Multimedia Systems 7: 55–69 (1999)

Multimedia Systems

Real-time spatial processing of sounds for music, multimedia and interactive human-computer interfaces

Jean-Marc Jot⋆

IRCAM, 1 place Igor-Stravinsky, F-75004 Paris, France

Abstract. This paper gives an overview of the principles and methods for synthesizing complex 3D sound scenes by processing multiple individual source signals. Signal-processing techniques for directional sound encoding and rendering over loudspeakers or headphones are reviewed, as well as algorithms and interface models for synthesizing and dynamically controling room reverberation and distance effects. A real-time modular spatial-sound-processing software system, called *Spat*, is presented. It allows reproducing and controling the localization of sound sources in three dimensions and the reverberation of sounds in an existing or virtual space. A particular aim of the Spatialisateur project is to provide direct and computationally efficient control over perceptually relevant parameters describing the interaction of each sound source with the virtual space, irrespective of the chosen reproduction format over loudspeakers or headphones. The advantages of this approach are illustrated in practical contexts, including professional audio, computer music, multimodal immersive simulation systems, and architectural acoustics.

Key words: 3D sound – Multimedia sound spatialization – Environmental audio

1 Introduction

1.1 Aims of spatial sound reproduction

The reproduction of complex sound scenes containing multiple sources at different positions in an existing or imaginary space has long been a major concern in professional recording and production of music and soundtracks. More recently, the evolution of computer technology has led to the development of "virtual reality" systems aiming at immersing an individual in an artificial scene through the reconstruction of multisensorial cues (particularly auditory, visual and haptic cues).

From an auditory point of view, the spatial cues to be reproduced can be divided into two categories: the *auditory localization* of sound sources (desirably in three dimensions), and the *room effect* resulting from indirect sound paths (reflections and reverberation on walls and obstacles). The benefits of spatial sound reproduction over mono reproduction are significant in a wide range of artistic, research, entertainment or industrial applications. These include professional audio and computer music, teleconferencing, simulation and virtual reality, telerobotics, and advanced machine interfaces for data representation or visually disabled users.

Spatial sound rendering is a key factor for improving the legibility, naturalness and telepresence of a virtual scene, restoring the ability for our auditory system to segregate sounds emanating from different directions (Blauert 1983; Begault 1994). It also allows manipulating the spatial attributes of sound events for creative purposes or augmented reality (Cohen and Wenzel 1995). However, as a general rule, the effectiveness of the reproduced auditory cues is influenced by their degree of coherence with concurrent visual or cognitive cues (or, also, by the absence of such additional cues).

1.2 Basic principles of spatial sound-processing

Spatial sound reproduction requires an electro-acoustic system (loudspeakers or headphones) which must be defined according to the context of application (e.g., concert performance, motion picture theater, domestic hi-fi installation, computer display, individual head-mounted display...). In association with this system, a technique or format must be defined for encoding directional localization cues on several audio channels for transmission or storage. A spatially encoded sound signal can be produced by two complementary approaches.

a) *Recording an existing sound scene* with a coincident or closely spaced microphone system (placed essentially at or near the virtual position of the listener within the scene). This can be, e.g., a stereo microphone pair, a dummy head, or a Soundfield microphone (Farrah 1979). Such a sound pickup technique can simultaneously encode, with varying degrees of fidelity, the spatial auditory cues associated to all

⋆ *New address*: Joint E-mu / Creative Technology Center, 1600 Green Hills Road, Suite 101, P.O. Box 660015. Scotts Valley, CA 95067, USA

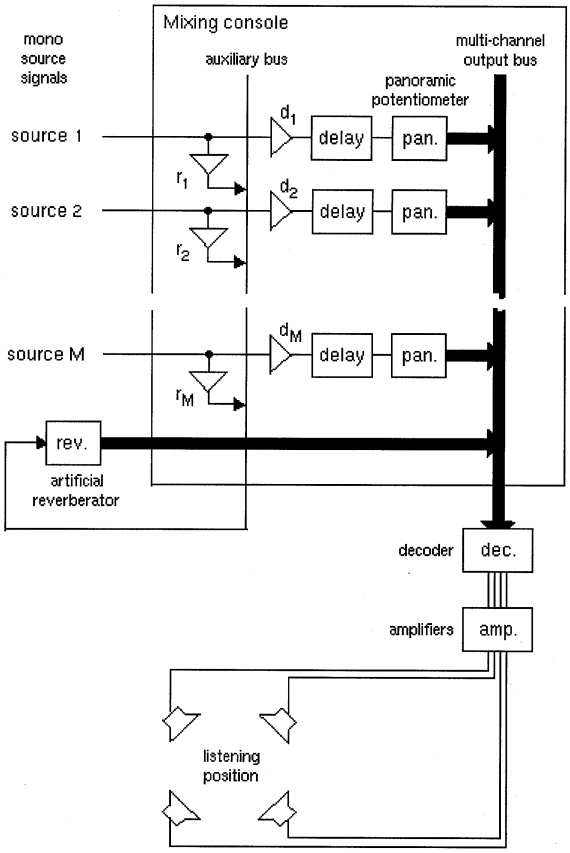

Fig. 1. Typical mixing architecture (here, assuming four-channel loudspeaker reproduction) combining a mixing console providing control over the directional effects and an external reverberation unit for synthesizing the temporal effects

the different sound sources in the recorded scene, as perceived from a given position. However, this approach will considerably limit the possibilities of future manipulations of the relative positions of the sources, modifications of room reverberation or adaptation to various reproduction setups or listening conditions.

b) *Synthesizing a virtual sound scene*. In this approach, the localization of each sound source and the room effect are artificially reconstructed by use of an electronic signal-processing system, which receives individual source signals and provides a control interface for describing the virtual sound scene. The control parameters may include the position, orientation and directivity of each source, along with an acoustic characterization of the virtual room or space. An example of this approach, taken from the professional audio field, is the post-processing of a multitrack recording using a stereo mixing console and peripherals such as artificial reverberators (Fig. 1).

1.2.1 Real-time processing

In an interactive application where elements of the sound scene can be dynamically modified by the user's or performer's actions (for instance, to track the movements of the sound sources or the listener), it is necessary to reconstruct a virtual sound scene and update its control parameters in real time. This will generally require local signal-processing resources within the audio display system, and involve a processing complexity increasing linearly with the number of sound events to be synthesized simultaneously. From a general point of view, the spatial synthesis parameters can be provided either by the user's actions (man-to-machine interface: mixing desk, graphic or gestual interface...), by a stand-alone process (musical sequencer or automation system, simulator, video game...), or by the analysis of an existing scene (via magnetic or ultrasound position trackers, cameras, etc.).

The spatial synthesis technique can be designed to simulate the directional encoding characteristics of a given microphone pickup technique. Such compatibility allows combining the two approaches (a) and (b) described above to produce a complex and realistic sound scene, while minimizing the signal-processing resources (by spatial processing of a limited number of source signals, and mixing with a prerecorded 'ambiance' signal). When a recording made with a stereo microphone pickup is to be mixed with monophonic signals recorded separately, the mixing console's panoramic potentiometers (or 'panpots') should, ideally, match the directional encoding characteristics of that stereo microphone system, in order to optimize the naturalness and coherence of the final mix (this is usually not possible, though, in current mixing environments).

1.3 Summary and outline

In the next section of this paper, the general principles and limitations of current spatial sound-processing and room simulation technologies are reviewed. Binaural, ambisonic, and conventional panning approaches are considered for encoding and rendering directional localization cues over loudspeakers or headphones. According to the context of application, performance and cost criteria dictate the choice of the most adequate directional encoding technique and rendering setup.

Recently developed zero-delay fast-convolution algorithms can be used for exact rendering of a room effect. However, artificial reverberation algorithms based on multichannel feedback delay networks can be more efficient for dynamic and tunable auralization of virtual rooms. The limitations of control interfaces based on physical propagation models are outlined in this context.

In the third section, a spatial sound processor developed by IRCAM and Espaces Nouveaux[1], the Spatialisateur, is introduced. Its modular signal-processing architecture and design are guided by computational efficiency, scalability and configurability considerations. This allows, in particular, straightforward adaptation to various multichannel output formats and reproduction setups over loudspeakers or headphones, while the control interface provides direct access to perceptually relevant parameters for specifying distance and reverberation effects, irrespective of the chosen reproduction format.

[1] IRCAM (Institut de Recherche et Coordination Acoustique/Musique) is a research, creation and education institute encouraging interaction between researchers and musicians. Espaces Nouveaux, also located in Paris, is a research and creation center focusing on sound in design and architecture.

In conclusion, current and prospective practical applications are outlined in several contexts: creation and production of music and soundtracks, multimedia and immersive simulation systems, post-production of recordings, live concert performance, sound reinforcement and architectural acoustics.

2 An overview of current 3D sound-processing techniques

2.1 The basic mixing architecture

In a natural situation, directional localization cues (perceived azimuth and elevation angles of a sound source with respect to the listener) are typically conveyed by the direct sound path from the source to the listener. However, the intensity of this direct sound is not a reliable distance cue in the absence of a room effect, especially in an electroacoustically reproduced sound scene (Blauert 1983; Begault 1994). Thus the typical mixing structure shown on Fig. 1 defines the minimum signal-processing system for conveying 3D localization cues simultaneously for M sound sources over P loudspeakers.

Each input channel of the mixing console receives a monophonic recorded or synthetic signal (preferably devoid of room effect) from an individual sound source, and contains a panning module which assigns a directional localization to this sound (this module is usually called a panoramic potentiometer, or 'panpot', in stereo mixing consoles). The role of the panning module is to encode, over P output channels, the acoustic information conveyed by a sound coming from a given direction in free field (i.e., in an anechoic environment). As shown on Fig. 1, an output-decoding stage may be included, depending on the spatial encoding technique used, before delivering the mix to the loudspeakers. The main output bus can carry an additional pre-recorded (spatially encoded) P-channel source signal, itself containing multiple sound sources at different positions with their associated environmental information (reverberation), to be mixed with the spatialized source signals.

Additionally, via an auxiliary output bus, all source signals can feed an artificial reverberator which delivers several uncorrelated reverberation signals to the main output channels, thus reproducing a diffuse immersive room effect, in which every sound source can contribute a different intensity. The direct sound level and the reverberation level can be adjusted individually in each source channel (gains d and r) in order to control the perceived distance of the corresponding sound source. Recent digital mixing consoles also include a tunable delay line in each channel, allowing the simulation of propagation delays in the virtual scene.

This mixing architecture can produce for the listener the illusion that the sound sources are located at different positions in a virtual room. Although it is usually implemented to produce a conventional two-channel stereo output, it can readily be extended to multichannel loudspeaker layouts in two or three dimensions. This essentially requires that an appropriate 'panpot' module be designed for a given encoding format or loudspeaker layout. Chowning (1971) designed a spatial processing system for computer music, which had an architecture similar to Fig. 1. This system allowed bidimensional control of the localization and movements of virtual sound sources over four loudspeakers. The localization of each source was parametrized by polar coordinates (distance and azimuth angle in the horizontal plane) referenced to the central listening position.

2.2 Approaches to directional encoding and rendering over loudspeakers

To reproduce the direction of each sound source over four loudspeakers, Chowning used a pairwise intensity-panning technique (sometimes referred to as 'discrete panning'), derived from the conventional stereo panning module (Chowning 1971; Theile and Plenge 1977). More generally, the current techniques for directional panning of sounds in two or three dimensions involving a reasonably limited number of loudspeakers are based on one of the following two approaches.

a) Some extension of the discrete surround panning technique, this approach being characterized by the fact that only a limited number of neighboring loudspeakers will be fed for rendering a given direction (except for sounds localized closer than the loudspeakers).

b) Simulation of the directional encoding characteristics of some arrangement of coincident or closely spaced microphones. The direction of an incident plane wave is encoded in the differences between the different recording channels: intensity differences are directly derived from the directivities and orientations of the microphones, and time differences (typically less than 1 ms) can be introduced by spacing the microphones.

As mentioned earlier, this second approach will offer the benefit of straightforward mixing compatibility with an actual recording made in an existing situation according to the same pickup technique.

This can be applied to any spatial sound pickup system, and particularly:

- conventional stereo recording techniques, using a pair of coincident or spaced microphones;
- binaural (dummy head) recording – actually a particular case of a spaced stereo microphone technique;
- the four-channel ambisonic 'B format', as produced by a Soundfield microphone, which is equivalent to the coincident association of one omnidirectional microphone and three orthogonal figure-of-eight microphones (Farrah 1979).

Unlike conventional two-channel stereo recording formats, which cannot encode sound directions in three dimensions, the B format encodes front-back and up-down cues, in addition to left-right cues and an omnidirectional pressure information. For reproduction over loudspeakers, a B-format recording must be processed through a decoding matrix (decoder in Fig. 1). Ambisonic decoders can accommodate various multichannel loudspeaker layouts of typically 4–8 loudspeakers (Gerzon 1985, 1992; Malham and Myatt 1995). As described below, a binaural recording should, similarly, be processed through a 'transaural' decoder in order to provide

faithful 3D reproduction over loudspeakers (Schroeder 1973; Cooper and Bauck 1989; Gardner 1997).

Because its directional encoding functions are spherical harmonics, the B format offers the particular feature of allowing manipulations of whole recorded or synthetic sound fields containing multiple sound sources at different positions, with reflections and reverberation. Such manipulations include rotations and symmetries (about an axis or a plane containing the listener's position), or 'dominance' (focus) in a given direction. These transformations, and combinations thereof, can be achieved simply by applying a 4×4 amplitude matrix to a B-format signal (Malham 1990; Malham and Myatt 1995). In another format, such transformations typically require reprocessing and mixing the elementary source signals.

2.2.1 Selecting a reproduction format over loudspeakers

In practice, all of the above directional reproduction techniques assume that the listener is located at a specific position with respect to the loudspeakers (the "sweet spot"), and some degradation of the auditory illusion must be expected for a non-centrally located listener. However, the type and annoyance of the degradation will depend on the technique used, the range of directions to be reproduced, and the listening conditions (particularly the number of loudspeaker channels, the dimensions of the loudspeaker layout and the size of the listening area, as well as the directivity characteristics of the loudspeakers and the acoustics of the listening room). In principle, wavefront reconstruction techniques or spatial sampling approaches – although essentially at a research and experimental stage at the time of this writing – can overcome these limitations (Berkhout et al. 1993; Nelson et al. 1996).

The consequence for the musician or sound engineer is that, for each particular listening or performance situation, the choice should be made of an appropriate loudspeaker layout and directional encoding technique. To achieve 3D sound reproduction, transaural techniques are appropriate in individual listening configurations over a small number of loudspeakers (2–4), while intensity panning or ambisonic techniques are recommended for larger audiences and concerts (using typically 6–8 loudspeakers in the horizontal plane, plus additional loudspeakers above and below for 3D localization). Generally speaking, ambisonics performs better at concealing the acoustic presence of the loudspeakers and maintaining a stable localization performance irrespective of the localization angle, whereas pairwise intensity panning performs perfectly for synthesizing the direction of a loudspeaker, but produces poorer rendering of lateral "phantom sources" (Thiele and Plenge 1977).

The already established or currently developing multichannel audio industry standards, if one excepts ambisonics and its derivatives, provide multichannel transmission/storage formats directly associated with specific loudspeaker layouts, leaving the choice of the encoding technique to the program producer. In particular, the widely used matrix encoding of a four-channel recording (left, center, right, surround) over two transmission/storage channels (Dolby stereo format) is applied downstream of the directional panning and mixing process over four channels, and thus does not perform in itself a directional encoding function in the sense used in this paper.

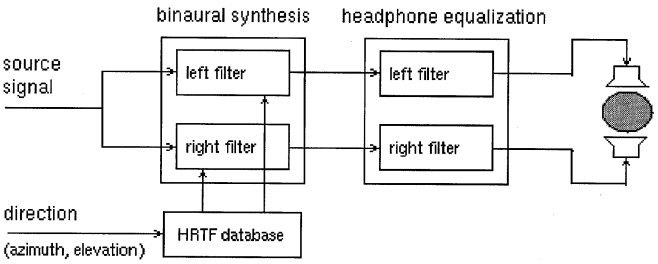

Fig. 2. Principle of binaural synthesis, simulating a free-field (i.e., anechoic) listening situation over headphones. The direction of the virtual sound source (azimuth and elevation) is reproduced by a pair of digital filters whose coefficients are loaded from a database of 'head-related transfer functions' (HRTFs)

In an effort to overcome some of the limitations of the conventional two-channel transmission format and address HDTV, multimedia and domestic entertainment applications, a '3/2-stereo' standard has been developed (see, e.g., Theile 1993). In addition to the left and right channels, the 3/2-stereo format provides a center channel for stabilizing frontal sounds, as well as two 'surround' channels (left and right) intended to feed rear lateral loudspeakers, essentially for rendering diffuse ambiance and reverberation. This format assumes a forward preference in the localization of primary sound sources in the reproduced scene, and does not address accurately controlled localization of lateral, rear or elevated sound sources.

2.3 Binaural and transaural processing

The binaural encoding format has the property of allowing 3D encoding of sounds (including elevation) over two channels only. A binaural recording (made with a dummy head or with two miniature microphones inserted in the ear canals of an individual) can logically be expected to provide exact spatial sound reproduction over headphones, since this technique aims at directly restoring the pressure signals at the two ears (provided that proper care is taken in equalizing the frequency response of the headphones used for playback).

The binaural panning module can be implemented in the digital domain as illustrated on Fig. 2. By use of a loudspeaker and a dummy head (or two miniature microphones inserted in the ear canals of an individual) in an anechoic room, a set of 'head-related transfer functions' (HRTFs) can be measured, subsequently allowing the simulation of any of a set of directions of incidence of a sound wave in free field (Blauert 1983; Moller 1992; Begault 1994).

2.3.1 Performance of binaural synthesis over headphones

The HRTFs encode the diffraction effects undergone by a sound wave on its way to the ear canals, which depend essentially on the morphology of the head and pinnae. Consequently, the applicability of binaural technology in broadcasting and recording is limited by the individual character

of the HRTFs. In order to ensure perfect reproduction over headphones, it is in theory necessary to carry out HRTF measurements for each listener and produce an individualized recording. A typical consequence of using non-individual HRTFs is the difficulty of rendering, over headphones, virtual sound sources localized in the frontal sector: these will often be heard above or behind, near or even inside the head (Wenzel et al. 1993; Begault 1994).

Natural rendering over headphones calls for the use of a head-tracking system in order to dynamically compensate for the movements of the listener's head in real time (especially head rotations) within the binaural synthesis process. Although this requirement may appear as a disadvantage of binaural reproduction, it also has the effect of reducing the problems associated with non-individual HRTFs and frontal source positions: the dynamic localization cues, restored by head-tracking, are naturally exploited by our auditory system to resolve front-back confusions (Blauert 1983; Begault 1994), and can also resolve the perceptual ambiguities resulting from the use of non-individual HRTFs. Consequently, binaural synthesis combined with headphone rendering provides a viable technology for the audio component of an immersive interactive system using a head-mounted visual display.

2.3.2 DSP implementation of binaural synthesis

Early binaural processors and binaural mixing consoles, developed in the late 1980s (Wenzel et al. 1988; Persterer 1989), used powerful digital signal processors in order to accurately implement the HRTF filters in real time. This would typically involve two 200-tap convolution (FIR) filters at a 48-kHz sample rate, requiring about 20 million multiply-accumulates per second (20 MIPS) to process a single static sound source.

Further research on the modeling of HRTFs has led to more efficient implementations, and addressed interpolation issues raised by the design of time-varying HRTF filters allowing the simulation of smooth dynamic movements of sound sources (Foster et al. 1991; Jot et al. 1995). A dynamic implementation essentially involves at least twice the computational cost of a static implementation, which would lead to about 40 MIPS with 200-tap FIR filters. However, with an implementation using minimum-phase pole-zero (IIR) filters and fractional delay filters, this cost can be reduced to about 7 MIPS, still at 48-kHz sample rate (Jot et al. 1995). This represents less than 20% of the computational capacity of a recent digital signal processor such as the Motorola DSP56002.

An alternative approach, proposed recently, consists of encoding directional cues in B format and decoding the B format for headphone playback (Malham 1993; Travis 1996). With this technique, the computationally intensive binaural synthesis process is concentrated at the decoding stage, using a static 4×2 matrix of HRTF filters whose computational cost can be evaluated to 20 MIPS at a 48-kHz sample rate, while the B-format panpots cost about 0.5 MIPS instead of 7 MIPS per source signal. In addition, B-format encoding offers the significant advantage of allowing compensation of the listener's head rotations (headtracking) at the playback stage, simply by inserting a dynamic 4×4 rotation matrix before the headphone decoder (about 2 MIPS).

Although it can probably not be expected to provide the same degree of fidelity over headphones as direct binaural encoding, B-format encoding is attractive for rendering multiple sound sources in a virtual reality context, and for playing back recorded sound fields or broadcast recordings over headphones with head-tracking. This approach of headphone rendering can be used with any loudspeaker reproduction technique (via binaural synthesis of 'virtual loudspeakers'), although the spherical harmonic structure of the B format makes it particularly elegant in this context.

2.3.3 Loudspeaker reproduction using transaural techniques

In order to render the 3D localization cues over a pair of loudspeakers, a binaural signal must be decoded through a 2×2 inverse matrix transfer function which attempts to cancel the cross-talk from each loudspeaker to the opposite ear (Schroeder 1973; Cooper and Bauck 1989). Although this technique implies a strong constraint on the position and orientation of the listener's head with respect to the loudspeakers, it is a viable approach in the recording industry for broadcasting 3D sound scenes over two channels, and it can also be used to improve audio reproduction in multimedia computer terminals.

Experience indicates that, with a carefully installed listening setup in the conventional stereophonic layout, transaural stereophony can produce effective localization cues outside of the frontal sector delimited by the two loudspeakers, although there remains a degree of uncertainty for virtual sound sources located in the rear sector or above and below the horizontal plane. On a less carefully installed hi-fi system, directional localization performance may be essentially reduced to that of conventional stereophony. Current research towards improved transaural reproduction involves the introduction of a head-tracking device and position-adaptive decoders (Gardner 1997) or multichannel extensions of the technique, possibly involving least squares optimization over a set of listening positions (Bauck and Cooper 1992; Nelson et al. 1996).

2.4 Artificial reverberation

Early digital reverberation algorithms based on delay lines with feedback, following Schroeder's pioneering studies (Schroeder 1962), evolved into more sophisticated designs during the 1980s (Gardner 1998). These improvements allowed shaping the early reflection pattern and simulating the later diffuse reverberation more naturally and accurately (Moorer 1979; Stautner and Puckette 1982; Kendall et al. 1986; Griesinger 1989; Jot and Chaigne 1991). An artificial reverberation algorithm including a multichannel feedback delay network, such as in Fig. 3, can mimic the reverberation decay characteristics of an existing room and deliver several uncorrelated channels of natural-sounding reverberation, while using only a fraction of the processing capacity of a typical programmable DSP (Jot 1992; 1997).

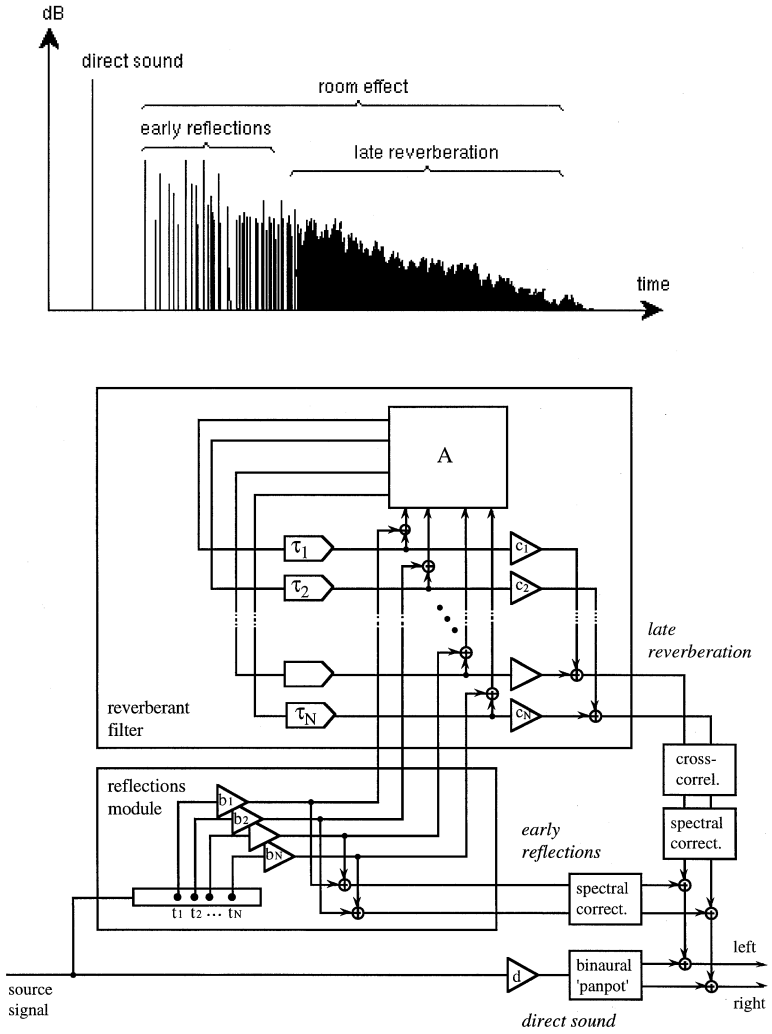

Fig. 3. Typical schematic echogram for a source and a receiver in a room, and cost-efficient real-time binaural room simulation algorithm based on a feedback delay network (Jot et al. 1993, 1995a)

In the algorithm of Fig. 3 (Jot et al. 1993; 1995a), the delay lengths t_i and gains b_i allow controling the time, amplitude and lateralization of each early reflection over headphones. The feedback matrix A is a unitary (energy-preserving) matrix, and the feedback delay lines t_i incorporate attenuation filters designed to provide accurate tuning of the reverberation decay time vs. frequency (Jot and Chaigne 1991; Jot 1992). The feedback delay network produces two uncorrelated channels of late reverberation, which are processed by a frequency-dependent matrix to restore the cross-correlation between the left and right ear signals in a diffuse sound field. Second-order IIR spectral correctors can be used for dynamic level adjustment of the direct sound, early reflections and late reverberation, independently, and continuous tuning of the reverberation decay time, all in three frequency bands (low, medium, high).

A typical implementation of a stereo version of this artificial reverberation algorithm, using eight feedback channels, requires about 200 multiply-accumulates per sample period (or 10 MIPS at a 48-kHz sample rate). This algorithm can also produce up to eight uncorrelated channels of artificial reverberation for essentially the same total processing cost. Changing the stereo implementation into a binaural reverberation algorithm essentially requires inserting a binaural panpot in the direct sound path, if diffuse-field equalization is adopted for the binaural output signal (Jot et al. 1995a). This brings the cost of the binaural reverberation algorithm of Fig. 3 to a total of 17 MIPS (less than half the capacity of a Motorola 56002). Despite this low processing cost, the control parameters of the algorithm can be tuned to provide a convincing simulation of an existing room, on the basis of a time-frequency analysis of a measured impulse response (Marin 1996; Jot et al. 1997).

2.4.1 The convolution approach

A new approach to real-time artificial reverberation was developed recently, based on hybrid convolution in the time and frequency domain (Gardner 1994; Reilly and McGrath 1995). Unlike earlier block-convolution algorithms, these hybrid algorithms allow implementing a very long convolution filter with no input-output delay, while still maintaining a reasonable computational cost. Convolution processing allows exact reproduction of reverberation on the basis of an impulse response measured in an existing room or derived from a computer model.

However, it is impractical to dynamically update the coefficients of the impulse response in a convolution proces-

sor in order to tune the artificial reverberation effect (e.g., modify the reverberation decay time or the early reflection parameters). As will be apparent in the next section, simulating moving sound sources may involve individual manipulation of each early reflection, which suggests an early reflections module implemented as in Fig. 3. In an interactive application, the convolution approach must typically be restricted to the rendering of the late reverberation, which can be achieved more efficiently by a feedback delay network as described earlier, especially over multiple output channels. Real-time convolution is thus essentially useful for static auralization, in the context of laboratory experiments or evaluation testing.

2.5 Dynamic distance and room effect control

2.5.1 Limitations of the conventional mixing architecture

When a reverberation processor is interfaced with a conventional mixing console as in Fig. 1, a heterogeneous system combining two unrelated control interfaces is obtained, where the direct sound and the room effect can only be manipulated independently. This is true both of the typical studio mixing architecture and of virtual mixing environments as found in modern digital direct-to-disk editing workstations or integrated digital mixing desks. The heterogeneity of the user interface makes the continuous adjustment of the perceived distance of a virtual sound source impractical for the musician or sound engineer, because this effect cannot be effectively rendered by simply attenuating the direct sound with no consideration of reverberation parameters. Furthermore, most current reverberation units offer poor control interfaces for tuning the acoustics of the virtual room, usually on the basis of factory-preset categories (large halls, chambers...) from which intuitive modifications are typically limited to adjustments of the decay time or the size of the virtual room.

In addition to these limitations in the control interface, these signal-processing architectures currently do not effectively address reproduction formats other than conventional stereophony. This makes traditional mixing structures inadequate for interactive and immersive audio simulation, or broadcasting and production of recordings in multichannel formats such as 3/2-stereo. Overcoming these limitations implies integrating the reverberation processing with the directional processing in each source channel of the mixing console, in association with an improved control interface, and the possibility of selecting between various multichannel output formats.

2.5.2 Chowning's model

Compared to traditional mixing environments, Chowning's initial design provided an improved distance control, simultaneously affecting the gains d and r in the architecture of Fig. 1: the intensity of the direct sound followed the natural inverse squared distance law, while the intensity of the reverberation decayed less rapidly with increasing distance, following an (empirically defined) inverse distance law (Chowning 1971). Additionally, the system incorporated pitch shifters for simulating Doppler effects accompanying dynamic variations of distance.

Despite these improvements, Chowning found it necessary to reinforce the perception of the direction of the sound event for large distances. An improvement was obtained by modifying the mixing architecture so that a fraction of the reverberation signal was directional (coming from the same direction as the direct sound). This implied that the late reverberation decay was stronger in the direction of the sound event, whereas, in a natural environment, the position of the sound source essentially affects the temporal and directional distribution of the early reflections, leaving the late reverberation evenly distributed over all directions of incidence. This fact suggests a modified mixing architecture as shown in Fig. 4, where each input channel includes an early reflection processor, while the late reverberation algorithm remains common to all sources (Moore 1983; Jot 1992).

2.5.3 Moore's model

Moore (1983) proposed a signal-processing architecture allowing the control of the amplitudes and time delays of the first reflections, for each source signal and each output channel, according to the following parameters:

- the position, directivity and orientation of each virtual sound source,
- the geometry of the virtual room and the absorption characteristics of the air and walls,
- the geometry of the loudspeaker system.

The general processing model proposed by Moore for concert performances consists of a polygonal 'listening' room (delimited by the loudspeaker positions and containing the audience), inserted in a larger room (the 'virtual' room) containing the virtual sound sources. The signals delivered to the loudspeakers are reconstructions of the signals captured by P notional microphones located at the positions of the P loudspeakers along the exterior perimeter of the 'listening' room. This directional encoding method simulates a recording technique using a non-coincident microphone system, according to the principles described in Sect. 2.2 (although the microphones are much more spaced in this model than in conventional non-coincident recording techniques).

The identification of indirect sound paths from each source to each notional microphone is based on a geometrical simulation of sound propagation, assuming specular reflections of sound waves on the walls of the virtual room (according to the image source model). The arrival time and frequency-dependent attenuation of each early reflection can be computed by simulating all physical phenomena along the corresponding 'sound ray' as a cascade of elementary linear filters (taking into account the directivities of the source and the microphone, as well as absorption by propagation through the air and by reflections on the walls).

Moore deals with the particular case of headphone reproduction by reducing the size of the 'listening room' to the size of a head, and placing the two notional microphones on its sides. The directional encoding model then becomes equivalent to an approximate implementation of the 'binaural

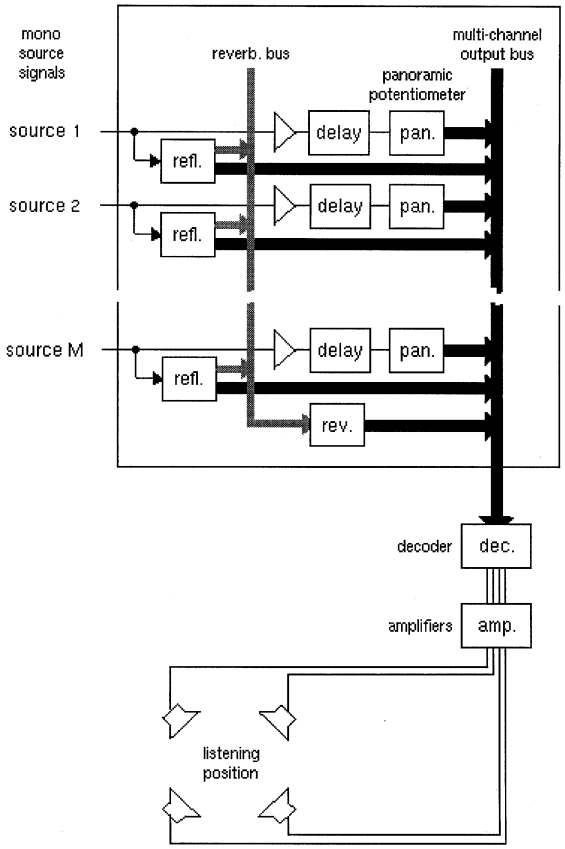

Fig. 4. Modified mixing architecture for reproducing several virtual sound sources located in the same virtual room, while controling the early reflection pattern for each individual source (Moore 1983; Jot 1992)

panpot' described in Sect. 2.3, and can be readily extended in order to simulate HRTF filtering more accurately (Persterer 1989; Foster et al. 1991; Jot 1992).

2.5.4 Limitations of the physical propagation model

The signal-processing structure underlying the above geometrical propagation model is similar to the basic mixing architecture of Fig. 1, except that each source signal must be fed in parallel to several input channels of the mixing console so that each additional channel reproduces one early reflection. The delay and gain can be adjusted in each channel to control the arrival time and amplitude of the reflection (as captured by a notional omnidirectional microphone placed at the reference listening position) and the panning module then derives P signals in order to encode the direction of incidence of the reflection.

Real-time digital audio processors performing binaural processing of both the direct sound and several early reflections have been proposed, where room reflection parameters are computed according to the image source model (Persterer 1989, Foster et al. 1991). Such systems involve a heavy real-time signal-processing effort, since a binaural panning module must be assigned to each early reflection, for each virtual sound source. For a typical sound scene containing several sources, this may imply about 30–50 binaural panpots. This means several hundred MIPS just for rendering directional effects, which is impractical for most real-world applications. In addition to this signal-processing cost, a significant computational effort is necessary for updating all reflection parameters dynamically whenever a sound source is displaced or the listener moves. As Moore noted, these parameters should be tracked fast enough to allow smooth dynamic variation of the delay times, and produce natural Doppler effects, both on the direct sound component and on each early reflection (see also Blauert and Lehnert 1995).

Fortunately, the signal-processing cost can be substantially reduced by introducing perceptually relevant simplifications in the spectral and binaural processing of early reflections (Jot et al. 1993; 1995a) or by using the B Format, for instance, as an intermediate encoding format for early reflections (Travis 1996), as mentioned in Sect. 2.3. Yet the overall processing complexity makes this exhaustive approach impractical, unless it is restricted to particularly simple geometries (such as rectangular rooms), a small number of source signals, and a small number of early reflections per source (limited, for instance, to first-order reflections). In contrast, a perceptual control paradigm, as described in the following, will allow drastically reducing the computational effort involved in the dynamic tracking of reverberation parameters, while simultaneously providing intuitive controls for tuning the reverberation quality of the virtual room.

3 Perceptually-based spatial sound-processing

In many applications involving real-time spatial sound-processing, the algorithms used for synthesizing the room effect need not reproduce the exact response of an existing room in a given situation, or model the physical propagation of sound in rooms. Although the statistical properties of room reverberation are relevant to the design of natural-sounding artificial reverberation algorithms (Schroeder 1962; Jot 1992; Jot et al. 1997), reference to the physics of the reverberation process should not impose constraints in the control interface as a consequence of the strategy implemented for synthesizing the room effect.

The following features are desirable, on the other hand, in a spatial reverberation processor.

Tunability, in real time, through perceptually relevant control parameters. The control parameters should include the azimuth and elevation of each virtual sound source, as well as descriptors of the room effect, separately for each source. The perceptual effect of each control parameter should be predictable and independent of the setting of other parameters. A *measurement and analysis procedure* should allow automatically deriving the settings of all control parameters to simulate an existing environment.

Configurability according to the reproduction setup and context. Since there is no single encoding or reproduction format that can satisfy all 3D sound applications, it should be possible, given a specification of the desired localization and reverberation effects, to configure the signal processor in order to allow reproduction of these effects in various formats over headphones or loudspeakers. This should include corrections for preserving the perceived effect, as much as possible, between different setups and different listening rooms.

Computational efficiency and scalability. The processor should make optimal use of the available computational resources. It should be possible, considering a particular application where the user or the designer can accept a loss of flexibility or independence between some control parameters, to further reduce the overall complexity and cost of the system by introducing relevant simplifications in the signal-processing and control architecture. One illustration is the system of Fig. 4, where the late reverberation algorithm is shared between several sources, assuming that these are constrained to move in the same virtual room, while an independent early reflection module is associated with each individual sound source.

3.1 The Spatialisateur

IRCAM and Espaces Nouveaux have developed since 1992 a spatial sound-processing software, the Spatialisateur, which incorporates earlier research on the perceptual characterization of room acoustical quality and on artificial reverberation and spatial processing of sounds (Jullien et al. 1992; Bloch et al. 1992; Jot 1992). The *Spat* software was developed in the *FTS/Max* object-oriented signal-processing environment, and runs in real time on the hardware platforms supported by the IRCAM Music Workstation (Puckette 1991; Déchelle and De Cecco 1995). At the time of this writing, *Spat* can run on Silicon Graphics workstations, Linux machines (including PCs), Power Macintosh computers, or NeXT workstations (the latter equipped with ISPW plug-in boards). Since its first release in April 1995, the software has been used regularly for musical composition and production of concerts and installations, and the post-production of CD recordings using 3D sound effects and 'virtual loudspeakers' surrounding the listener. Other current applications include assisted reverberation systems for auditoria and research on human-computer interfaces, virtual reality, and room acoustics perception.

Spat appears as a library of signal-processing and control interface modules for real-time spatial processing of sounds. The elementary objects include panpots, artificial reverberators and parametric spectral correctors. The signal-processing operations for reconstructing localization and room effect cues associated to one source signal can be integrated in a single compact processor (an object named *Spat*). Several *Spat* processors can be associated in parallel in order to process several source signals simultaneously, and each processor can be easily configured for a chosen encoding technique and loudspeaker layout, or heaphones.

The design approach adopted in the Spatialisateur project focuses on giving the user the possibility of specifying the desired effect from the point of view of the listener, rather than from the point of view of the technological apparatus or physical process which generates that effect. A higher level user interface controls the different signal-processing sub-modules of a *Spat* processor simultaneously, and allows specifying the reproduced effect, for one source signal, through a set of control parameters whose definitions do not depend on the chosen reproduction format or setup (Fig. 5). These parameters include the azimuth and elevation of the virtual sound source, as well as descriptors of the room acoustical quality associated with the sound source.

3.1.1 Perceptual control of room acoustical quality

Spat provides a higher level control interface, in which room acoustical quality is not controlled through a model of the geometry and wall materials of the virtual room, but via a formalism directly related to the perception of the reproduced sound event by the listener, involving a small set of mutually independent 'perceptual factors' (Fig. 5).

- Source perception: *source presence*, *brilliance* and *warmth* (energy and spectrum of direct sound and early reflections).
- Source/room interaction: *envelopment* and *room presence* (relative energies of direct sound, early and late room effect), *running reverberance* (early decay time).
- Room perception: *late reverberance* (late decay time), *heaviness* and *liveness* (variation of late decay time vs frequency).

This control interface also provides parameters for controling the directivity and the orientation of the virtual sound source. The directivity is specified by an axis spectrum, a directivity index and aperture angles (all in three frequency bands with adjustable cross-over frequencies), while the orientation is expressed by *yaw*, *pitch* and *roll* angles.

The definition of the above perceptual factors is derived from psycho-experimental research carried out at IRCAM on the perceptual characterization of room acoustical quality in concert halls, opera houses and auditoria (Jullien et al. 1992; Jullien 1995). In the graphic user interface shown in Fig. 5, each slider is scaled according to the average sensitivity of listeners with respect to the perceptual factor it controls, and each perceptual factor is related to a measurable acoustical index characterizing the sound transformation. These relations are implemented in the Spatialisateur's perceptual control module in order to map this representation of room acoustical quality into low-level signal-processing parameters.

Some of these acoustical indexes are similar to well-established indexes used for characterizing concert hall acoustics (although not explicitly implemented in current commercial reverberation units), such as the envelopment or the early decay time. The unicity of the particular set of indexes used in the Spatialisateur follows from an attempt to provide an exhaustive characterization of room acoustical quality through a minimal set of mutually independent parameters (Lavandier 1989; Jullien et al. 1992; Jullien 1995). These indexes can be computed by analyzing an impulse response measured in an existing room, which allows setting the Spatialisateur's controls in order to mimic a real situation. As a result, virtual and real acoustical qualities can be manipulated within a unified framework.

Simplified versions of the generic room effect model can be invoked, resulting in both a reduction in DSP computation cost and a simplification of the high-level control interface (as it becomes impossible to adjust some of the perceptual factors independently from the others). The model can also

be driven by a geometrical description of the sound scene involving the relative position coordinates of the sources and the listener and a global description of the room in terms of its volume and its statistical absorption coefficient. This makes use of a statistical energetic model of room reverberation decays as a function of time and source-receiver distance (Jot et al. 1997), still not calling for an exhaustive description of the geometry and absorption characteristics of room boundaries.

3.2 User interface: physical vs perceptual approach

The synthesis of a virtual sound scene relies on a description of the positions and orientations of the sound sources and the acoustical characteristics of the space. This description is then translated into parameters of a signal-processing algorithm. From a general point of view, the space can be described either by a physical and geometrical model, or by a set of attributes describing the perceived acoustical quality associated with each sound source (Jullien and Warusfel 1994; Jot 1992). The first approach typically suggests a graphic user interface representing the room geometry and the positions of the sources and listener, associated to a computer algorithm simulating the propagation of sound in rooms (such as the image source model). The second approach relies on a model of the perception of room acoustical quality, suggesting a graphic user interface such as shown on Fig. 5, and forming a basis for a wide range of multidimensional control interfaces (this will be developed further in Sect. 4).

As discussed by Jullien and Warusfel (1994), a physically based user interface will not allow direct and effective control of the sensation perceived by the listener. Although localization is naturally specified via a geometrical user interface, many aspects of room acoustical quality (such as envelopment or early reverberance) will be affected simultaneously by a change in the position of the source or the listener, in a manner that is not easily predictable and depends on room geometry and wall absorption characteristics. On the other hand, adjustments of the room acoustical quality can only be achieved by modifying these geometry and absorption parameters, although the effects of such modifications are often unpredictable or imperceptible. Additionally, a physically based user interface will only allow the reproduction of physically realizable situations: source positions will be constrained by the geometry of the space and, even if the modeled room is imaginary, the laws of physics will limit the range of realizable acoustical qualities. For instance, in a room of a given shape, modifying wall absorption coefficients in order to obtain a longer reverberation decay will cause a simultaneous increase in reverberation level.

In contrast to a physical approach, a perceptual approach leads to a more intuitive and effective user interface, because the control parameters are directly related to audible sensations. Additionally, a perceptually based specification of the room effect or a statistical model the reverberation decay essentially prescribe a time-frequency energy distribution in the impulse response, which can be efficiently mapped to the signal-processing parameters of an artificial reverberation algorithm such as described in Sect. 2.4 and Fig. 3 (Jot 1997; Jot et al. 1997). The specification in terms of energy distribution leaves some freedom in the determination of the microscopic structure of the impulse response, allowing perceptually based simplifications to be made in the implementation of the signal-processing module (provided that its design satisfies the criteria for ensuring the naturalness of the artificial reverberation).

By using control models which do not require an exhaustive description of the room geometry and physical properties of the walls, an efficient and scalable implementation is made possible, both for the signal-processing module itself and for the control process which dynamically updates the low-level signal-processing parameters according to the higher level control interface parameters, without compromising the naturalness and plausibility of the simulated sound scene.

3.3 A modular signal-processing architecture

The modularity of the *Spat* software makes it possible to configure a spatial processing architecture according to various applications or with different computational costs, depending on the reproduction format or setup, the desired flexibility in controling the room effect, and the available computational resources. As shown in Fig. 5, a *Spat* processor can be formed by cascade connection of four configurable sub-modules: *Source*, *Room*, *Pan*, *Out*. Configuring a *Spat* module is done in a straightforward way via arguments calling appropriate versions of these sub-modules from the *Spat* library.

3.3.1 Artificial reverberation modules

The *Room* module is a computationally efficient multichannel reverberator based on a feedback delay network, designed to ensure the necessary degree of naturalness and accuracy for professional audio or virtual-reality applications (Jot 1992; 1997). The input signal (assumed devoid of reverberation) can be pre-processed by the *Source* module, which can include a dynamically variable low-pass filter and delay line to reproduce air absorption and Doppler effects, as well as spectral equalizers allowing additional corrections according to the nature of the input signal. The *Room* module can itself be broken down to elementary reverberation modules (e.g., an early reflection module or a late reverberation module), which allows building a variety of mixing architectures such as those of Fig. 1 or Fig. 4. The reverberation modules are provided in several versions differing in complexity (number of feedback or feedforward channels), so that computational efficiency can be traded off for time or frequency density of the synthetic reverberation.

3.3.2 Directional encoding modules

The multichannel output format of the *Room* module is directly compatible with the reproduction of frontal sounds in the 3/2-stereo format, and comprises seven channels: a ‘center’ channel conveying the direct sound component, a

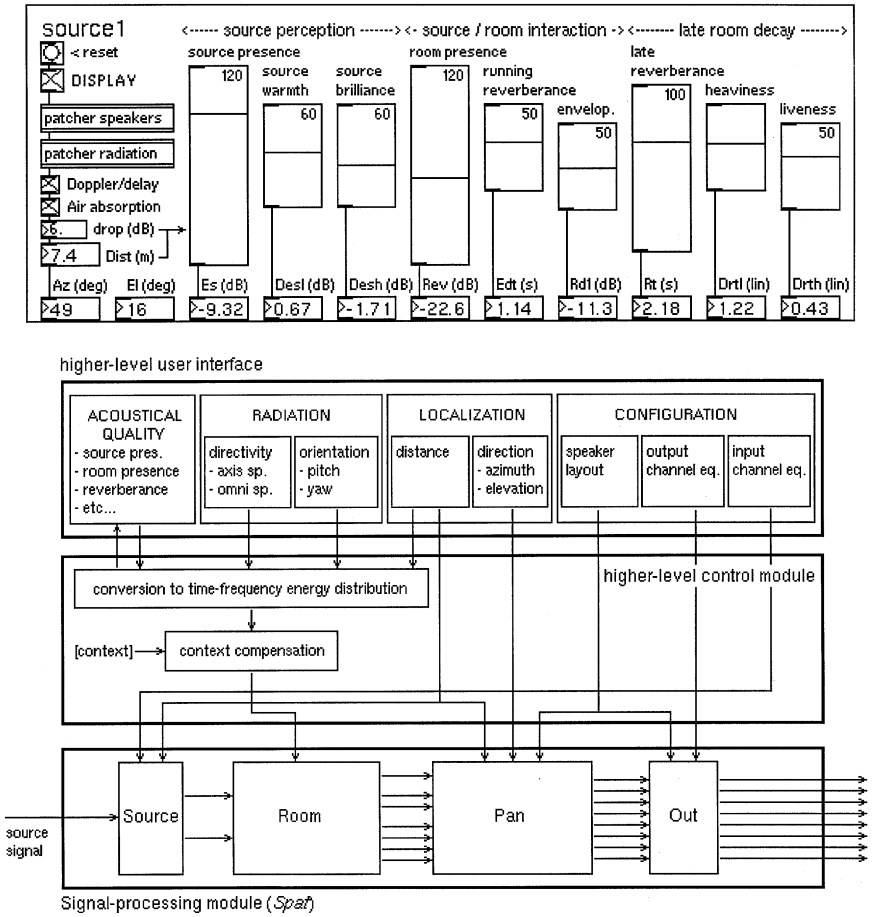

Fig. 5. Higher level user interface and structure of a *Spat* processor (for one source signal). The user interface includes perceptual attributes for tuning the desired effect, as well as configuration parameters which can be set at the beginning of a performance or work session, according to the reproduction format and the characteristics of the listening setup (Jot et al. 1995b)

'left' and 'right' pair of channels conveying the early reflections, and four uncorrelated 'surround' channels conveying the diffuse reverberation. The output of the *Room* module can optionally be post-processed by the directional distribution module *Pan*, a $7 \times P$ matrix which converts the above '3/4-stereo' format to the directional encoding format chosen by the user, and simultaneously encodes the perceived direction of the sound event (Jot et al. 1995b; Jot 1997).

The *Pan* module can be configured for the following encoding formats:

- discrete intensity panning over various 2D or 3D loudspeaker layouts typically comprising 4–8 channels (although the structure of the *Pan* module can be readily extended to a higher number of channels if necessary);
- binaural encoding for 3D sound reproduction over headphones (which can be further decoded for reproduction over 2–4 loudspeakers);
- B-format encoding (which can be decoded for rendering over various 2D or 3D loudspeaker layouts);
- conventional two-channel stereophony (simulating various coincident or non-coincident techniques such as MS, XY, AB, etc.), or Dolby-stereo (Lt, Rt)-compatible encoding.

3.3.3 Adaptation to the listening conditions

The reproduced effect can be specified perceptually in the higher level control interface, irrespective of the reproduction context, and this effect is, as much as possible, preserved from one reproduction mode or listening room to another. The optional *Out* module can be used as a decoder for adapting the output of the *Pan* module to the geometry and acoustical response of the loudspeaker system: it can be configured to provide spectral and time delay correction (in each output channel), perform headphone-to-loudspeaker conversion of a binaural signal, or decode a B-format signal for reproduction over loudspeakers or headphones. In a mixing application, only one *Out* module will generally be necessary (as illustrated by the placement of the decoder in Fig. 1 or Fig. 4).

In a typical multichannel reproduction setup, the *Out* module is used for equalizing the direct sound path from each loudspeaker to a reference listening position, without attempting to compensate for the effects of the reflections and reverberation in the listening room. However, in order to correct for the temporal and spectral effects of listening room acoustics, the high-level control processing module includes a 'context compensation' algorithm which automatically adjusts the control parameters of the *Room* module, so that the perceived effect at a reference listening position be as close as possible to the specification given via the higher level user interface (Fig. 5). The originality of this compensation algorithm lies in that it is based on a deconvolution of the *energy* distribution in the impulse response (Jot et al. 1995b; Jot 1997), instead of a more traditional amplitude deconvolution approach.

The compensation of listening room reverberation by 'echogram deconvolution' does not allow exact signal reconstruction of a given impulse response at a given listening position, and suffers from the general limitation that the desired virtual acoustical quality should be sufficiently reverberant compared to the listening conditions. However, unlike the amplitude deconvolution method, this approach does not involve a prohibitive constraint on the listener's position, and yields an efficient real-time compensation process allowing, for instance, the reproduction of the perceived acoustical quality of a given room in another one, with recorded or live source signals.

4 Applications and perspectives

4.1 Signal-processing architectures for professional audio or interactive multimodal interfaces

Even in a computationally demanding reproduction format such as binaural stereophony, a complete implementation of a *Spat* processor requires less than 500 multiply-accumulates per sample period, i.e., about 25 MIPS at a sample rate of 48 kHz (Jot et al. 1995a; Jot 1997). This can be handled by a single programmable digital signal processor, such as the Motorola DSP56002 or Texas Instruments TMS320C40. It is thus economically viable to insert a full spatial processor (performing both directional panning and artificial reverberation) in each channel of a digital mixing environment by devoting one DSP to each source channel. More economical mixing structures can also be designed, where a single reverberation processor is shared by all source signals, with the only constraint that the late reverberation then receives the same decay time settings for all sound sources (which is natural if they are assumed to be located in the same room).

A configurable mixing console can be designed, capable of producing recordings in traditional or exotic formats, as well as currently developing industry standards: conventional two-channel stereo, 3D two-channel stereo (over headphones or loudspeakers), 3/2-stereo or various multichannel 2D or 3D loudspeaker layouts. The increased processing flexibility of this type of mixing architecture calls for a new generation of user interfaces for studio recording and computer music applications. Providing a reduced set of independent perceptual attributes for each virtual source, as discussed in this paper, seems promising from the point of view of ergonomy and automation.

4.1.1 Virtual reality and multimedia

Spatial sound processors for virtual reality and multimedia (video games, simulation, teleconference, etc.) also rely on a real-time mixing architecture and can benefit substantially from the reproduction of a natural-sounding room effect allowing effective control of the perceived distance of sound events. Many applications involve the simulation of several sources located in the same virtual space, which allows incorporating artificial reverberation efficiently. The RISC architectures of current general-purpose personal computers and workstations offer sufficient computational resource to handle this task in real time, without dedicated signal-processing hardware. It is possible to further reduce the processing cost in applications which can accommodate a less refined reproduction or control of the room effect (e.g., for video games or augmented reality applications where an artificial sensation of distance must be controlled, while means for refined control of the virtual room's signature may be less necessary than for professional recording or computer music).

Binaural reproduction over headphones is particularly suited to virtual reality or multimedia applications, and can be combined with real-time image synthesis in order to immerse a spectator in a virtual environment. *Spat* is designed to allow remote control through pointing or tracking devices and ensure a high degree of interactivity, with low latency and a typical localization control rate of 33 Hz (fast enough for video synchronization or operation with a head-tracking system). An alternative reproduction environment for simulators is a booth equipped with a multichannel loudspeaker system (such as Espaces Nouveaux's "Audiosphere"). Current directions of research include modeling individual differences in HRTFs and individual equalization of binaural recordings, as well as improved techniques for multichannel reproduction over a wide listening area.

4.2 Live computer music performances and architectural acoustics

The perceptual approach adopted in the Spatialisateur project allows the composer to immediately take spatial effects into account at the early stages of the compositional process, without a prescribed reference to a particular electro-acoustical apparatus or performing space. Executing the spatial processing in real time during the concert performance allows introducing automatic corrections according to the reproduction setup and acoustical context. Localization effects, often manipulated in contemporary electro-acoustic music, can thus be more reliably preserved from one performance situation to another. Spatial reverberation processing allows more convincing illusions of remotely located virtual sound sources and helps concealing the acoustic signature of the loudspeakers for a wider listening area. This makes it possible to improve the perceptual integration of live sources with synthetic or pre-recorded signals in the virtual sound scene, which is a significant challenge in the field of computer music (Warusfel 1990; Jullien and Warusfel 1994).

Consequently, a computer music work need not be written *a priori* for a specific number of loudspeakers in a specific geometrical layout. As an illustration, consider an electro-acoustic music piece composed in a personal studio equipped with four loudspeakers. Rather than producing a four-channel mix to be used in all subsequent concert performances, a score describing all spatial effects applied to each sound source can be recorded with a MIDI sequencer software. A new mix can then be produced automatically for a concert performance or installation using eight loudspeaker channels, or a transaural CD recording preserving 3D effects in domestic playback over two loudspeakers. This only implies reconfiguring the signal-processing structure by calling adequate versions of the *Pan* and *Out* modules, and

adjusting the loudspeaker layout and equalization parameters appropriately.

A related application is the design of an electro-acoustic system allowing the modification of the acoustical quality of an existing room, for sound reinforcement or reverberation enhancement purposes, with live sources or pre-recorded signals. In the case of sound systems addressing relatively large audience areas (such as large concert halls or multipurpose halls), the signal-processing architecture can be configured specifically (by interconnecting sub-modules of *Spat* and *Room*), according to a division of the audience and stage areas into adjacent zones, in order to ensure effective control of the perceptual attributes related to the temporal distribution of the direct sound and the early reflections, over all seats in the audience area and for all sources on the stage.

4.3 Musical and multidimensional interfaces

Spat is used in musical creation projects, in concert performances, and in the post-production of CD recordings. At the compositional stage, the perceptual paradigm allows manipulating spatial attributes of sounds as natural extensions of the musical language. Providing perceptually relevant attributes for describing the room effect can encourage the composer to manipulate room acoustical quality as a musical parameter, together with the localization of sound events (Bloch et al. 1992; Jullien and Warusfel 1994).

4.3.1 Automation of spatial parameters

In one approach, the spatial processor's score is recorded in successive passes on the tracks of a sequencer. During each pass, additional manipulations of the spatial attributes of one or several sound sources can be introduced in the score (and monitored simultaneously in real time, combined with spatial manipulations already written in the score). This technique was initiated in 1993 with an early Spatialisateur prototype in a composition by Georges Bloch (*Palmipèdes d'agrément de la rue Morère* for two violas – one live and one recorded).

This is similar to operating an automation system in a mixing console, albeit allowing the manipulation of a coherent set of spatial and room acoustical parameters, which is not possible in current mixing architectures. In this procedure, it is critical that the control parameters be *mutually independent* from a perceptual point of view, i.e., that the manipulation of a spatial attribute may not destroy or modify the perceived effect of previously stored manipulations of other spatial attributes (except possibly in extreme and straightforward cases: for instance, an extremely low setting of the room presence will make adjustments of the late reverberance hardly perceptible). For operational efficiency, it is desirable that the perceived effect of each parameter be *predictable*, particularly when it is desired to edit the score or write it directly without real-time monitoring. As discussed earlier in this paper, such modes of operation are quite impractical within a physically based framework, or with the traditional combination of mixing consoles and reverberation units.

4.3.2 Multidimensional control interfaces

Besides a sequencing or automation process, another approach for creating simultaneous variations of several spatial attributes for one or several virtual sound sources consists of mapping a sub-set of these perceptual attributes to the coordinates of a multidimensional graphic or gestual interface. A basic illustration of this approach is included in the higher level user-interface of the Spatialisateur (Fig. 5), in order to allow straightforward control of a *Spat* processor with a bidimensional or 3D control interface delivering polar localization coordinates to the processor: the 'distance' control is mapped logarithmically to the 'source presence' perceptual factor, with the 'drop' parameter defining the drop of the source presence in dB for a doubling of the distance (setting 'drop' to 6 dB simulates the natural attenuation of a sound with distance). An extension is a bidimensional control map of the virtual sound scene, representing the sound sources at different positions and orientations in the horizontal plane around the listener.

This mapping principle can, of course, be implemented in many other fashions. Because of the nature of the multidimensional scaling analysis procedure from which the definition of the perceptual factors was derived (Lavandier 1989; Jullien et al. 1992; Jullien 1995), this set of factors forms an orthogonal system of perceptual coordinates, defining a Euclidean norm to measure the perceptual dissimilarity between acoustical qualities. This implies that linear weighting along one perceptual factor or a set of perceptual factors provides a general and perceptually relevant method for interpolating between different acoustical qualities (Jullien and Warusfel 1994). For instance, it allows implementing a gradual and natural-sounding transition from the sensation of listening to a singer 20 m away from the balcony of an opera house to the sensation of standing 3 m behind the singer in a cathedral (possibly based on acoustical impulse response measurements made in two existing spaces). In this example, a physically based control paradigm would require implementing an arguable geometrical and physical "morphing" process between the two situations.

These perspectives suggest the development of new multidimensional interfaces for music and audio components of virtual reality. An additional direction of research is the extension of the perceptual control formalism to spaces such as small rooms, chambers, corridors or outdoor spaces. In the current implementation of *Spat*, such spaces can be dealt with by manipulating, in addition to the higher level perceptual factors, the lower level processing parameters provided in the control interface of the *Room* module.

Acknowledgements. The perceptual control interface developed in this project for controling room acoustical quality is derived from psychoacoustic research carried out at IRCAM under the direction of Jean-Pascal Jullien and Olivier Warusfel. The basic research on artificial reverberation algorithms was carried out at the Ecole Nationale Supérieure des Télécommunications, under the supervision of Prof. Antoine Chaigne and in collaboration with Studer Digitec. Research on binaural techniques and processing was carried out in collaboration with the Centre National d'Etudes des Telecommunications, and includes contributions by Martine Marin and Véronique Larcher. 3D loudspeaker reproduction modules include contributions by Philippe Dérogis. Musical/graphical user interfaces were developed in collaboration with Georges Bloch, Gerhard Eckel, Tom Mays, Gilbert

Nouno, Laurent Cerveau and Olivier Warusfel. Spatialisateur technology is covered by issued and pending international patents.

References

1. Bauck JL, Cooper DH (1992) Generalized transaural stereo. Proc. 93rd Convention Audio Eng Soc, 1992, preprint 3401
2. Begault D (1994) 3-D Sound for virtual reality and multimedia. Academic Press, London
3. Berkhout AJ, Vries D de, Vogel P (1993) Acoustic control by wave field synthesis. J Acoust Soc Am 93: 2764–2778
4. Blauert J (1983) Spatial Hearing: the Psychophysics of Human Sound Localization. MIT Press, Cambridge, Mass.
5. Blauert J, Lehnert H (1995) Binaural technology and virtual reality. In: Proc. 2nd International Conf. on Acoustics and Musical Research, 1995, Ferrara, Italy, pp 3–10
6. Bloch G, Assayag G, Warusfel O, Jullien J-P (1992) Spatializer: from room acoustics to virtual acoustics. In: Proc. International Computer Music Conference, 1992
7. Chowning J (1971) The simulation of moving sound sources. J Audio Eng Soc 19(1): 2–6
8. Cohen M, Wenzel E (1995) The design of multidimensional sound interfaces. Technical Report 95-1-004. University of Aizu, Japan
9. Cooper DH, Bauck JL (1989) Prospects for transaural recording. J Audio Eng Soc 37(1/2): 3–19
10. Dechelle F, De Cecco M (1995) The IRCAM real-time platform and applications. In: Proc.1995 International Computer Music Conference
11. Farrah K (1979) The Soundfield microphone. Wireless World 85(99): 48–50
12. Foster S, Wenzel EM, Taylor RM (1991) Real-time synthesis of complex acoustic environments. In: Proc.1991 IEEE Workshop on Applications of Digital Signal Processing to Audio and Acoustics. IEEE CS Press, Piscataway, N.J.
13. Gardner WG (1994) Efficient convolution without input-output delay. J Audio Eng Soc 43(3): 127–136
14. Gardner WG (1997) 3-D Audio Using Loudspeakers. Ph.D. Thesis. Massachussets Institute of Technology, Media Lab, Cambridge, Mass
15. Gardner WG (1998) Reverberation algofithms. In: Kahrs M (ed) Applications of Signal Processing to Audio and Acoustics, 1998. Kluwer Academic Publishers, Dordrecht
16. Gerzon MA (1985) Ambisonics in multichannel broadcasting and video. J Audio Eng Soc 33(11): 859–871
17. Gerzon MA (1992) Psychoacoustic decoders for multispeaker stereo and surround sound. In: Proc. 93rd Convention Audio Eng. Soc, 1992, preprint 3406
18. Griesinger D (1989) Practical processors and programs for digital reverberation. In: Proc. 7th Audio Eng. Soc. International Conference, 1989, pp 187–195
19. Jot J-M (1992) Etude et réalisation d'un spatialisateur de sons par modèles physiques et perceptifs. Doctoral dissertation. Télécom, Paris, France
20. Jot J-M (1997) Efficient models for reverberation and distance rendering in computer music and virtual audio reality. In: Proc. 1997 International Computer Music Conference, Thessaloniki, Greece, pp 236–243
21. Jot J-M, Chaigne A (1991) Digital delay networks for designing artificial reverberators. In: Proc. 90th Convention Audio Eng. Soc, 1991, preprint 3030
22. Jot J-M, Warusfel O, Kahle E, Mein M (1993) Binaural concert hall simulation in real time. In: Proc. 1993 IEEE Workshop on Applications of Digital Signal Processing to Audio and Acoustics, New Paltz, NY
23. Jot J-M, Larcher V, Warusfel O (1995a) Digital signal processing issues in the context of binaural and transaural stereophony. Proc. 98th Convention Audio Eng. Soc, 1995, Paris, France, preprint 3980
24. Jot J-M, Jullien J-P, Warusfel O (1995b) Method for simulating room acoustical quality and associated digital audio processor. French Patent No. 95 10111. US patent No. 5,812,674
25. Jot J-M, Cerveau L, Warusfel O (1997) Analysis and synthesis of room reverberation based statistical time-frequency model. In: Proc. 103rd Convention of the Audio Eng. Soc, 1997, preprint 4629
26. Jullien J-P (1995) Structured model for the representation and the control of room acoustical quality. In: Proc. 15th International Conf. on Acoustics, 1995, pp 517–520
27. Jullien J-P, Kahle E, Winsberg S, Warusfel O (1992) Some results on the obiective and perceptual characterization of room acoustical quality in both laboratory and real environments. Proc Inst Acoust XIV(2)
28. Jullien J-P, Warusfel O (1994) Technologies et perception auditive de l'espace. Cahiers de l'IRCAM, vol. 5 (L'espace)
29. Kendall G, Martens W, Freed D, Ludwig D, Karstens R (1986) Image-model reverberation from recirculating delays. In: Proc. 81st Convention Audio Eng. Soc, 1986, preprint 2408
30. Lavandier C (1989) Validation perceptive d'un modèle objectif de la qualité acoustique des salles. Doctoral dissertation. Université du Maine, Le Mans, France
31. Malham DG (1990) Ambisonics – a technique for low-cost, high-precision, three-dimensional sound diffusion. In: Proc. 1990 International Computer Music Conference, Glasgow, pp 118–120
32. Malham DG (1993) 3-D sound for virtual reality using ambisonic techniques. In: Proc. 3rd Annual Conf. on Virtual Reality, 1993, London
33. Malham DG, Myatt A (1995) 3-D sound spatialization using ambisonic techniques. Comput Music J 19(4): 58–70
34. Marin M (1996) Etude de la Localisation en Restitution du Son pour la Téléconference de Haute Qualité. Doctoral dissertation. Université du Maine, Le Mans, France
35. Moller H (1992) Fundamentals of binaural technology. Appl Acoust 36: 171–217
36. Moore FR (1983) A general model for spatial processing of sounds. Comput Music J 7(3): 6–15
37. Moorer JA (1979) About this reverberation business. Comput Music J 3(2): 13–18
38. Nelson PA, Orduña-Bustamante F, Hamada H (1996) Multichannel signal processing techniques in the reproduction of sound. J Audio Eng Soc 44(11): 973–989
39. Persterer A (1989) A very high performance digital audio processing system. In: (eds) Proc. 13th International Conf. on Acoustics, 1989, Belgrade
40. Puckette M (1991) Combining event and signal processing in the Max graphical programming environment. Comput Music J 15(3)
41. Reilly A, McGrath D (1995) Convolution processing for realistic reverberation. In: Proc. 98th Convention Audio Eng. Soc, 1995, preprint 3977
42. Schroeder MR (1962) Natural-sounding artificial reverberation. J Audio Eng Soc 10(3): 219–223
43. Schroeder MR (1973) Computer models for concert hall acoustics. Am J Phys 41: 461–471
44. Stautner J, Puckette M (1982) Designing multi-channel reverberators. Comput Music J 6(1): 52–65
45. Theile G, Plenge G (1977) Localization of lateral phantom sources. J Audio Eng Soc 25(4): 196–200
46. Thiele G (1993) The new sound format 3/2-stereo. In: Proc. 94th Convention Audio Eng. Soc, 1993, preprint 3550a
47. Travis C (1996) A virtual-reality perspective on headphone audio. In: Proc. 101st Conv. Audio Eng. Soc., Preprint 4354
48. Warusfel O (1990) Etude des paramètres liés a la prise de son pour les applications d'acoustique virtuelle. In: Proc. 1st French Congress on Acoustics, 1990, Lyon, France, pp 877–880
49. Wenzel EM, Arruda M, Kistler DJ, Whightman FL (1993) Localization using nonindividualized head-related transfer functions. J Acoust Soc Am 94: 111–123
50. Wenzel EM, Whightman FL, Foster SH (1988) A virtual display system for conveying three- dimensional acoustic information. In: Proc. Human Factors Society 32nd Annual Meeting, 1988, pp 86–90

JEAN-MARC JOT was born in Saint Dizier, France, in 1965. He graduated from Ecole Nationale Supérieure des Techniques Avancées, Paris, in 1987, and received his Ph. D. from Ecole Nationale Supérieure des Télécommunications (ENST), Paris, in 1992. From 1988 to 1992 he worked with Studer-Digitec (Chatou, France) as a DSP research engineer within the digital mixing console project. During this period, he carried out research in the Acoustics group at ENST, under the supervision of Prof. Antoine Chaigne, on real-time DSP algorithms for professional audio applications, particularly artificial reverberation and spatial sound processing. Since 1992, he has been a researcher in the Room Acoustics team at IRCAM, Paris, heading the R & D project "Spatialisateur". His research interests include digital audio signal-processing algorithms, spatial/binaural recording and processing of sounds, artificial reverberation, models and measurement techniques in room acoustics.