel ($z = -3.16,\ p = 0.0016$).

7.5.2.3 Discussion and Conclusions

The histograms for concrete and wood in Figure 7.6 show that subjects tolerate swapping between the temporal features of C and W, both belonging to the solid category, conversely the substitution in the same signals with temporal features extracted from aggregate materials (i.e. G and T) harms the recognition. This result is in favor of the initial hypothesis. The effect of spectral manipulations of C and W is more articulate. In this case the hypothesis is essentially confirmed with wood, whose distinctive color cannot be changed using any other spectrum. In parallel, subjects are tolerant to substitutions in C with spectra from aggregate materials. This greater tolerance may be due to the basic lack of distinct color of concrete floors, especially for listeners who usually wear rubber sole shoes such as snickers (indeed the majority of our sample). The same conclusion finds partial confirmation

by the greater confidence shown by subjects in recognizing aggregate materials, in the limits of the significance of these data.

The histograms in Figure 7.6 regarding gravel and twigs partially support the initial hypothesis. Time swaps between G and T are tolerated to a lesser extent compared to solid floors. Like before, substituting the temporal features of solid materials in an aggregate sound is not tolerated. Spectral substitutions are not as destructive as they were for solid materials, especially in the case of gravel. The worst situation is when the spectrum of W is substituted in T, again probably due to the distinct color that wood resonances bring into the sound.

Figure 7.7 would further support this discussion. In fact, in spite of the low significance of the data from time manipulations (i.e. black bars), it shows that subjects are primarily sensitive to temporal substitutions between solid and aggregate materials. In parallel, spectral changes are more tolerated during the recognition of aggregate material compared to solid floors.

The proposed experiment has confirmed that solid and aggregate floor materials exhibit precise temporal features, that cannot be interchanged while designing accurate walking sounds. Within such respective categories, spectral color represents an important cue for the recognition of solid materials, conversely sounds of aggregate materials seem to tolerate larger artefacts in their spectra.

7.5.3 Evaluation of soundscapes: interactivity

In two preliminary experiments [300] it was investigated how subjects react to different soundscape dynamics while walking in a virtual auditory scene.

The task of the first experiment was to walk across a circular perimeter inside the walking area in a room. Eight loudspeakers were placed at the angles and middle points of each side of a rectangular floor. The user's position was tracked by a MoCap system, and then used to localize synthetic footsteps along the task through the speakers. Localization was performed using the ambisonics tools for Max/MSP, allowing to move virtual sound sources on a three-dimensional space (refer to Section 7.4.1). During the experiment the loudspeakers were hidden by opaque, acoustically transparent curtains.

During the walk, subjects were exposed to six conditions—again, refer to Section 7.4.1: i) static soundscape; ii) coherent interactive soundscape; iii) incoherent interactive soundscape; iv) static soundscape with static distractors; v) coherent interactive soundscape with dynamic distractors; vi) incoherent interactive soundscape with dynamic distractors. Incoherent means that the footstep sounds were localized opposite to the actual position of the moving subject in the walking area.

Distractors consisted of footstep sounds of a phantom subject walking in the same area. Specifically, static distractors were rendered by displaying their sound with equal intensity from all room loudspeakers, whereas dynamic distractors covered a triangular trajectory inside the circular perimeter.

Participants were exposed to twelve trials, where the six conditions were presented twice in randomized order. Performing each trial took about one minute. Each condition was presented using virtual floors made of wood and forest underbrush. The reason for choosing one solid and one aggregate material was to discern whether the surface type affected the results.

The distractors displayed the same virtual surfaces, but at lower intensity, slightly different timbre, and moderately faster gait cycle.

After the presentation of each stimulus, participants were required to answer the following questions on a seven-point Likert scale:

- How well could you localize under your feet the footstep sounds you produced?
- How well did the sounds of your footsteps follow your position in the room?
- How much did your walk in the VE seem consistent with your walk in the real world?
- How natural did your interaction with the environment seem?
- To what degree did you feel confused or disoriented while walking?

Our hypotheses were that a coherent interactive soundscape would result in improved subjective appreciation and consequent rating; that the incoherent dynamic condition would have been judged as the worst; and that the use of distractors would have decreased the subjective appreciation of the perceived auditory scene, hence its rating.

First, a significant difference was found between the surface materials for what concerns the coherent interactive and static condition, in the case of absence of distractors: the difference between such conditions is negligible in the case of wood, whereas this difference becomes significant in the case of forest underbrush ($p < 0.0001$). Conversely, the same difference was not significant in presence of the distractors.

Secondly, for both materials the incoherent interactive condition gave rise to poorer evaluations in terms of localization, following, consistency and naturalness as well as to less confidence on orientation, both in presence and in absence of distractors. In detail, for both materials significance was found concerning the difference between the coherent and incoherent interactive conditions (for wood: $p < 0.001$ and $p < 0.01$; for forest: $p < 0.000001$ and $p < 0.05$, both respectively with and without distractors), as well as between the static and incoherent interactive conditions (for wood: $p < 0.01$ and $p < 0.01$; for forest: $p < 0.000001$, both respectively with and without distractors).

Thirdly, for both materials the evaluations in absence of distractors were almost always better than in presence of them concerning localization, following, consistency, naturalness and orientation. This difference was significant for the forest underbrush case ($p < 0.05$), whereas for the wood it was not.

For both materials, the disorientation was higher in presence rather than in absence of distractors, but significant differences between these two conditions were found only for the forest underbrush case ($p < 0.05$). The incoherent interactive with distractors condition was rated as the most disorienting for both materials.

Conversely, for the forest underbrush case the coherent interactive condition was rated as the least disorienting.

Overall, the coherent interactive condition gave rise to significantly better results than the static one concerning the forest underbrush case in absence of distractors. A subsequent analysis for each of the investigated parameters revealed significant differences between the two conditions only for the "naturalness" parameter ($p < 0.05$).

It is therefore possible to conclude that users can perceive that their interaction with the VE is neither realistic nor natural when the source is not moving coherently with their position. The hypothesis concerning the distractors was confirmed: for both materials, almost always the evaluations in absence of distractors were better than when the distractors were present, although significant differences were found only for the forest case. In addition, the disorientation was higher in presence of distractors (but significant only for the forest case). This evidence suggests that the use of distractors, i.e., walking sounds evoking the presence of another person walking in the same room as the subject, is likely to influence the perception of self-produced footstep sounds.

Starting from the results of the first experiment we designed a second experiment, investigating in more detail the subjective perception of the static and coherent interactive soundscapes. The task consisted of walking freely inside the walking area. Participants were exposed to fourteen trials, where seven virtual surface materials were randomly presented in presence of both types of soundscapes. Such materials, five aggregate and two solid, consisted of gravel, sand, snow, dry leaves, forest underbrush, wood and metal.

Each trial lasted about one minute. After the presentation of each stimulus participants were required to evaluate, on a seven-point Likert scale, the same questions presented in the first experiment.

The goal of this experiment was to assess whether participants showed a preference for either display method, while exploring the VE during a free walk (i.e., without any predefined trajectory like in the first experiment). Furthermore we were interested in assessing whether the surface property affected the results.

Results show that participants did not show any preference for either method. The answers to the questionnaire were very similar for all the surfaces, with no significant differences. This result suggests that both methods could be used in a VE, to deliver interactively generated footstep sounds. However, other tests should be conducted to add quantitative elements to this conclusion.

7.5.4 Evaluation of soundscapes: ecology

An experiment was conducted aiming at understanding the role of soundscapes in creating a sense of place and context when designing a virtual walking experience [300]. More in detail, the goal of the experiment was to investigate the ability of subjects to recognize the different walking sounds they were exposed to in three

conditions: without soundscape, with ecologically coherent soundscape (e.g. footsteps sounds on a soundscape reporting of a beach) and with ecologically incoherent soundscape (e.g. footsteps sounds on a soundscape reporting of a ski slope).

The interactive footsteps were synthesized in real time while subjects were walking using the system described in Section 7.4.1. Offline, the following soundscapes were built: a crowded beach, the courtyard of a farm, a ski slope, a forest, and a garden with trees during fall. All soundscapes were diffused as static.

The task was to recognize the surface material, as well as to evaluate the realism and quality of the footstep sounds. In the conditions with soundscape, participants were also asked to recognize the surrounding environment in which they were walking.

Results showed that the addition of a coherent soundscape resulted in a better recognition of the surfaces, along with a higher realism and quality of the proposed sound compared to the conditions without and with incoherent soundscape.

For some surface materials, adding a coherent soundscape significantly improved the surface recognition compared to the case in which the soundscape was not provided, and this happened especially with materials whose recognition was difficult without soundscape. Similarly, the percentages of correct answers were higher in the condition with coherent soundscape compared to the condition with incoherent soundscape, significantly for some materials. Furthermore, the same percentages were higher in the condition without soundscape compared to the condition with incoherent soundscape. As expected, adding an incoherent soundscape created an ecological mismatch which often confused the subjects.

The analysis of the wrong answers reveals that in all the experiments none of the proposed aggregate surfaces was confused with a solid one. This means that subjects were able to robustly identify the type of surface.

Regarding the evaluations on the realism and quality of the footsteps sounds in the three conditions, higher evaluations were found in the condition with coherent soundscape compared to the condition without and with incoherent soundscape, as well as for the condition without soundscape compared to that with incoherent soundscape. For some materials these evaluations were statistically significant.

Additionally, the percentages of correct guess of the soundscape were higher with coherent rather than incoherent soundscape.

Overall, subjects observed that soundscapes play an important role in ground surface recognition, precisely for their ability to create a context. Especially in presence of conflicting information, as it was the case with incoherent soundscapes, subjects tried to identify the strongest ecological cues in the auditory scene while performing their recognitions.

This experiment gives strong indication of the importance of context in the recognition of a virtual auditory scene, where walking sounds generated by subjects and soundscapes are combined. Though, it is only a preliminary investigation: further experiments are needed to gain a better understanding of the cognitive factors involved when subjects are exposed to different sound events, especially when a situation of semantic incongruence is present.

7.6 Conclusions

This chapter provided a description of how to synthesize walking sounds using physics-based and physically inspired models, including recipes for constructing and parameterizing model compositions dictated by the designer's experience and taste. Several surfaces have been simulated, both solid and aggregate. The simulations work in real time and are controlled by kinds of input devices such as those described in Chapter 2.

After reporting on current auditory display possibilities, we also described experiments whose aim was to assess the ability of subjects to recognize the simulated surfaces, the saliency of temporal and frequency cues in footstep sounds, the role of soundscapes in enhancing the interactivity and ecological realism of an auditory scene. These experiments validate the quality of the proposed synthesis engines, and testify their possibilities and limits to faithfully recreate virtual walking experiences.