Chapter 2
Procedural Modeling of Interactive Sound Sources in Virtual Reality

Federico Avanzini

Abstract This chapter addresses the first building block of sonic interactions in virtual environments, i.e., the modeling and synthesis of sound sources. Our main focus is on procedural approaches, which strive to gain recognition in commercial applications and in the overall sound design workflow, firmly grounded in the use of samples and event-based logics. Special emphasis is placed on physics-based sound synthesis methods and their potential for improved interactivity. The chapter starts with a discussion of the categories, functions, and affordances of sounds that we listen to and interact with in real and virtual environments. We then address perceptual and cognitive aspects, with the aim of emphasizing the relevance of sound source modeling with respect to the senses of presence and embodiment of a user in a virtual environment. Next, procedural approaches are presented and compared to sample-based approaches, in terms of models, methods, and computational costs. Finally, we analyze the state of the art in current uses of these approaches for Virtual Reality applications.

2.1 Introduction

Takala and Hahn [86] were possibly the first scholars who proposed a sound rendering pipeline, in analogy with the image rendering pipeline, aimed at producing an overall "soundtrack" starting from a description of the objects in an audio-visual scene. Their pipeline included sound modeling and sound rendering stages, running in parallel with the image rendering pipeline. Figure 2.1 proposes an updated picture, which considers several aspects investigated by researchers throughout the last three decades and may represent a general pipeline for sound simulation in Virtual Reality (hereinafter, VR).

Much of recent and current research is concerned with aspects related to the "Propagation" and "Rendering" blocks represented in this figure, as well as the

F. Avanzini (✉)
Laboratory of Music Informatics, Department of Computer Science, University of Milano,
Via G. Celoria 18, IT-20135 Milano, Italy
e-mail: federico.avanzini@di.unimi.it

M. Geronazzo and S. Serafin (eds.), *Sonic Interactions in Virtual Environments*,
Human—Computer Interaction Series, https://doi.org/10.1007/978-3-031-04021-4_2

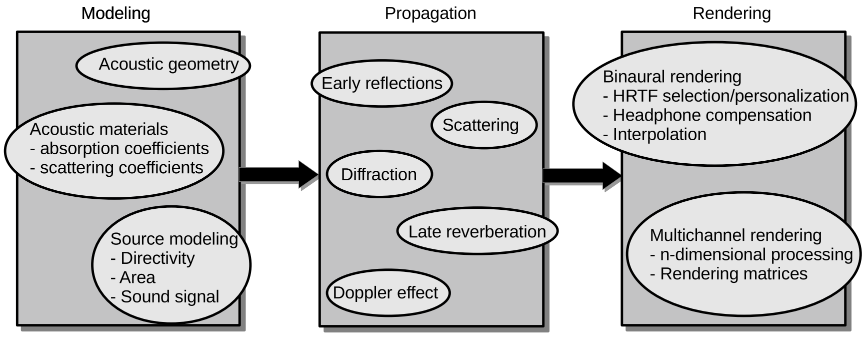


Fig. 2.1 A general pipeline for sound simulation in Virtual Reality (figure based on [51])

geometrical and material properties of acoustic enclosures in the "Modeling" block. This chapter focuses instead on the remaining balloon of the "Modeling" block, the modeling of *sound sources*.

One obvious motivation for looking into sound source modeling is that all sounds occurring in a virtual (and in a real) environment originate from some sources, before propagating into the environment and finally reaching the listener. Secondly, many of the sonic interactions occurring in a virtual environments are interactions between the subject's avatar and sound sources. Here, our definition of *interactive* is analogous to the one given by Collins [20] for video-game audio: whereas adaptive audio generically refers to audio that reacts appropriately to events and changes occurring in the simulation, interactive audio refers to sound events occurring directly in reaction to avatar's gestures (ranging from pressing a button to walking or hitting objects in the virtual scene).

The current dominant paradigm in VR audio, largely based on sound samples[1] triggered by specific events generated by the avatar or the simulation, is minimally adaptive and interactive. This is the main motivation for looking into *procedural* approaches to sound generation.

2.2 What to Model

The first question that should be asked is as follows: what are the sound sources that need to be modeled in a virtual environment, and how can these be organized into a coherent and comprehensive taxonomy? Such a taxonomy would provide a useful tool to analyze in a systematic way the state of the art of the research in this field and possibly to spot research directions that are still under-explored.

[1] For the sake of clarity, in this chapter, we use the term "sample" in its commonly accepted meaning of pre-recorded/pre-processed sound excerpt, rather than that of a single value of a digital signal.

2.2.1 Diegetic Sounds

One first possible and often used distinction can be mutated from narrative theory. The term *diegesis* has been used in film theory to refer to the fictional world of the film story, and correspondingly the adjective *diegetic* refers to elements that are part of the depicted fictional world. By contrast, non-diegetic elements are those which should be considered non-existent in the fictional world.

As far as sound in particular is concerned, three main categories are traditionally used in films: speech and dialogue, sound effects, and music [80]. The first two categories comprise diegetic sounds, while music is a non-diegetic element having mostly an affective and emotional role, a distinction that may be related to the motto "Sound effects make it real, music makes you feel" [49].

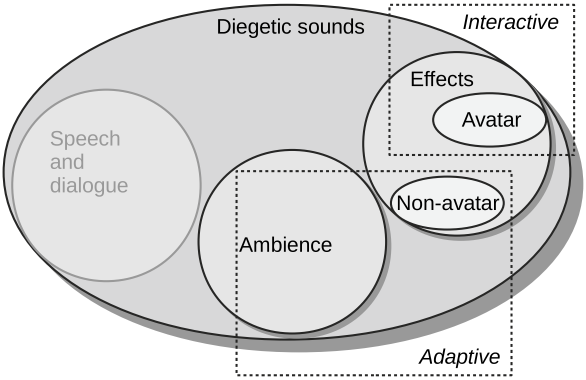
Several taxonomies for sounds in video-games have been proposed and are typically based on similar categories [42]. These may be employed in the context of VR as well, with the additional caveat that VR applications only partly overlap with video-games. In particular, VR, and immersive VR specifically, may be defined as "a medium in which people respond with their whole bodies, treating what they perceive as real" [77]. In light of this definition, in this chapter, we focus on diegetic sounds, those that "make it real": in other words, those that contribute most to the overall sense of the presence of a user within a virtual environment, which we will discuss in Sect. 2.3.

An interesting example of a taxonomy for sound in games is provided by Stockburger [84], who considers five different types of sound objects. Non-diegetic elements include (i) music, but also (ii) interface sounds, which may sometimes be included into the diegetic part of the game environment; proper diegetic elements instead comprise the three categories of (iii) speech and dialogue, (iv) ambience (or "zone" sounds in Stockburger's definition), and (v) effects.

Speech and dialogue are very relevant components of a virtual environment; however, our focus in this chapter is on non-verbal sound. The distinction between ambience and effect sounds is mainly a perspectival one: the former are background sounds, connected to locations or zones (understood both as different spatial locations in an environment and different levels in a game) and having distinct auditory qualities; the latter are instead foreground sounds other than speech, that are cognitively linked to objects or events, and are therefore perceived as being produced by such objects and events. Sound-producing objects may be moving or static elements, may be directly interactable by the avatar or just synchronized to the visual simulation, or may be even outside the visual field of view.

Stockburger [84] proceeds in distinguishing effect subcategories, depending on the elements of the environment they are linked to. His classification is heavily tailored to games, but serves as an inspiration to further inspect and subdivide effect sounds. For the purpose of the present discussion, we only make a distinction between two subcategories: (i) effects linked to the avatar, and (ii) all remaining effects in the environment. Effects linked to the avatar are related to sounds produced by the avatar's movement or object manipulation: footsteps, swishing of an object cutting

Fig. 2.2 Categories and interactivity of diegetic sounds in a virtual environment



through the air, knocking on a wall, clothes, etc. They can also include sounds produced by the avatar's own body, such as breathing or scratching. The remaining effects in the environment may include non-verbal human sounds, sounds produced by human activities, machine sounds, and so on. A visual summary is provided in Fig. 2.2. The categories and subcategories identified here can be usefully mapped into interactive and adaptive sound sources.

2.2.2 Everyday Sounds

An orthogonal approach with respect to the previous one amounts to characterizing sound sources in terms of the physical mechanisms and events that are associated to those sources.

Typical lists of audio assets for games or VR include, at the second level of classification (after the branch between ambience and sound effects), such categories as footsteps, doors, wind and weather, and cars and engines, with varying degrees of detail. These categories in fact refer to objects and events that are physically responsible for the corresponding sounds; however, such classifications follow common practices rather than a standardized taxonomy. A more systematic categorization can be found in the classic works by Gaver [33, 34], who proposed an "ecological" categorization of everyday sounds (the ecological approach to auditory perception will be discussed in more detail in Sect. 2.3.2). Gaver derived a tentative map of everyday sounds, which is shown in Fig. 2.3 and discussed in the remainder of this section.

At the highest level, Gaver's taxonomy considers three broad classes of sounds: those involving vibrating solids, liquids, and aerodynamics in sound generation, respectively. Sounds generated by solid objects have patterns of vibrations structured by a number of physical attributes: those of the *interaction* that has produced the vibration, those of the *material* of the vibrating objects, and those of the *geometry* and configuration of the objects. Sounds involving liquids (e.g., dripping and splashing) also depend on an initial deformation that is counter-acted by restoring forces in

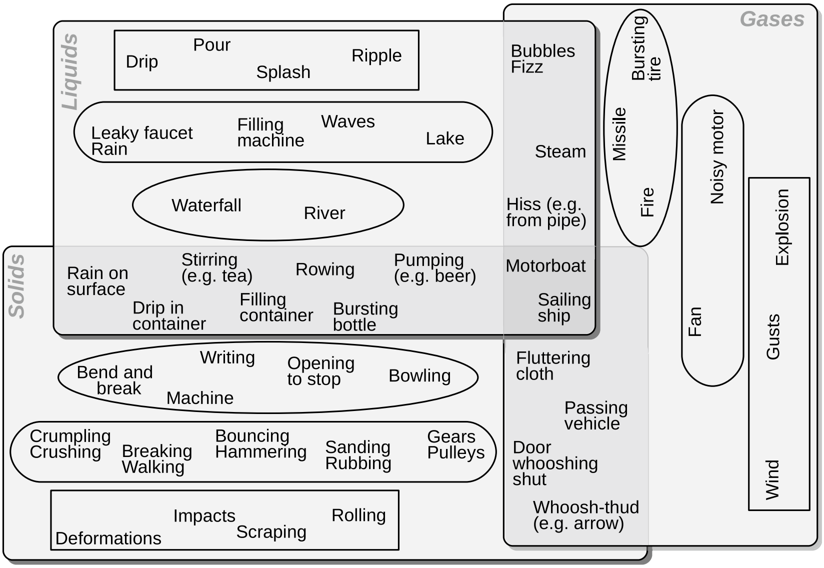


Fig. 2.3 A taxonomy of everyday sounds that may be present in a virtual environment. Within each class (solids, liquids, and gases), rectangles, rounded rectangles, and ellipses represent basic, patterned, and compound sounds, respectively. Intersections between classes represent hybrid sounds. Figure based on the taxonomy of everyday sounds by Gaver [34, Fig. 7]

the material, but no audible sound is produced by the vibrations of the liquid and instead the resulting sounds are created by the resonant cavities (bubbles) that form and oscillate in the liquid. Aerodynamic sounds are caused by the direct modification of atmospheric pressure differences from some source, such as those created by an exploding balloon or by the noise of a fan, or even events in which such changes in pressure transmit energy to objects and set them into vibration (e.g., when wind passes through a wire).

At the next level, sounds are classified along layers of complexity, defined as follows. “Basic” sound-producing events are identified for solids, liquids, and gases: sounds made by vibrating solids may be caused by impacts, scraping, or other interactions; liquid sounds may be caused by discrete drips, or by more continuous splashing, rippling, or pouring events; and aerodynamic sounds may be made by discrete, sudden changes of pressure (explosions), or by more continuous introductions of pressure variations (gusts and wind). “Patterned” sounds are situated at a higher level of complexity, as they are produced through temporal patterning of basic events. As an example, walking, breaking, bouncing, and so on are all complex events involving patterns of simpler impacts. Similarly, crumpling or crushing are examples of patterned deformation sounds. “Compound” sounds occupy the third level of complexity and involve more than one type of basic and patterned events. An example may be provided by the sound of a door slam, which involves the squeak of scraping hinges and the impact of the door on its frame, or a complex activity such as writing, which involves irregular temporal patterns of both impacts and scrapes. Compound sounds involve mutual constraints on their building components: as an example, concatenating the creak of a heavy door closing slowly with the slap of a light door slammed shut would arguably not sound natural.

Finally, Gaver's taxonomy also considers "hybrid" events, in which two or three types of material are involved. An example of a hybrid sound involving solids and liquids is the one produced by raindrops hitting a window glass, which involves attributes of both liquid and vibrating solid sounds.

A taxonomy such as the one discussed here has at least two very attractive features. First, it provides a comprehensive framework for classifying any everyday sound potentially encountered in our world (and thus in a virtual world as well), with a fine level of detail. Secondly, its hierarchical structure provides a theoretical framework that can aid not only the sound design process but also the development of sound design tools. An example of an ecologically inspired software library for procedural sound design will be discussed in Sect. 2.5.3.

2.3 Perceptual and Cognitive Aspects

In this section, we critically review and discuss some relevant aspects related to the perception and cognition of sonic interactions and provide links between these aspects and central concepts of VR, such as the plausibility illusion, the place illusion, the sense of embodiment, and the sense of agency. Nordahl and Nillson [57] also consider how sound production and perception relate to plausibility illusion, place illusion, and the sense of body ownership, although from a somewhat different angle.

Our main claim is that interactive sound sources in a virtual environment contribute in particular to the plausibility illusion, the sense of agency, and the sense of body ownership. In addition, our analysis of perceptual and cognitive aspects provides requirements and guidelines for the development and the implementation of sound models.

2.3.1 Latency, Causality, and Multisensory Integration

In any interactive system, latency and its associated jitter have a major perceptual impact. High latency or jitter may impair the user's performance or, at least, provide a frustrating and tiring experience. Perceptually acceptable limits for latency and jitter in an interactive system should therefore be determined. However, such limits depend on several factors which are not easily disentangled.

Characterizing latency and jitter in the sound rendering pipeline can be restated as a problem of perceived synchronization between pairs of events [46], which in turn may be divided into three categories: (i) an external and an internal temporal pattern (such as those occurring in a collaborative activity, e.g., music playing, between two persons in a virtual environment); (ii) pairs of external events (which may or may not pertain to the same sensory modality, such as pairs of sounds or a visual flash and a sound); (iii) actions of the user and their effects (e.g., the pressing of a button and the corresponding feedback sound).

The latter case in particular is tightly connected to the definition of interactive sound adopted in this chapter. It is inherently a problem of multimodal synchronization, as it involves a form of extrinsic (auditory) feedback and a form of intrinsic (tactile, proprioceptive, and kinesthetic) feedback generated by the user's action [53]. The complex interaction occurring between these modalities influences their perceived synchronization (and thus the acceptable latency). High latencies can deteriorate the quality of the interaction, impair the performance on a given task, and even disrupt the perceived link of causality between the user's action and the resulting sonic outcome.

The task at hand also influences the acceptable latency. As an example, it has been traditionally accepted that music performance is a task requiring extremely low ($\leq$ 10 ms) latencies between the player's actions and the response of a digital musical instrument [99]. Similarly, it has been shown that even small amounts of jitter can be detrimental to the perceived quality of the interaction [41]. In this respect, music provides a good "worst case" and a lower bound for latency in other, non-musical tasks, where various studies suggest that higher latencies may be acceptable or even unperceivable [43, 93].

The type of interaction must be considered as well. Impulsive interactions (either musical, such as playing a drum, or non-musical, such as knocking on a door) are likely to require lower latencies than continuous ones (bowing a violin string, or accompanying a closing door). As an example, it has been shown that the continuous interaction involved in playing a theremin allows for relatively high ($>$ 30 ms) latencies, despite this being a musical task [54]. Finally, cognitive aspects also play a role: humans create expectations for the latency between their actions and the resulting feedback, detect disturbances to such expectations, and compensate for them. A study on the latency in live musical sound monitoring [48] showed significant discrepancies between different instruments, suggesting that certain players (e.g., pianists) are more tolerant to latency as they are accustomed to the inherent mechanical latency of their instrument, while others (e.g., drummers) are less so.

We conclude this section with a hint at the second type of synchronization mentioned at the beginning, i.e., that between pairs of external (possibly multimodal) events. Humans achieve robust perception through both the combination and the integration of information from multiple sensory modalities: the former strategy refers to interactions between non-redundant and complementary sensory signals aimed at disambiguating the sensory estimate, while the latter describes interactions between redundant signals aimed at reducing the variance in the sensory estimate and increasing its reliability [28]. The temporal relationships between inputs from different senses play an important role in multisensory combination and integration, which can be realized only within a window of synchrony between different modalities (e.g., auditory and visual, or auditory and haptic feedbacks) where a single percept is produced. Many studies [19, 83, 96] report quantitative results about "integration windows" between modalities, which can be used as constraints for the synchronization of the sound simulation pipeline with the visual (and possibly the haptic) modality. For more details regarding these issues, please refer to Part IV in this book, and in particular to Ch. 10.

2.3.2 Everyday Listening and the Plausibility Illusion

Human listeners are extremely good at interpreting sounds in terms of the events that produced them. The patterns of mechanical or aeroacoustic vibrations generated by sound-producing events depend on (and thus carry information about) contact forces, duration of contact, time-variations of the interaction, sizes, shapes, materials, and textures of the involved objects. We are immersed in a landscape of everyday sounds since the day we are born, and we have learned to extract meaning from this continuous and omnidirectional flow of information.

Gaver [34] introduced the concept of *everyday listening*, as opposed to *musical listening*. When a listener hears a sound, she might concentrate on attributes like pitch, loudness, and timbre, or she might notice its masking effect on other sounds. These are examples of musical listening, meaning that the considered perceptual dimensions and attributes have to do with the sound itself, and are those used in the creation of music. On the other hand, the listener might concentrate on the characteristics of the sound source and possibly the surrounding environment. When hearing an approaching car, she might notice that the engine is powerful, that the car is approaching quickly from behind, or even that the road is a narrow alley with echoing walls on each side. This is an example of everyday listening.

The two perceptual processes associated to musical and everyday listening cannot be completely disentangled and may occur simultaneously. Still, the idea that in our everyday listening experience the physical characteristics of sound-producing objects can be linked to the corresponding acoustic features is a powerful one. The literature of ecological acoustics provides several quantitative results on such links. The underlying assumption is that the flow of acoustic energy reaching our ears, the *acoustic array*, contains specific patterns, or *invariants*, which the listener exploits to infer information about the environment and guide her action. These concepts and terminology originate in the framework of ecological perception, rooted in Gibson's works on visual perception in the 1950s [35, 55].[2]

Acoustic invariants associated to sound events may include several attributes of a vibrating solid, such as its size, shape, and density, as these attributes contribute differently to characteristics of the resulting sound such as pitch, spectrum, amplitude envelope, and so on. In patterned sounds (see Sect. 2.2.2), the relevant information is also carried by the timing of successive events: footstep sounds must occur within

[2] In this context, the label "ecological" is associated to two main concepts: 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, so a theory of perception cannot ignore what animals do.

a range of rates and regularities in order to be perceived as walking; the regularity in the temporal pattern of a bouncing sound provides information about the shape of the object (e.g., a sphere versus a cube).

The mapping between physical parameters and acoustic features is in general many-to-many. A single physical parameter can influence simultaneously many characteristics of the sound, and different physical parameters influence the same characteristics in different ways. As an example, changing the size of an object will scale the sound spectrum, i.e., will change the frequencies of the sound but not their pattern. On the other hand, changing the object's shape results in a change in both the frequencies and their relationships. Acoustic invariants are thus the result of these complex patterns of change. Surveys of classic studies in ecological acoustics and acoustic invariants have been provided in previous works [5, 36].

The above discussion provides a solid theoretical framework to reason on the importance of ecologically valid acoustic information in eliciting the qualia of *presence* [72] in an immersive VR system. Among the many definitions proposed in the literature, we follow Skarbez et al. [76] in defining presence broadly as "the perceived realness of a mediated or virtual experience". Slater et al. [77] introduced the concepts of plausibility illusion and place illusion, to refer to two distinct subjective internal feelings, both of which contribute to eliciting the sense of presence in a subject experiencing an immersive VR scenario. This conceptual model of presence is depicted in Fig. 2.4.[3]

In this section we are particularly interested in the plausibility illusion, i.e., the illusion that the scenario being depicted is actually occurring (we will discuss the place illusion in Sect. 2.3.3 next). This is determined by the overall credibility of a virtual environment in comparison with subjective expectations. Slater argued that an important component of the plausibility illusion is "for the virtual reality to provide correlations between external events not directly caused by the participant and his/her own sensations" [77]. Skarbez et al. [76] proposed the construct of coherence, an objective characteristic of a virtual scenario that gives rise to the plausibility illusion (see Fig. 2.4, right) and depends on the internal logical and behavioral consistency of the virtual experience, with respect to prior knowledge. Building on these definitions, we argue that sound will contribute to the plausibility illusion of a virtual scenario as long as coherence is ensured for the auditory modality, i.e., as long as sound carries relevant ecological information expected by the user's everyday listening experience.

It shall be noted that coherence makes no assumptions about the high fidelity of a virtual environment to the real world. Consequently, the plausibility illusion "does not require physical realism" [77]: several studies show that virtual characters or objects displayed with low visual fidelity in the virtual environment do not disrupt the illusion. With regard to the auditory domain, this observation may be related to the concept of cartoon sounds [69], i.e., simplified descriptions of sounding phenomena with exaggerated features. We argue that cartoon sounds do not disrupt the

[3] Skarbez et al. [76] consider a third component, the social presence illusion, which we do not address here.

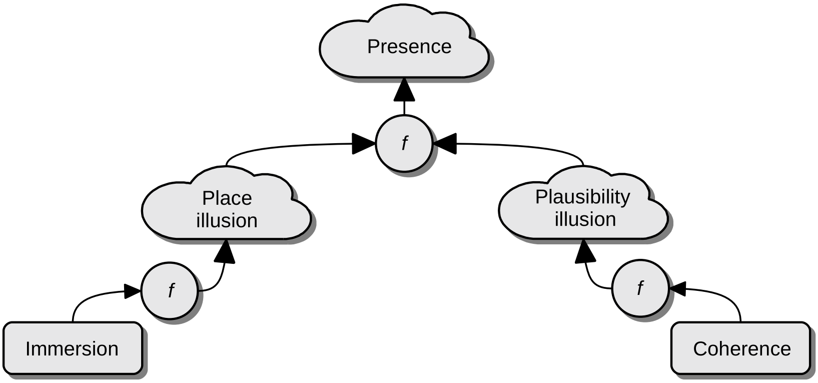


Fig. 2.4 A conceptual model of presence: cloud boxes represent subjective internal feelings (qualia), circles represent functions affected by individual differences, and rounded rectangles represent objective characteristics of the virtual experience. Figure based on Skarbez [76, Fig. 2]

plausibility illusion as long as they still carry relevant ecological information. This is fundamentally the same principle exploited in the empirical science of Foley Art for creating ecologically plausible sound effects [2].

2.3.3 Active Perception, Place Illusion, Embodiment

The "enactive" approach to experience posits that it is not possible to disassociate perception and action schematically and that every kind of perception is intrinsically active and thoughtful. One of the most influential contributions in this direction is due to Varela et al. [94]. In the authors' conception, experience does not occur inside the perceiver, but rather it is enacted by the perceiver while exploring the environment. In this view, the subject of mental states is the embodied, environmentally situated perceiver. The term "embodied" highlights two main points: (i) perception depends upon the kinds of experience that are generated from specific motor capabilities, and (ii) these capabilities are themselves embedded in a biological, psychological, and cultural context. Sensory and motor processes are fundamentally inseparable, and perception consists in exercising an exploratory skill. As an example [58], the sensation of softness experienced when holding a sponge consists in being aware that one can exercise certain skills: one can press the sponge, and it will yield under the pressure. The experience of the softness of the sponge is characterized by a variety of such possible patterns of interaction. Sensorimotor dependencies, or contingencies, are the laws that describe these interactions. When a perceiver knows that he is exercising the sensorimotor contingencies associated with softness, then he is experiencing the sensation of softness.

Embodied theories of perception provide the ground for discussing further central concepts for VR, such as immersion, place illusion, sense of embodiment, and their relation to interactive sound. As depicted in Fig. 2.4 (left), immersion is an objective property of a VR system. Research has concentrated largely on characteristics such as latency, rendering frame rate, and tracking [22]. However, immersive systems can be also characterized in relation to the supported sensorimotor contingencies, which in turn define a set of valid actions that are perceptually meaningful (for instance, with a head-mounted display and head-tracking, it is possible to turn your head or bend forward producing changes in the rendered visual images). When a system supports sensorimotor contingencies that approximate those of physical reality, it can give rise to the place illusion, a specific subjective internal feeling which is the illusion of being located inside the rendered virtual environment, of "being there" [77]. Whereas the plausibility illusion is based on what a subject perceives in the virtual environment, the place illusion is based on how she is able to perceive it.

The great majority of studies addressing explicitly the effect of sound on the place illusion are concerned with spatial attributes: this is not entirely surprising, since many of these attributes are perceived by exercising specific motor actions (e.g., rotating the head to perceive the distance or the direction of a sound source or a reflecting surface). In this respect, directivity is possibly the only sound source attribute contributing to the place illusion, while other ecological attributes are more likely to contribute to the plausibility illusion only, as discussed in Sect. 2.3.2. In accordance with this picture, over the years, various authors [11, 38, 60] found that spatialized sound positively influences presence as being there when compared to no-sound or non-spatialized sound conditions, but does not affect the perceived realism of the environment. A comprehensive survey up to 2010 is provided by Larsson [47].

The sense of embodiment refers to yet another subjective internal feeling. Specifically, the sense of embodiment in an immersive virtual environment is concerned with the relationship between one's self and one's body, whereas the sense of presence refers to the relationship between one's self and the environment (and may occur even without the sensation of having a body). Kilteni et al. [45] provide a working definition of a sense of embodiment toward an artificial body, as the sense that emerges when that artificial body's properties are processed as if they were the properties of one's own biological body. Further, the authors associate it to three main components: (i) the sense of self-location, (ii) of body ownership, and (iii) of agency, the latter being investigated as an independent construct by other researchers [17].

The sense of self-location refers to one's spatial experience of being inside a body, rather than being inside a world (with or without a body), and is highly determined by the visuospatial perspective, proprioception, and vestibular signals, as well as tactile sensations at the border between our body and the environment. The sense of body ownership refers to one's self-attribution of an artificial body perceived as the source of the experienced sensations and emerges as a complex combination of afferent multisensory information and cognitive processes that may modulate the processing of sensory stimuli, as demonstrated by the well-known rubber hand illusion [13]. The sense of agency refers to the sense of having global motor control in relation to one's own body and has been proposed to result from a comparison between the predicted and the actual sensory consequences of one's actions [24]: when the two match by, for example, the presence of synchronous visuomotor correlations under active movement, one feels oneself to be the agent of those actions.

The above discussion suggests that interactive sounds occurring directly in reaction to the avatar's gestures in a virtual scenario, and coherently with the available sensorimotor contingencies, can positively affect the sense of agency in particular. One relevant example is provided by footsteps: several studies have addressed the issue of generating footstep sounds [14, 85, 95] however without assessing their specific relevance to the sense of agency. Other studies have shown that interactively generated sound can support haptic sensations, as in the case of impact sounds reinforcing or modulating the perceived hardness of an impulsive contact [6], or friction sounds affecting the perceived effort in dragging an object [4] (refer to Chap. 12 for other audio-haptic case studies). Yet, no attempt was made in these studies to specifically address the issue of agency.

Even less research seems to have been conducted on the effects of interactive sound on the sense of body ownership. Footsteps provide a relevant example also in this case, as the sound of steps can be related to the perceived weight of one's own body [85] or that of an avatar [74]. Sikström et al. [73] evaluated the role of self-produced sounds in participants' sensation of ownership of virtual wings in an immersive scenario. A related issue is that of the sound of one's own voice in a virtual environment [61].

2.4 Events Versus Processes

Having discussed the perceptual and cognitive aspects involved in interactive sound generation, we now jump back to the pipeline of Fig. 2.1 and look specifically at the "source modeling" box.

When creating sound sources in a virtual environment, approaches based on sample playback are still the most common ones [12], taking advantage of sound design techniques that have been refined through a long history, and being able to yield perfect realism, "at least for single sounds triggered only once" [21]. From a completely different perspective, procedural approaches defer the generation of sound signals until runtime, when information on sound-producing event is available and can be used to yield more interactive sonic results. This section discusses these two dichotomical approaches.

2.4.1 Event-Driven Approaches

Approaches based on sample playback follow an event-driven logics, in which a specific sound exists as a waveform stored in a file or a table in memory and is

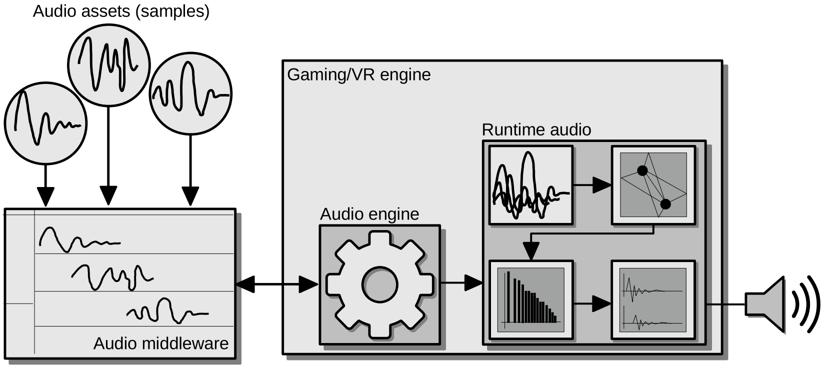


Fig. 2.5 Event-driven logics for VR sound design using samples and audio middleware software

bound to some event occurring in the virtual world. Borrowing an example from Farnell [31]: `if (moves(gate)) play(scrape.wav)`.

One immediate consequence of this is that the playback and the post-processing of samples are dissociated from the underlying physics engine and the graphical rendering engine. In the case of a sound played back once, the length of the sound is predetermined and thus any timing relationship between auditory and visual elements must also be predefined. In the case of a looped sound, the endpoint must be explicitly given, e.g., as a response to a subsequent event. More in general, the playback of sound is controlled by a finite and small set of states (such as in the case of an elevator that can be starting, moving, stopping, or stopped). Correspondingly, any event is bound to a sound "asset", or to some post-processing of that asset.

Current practices of sound design for VR are deeply and firmly rooted in such event-driven logics, depicted in Fig. 2.5. One clear example of this is provided by "audio middleware" software [12], which are tools that facilitate the work of the sound designer by reducing programming time and testing the sound design in real time along with the game engine. The most commonly adopted middleware solutions, such as FMOD Studio (Firelight Technologies)[4] and Wwise (Audiokinetic),[5] largely follow the traditional paradigm of DAWs (Digital Audio Workstations) and include GUIs for adding, controlling, and processing samples; linking them to objects, areas, and events of the virtual environment; and imposing rules for triggering and playback.

One of the main acknowledged limitations of samples is that they are static, and they are just single, atomic instances of events. The repetitiveness involved in multiple playbacks of the same sounds has the potential to disrupt many of the perceptual and cognitive effects discussed in Sect. 2.3, and even to lead to fatigue. Partial remedies to this problem include the use of multiple samples for the same event, as well as the

[4] https://www.fmod.com/.

[5] https://www.audiokinetic.com/products/wwise/.

use of various post-processing operations, the most common being modifications to pitch, time, dynamics, as well as sound layering and looping [75].

Well-established time-stretching and pitch-shifting algorithms exist; however, the quality of the processing is in general guaranteed only for relatively small shifting and stretching factors. Concerning dynamics, typical approaches are based on blending, cross-fading, and mixing of different samples, similarly to a musical sampler (and with similar limitations as well). Layering and looping are especially useful for the construction of ambiences: multiple sounds can be individually looped and played concurrently to create complex and layered ambiences. Repetitiveness can be reduced by assigning different lengths to different loops, and immersion can be enhanced by rendering individual layers at different virtual spatial locations. All this requires manual operations by the sound designer, such as splitting, cross-fading, and so on.

Further countermeasures to repetition and listener fatigue include the use of techniques based on randomization. These can be applied to many aspects of sound, including, but not limited to (i) pitch and amplitude variations, (ii) sample selection, (iii) sample concatenation, (iv) looping, and (v) location of sound sources. As an example, randomized sample selection amounts to performing randomizations of alternative samples associated to the same event, e.g., a collision: a different sample is played back at each occurrence of the event, mimicking the differences occurring due to slightly different contact points and velocities. In randomized concatenation, different samples are concatenated to build a composite sound in response to a repetitive sequence of events, such as in the case of footsteps, weapon sounds, and so on. Triggering different points with different probabilities can also be used to reduce the repetitiveness of looped layers in ambience sounds. The audio middleware solutions mentioned above typically implement several of these techniques.

Randomization techniques hint at another issue with samples, which is the need for very large amounts of data. Putting together a large sample library is a slow and labor-intensive process. Moreover, data need to be stored in memory, possibly in secondary storage, from which they then have to be prefetched before playback.

2.4.2 Procedural Approaches

Techniques based on the randomization of several sample-processing parameters, such as those discussed above, are sometimes loosely referred to as *procedural* in the sound design practice [75, Chap. 2]. Here, we favor a stricter definition. In Farnell's words [30], procedural audio is *"sound as a process, rather than sound as data"*. This definition shifts the focus onto the creation of audio assets, as opposed to the manipulation of existing ones.

Procedural audio is thus synthetic sound, is real time, and most importantly is created according to a set of programmatic rules and live input. This implies that procedurally generated sound is synthesized at runtime, when all the needed input

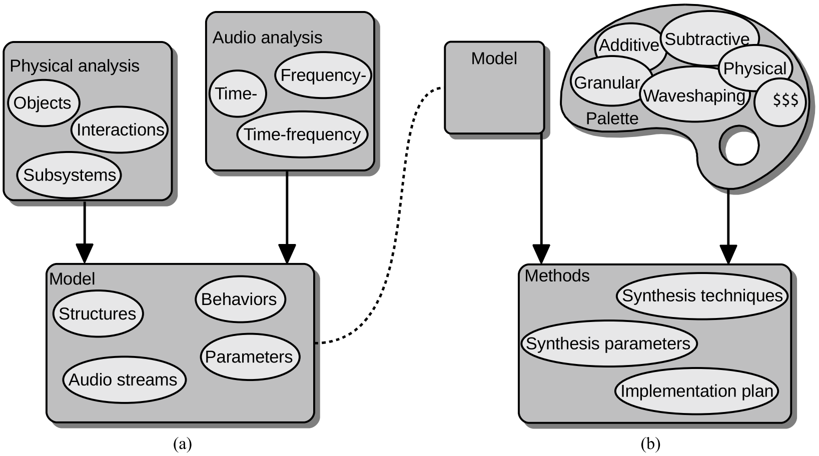


Fig. 2.6 Procedural sound design: **a** model building, and **b** method analysis stages (figures loosely based on Farnell [30, Figs. 16.4–5])

and contextual information are available, whereas in a sample-based approach, most of the work is performed offline prior to execution, implying that "many decisions are made in advance and cast in stone" [31].

The stages involved in the process of procedural sound design may be loosely based on those of software life-cycle, including (i) requirements analysis, (ii) research and acquisition, (iii) model building, (iv) method analysis, (v) implementation, (vi) integration, (vii) test and iteration, and (viii) maintenance. Figure 2.6 provides a graphical summary of the two central stages, i.e., model building and method analysis.

Building a model (Fig. 2.6a) provides a simplification of the properties and behaviors of a real object, which starts from the analysis of sound data (including time- and/or frequency-domain analysis, extraction of relevant audio features, etc.), as well as a physical analysis of the involved sound-generating mechanisms, and results into a set of parametric controls and behaviors. The hierarchy of everyday sounds depicted in Fig. 2.3 provides a useful reference framework: the model at hand can be positioned inside this hierarchy. Moreover, following the discussion on everyday listening of Sect. 2.3.2, the choice of the model parametrization can be informed by the knowledge of relevant acoustic invariants carrying information about sound-generating objects and events.

The method analysis stage is where the most appropriate sound synthesis and processing techniques are chosen, starting from a palette of available ones, and based on the model at hand. Figure 2.6b shows a set of commonly employed sound synthesis techniques (in Sect. 2.4.3, we will explore physics-based techniques in particular). As a result of this stage, an implementation plan is produced that includes a set of techniques and corresponding low-level synthesis parameters, as well as the involved audio streams.

Based on this discussion, we can identify two main qualities of procedural approaches with respect to sample playback. The first one is their intrinsic adaptability and interactivity (according to the definitions given in Sect. 2.1), which derive from the deferring of sound generation at runtime based on programmatic rules and user input, and result in ever-changing sonic results in response to real-time control. The second one is flexibility, where a single procedural model can be parametrized to produce a variety of sound events within a given class of sounds: this contrasts with sample-based, event-driven approaches, where ever-increasing amounts of data and assets are needed in order to cope with the needs of complex virtual worlds.

2.4.3 Physics-Based Methods

Looking back at Fig. 2.6b, one of the available paints in the palette of sound synthesis techniques is that of physics-based methods.

The boundaries between what can be considered physical (or physics-based, or physics-informed) sound synthesis are somewhat blurry in the scientific literature. Here, we adopt the definition given by Smith [78] and refer to synthesis techniques where " [...] there is an explicit representation of the relevant physical state of the sound source. For example, a string physical model must offer the possibility of exciting the string at any point along its length. [...] All we need is Newton." The last claim refers to the idea that physical modeling always starts with a quantitative description of the sound sources based on Newtonian mechanics. Such description may be approximate and simplified to various extents, but the above definition provides an unambiguous—albeit broad—characterization in terms of physical state access. Resorting to a simple (yet historically relevant [68]) example, we can say that additive synthesis of bell sounds is not physics-based, as additive sinusoidal partials only describe the time-frequency characteristics of the sound signal without any reference to the physical state of the bell. On the other hand, modal synthesis [1] of the same bell, with modal oscillators tuned to the sound partials, is only apparently a similar approach: a linear combination of the modes can provide the displacement and the velocity at any point of the bell, and each modal shape defines to what extent an external force applied at a given point affects the corresponding mode.

The history of physics-based synthesis is rooted in studies on the acoustics of the vocal apparatus [44] and of musical instruments [39, 40], where numerical models were initially used for simulation rather than synthesis purposes. Current techniques are based on several alternative formulations and methods, including ordinary or partial differential equations, equivalent circuit representations, modal representations, finite-difference and finite-element schemes, and so on [78]. Comprehensive surveys of physical modeling approaches have been published [79, 89]. Although these deal with musical sound synthesis mostly, much of what has been learned in that domain can be applied to the physical modeling of any sounding object.

Although physics-based synthesis is sometimes made synonymous with procedural audio, Fig. 2.6b provides a clear picture of the relation between the two. In this perspective, “procedural audio is more than physical modeling,” [31] and the latter can be seen as one of the tools at the disposal of the sound designer to reduce a sound to its behavioral realization. Combining physics-based approaches with knowledge of auditory perception and cognition often results in procedural models in which the physical description has been drastically simplified while retaining the ecological validity of sounds and the realism of the interactions, thus preserving the plausibility illusion of the resulting sonic world and the sense of agency of the subject (see related discussions in Sects. 2.3.2 and 2.3.3).

2.4.4 Computational Costs

Event-driven and procedural approaches must be analyzed also in terms of the involved computational requirements. In case of insufficient resources, excessive computational costs may introduce artifacts in the rendered sound or in alternative may require to increase the overall latency of the rendering up to a point where the perception of causality and multisensory integration are disrupted (see Sect. 2.3.1).

With reference to Fig. 2.1, it can be stated that one main computational bottleneck in the sound simulation rendering pipeline [51] is the “per sound source” cost. This relates in particular to the sound propagation stage (see Chap. 3), as reflections, scattering, occlusions, Doppler effects, and so on must be computed for each sound source involved in the simulation. But it also includes the source modeling stage, with particular reference to the generation of the sound source signals.

Sample playback has a fixed cost, irrespective of the sound being played. Moreover, the cost of playback is very small. However, samples must be loaded in memory before being played. As a consequence, when a sound is triggered, the playback may involve a prefetch phase where a soundbank is loaded from the secondary memory. Moreover, some management of polyphony must be set in place in order to prioritize the playback in case of several simultaneously active sounds. This can use policies similar to those employed in music synthesizers: typically, sounds falling below a certain amplitude threshold are dropped, leaving place for other sounds. The underlying assumption is that louder sounds mask softer ones, so that dropping the latter has no or minimal perceptual consequences. Although modern architectures allow for the simultaneous playback of hundreds of audio assets, generating complex soundscapes may exceed the amount of available channels.

On the other hand, procedural sound has variable costs, which depend on the complexity of the corresponding model and on the employed methods. This is particularly evident in the case of physics-based techniques: for large-scale, brute-force approaches, like higher dimensional finite-element or finite-difference methods, real time is still hard to achieve. On the other hand, techniques like modal synthesis can be implemented very efficiently, albeit at the cost of reduced flexibility of the models (e.g., interaction with sounding objects limited to single input-output), which in turn can have a detrimental effect on the plausibility illusion. Some non-physical methods are very cheap in terms of computational requirements, as in the case of subtractive synthesis for generating wind or fire sounds. Section 2.5.1 provides several examples of procedural methods for various classes of everyday sounds.

Although it is generally true that sample-based methods outperform procedural audio for small amounts of sounds, it has been noted [30] that this is not necessarily true in the limit of larger numbers: whereas the fixed cost of sample playback results in a computational complexity that is linear in the number of rendered sources, the availability of very cheap procedural models can produce the result that for high numbers of sources the situation reverses and procedural sound starts to outperform sample-based methods.

2.5 Procedural and Physics-Based Approaches in VR Audio

Given these premises, what is the current development of procedural and physics-based approaches in audio for VR? In this section, we show that, despite a substantial amount of research, these approaches are still struggling to gain popularity in real-world products and practices.

2.5.1 Methods

Far from providing a comprehensive survey of previous literature in the field, which would go way beyond the scope of this chapter, this section aims at assessing to what extent the taxonomy of everyday sounds provided in Fig. 2.3 has been covered by existing procedural approaches. This exercise also serves as a testbed to verify the generality of that taxonomy. For a recent and broad survey, see Liu and Manocha [51].

Solid sounds are by far the most investigated category. For basic models, modal synthesis [1] is the dominant approach. There are several works investigating the use of modal methods for the procedural generation of contact sounds between solid objects, including the optimization of modal synthesis through specialized numerical schemes and/or perceptual criteria, as in the work by Raghuvanshi et al. [63]. Procedural models of surface textures have been proposed by several scholars [66, 91] and applied to scraping and rolling sounds [64]. Basic interaction forces (impact and sliding friction) can be modeled with a variety of approaches that range from qualitative approximations of temporal profiles of impulsive force magnitudes [92] to the physical simulation of stick-slip phenomena in friction forces [7].

At the next level of complexity, models of patterned solid sounds have also been widely studied. Stochastic models of crumpling phenomena have been proposed, with applications to cloth sound synthesis [3], crumpling paper sounds, or sounds produced by deformations of aggregate materials, such as sand, snow, or gravel [15]. The latter have also been used in the context of walking interactions [81] (see also Sect. 2.3.3) to simulate the sound of a footstep onto aggregate grounds. Breaking sounds have been modeled especially with the purpose of synchronizing animations of brittle fractures produced by a physics engine [59, 100].

The category of aerodynamic sounds is less studied. Within the basic level of complexity, the sound produced by wind includes those resulting from interaction with large static obstructions, aeolian tones, and cavity tones: these have been procedurally modeled with techniques ranging from computationally intensive fluid-dynamics simulations [26] to simple (yet efficient and effective) subtractive schemes using noisy sources and filters [30]. These can be straightforwardly employed to construct patterned and compound sonic events, including windy scenes, swinging objects, and so on [71]. Other basic aeroacoustic events include turbulences, most notably explosions, which are a key component of more complex sounds such as gunshots [37] and fire [18]. Yet another relevant patterned sonic event is that produced by combustion engines [10].

Liquid sounds appear to be the least addressed category. Basic procedural models include sounds produced by drops in a liquid [90] or by pouring a liquid [65], whereas patterned and compound sonic events have been more often simulated using concatenative approaches relying on the output of the graphical procedural simulation [98]. A relevant example of hybrid solid-liquid sounds is that of rain [50].

2.5.2 *Optimizations*

We have provided in Sect. 2.4.4 a general discussion on computational costs associated to procedural approaches, in comparison to sample-based methods. Since the former typically results in higher "per sound source" costs than the latter, various studies have proposed strategies for reducing the load of complex procedural audio scenes in virtual environments.

One attractive feature of procedural sound in terms of computational complexity is the possibility of dynamically adapting the level of detail (LOD) of the synthesized audio. The concept of LOD is a long-established one in computer graphics and encompasses various optimization techniques for decreasing the complexity of 3D object rendering [52]. The general goal of LOD techniques is to increase the rendering speed by reducing details while minimizing the perceived degradation of quality. Most commonly, the LOD is varied as a function of the distance from the camera, but other metrics can be used, including size, speed of motion, priority, and so on. Reducing the LOD may be achieved by simplifying the 3D object mesh, or by using impostors (i.e., replacing mesh-based with image-based rendering), and other approaches can be used to dynamically control the LOD of landscape rendering, crowd simulation, and so on.

Similar ideas may be applied to procedural sound, achieving further reductions of computational costs for complex sound scenes with respect to sample playback. However, very few studies explored the concept of LOD in the auditory domain, and there is not even a commonly accepted definition in the related literature: some scholars have coined the term Sound Level Of Detail (SLOD) [70], while others use

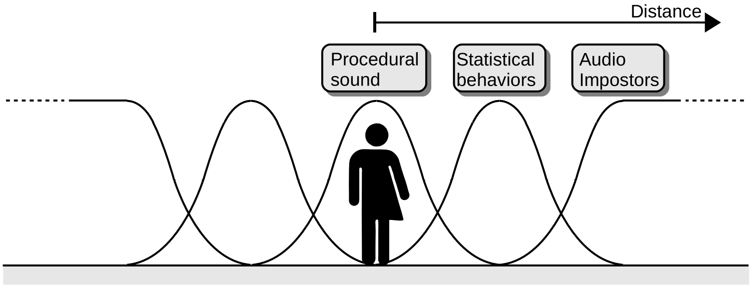


Fig. 2.7 Example of dynamic LOAD based on the radial distance from the listener, where levels of details are associated to three overlapping proximity profiles. Figure partly based on Schwarz et al. [70, Fig. 3]

Level Of Audio Detail (LOAD) [27], both generically referring to varying sound resolution according to the required perceived precision. Here, we stick to the latter definition (LOAD), since this seems to be more frequently adopted in recent literature.

Strategies for dynamic LOAD can be partly derived from graphics. Simple approaches amount to fade out and turn off distant sounds based on radial distance or zoning. Depending on their distance, sound sources may be also clustered or activated according to some predefined behavior. Techniques based on impostors can be used as well: as an example, when rendering the sound of a crowd, individual sounds emitted by several characters can be replaced by a global sample-based ambience sound. However, one should be aware of the differences between visual and auditory perception and exploit the peculiarities of the latter to develop more advanced strategies for dynamic LOAD. Figure 2.7 depicts an example of a dynamic LOAD strategy based on radial distance, in which levels of details are associated to three overlapping proximity profiles around the listener (foreground, middle ground, and background): sounds in the foreground are rendered individually through procedural approaches; those that fall into the middle ground can be rendered through some simplifying approaches (clustering, grouping, and statistical behaviors); and finally, sounds in the background may be substituted by audio impostors such as audio files.

Pioneering work in this direction was carried out by Fouad et al. [32], although the authors did not explicitly refer to the concept of LOD. This work proposes a set of “perceptually based scheduling algorithms”, that allows a scheduler to assign execution time to each sound in the scene minimizing some perceptually motivated error metric. In particular, sounds are prioritized depending on the listener’s gaze, the loudness, and the age of the sound. Tsingos and coworkers [56, 88] proposed an approach to reduce the number of (sample-based) sound sources in a complex scenario, by combining perceptual culling and perceptual clustering. The culling stage removes perceptually inaudible sources based on a global masking model, while the clustering stage groups the remaining sound sources into a predefined number of clusters: as a result, a representative point source is constructed for each cluster and a set of equivalent source signals is generated. Schwarz et al. [70] proposed a design with three LOADs based on proximity and smooth transitions between proximity levels, very much like those depicted in Fig. 2.7: (i) foreground, i.e., individually driven sound events (e.g., individual raindrops on tree leaves); (ii) middle ground, i.e., group-driven sound events, at the point where individual events cannot be isolated and can be replaced by stochastical behaviors; (iii) background, i.e., sound sources that are further away and can be rendered by audio impostors such as audio files or dynamic mixing of groups of procedural impostors. More recently, Dall'Avanzi et al. [23] analyzed the effect on player's immersion in response to soundscapes with two applied LOADs. Two groups of participants played two different versions of the same game, and the player's immersion was measured through two questionnaires. However, results in this case showed no considerable difference between the two groups.

Other researchers proposed or evaluated LOAD techniques specifically tailored to certain synthesis methods. Raghuvanshi et al. [63] addressed modal synthesis and investigated various perceptually motivated techniques for improving the efficiency of the synthesis. These include a "quality scaling" technique that effectively controls the dynamic LOAD: briefly, in a scene involving many sounding objects, the number of modes assigned to individual objects scales with objects location from foreground to background, without significant losses in perceived quality. Durr et al. [27] evaluated through subjective tests various procedural models of sound sources with three applied LOADs. Specifically, three procedural models proposed by Farnell [30] (see also Sect. 2.5.1) were chosen for investigation: (i) fire sounds employ subtractive synthesis to generate and combine hissing, crackling, and lapping features; (ii) bubbles sounds use a form of additive synthesis with frequency- and amplitude-controlled sinusoidal components representing single bubbles; (iii) wind sounds are again produced using subtractive synthesis (amplitude-modulated noise and various filtering elements to represent different wind effects). A different approach to applying LOAD was implemented for each model. Correspondingly, listening tests provided different results for each model in terms of perceived quality at different LOADs.

The reader interested in further discussion about audio quality should also refer to Chap. 5.

2.5.3 Tools

In spite of all the valuable research results produced so far, there is still a lack of software tools that assist the sound designer in using procedural approaches.

Designers working with procedural audio use a variety of audio programming environments. Popular choices include (but are not limited to) Pure Data,[6] Max/MSP,[7] or CSound.[8] The first two in particular implement a common, dataflow-

[6] https://puredata.info/.

[7] https://cycling74.com/.

[8] https://csound.com/.

oriented paradigm [62] and use a visual patch language where “the diagram is the program”: Farnell [31] argues that this paradigm is particularly suited for procedural audio as it has a natural congruence with the abstract model resulting from the design process. On the other hand, integrating these environments into the most widespread gaming/VR engines is not straightforward: at the time of writing, some active open-source projects include libpd [16], a C library that turns Pure Data into an embeddable audio synthesis library and provides wrappers for a range of languages, and Cabbage [97], a framework for developing audio plugins in Csound, including plugins for the FMOD middleware. Commercial gaming/VR engines typically provide limited functionalities to support procedural sound design, although some recent developments may hint at an ongoing change of perspective: as an example, the Blueprint visual scripting system within the Unreal Engine has been used for dataflow-oriented procedural audio programming, also using some native synthesis (subtractive, etc.) capabilities.

All of the tools mentioned above still require to work at a low level of abstraction, implying that the sound designer must have the technical skills needed to deal with low-level synthesis methods and parameters, and at the same time limiting productivity. There is a clear need for tools that allow the designer to work at higher levels of abstraction. One instructive example is provided by the Sound Design Toolkit (SDT), an open-source software package developed over several years [9, 25] which provides a set of sound models for the interactive generation of several acoustic phenomena. In its current embodiment, SDT is composed of a core C library exposing an API, plus a set of wrappers for Max and Pure Data, and a related collection of patches and help files. Interestingly, the collection is based on a hierarchical taxonomy of everyday sound events which follows very closely the one depicted in Fig. 2.3 and implements a rich subset of its items. The designer has access to both low-level parameters (e.g., the modal frequencies of a basic solid resonator) and to high-level ones (e.g., the initial height of a bouncing object).

Commercial products facilitating the designer’s workflow are also far from abundant: Lesound[9] (formerly Audiogaming) sells a set of plugins for FMOD and Wwise that include procedural simulations of wind, rain, motor, and weather sounds, while for its part AudioKinetic (developer of Wwise) develops the soundseed plugin series, which include procedural generation of wind and whooshing sounds as well as impact sounds. Nemisindo[10] provides a web-based platform for real-time synthesis and manipulation of procedural audio, which stems from the FXive academic project [8], but no plugin-based integration with VR engines or audio middleware software is available at the time of writing.

A much-needed facilitating tool for the sound designer is one that automates part of the design process, allowing in particular for automatic tuning of the parameters of a procedural model starting from a target (e.g., recorded) sound. This would provide a means to recreate procedurally a desired sound and more in general to ease the design by providing a starting set of parameter values that can be further edited.

[9] https://lesound.io/.

[10] https://nemisindo.com.

In the context of modal synthesis, various authors have proposed automatic analysis approaches for determining modal parameters from a target signal (e.g., an impact sound). In this case, the parametrization of the model is relatively simple: every mode at a given position is fully characterized by a triplet of scalars representing its frequency, decay coefficient, and gain. This generalizes to an array of gains if multiple points on the object are considered, or to continuous modal shapes as functions of spatial coordinates on the object. Ren et al. [67] proposed a method that extracts perceptually salient features from audio examples and a parameter estimation algorithm searches for the best material parameters for modal synthesis. Based on this work, Sterling et al. [82] added a probabilistic model for the damping parameters in order to reduce the effect of external factors (object support, background noise, etc.) and non-linearities on the estimate of damping. Tiraboschi et al. [87] also presented an approach to the automatic estimation of modal parameters based on a target sound, which employs a spectral modeling algorithm to track energy envelopes of detected sinusoidal components and then performs linear regression to estimate the corresponding modal parameters.

While the case of solid objects and modal synthesis is a relatively simple one, the issue of automatic parameter estimation has been largely disregarded for other classes of sounds and models.

2.6 Conclusions

Our discussion in this chapter has hopefully shown that procedural approaches offer extensive possibilities for designing sonic interactions in virtual environments. And yet as of today the number of real-world applications and tools utilizing these approaches is very limited. In fact, not much has changed since ten or fifteen years ago, when other researchers observed a similar lack of interest from the industry [12, 29], with the same technical and cultural obstacles to adoption still in place. In a way recent technological developments have further favored the use of sample-based approaches: in particular, decreasing costs of RAM and secondary storage, as well as optimized strategies to manage caching and prefetching of sound assets, have made it possible to store ever larger amounts of data. This state of affairs mimics closely what happened in the music industry during the last three decades: physics-based techniques in particular have been around for a long time, but the higher sound quality and accuracy of samples are still preferred over the flexibility of physical models for the emulation of musical instruments.

Perhaps then the question is not whether procedural approaches can overcome sample-based audio, but when, i.e., under what specific circumstances. In this chapter, we have provided some elements, particularly links to a number of relevant perceptual and cognitive aspects, such as the plausibility and place illusions, the sense of embodiment, and the sense of agency. We argue that procedural audio can compete with samples in cases where either (i) very large amounts of data are needed to minimize repetition and support the plausibility illusion, or (ii) interactivity is needed beyond an event-driven logics, in order to provide tight synchronization and plausible variations with user actions, and to support her sense of agency and body ownership.

One example of the first circumstance is provided by wind sounds: good recordings of real wind effects are technically difficult to come by and long recordings are required to create convincing ambiences of windy scenes using looping, while on the other hand procedurally generated wind sounds achieve high levels of realism. It is therefore no surprise that the few commercially available tools for procedural sound all include wind (see Sect. 2.5.3) and have been successfully employed also in large productions.[11] While wind falls in the category of adaptive, rather than interactive sounds, two relevant examples for the second circumstance may be provided by footsteps and sliding friction (bike breaking, hinges squeaking, rubbing, etc.): beside requiring large amounts of data and randomization to avoid repetition, these sounds arise in response to complex and continuous motor actions by the user, which cannot be fully captured by an event-driven logics.

Future research and development should therefore focus on cases where procedural models can compete with samples, looking more deeply into the effects on the plausibility illusion, sense of agency, and sense of body ownership. From a more technical perspective, promising directions for future research include the development of dynamic LOAD techniques, as well as high-level authoring tools and automation.

Acknowledgements This chapter is partly based on the ideas and materials which I developed for my course "Sound in Interaction", held at the University of Milano for the MSc degree in Computer Science.

References

1. Adrien, J.-M. in Representations of Musical Signals (eds De Poli, G., Piccialli, A., Roads, C.) 269-297 (MIT Press, Cambridge, MA, 1991).
2. Ament, V. T.: The Foley grail: The art of performing sound for film, games, and animation Second edition (CRC Press, New York, 2014).
3. An, S. S., James, D. L., Marschner, S.: Motion-driven Concatenative Synthesis of Cloth Sounds. ACM Trans. Graphics **31** (July 2012).
4. Avanzini, F., Rocchesso, D., Serafin, S.: Friction sounds for sensory substitution, in Proc. Int. Conf. Auditory Display (ICAD04) (Sidney, July 2004).
5. Avanzini, F. in Sound to Sense, Sense to Sound. A State of the Art in Sound and Music Computing (eds Rocchesso, D., Polotti, P.) 345–396 (Logos Verlag, Berlin, 2008).
6. Avanzini, F., Crosato, P. in Haptic and audio interaction design (eds Mc-Gookin, D., Brewster, S.) 24–35 (Lecture Notes in Computer Science 4129/2006, Springer Verlag, Berlin/Heidelberg, 2006).

[11] As an example, the procedural wind simulator by Lesound has been reportedly used for generating ambiences in Quentin Tarantino's Django Unchained, see http://lesound.io/product/audiowind-pro/.

7. Avanzini, F., Serafin, S., Rocchesso, D.: Interactive simulation of rigid body interaction with friction-induced sound generation. IEEE Trans. Speech Audio Process. **13**, 1073–1081 (2005).
8. Bahadoran, P., Benito, A., Vassallo, T., Reiss, J. D.: FXive: A web platform for procedural sound synthesis, in Proc. 144 Audio Engin. Soc. Conv. (Milano, 2018).
9. Baldan, S., Delle Monache, S., Rocchesso, D.: The sound design toolkit. SoftwareX **6**, 255–260 (2017).
10. Baldan, S., Lachambre, H., Delle Monache, S., Boussard, P.: Physically informed car engine sound synthesis for virtual and augmented environments, in Proc. IEEE Int. Workshop on Sonic Interactions for Virtual Environments (SIVE2015) (Arles, 2015), 21–26.
11. Bormann, K.: Presence and the utility of audio spatialization. Presence: Teleoperators and Virtual Environment **14**, 278–297 (2005).
12. Böttcher, N.: Current problems and future possibilities of procedural audio in computer games. Journal of Gaming & Virtual Worlds **5**, 215–234 (2013).
13. Botvinick, M., Cohen, J.: Rubber hands ’feel’ touch that eyes see. Nature **391**, 756–756 (1998).
14. Bresin, R., Papetti, S., Civolani, M., Fontana, F.: Expressive sonification of footstep sounds, in Proc. Interactive Sonification Workshop (Stockholm, 2010), 51–54.
15. Bresin, R. et al.: Auditory feedback through continuous control of crumpling sound synthesis, in Proc. Workshop Sonic Interaction Design (CHI2008) (Firenze, 2008), 23–28.
16. Brinkmann, P., Wilcox, D., Kirshboim, T., Eakin, R., Alexander, R.: Libpd: Past, Present, and Future of Embedding Pure Data, in Proc. Pure Data Convention (New York, 2016).
17. Caspar, E. A., Cleeremans, A., Haggard, P.: The relationship between human agency and embodiment. Consciousness and cognition **33**, 226–236 (2015).
18. Chadwick, J. N., James, D. L.: Animating Fire with Sound. ACM Trans. Graphics **30** (2011).
19. Chen, L., Vroomen, J.: Intersensory binding across space and time: a tutorial review. Attention, Perception, & Psychophysics **75**, 790–811 (2013).
20. Collins, K. in Essays on Sound and Vision (eds Richardson, J., Hawkins, S.) 263–298 (Helsinki University Press, Helsinki, 2007).
21. Cook, P. R.: Real sound synthesis for interactive applications (CRC Press, 2002).
22. Cummings, J. J., Bailenson, J.N.: How immersive is enough? Ameta-analysis of the effect of immersive technology on user presence. Media Psychology **19**, 272–309 (2016).
23. Dall’Avanzi, I., Yee-King, M.: Measuring the impact of level of detail for environmental soundscapes in digital games, in Proc. 146 Audio Engin. Soc. Conv. (London, 2019).
24. David, N., Newen, A., Vogeley, K.: The “sense of agency” and its underlying cognitive and neural mechanisms. Consciousness and cognition **17**, 523–534 (2008).
25. Delle Monache, S., Polotti, P., Rocchesso, D.: A toolkit for explorations in sonic interaction design, in Proc. Int. Conf. Audio Mostly (AM2010) (Piteå, 2010), 1–7.
26. Dobashi, Y.,Yamamoto, T., Nishita, T.: Real-time Rendering of Aerodynamic Sound using Sound Textures based on Computational Fluid Dynamics, in Proc. ACM SIGGRAPH 2003 (San Diego, 2003), 732–740.
27. Durr, G., Peixoto, L., Souza, M., Tanoue, R., Reiss, J. D.: Implementation and evaluation of dynamic level of audio detail, in Proc. 56th AES Int. Conf. Audio for Games (London, 2015).
28. Ernst, M. O., Bülthoff, H. H.: Merging the senses into a robust percept. TRENDS in Cognitive Sciences **8**, 162–169 (2004).
29. Farnell, A.: An introduction to procedural audio and its application in computer games (2007). URL http://obiwannabe.co.uk/html/papers/proc-audio/proc-audio.pdf. Accessed March 29, 2021.
30. Farnell, A.: Designing sound (MIT Press, 2010).
31. Farnell, A. in Game sound technology and player interaction: Concepts and developments (ed Grimshaw, M.) 313–339 (Information Science Reference, 2011).
32. Fouad, H., Hahn, J. K., Ballas, J. A.: Perceptually Based Scheduling Algorithms for Real-time Synthesis of Complex Sonic Environments, in Proc. Int. Conf. Auditory Display (ICAD97) (Palo Alto, 1997).

33. Gaver, W. W.: How do we hear in the world? Explorations of ecological acoustics. Ecological Psychology **5**, 285–313 (1993).
34. Gaver, W. W.: What in the world do we hear? An ecological approach to auditory event perception. Ecological Psychology **5**, 1–29 (1993).
35. Gibson, J. J.: The ecological approach to visual perception (Lawrence Erlbaum Associates, Mahwah, NJ, 1986).
36. Giordano, B., Avanzini, F. in Multisensory Softness (ed Luca, M. D.) 49–84 (Springer Verlag, London, 2014).
37. Hacıhabiboğlu, H. in Game Dynamics: Best Practices in Procedural and Dynamic Game Content Generation (eds Korn, O., Lee, N.) 47–69 (Springer International Publishing, Cham, 2017).
38. Hendrix, C., Barfield, W.: The Sense of Presence within Auditory Virtual Environments. Presence: Teleoperators and Virtual Environment **5**, 290–301 (1996).
39. Hiller, L., Ruiz, P.: Synthesizing Musical Sounds by Solving the Wave Equation for Vibrating Objects: Part I. J. Audio Eng. Soc. **19**, 462–470 (1971).
40. Hiller, L., Ruiz, P.: Synthesizing Musical Sounds by Solving the Wave Equation for Vibrating Objects: Part II. J. Audio Eng. Soc. **19**, 542–551 (1971).
41. Jack, R. H., Stockman, T., McPherson, A.: Effect of latency on performer interaction and subjective quality assessment of a digital musical instrument, in Proc. Int. Conf. Audio Mostly (AM'16) (Norrköping, 2016), 116–123.
42. Jørgensen, K. in Game sound technology and player interaction: Concepts and developments (ed Grimshaw, M.) 78–97 (Information Science Reference, 2011).
43. Kaaresoja, T., Brewster, S., Lantz, V.: Towards the temporally perfect virtual button: touch-feedback simultaneity and perceived quality in mobile touchscreen press interactions. ACM Trans. Applied Perception **11**, 1–25 (2014).
44. Kelly, J. L., Lochbaum, C. C.: Speech synthesis, in Proc. 4th Int. Congr. Acoustics (Copenhagen, 1962), 1–4.
45. Kilteni, K., Groten, R., Slater, M.: The sense of embodiment in virtual reality. Presence: Teleoperators and Virtual Environments **21**, 373–387 (2012).
46. Lago, N. P., Kon, F.: The quest for low latency, in Proc. Int. Computer Music Conf. (ICMC2004) (Miami, 2004).
47. Larsson, P., Väljamäe, A., Västfjäll, D., Tajadura-Jiménez, A., Kleiner, M. in The engineering of mixed reality systems (eds Dubois, E., Gray, P., Nigay, L.) 143–163 (Springer, 2010).
48. Lester, M., Boley, J.: The effects of latency on live sound monitoring, in Proc. 123 Audio Engin. Soc. Convention (New York, 2007).
49. Liljedahl, M. in Game sound technology and player interaction: Concepts and developments (ed Grimshaw, M.) 22–43 (Information Science Reference, 2011).
50. Liu, S., Cheng, H., Tong, Y.: Physically-Based Statistical Simulation of Rain Sound. ACM Trans. Graphics **38** (2019).
51. Liu, S., Manocha, D.: Sound Synthesis, Propagation, and Rendering: A Survey. arXiv preprint. 2020.
52. Luebke, D. et al.: Level of detail for 3D graphics (Morgan Kaufmann, 2003).
53. Magill, R. A., Anderson, D. I.: Motor learning and control: Concepts and applications. Eleventh edition (McGraw-Hill New York, 2017).
54. Mäki-Patola, T., Hämäläinen, P.: Latency tolerance for gesture controlled continuous sound instrument without tactile feedback, in Proc. Int. Computer Music Conf. (ICMC2004) (Miami, 2004).
55. Michaels, C. F., Carello, C.: Direct Perception (Prentice-Hall, Englewood Cliffs, NJ, 1981).
56. Moeck, T. et al.: Progressive perceptual audio rendering of complex scenes, in Proc. Symp. on Interactive 3D Graphics and Games (I3D'07) (Seattle, 2007), 189–196.
57. Nordahl, R., Nilsson, N. C. in The Oxford handbook of interactive audio (eds Collins, K., Kapralos, B., Tessler, H.) (Oxford University Press, 2014).
58. O'Regan, J. K., Noë, A.: A sensorimotor account of vision and visual consciousness. Behavioral and Brain Sciences **24**, 883–917 (2001).

59. Picard, C., Tsingos, N., Faure, F.: Retargetting Example Sounds to Interactive Physics-Driven Animations, in Proc. AES Conf. Audio in Games (London, 2009).
60. Poeschl, S., Wall, K., Doering, N.: Integration of spatial sound in immersive virtual environments an experimental study on effects of spatial sound on presence, in Proc. IEEE Conf. Virtual Reality (Orlando, 2013), 129–130.
61. Pörschmann, C.: One's own voice in auditory virtual environments. Acta Acustica un. w. Acustica **87**, 378–388 (2001).
62. Puckette, M.: Max at seventeen. Computer Music J. **26**, 31–43 (2002).
63. Raghuvanshi, N., Lin, M. C.: Physically Based Sound Synthesis for Large-Scale Virtual Environments. IEEE Computer Graphics and Applications **27**, 14–18 (2007).
64. Rath, M., Rocchesso, D.: Continuous sonic feedback from a rolling ball. IEEE MultiMedia **12**, 60–69 (2005).
65. Rath, M., Fontana, F. in The Sounding Object (eds Rocchesso, D., Fontana, F.) 173–204 (Mondo Estremo, Firenze, 2003).
66. Ren, Z., Yeh, H., Lin, M. C.: Synthesizing contact sounds between textured models, in Proc. IEEE Conf. Virtual Reality (Waltham, 2010), 139–146.
67. Ren, Z., Yeh, H., Lin, M. C.: Example-guided physically based modal sound synthesis. ACM Trans. on Graphics **32**, 1 (2013).
68. Risset, J.-C., Wessel, D. L. in The psychology of music (ed Deutsch, D.) Second edition, 113–169 (Elsevier, 1999).
69. Rocchesso, D., Bresin, R., Fernstrom, M.: Sounding objects. IEEE MultiMedia **10**, 42–52 (2003).
70. Schwarz, D., Cahen, R., Brument, F., Ding, H., Jacquemin, C.: Sound level of detail in interactive audiographic 3D scenes, in Proc. Int. Computer Music Conf. (ICMC2011) (Huddersfield, 2011), 312–315.
71. Selfridge, R., Moffat, D., Reiss, J. D.: Sound synthesis of objects swinging through air using physical models. Applied Sciences **7**, 1177 (2017).
72. Sheridan, T. B., Furness, T. A. (eds.): Premier Issue, Presence: Teleoperators and Virtual Environment, vol. 1 (1992).
73. Sikström, E., De Götzen, A., Serafin, S.: The role of sound in the sensation of ownership of a pair of virtual wings in immersive VR, in Proc. Int. Conf. Audio Mostly (AM'14) (Aalborg, 2014), 1–6.
74. Sikström, E., De Götzen, A., Serafin, S.: Self-characterstics and sound in immersive virtual reality - Estimating avatar weight from footstep sounds, in Proc. IEEE Conf. Virtual Reality (Arles, 2015), 283–284.
75. Sinclair, J.-L.: Principles of Game Audio and Sound Design: Sound Design and Audio Implementation for Interactive and Immersive Media (CRC Press, 2020).
76. Skarbez, R., Brooks Jr, F. P., Whitton, M. C.: A survey of presence and related concepts. ACM Computing Surveys **50**, 1–39 (2017).
77. Slater, M.: Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. Phil. Trans. R. Soc. B **364**, 3549–3557 (2009).
78. Smith, J. O.: Physical Audio Signal Processing. Online book. 2010. URL http://ccrma.stanford.edu/Ëœjos/pasp/. Accessed March 11, 2021.
79. Smith, J. O.: Virtual acoustic musical instruments: Review and update. J. New Music Res. **33**, 283–304 (2004).
80. Sonnenschein, D.: Sound design: The expressive power of music, voice, and sound effects in cinema (Michael Wiese Productions, 2001).
81. Human Walking in Virtual Environments: Perception, Technology, and Applications (eds Steinicke, F., Visell, Y., Campos, J., Lecuyer, A.) (Springer Verlag, New York, 2013).
82. Sterling, A., Rewkowski, N., Klatzky, R. L., Lin, M. C.: Audio-Material Reconstruction for Virtualized Reality Using a Probabilistic Damping Model. IEEE Trans. on Visualization and Comp. Graphics **25**, 1855–1864 (2019).
83. Stevenson, R. A. et al.: Identifying and quantifying multisensory integration: a tutorial review. Brain Topography **27**, 707–730 (2014).

84. Stockburger, A.: The game environment from an auditory perspective, in Proc. Level Up: Digital Games Research Conference (eds Copier, M., Raessens, J.) (Utrecht, 2003).
85. Tajadura-Jiménez, A. et al.: As light as your footsteps: altering walking sounds to change perceived body weight, emotional state and gait, in Proc. ACM Conf. on Human Factors in Computing Systems (Seoul, 2015), 2943–2952.
86. Takala, T., Hahn, J.: Sound Rendering. Computer Graphics **26**, 211–220 (1992).
87. Tiraboschi, M., Avanzini, F., Ntalampiras, S.: Spectral Analysis for Modal Parameters Linear Estimate, in Proc. Int. Conf. Sound and Music Computing (SMC2020) (Torino, 2020), 276–283.
88. Tsingos, N., Gallo, E., Drettakis, G.: Perceptual audio rendering of complex virtual environments. ACM Trans. on Graphics (TOG) **23**, 249–258 (2004).
89. Välimäki, V., Pakarinen, J., Erkut, C., Karjalainen, M.: Discrete-time modelling of musical instruments. Rep. Prog. Phys. **69**, 1–78 (2006).
90. Van den Doel, K.: Physically based models for liquid sounds. ACM Trans. Applied Perception **2**, 534–546 (2005).
91. Van den Doel, K., Kry, P. G., Pai, D. K.: FoleyAutomatic: Physically-based Sound Effects for Interactive Simulation and Animation, in Proc. ACM SIGGRAPH 2001 (Los Angeles, 2001), 537–544.
92. Van den Doel, K., Pai, D. K. in Audio Anecdotes (ed Greenebaum, K.) (AK Peters, Natick, MA, 2004).
93. Van Vugt, F. T., Tillmann, B.: Thresholds of auditory-motor coupling measured with a simple task in musicians and non-musicians: was the sound simultaneous to the key press? PLoS One **9**, e87176 (2014).
94. Varela, F., Thompson, E., Rosch, E.: The Embodied Mind (MIT Press, Cambridge, MA, 1991).
95. Visell, Y. et al.: Sound design and perception in walking interactions. Int. J. Human-Computer Studies **67**, 947–959 (2009).
96. Vroomen, J., Keetels, M.: Perception of intersensory synchrony: a tutorial review. Attention, Perception, & Psychophysics **72**, 871–884 (2010).
97. Walsh, R.: Audio plugin development with cabbage, in Proc. Linux Audio Conf. (Maynooth, 2011), 47–53.
98. Wang, K., Liu, S.: Example-based synthesis for sound of ocean waves caused by bubble dynamics. Comput. Anim. and Virtual Worlds **29**, e1835 (2018).
99. Wessel, D.,Wright, M.: Problems and prospects for intimate musical control of computers. Computer Music J. **26**, 11–22 (2002).
100. Zheng, C., James, D. L.: Rigid-body fracture sound with precomputed soundbanks. ACM Trans. Graphics **29** (2010).