Perceptual Spatial Audio Recording, Simulation, and Rendering

Hüseyin Hacıhabiboğlu, Senior Member, IEEE, Enzo De Sena, Member, IEEE, Zoran Cvetković, Senior Member, IEEE, James Johnston Fellow, IEEE, and Julius O. Smith Member, IEEE

Abstract—Developments in immersive audio technologies have been evolving in two directions: physically-motivated and perceptually-motivated systems. Physically-motivated techniques aim to reproduce a physically accurate approximation of desired sound fields by employing a very high equipment load and sophisticated computationally intensive algorithms. Perceptually-motivated techniques, on the other hand, aim to render only the perceptually relevant aspects of the sound scene by means of modest computational and equipment load. This article presents an overview of perceptually motivated techniques, with a focus on multichannel audio recording and reproduction, audio source and reflection culling, and artificial reverberators.

I. INTRODUCTION

Since Blumlein introduced the original concept of stereophonic recording using a pair of “figure-of-eight” microphones, spatial sound technologies have steadily grown in sophistication, complexity and capabilities. Delivering a convincing illusion of a desired sound field requires finding solutions to several problems lying at the intersection of physics, psychoacoustics, and engineering. First, the relevant sound field information needs to be identified, and methods for its acquisition devised, which amounts to designing an array of microphones. Then, methods for rendering the identified spatial audio information in some optimal way need to be developed. This requires the design of a playback system, including hardware configuration, along with necessary signal processing algorithms. If the spatial sound field is virtual, as opposed to generated by an actual acoustic event, the required playback signals need to be synthesized rather than recorded. To that end, ideally an accurate approximation of the desired sound field would be computed and then “recorded” using a virtual microphone array to be played via the corresponding actual loudspeaker array; this process is referred to as auralization. However, due to the very high numerical complexity of sound field simulation methods, typically the auditory perspective is first rendered via level differences between pairs of loudspeakers, and then overlaid by room effects. The past nine decades of spatial audio reproduction and synthesis has seen innovations and developments in all of these directions.

Generating the experience of a spatial sound scene can be achieved in a number of ways. Comparing different methods, at one extreme there are binaural techniques [1], which provide a convincing experience over two channels, by presenting stereophonic audio cues, that is, interaural time, level and spectral differences, referred to as “ear signals”. Binaural presentations work best over headphones, however, with cross-talk cancellation [2], they can be successfully used also with a pair of loudspeakers, although the effect is confined to a very narrow listening area. For a listener who is not static, the auditory illusion can be maintained via head tracking mechanisms combined with the real-time adaptation of the binaural signals. The advent of virtual and augmented reality systems has recently revived interest in binaural systems. However, some inherent problems of binaural audio such as individualization remain [3], limiting the spatial quality of the auditory experience they provide.

At the other extreme there are systems that aim to reconstruct an accurate physical approximation of a sound field. Notable examples include wave field synthesis (WFS) [4] and higher-order Ambisonics (HOA) [5]. WFS is based on Huygens principle and Kirchhoff-Helmholtz integral, which together state that the sound field due to a primary source can be exactly synthesized by infinitely many secondary sources on the surface.
enclosing a reproduction volume. Such a system can achieve a spatially extensive listening area and can be used in large auditoria such as film theaters. Ambisonics is based on sound field approximation using its spherical harmonics at the center of the listening area. Higher-order Ambisonics (HOA) is capable of achieving results comparable to wave field synthesis close to the center of the reproduction rig. While both WFS and HOA provide elegant solutions to the spatial recording and reproduction problem, they have high equipment load requirements, which can reach several hundreds of carefully positioned loudspeakers. For this reason, their application domain has so far been confined to specialist high-end systems. WFS and HOA can also run on systems with a more practical equipment load by including perception-inspired corrections. Comprehensive reviews of WFS and HOA have recently been published [6], [7].

In between these two extremes are systems with five to ten channels which are suitable for use in small to medium size listening rooms. Such systems do not possess a sufficient number of channels to physically reconstruct a sound field in a wide listening area, neither are they capable of accurately reconstructing the ear signals for listeners in multiple locations. Therefore, they must rely to a large degree on perceptual effects, similar to those used for binaural systems, to generate the illusion of a desired sound field within not overly confined areas.

As with recording and reproduction technologies, there is a variety of techniques for sound field simulation. At one extreme there are physically-motivated methods which aim to calculate an approximate solution of the wave equation. For that purpose several numerical methods have been developed that achieve a very high level of accuracy. However, they typically have prohibitively high computational costs. Examples include finite-difference time domain (FDTD), finite element method (FEM), and boundary element method (BEM) [8]. While these methods lend themselves to parallelization, the associated computational cost is still too high for real-time operation at interactive rates and on low-cost devices.

At the other extreme there are methods which try to render only some higher level perceptual effects. These methods, called artificial reverberators, require only a fraction of the computational load associated with physically-motivated room simulators and typically aim to mimic only certain characteristics of the tail of typical room impulse responses, such as modal density, echo density, and timbral quality [9]. They do not model explicitly a given physical space, but rather are used to obtain a pleasing reverberant effect and have been widely used for artistic purposes in music production.

In between these two extremes are methods that aim to render a certain physical sound scene, but only model its most perceptually relevant aspects. Full-blown room auralization systems typically aim to render each and every reflection and diffraction up to a given order for each source [10], [11]. More recent methods achieve remarkable computational savings by rendering accurately only first order reflections, while replacing higher order reflections with their progressively coarser approximations [12]. Further computational savings are possible by eliminating sources whenever they are inaudible, a process referred to as audio source culling.

This article is concerned with spatial audio systems and methodologies which substantially rely on psychoacoustics. We will first present a concise summary of spatial auditory perception, followed by a brief history of audio reproduction methods which rely on human auditory perception. More specifically, an overview of binaural audio, stereophony and multichannel audio systems will be given. Then, perceptually-motivated multichannel audio reproduction systems, such as vector-based amplitude panning (VBAP), directional audio coding (DirAC), perceptual sound field reconstruction (PSR), and their extensions will be reviewed. Finally, application of perceptual knowledge in the contexts of artificial reverberation, audio source culling and room auralization will be discussed.

II. SPATIAL AUDITORY PERCEPTION

The primary mechanism which humans use to localize sound sources in the horizontal plane is based on the differences between the signals received by the two ears. Due to the spatial separation between the ears, the sound wave generated by a sound source reaches the two ears with a different delay, called interaural time difference (ITD). Moreover, the sound wave is scattered by the head causing the level of the signal at the ear further away from the source, contralateral ear, to be reduced in comparison with the level of the signal at the ear closer to the source, ipsilateral ear. This level difference is called interaural level difference (ILD).

The interaural time delay for a typical human head can vary between ±750 µs in the acoustic free field. Humans can detect ITDs as low as 10-20 µs at the front direction corresponding to about 1° in the horizontal plane. Similarly, the ILD is frequency dependent and can be as high as 21 dB at 10 kHz. Sensitivity to changes

1Psychophysics of spatial hearing has been an active research area for the past century. Most of the information given in this section has been thoroughly reviewed by Blauert in [13]. Interested reader is referred to this excellent volume for more information and an extensive set of further references.
in ILD is also frequency-dependent, and for instance for pure tones it varies between 0.5 and 2.5 dB. In contrast with ITD which is the primary localization cue at low frequencies, ILD cues are more important in sound source localization at higher frequencies. This is due to the low level of scattering at low frequencies when the wavelength is close to or larger than the size of the head. ITD and ILD cues also change with the distance of a sound source and the size of the head.

Note that ITD and ILD pairs do not uniquely specify the source direction. If, for the purpose of illustration, we assume a spherical head, binaural cues will be identical for sound sources placed on conic shaped surfaces at each side of the head. These surfaces are called cones of confusion. In the horizontal plane, sources on the conic section which is the intersection of the horizontal plane with the cone of confusion will have front-back ambiguity. Humans can typically resolve this ambiguity by small head movements.

The elevation of a sound source is perceived based primarily on the spectral shaping of its signal which occurs as a result of the scattering of the sound around the head. This spectral shaping depends on the elevation in a manner which is determined by the sizes and shapes of the pinnae, head and torso. Consequently, the frequency content of the sound itself also affects the perception of the elevation of its source.

Subjective localization of sound sources involves a significant level of uncertainty. Localization blur is the smallest change in the direction of a source that will result in its perceived direction to change. For sources in the horizontal plane, localization blur is generally lower than around 10°. For sources in the median plane, localization blur in the order of 20° can be observed. A related concept is locatedness, which refers to the perception of the spatial extent of a sound source. This is an important attribute because the center of mass of a sound source can be localized accurately yet the source can still be diffusely located. Two other measures of spatial resolution of hearing are the minimum audible angle (MAA) and minimum audible movement angle (MAMA). MAA correspond to the minimum change in the direction of a static source in order for a listener to discriminate it as being to the left or to the right of the original direction. MAMA, on the other hand, is a measure of spatial resolution for moving sources, which quantifies the smallest arc that a moving sound source must travel to be discriminable from a stationary source [14].

Perception of the distance of a sound source is both less reliable and less well-understood than the perception of the direction of a sound source. Several cues affect the perception of distance. Among these, intensity is the only cue which is inherently related to the sound source and also the only absolute cue. The other distance cues are related either to the environment (direct-to-reverberant energy ratio, lateral reflections), to the physical properties of the listener (e.g. auditory parallax) or to cognitive aspects (e.g. familiarity) [15].

An interesting property of distance perception is the overestimation and underestimation of distance at different ranges and for different sounds. Apparent distances of sources far away from a listener are underestimated and those closer than around 1-2 m are overestimated [16]. Familiarity, which is a cognitive cue related to prior exposure to and knowledge of the characteristics of the sound source, also has a similar effect. For example, distance of whispered speech is underestimated while that of shouted speech is overestimated.

An important capability of the human auditory perception mechanism lies in its ability to localize sources in reverberant environments such as rooms and other enclosed spaces. This is made possible by suppressing reflections that come immediately after the direct sound. When a broadband impulse and a delayed copy of it are presented from different directions with a short delay of less than 1 ms in between, a single auditory event is perceived at a direction between the directions of the two sources, gradually shifting towards the leading source as the lag in the time of arrival increases. This effect is called “summing localization” and both sources contribute to the perceived direction of the auditory event. When the delay is between 1 ms and 5 ms, a single fused auditory event close to the leading source can be heard. Within this delay range, the presence of the lagging source is audible since it changes the timbre of the auditory event, however its direction cannot be discriminated easily. Above 5 ms, the broadband click and its echo are perceived as distinct sound events. The time delay above which two distinct events are heard is called the “echo threshold”. While the classic demonstration of these effects involve broadband click pairs, different signals will have different echo thresholds. For example, the echo threshold can be as high as 20 ms for speech and music signals.

The effect that the direction of the auditory event depends predominantly on the leading source is called localization dominance, whereas the effect that a single auditory event is perceived when there are two sound events is called fusion. The effect that the discrimination of the direction of the lagging sound source is sup-

---

2 An auditory event is defined as an event perceived by a listener typically (but not necessarily) in response to a sound event.
pressed is called lag discrimination suppression. These three effects are collectively known as the precedence effect [17].

Another important binaural cue is interaural coherence (IC) which is a measure of the coherence of signals received by the two ears. IC is high for sounds coming directly from the source where the two ear signals are highly correlated and low in the diffuse sound field where the correlation is low. Therefore, IC provides the information about the level of reverberation and thus about the spaciousness of the environment.

III. HISTORY OF PERCEPTUALLY MOTIVATED SPATIAL AUDIO

Binaural audio and multichannel stereophony are two of the most common spatial audio technologies predating more recent technologies such as WFS by more than half a century. Binaural audio has found extensive use and has received renewed interest especially for virtual reality applications, while multichannel systems have been the de facto standard for home entertainment and automotive audio systems. Simultaneously, due to the popularity and market dominance of two-channel audio formats, stereophony which uses two loudspeakers is still commonly used.

A. Binaural audio

Binaural audio\(^3\) is based on a simple assumption: if the signals that would be received at the ears of a listener as a result of an acoustic event, are provided to the listener with sufficient accuracy, he or she will feel perceive an auditory event which would correspond to the original acoustic event. These ear signals can be either recorded with microphones implanted in the ear canals of an artificial human head, such as KEMAR or Neumann KU-100, or synthesized using signal processing methods. In both cases, the signals are usually presented over a pair of headphones.

The microphones used for recording binaural audio are also known as “dummy head” microphones and are manufactured to resemble a typical human head. The external ears of these microphones are typically modeled after external ears of humans who have exceptional spatial hearing acuity and molded in silicon. The recorded signals need to be played back by using headphones equalized appropriately using free-field or diffuse field equalization, depending on the type of the environment in which the recording was made [1].

Binaural synthesis is based on the knowledge of the acoustic transfer paths between the source and the two ears. These paths are characterized by their impulse responses, referred to as the head-related impulse response (HRIR), or head-related transfer function (HRTF) in the frequency domain; for each source position there will be two of them, one for the left and one for the right ear. When HRIRs are convolved with dry source signals, the resulting signals will incorporate the necessary binaural cues for the given source position. In the case of a sound field created by \(P\) sources in the far field, right and left ear signals can be synthesized as:

\[
x_L(n) = \sum_{p=1}^{P} x_p(n) \ast h_{L,\theta_p,\phi_p}(n),
\]

\[
x_R(n) = \sum_{p=1}^{P} x_p(n) \ast h_{R,\theta_p,\phi_p}(n).
\]

where \(x_p(n)\) is the pressure signal due to source \(p\), \(h_{L,\theta_p,\phi_p}(n)\) and \(h_{R,\theta_p,\phi_p}(n)\) represent the HRIRs for the left and the right ears for a source at a direction \((\theta, \phi)\) where \(\theta\) and \(\phi\) are the azimuth and elevation angles, respectively, and \(\ast\) denotes convolution. This approach assumes that the acoustical system consisting of these sources and the listener is linear and time-invariant and that the resulting left and right ear signals provide the necessary spatial hearing cues pertaining to the acoustic field that would be generated by these \(P\) sources.

In free-field, HRIRs can be considered to be finite and are typically up to 12 ms long, corresponding to approximately around 512 samples at the 44.1 kHz sampling rate. This does not present a significant computational cost for a single component. However, as the number of components increases, such as when a source and its reflections in a room are being rendered, the computational cost of convolution becomes an important bottleneck. In order to overcome this limitation, different filter design approaches have been proposed (e.g. \([18]\)). These filters are designed to capture salient binaural cues while reducing the computational cost significantly.

Two essential requirements of binaural synthesis are the availability of a set of HRIR measurements densely sampled on a spherical shell and the match between these HRIRs and the actual HRIRs of the listener. Regarding the first requirement, interpolation methods such as kernel regression \([19]\) can be used in order to increase the granularity of the available directions. The second requirement necessitates the measurement
of individualized HRIRs which is both time-consuming and costly. For that reason, many existing research-grade and commercial solutions use generic HRIRs. This, however, is not an ideal solution since there are significant differences between the spectra of the generic HRIRs and individual HRIRs of the listener and these cues are essential for elevation perception [13]. Practical setups that allow quick measurements of HRIRs around a geodesic sphere around the listener’s head have recently been developed [20]. There also exist commercial products which allow tailoring a stored set of HRIRs based on the head size 4. However, head size alone can improve only the ITD and ILD cues provided by the system, but not the spectral cues used in the perception of source elevation.

Binaural synthesis also allows interactivity if the position and orientation of the listener’s head can be tracked [21]. High-precision and high-accuracy magnetic trackers had been the de facto method for tracking a listener’s head. Recent developments made it possible to track a user’s head with inexpensive devices 5. These developments make binaural synthesis an excellent solution for virtual reality applications.

For binaural synthesis, a side effect of system errors, such as a pair of improperly equalized headphones, or an HRIR set which does not match well the HRIRs of the user, or inaccurate head tracking, is inside-the-head localization [22]. This undesirable effect can be partly alleviated by adding simulated reflections and artificial reverberation.

Binaural audio can also be presented via a pair of loudspeakers, however each ear then receives not only the signal intended for it, but also the signal intended for the other ear, which impairs the coherence of binaural cues. This effect is known as cross-talk [23]. There are methods for cross-talk cancellation based on predicting the response of the cross-talk path and inverting it [24]. Such methods pre-process the left and right channels using a $2 \times 2$ cross-talk cancellation filter matrix, which is obtained as the inverse of the matrix containing the direct and cross-talk acoustic transfer paths. When a listener is sitting still at the position for which cross-talk cancellation is made, a single set of cross-talk cancellation filters can be very effective. However even small head movements require filter adaptation which increases the computational overhead associated with cross-talk cancellation.

Cross-talk canceled binaural audio, also known as transaural audio, has distinct benefits in comparison with two-channel stereophony and headphone-based binaural presentation: 1) it has the capability to simulate sources behind the listener even when there is no corresponding physical source (i.e. loudspeaker), and 2) it provides a better externalization of the simulated sources due to the presentation being made over loudspeakers [25].

B. Two-channel Stereophony

Two-channel stereophony is an alternative spatial audio technology which requires the minimal number of channels to produce the impression of spatial sound. In the usual implementation, two-channel stereophony uses two loudspeakers, at the same distance from the listener, positioned $30^\circ$ to either side of the front direction, to provide a frontal auditory scene within a base angle of $60^\circ$. The ideal listening position, referred to as the sweet spot, thus forms an equilateral triangle with the loudspeakers. Two-channel stereophony creates the illusion of a sound source in a given direction within the base angle by means of interchannel time differences (ICTD) and interchannel level differences (ICLD) of the two channels over which the source signal is presented. Fig. 1a shows the standard stereophonic setup and illustrated how the gains and delays of each channel are linked to ICTD and ICLD.

Although it is intuitively clear that the direction of the virtual source is pulled towards the loudspeaker which produces the louder and earlier version of the signal, knowing the precise relationship between the perceived source direction and presented (ICTD, ICLD) pairs requires extensive psychoacoustic measurements.

The first comprehensive study of the relationships between ICTD and ICLD, referred to as stereophonic panning laws, was conducted by Franssen [26]. Another study by Williams [27] combined earlier studies on ICTD and ICLD, and the panning curves presented in that study are now known as Williams’ curves. Williams’ psychoacoustic curves are illustrated in Fig. 1b, which shows curves of (ICTD, ICLD) pairs that create a virtual source in the direction of the left loudspeaker (the blue curve) and the right loudspeaker (the orange curve). Pairs of (ICTD, ICLD) that are below or above the two curves are also localized at the left and right loudspeaker, respectively. Virtual sources in directions between the loudspeakers are then created by means of (ICTD, ICLD) pairs which evolve along a line that connects two points on the psychoacoustic curves.

Note that there are many different (ICTD, ICLD) pairs that can create a virtual source in the same direction. Intensity stereophony is achieved when the ICTD is zero.

---

4 Available online: https://www.ossic.com/
5 Available online: http://www.3dsoundlabs.com
ICLD(dB) = 20 \log_{10} \left( \frac{g_L}{g_R} \right)

ICTD(ms) = d_R - d_L

\begin{align*}
\Gamma_L(\theta) &= (1 - \alpha_L) + \alpha_L \cos(\theta - \theta_L) \\
\Gamma_R(\theta) &= (1 - \alpha_R) + \alpha_R \cos(\theta - \theta_R)
\end{align*}

where \( \Gamma_L(\theta) \) and \( \Gamma_R(\theta) \) are the directivity patterns which represent the directional sensitivity of the left and the right microphones, respectively, \( \theta \) is the angle defined counter-clockwise from the acoustic axis of the corresponding microphone, and \( \theta_L \) and \( \theta_R \) are the rotation angles of the left and the right microphones. Designing stereophonic microphone pairs then requires optimizing the ICTD and ICLD by a careful selection of i) \( \alpha_L \) and \( \alpha_R \), ii) \( \theta_L \) and \( \theta_R \), and iii) the distance \( D \) between the two microphones.

One of the first stereophonic recording microphone pairs was developed by Alan Blumlein, and consisted of two coincident bidirectional microphones (i.e. \( \alpha_L = \alpha_R = 1 \)) positioned at right angles with each other. Many different microphone configurations have been devised since then. These methods can roughly be categorized into three groups: coincident, near-coincident and spaced [29]. Coincident pairs have two co-located \( (D = 0) \) directional microphones, resulting in the recorded left and right channel signals that have only amplitude differences. Examples of coincident microphone pairs are the Blumlein pair, the XY stereo pair, and the M/S pair [29]. Spaced arrays, such as the AB pair [29], typically use omnidirectional microphones (\( \alpha_L = \alpha_R = 0 \)) with a separation \( D \) that is many multiples of the
functions are used to pan the sound source. Near-incident recording techniques on the other hand use direction microphones separated by a small distance comparable to the size of a human head and record both ICTD and ICLD. Two notable examples are the Nederlandse Omroep Stichting (NOS) and Office de Radiodiffusion Télévision Française (ORTF) pairs which both use cardioid \((\alpha_L = \alpha_R = 0.5)\) microphones and have separations of \(D_{ORTF} = 17 \text{ cm}\) and \(D_{NOS} = 30 \text{ cm}\), respectively [29].

Synthetic stereophony has been predominantly based on intensity panning, since it is considered to provide the most stable virtual sound imaging. Indeed, the inclusion of ICTDs is sometimes considered to yield audible artifacts, such as tonal coloration due to comb filter effects. Another often cited reason to avoid using ICTDs is the difficulty of controlling the direction of a virtual source by means of time delays. This view has been recently challenged [30], as it will be discussed in Section IV-C.

The general form of an intensity panning law relates the gains \(g_L\) and \(g_R\) of the left and right loudspeakers, respectively, to a function of the source direction \(\theta_s\) and the stereophonic base angle \(\theta_B\), between the loudspeakers. More specifically, a panning law has the form:

\[
\frac{g_L(\theta_s) - g_R(\theta_s)}{g_L(\theta_s) + g_R(\theta_s)} = \frac{f(\theta_s)}{f(\theta_B)}.
\] (5)

The total power can be maintained via the constant power constraint \(g_L(\theta_s)^2 + g_R(\theta_s)^2 = 1\). Two commonly used functions are \(f(\theta) = \sin(\theta)\) and \(f(\theta) = \tan(\theta)\) which give rise to so called the sine panning law and tangent panning law, respectively.

The tangent panning law has been derived based on perceptual considerations independent from known psychoacoustic curves [31]. In the context of Williams’ psychoacoustic curves (see Fig. 1b), the tangent panning law operates along the vertical axis, i.e. zero ICTDs, and connects two points with \(\pm \infty\) level differences. Thus, as opposed to panning laws described by Williams’ curves, which specify minimal level differences needed to create virtual sources in loudspeaker directions, the tangent law achieves the same affect by employing maximal level differences.

C. Multichannel Stereophony

An early work by Steinberg and Snow [32] in 1934 suggested that better auditory perspective is possible if at least three independent microphones are used to capture a frontal sound field and these signals are played back via three loudspeakers. Due to the hardware requirements and technical difficulties in the integration of a three-channel system in the radio broadcast, however, this finding has been obscured by the success and widespread adoption of two-channel stereophony.

The advent of quadrophony and cinematic sound spurred interest in multichannel systems. Traditionally there are two different types of multichannel audio formats: discrete and matrix [33][34]. In discrete multichannel audio, there is one-to-one correspondence between channels and speakers. The storage and transmission of multichannel audio are all made using the same number of channels. In matrix multichannel, the original channels are encoded to a smaller number of (e.g. two) channels for transmission/storage over common channels/media and then decoded back to the original channel multiplicity prior to playback. This requires appending auxiliary information to the encoded audio to be used at the decoding stage. More recently, object-based formats have appeared where content and context are encoded separately.

Surround sound is the more commonly known name for multichannel stereophony. There exist several reproduction setups such as 5.1, 7.1, 10.2, and 22.2, which use 5, 7, 10, and 22 main channels, respectively and 1 or 2 low-frequency channels, as described in an ITU report (ITU-R BS.2159-4). There are also commercial, object-based formats such as Dolby ATMOS\(^6\), DTS-X\(^7\) and Auro-3D\(^8\) which are very flexible and are likely to dominate the cinematic sound industry in the foreseeable future considering the new ISO/IEC standards such as MPEG-D and MPEG-H.

Commercial microphone arrays for multichannel recording exist, but these arrays are based more on practice in the field than on a solid theory and understanding of the underlying acoustic processes. The microphone arrays used for recording 5.1 multichannel audio typically include cardioid, supercardioid or hypercardioid microphones positioned on a tree structure [35], [36]. These arrays can in general be separated into two groups: i) five-channel main microphone techniques, and ii) front-rear separation techniques. The former uses five closely positioned microphones which are mapped directly to the five main channels of a 5.1 reproduction setup. The latter uses two separate arrays to record direct field and ambience separately. For some arrays (such as INA-5 [37]), there is a one-to-one correspondence between the microphone and loudspeaker channels. For some other arrays (such as Soundfield Microphone [38], Fukada Tree [39] or Hamasaki Tree [40]) the signals

\(^7\)Available online: http://dts.com/dtsx
\(^8\)Available online: http://www.auro-3d.com
obtained from individual microphone channels need to be mixed.

Some well-known multichannel arrays used for recording multichannel audio are shown in Fig. 2. It may be observed that a variety of microphone arrangements exist that try to address the common objective of obtaining an authentic auditory perspective and a high level of envelopment and immersion using existing first-order microphone directivity patterns.

The microphone arrays for recording 10.2 multichannel stereophony are still rather experimental (see ITU-R BS.2159-4 report). Similarly, recording for a 22.2 reproduction system will depend strongly on the venue and context. In fact, multichannel stereophonic systems with higher channel counts, by virtue of the degrees of design freedom they provide, allow for more flexibility, but also make it more difficult to design recording setups with strict perceptual rationale.

Recommended reproduction setups for multichannel systems are either standardized (e.g. ITU-R BS.775-1) or in the process of standardization by different standardization bodies [41]. These setups mainly rely on the frontal channels for the presentation of audio content which accompany visual content (usually films or games). The left and right front channels typically correspond exactly to the two-channel stereophonic setup for cross- and backwards compatibility. The difference in these setups is mainly about how ambience is played back. Some of the standards like ITU-R BS.1116.1 and ITU-R BS.1534.1 define formal procedures for the subjective evaluation of these systems.

IV. PERCEPTUALLY-MOTIVATED MULTICHANNEL RECORDING AND REPRODUCTION

There has been some recent work in the direction of developing systematic frameworks for the design of multichannel stereo systems, most notably Vector-base Amplitude Panning (VBAP), Directional Audio Coding (DirAC), and Perceptual Sound Field Reconstruction (PSR). We review them in this section.

A. Vector-base Amplitude Panning (VBAP)

It was shown as early as 1973 that tangent panning provides a stereophonic image that is more robust to head rotations than sine panning for the standard stereophonic loudspeaker setup [31]. Pulkki showed that tangent panning can be expressed using an equivalent, vector-based formulation in the horizontal plane and also proposed a three-dimensional extension to two-channel intensity panning which allows rendering elevated virtual sources over flexible loudspeaker rigs [42]. This method is called vector-base amplitude panning (VBAP).

Originally, VBAP was designed for a loudspeaker array with elements placed on the vertices of a geodesic dome that are situated at the acoustic far field of the listener. Fig. 3 shows a section of such a sphere with three loudspeakers with a listener positioned at the center of the array. The directions of the three loudspeakers are indicated as $v_1$, $v_2$, and $v_3$, and the corresponding gains as $g_1$, $g_2$, and $g_3$. A virtual source in a direction $v_s$ between the loudspeakers can be generated by selecting the gains that satisfy $v_s = Vg$ where $V$ is a matrix whose columns are the directions of the loudspeakers and $g = [g_1 g_2 g_3]^T$. In addition, the calculated loudspeaker gains are normalized in order to keep the total power constant.

On the full geodesic sphere, active regions are selected based on the closest three points on the grid and only those loudspeakers are used for source rendition. This is in contrast with physically-based approaches such as Ambisonics where even for a single source from a single direction, all loudspeakers are potentially active. A major assumption behind VBAP in three dimensions is that summing localization would occur not only with two, but also with three sources. This assumption was subjectively tested for different setups and virtual source directions and it was shown to result in a good subjective localization accuracy for elevated virtual sources [43], [44].

An issue resulting from utilization of intensity panning in VBAP is the nonuniformity of the spatial spread of the panned source. More specifically, sources panned closer to the actual loudspeakers in the reproduction rig have a smaller spatial spread, while virtual sources panned to directions between loudspeakers have a larger spatial spread. The main cause of this issue is the usage of a single loudspeaker when the virtual source direction coincides with the direction of that loudspeaker.

This issue was addressed by panning the virtual source to multiple directions by using three loudspeakers (instead of two) for all source directions in the horizontal plane or four loudspeakers (instead of three) in the 3D case. This approach was called as multiple-speaker amplitude panning (MDAP) [45]. In a study comparing VBAP with MDAP it was shown that both VBAP and MDAP provide good subjective localization accuracy with MDAP being more accurate than VBAP [46]. In another, more recent evaluation carried out within the context of the MPEG-H standard, VBAP resulted in very good subjective localization accuracy including not only the source azimuth but also its distance [47]. In yet another study, VBAP was shown to provide good
localization performance also for sources in the median plane [48].

Note that VBAP is a technology for sound field synthesis, and in the context of sound field recording and reproduction it is used at the reproduction end of schemes such as Directional Audio Coding (DirAC).

### B. Spatial encoding methods

A class of multichannel audio methods involves dividing recorded signals into time or time-frequency bins and estimating certain spatial attributes within each bin. One of these methods is the spatial impulse response rendering (SIRR) method [49], [50]. At the recording stage, SIRR records the impulse response of a room using a B-format microphone, i.e., a microphone that provides the omnidirectional sound pressure component as well as the three axial pressure gradient components of the sound field [29]. The impulse response is first transformed into a time-frequency representation and then processed to obtain estimates of the acoustic intensity vectors at each time-frequency bin. It is assumed that each time-frequency bin corresponds to a single plane wave and thus that the direction of acoustic intensity vector also represents the direction of that plane wave. A diffuseness estimate is also obtained for each time-frequency bin using the ratio of the real part of acoustic intensity to the total energy. These parameters along with the sound pressure component obtained from the B-format recording form the basis for the reproduction stage.

At the reproduction stage, direct and diffuse parts of the signal are treated differently. For the direct part, azimuth and elevation estimations in each time-frequency bin are used to pan portions of the B-format omnidirectional component accordingly using VBAP. The diffuse part is reproduced by generating multiple decorrelated copies of the recorded sound played back from all loudspeakers. The so-obtained channel impulse responses are then convolved with the desired anechoic sound sample.

A similar method called spatial decomposition method (SDM) was recently proposed in [51]. Instead of using a time-frequency representation of the room impulse response, SDM simply divides it into time frames. Sim-
Fig. 3. Arrangement of three loudspeakers and a phantom image panned using VBAP. The vectors used in the formulation of VBAP are also shown.

Similarly to SIRR, SDM assumes that within each time frame there is at most a single acoustic event (e.g. a reflection from the room walls), the direction of which is calculated using available direction-of-arrival estimation algorithms. Using this estimate, each time frame of the impulse response is panned between the loudspeakers using VBAP. The loudspeaker signals are then convolved with the desired anechoic sound sample.

Notice that, as opposed to SIRR, SDM does not explicitly differentiate direct and diffuse components. However, the later part of the room impulse response is still rendered as diffuse. This is due to the fact that, as time progresses, a progressively larger number of echoes appear within each time frame, and, as a consequence, the direction-of-arrival algorithm tends to provide random estimates. In a formal listening experiment using synthesized room impulse responses, SDM was shown to outperform SIRR [51].

SIRR and SDM are not designed for continuous signals but for spatial room impulse responses, which are then convolved with an anechoic signal. In other words, they cannot be used for actual recordings of dynamic sound scenes. Directional Audio Coding (DirAC) is a flexible spatial audio system for recording, coding, compression, transmission and reproduction based on SIRR that overcomes this limitation [52]. Similarly to SIRR, DirAC starts with an energy analysis of the recorded sound, to assign a direction and a diffuseness level to each instant of the output channels of a filter bank that approximates the equivalent rectangular bandwidth (ERB) scale. The direction predictions are then smoothed to imitate the temporal resolution of the auditory system. At the reproduction stage, these components are panned using VBAP. Fig. 4 shows the recording, processing and reproduction stages of DirAC.

DirAC was evaluated and compared with Ambisonics (with different decoders) for reproduction quality using listening tests similar to MUSHRA [53]. The evaluation included different loudspeaker rigs (with 4, 5, 8, 12, and 16 loudspeakers), different audio material (music, speech, singing voice, percussion), different simulated reverberation characteristics, and different listener positions. It was found that DirAC provides an excellent reproduction quality (better than 80 on average over a maximum rating of 100) for the central listening position, and acceptable reproduction quality (better than 60 on average over a maximum rating of 100) for the off-center positions. Ambisonics reproductions obtained using both decoders were rated consistently below DirAC. These results provide an instructive example of a perceptually-motivated reproduction method achieving better subjective performance than a physically-motivated approach.

C. Perceptual Sound Field Reconstruction

Perceptual Sound Field Reconstruction (PSR) [30], [54] is a recently developed flexible multichannel recording, reproduction and synthesis technology. Similarly to DirAC, it provides a systematic framework for recording and reproduction of sound scenes. However in contrast to DirAC, it performs panning of individual time-frequency components and renders the diffuse sound field via all channels, relying on extensive processing of microphone array recordings to extract the necessary directional information and components, PSR relies on designing underlying microphone arrays in a way which captures the required directional cues. When the recorded signals are played back, with no additional processing, directions of wave fronts of all sound sources and all reflections are rendered accurately.

A block diagram of a five-channel PSR system with uniform distribution of channels is shown in Fig. 5. Another difference between DirAC and PSR is in that while DirAC uses only ICLDs for rendering auditory perspective, PSR employs both ICTDs and ICLDs, and allows for trading one for the other while designing the directivity patterns of the microphones used in the arrays.

PSR uses near-coincident circular microphone arrays to capture time differences between channels. The difference in the time of arrival of a sound wave propagating...
from a direction \( \theta \) between microphones at angles \( \phi_l \) and \( \phi_{l+1} \) (see Fig. 6) is

\[
\tau_l(\theta) = \frac{2r_a}{c} \sin \left( \frac{\phi_{l+1} - \phi_l}{2} \right) \sin \left( \frac{\phi_{l+1} + \phi_l}{2} - \theta \right),
\]

where \( r_a \) is the radius of the microphone array, and \( c \) is the speed of sound. Microphone directivity patterns \( \Gamma_l(\theta) \) are designed such that level differences with which the sound wave is recorded are equal to the level difference which in combination with the time differences, as given in the above, creates a perception of the sound source in the direction, \( \theta \). One way to achieve this is by designing \( \Gamma_l(\theta) \) to satisfy the following relationship:

\[
\frac{\Gamma_{l+1}(\theta)}{\Gamma_l(\theta)} = \frac{\sin (\theta - (\phi_l - \beta))}{\sin ((\phi_{l+1} + \beta) - \theta)},
\]

where \( \beta \) is selected in such a way that for \( \theta \) which coincides with the direction of one of the microphones, the level difference is equal to the level difference needed to create the perception of the sound wave in the direction of the corresponding loudspeaker. That level difference is labeled as EL (or FL with the reversed sign) in Fig. 1b for the case where maximal ICTD is 0.6 ms. Thus, as the direction of the sound source moves between the two microphones the captured time and level difference traverse a curve connecting two end points, illustrated by the straight line between points EL and FL in Fig. 1b, which correspond to virtual sources in the directions of two corresponding loudspeakers. Microphones are additionally required to satisfy the constant power condition and

\[
|\Gamma_l(\theta)|^2 + |\Gamma_{l+1}(\theta)|^2 = 1, \quad \phi_l \leq \theta \leq \phi_{l+1}
\]

sufficiently high attenuation outside the sector between
the axes of the two adjacent microphones, so that every sound wave is effectively recorded and rendered only by the pair of two closest channels.

The degree with which time differences are present is controlled by the radius of the microphone array, which is present implicitly in the $\beta$ factor in (6). In the special case when $\beta = 0$ the system implements intensity stereophony based on the tangent panning law.

An example of a directivity pattern designed according to PSR principles for a five-channel uniformly spaced system, for an array of with $r_a = 15$ cm, is shown in Fig. 7, along with its 2nd-order approximation and the polar pattern which corresponds to $\beta = 0$, that is the pattern designed for intensity stereophony according to the tangent law.

The five channel PSR system design based on intensity and time-intensity principles, as specified above, was subjectively evaluated and compared with 2nd-order Ambisonics in terms of subjective localization accuracy [30]. Fig. 8 shows the results of a localization test carried out using different recording/reproduction systems. The time-intensity PSR technology performed well especially at off-center listening positions while it performed worse for localization at lateral source directions, which is due to the fact that psychoacoustic curves for frontal presentation were used for all pairs of loudspeakers. Another set of tests considered the locatedness of generated phantom sources, and showed that time-intensity based PSR provides better locatedness of phantom sources than techniques based on intensity only (shown in the bar charts in Fig. 8), which is attributed to the higher naturalness of the presented binaural cues [55].

The issue of intensity versus time-intensity techniques is a matter of debate among audio engineers and recording artists, and the widely held view is that although time-intensity stereophony provides more naturally sounding sources, intensity stereophony provides more stable imaging. This result provides a new insight into the issue and demonstrates that time-intensity techniques, if designed with a careful consideration of underlying psychoacoustical requirements, are capable of actually providing stable auditory perspective.

The development of techniques for higher-order differential microphone arrays (DMA) [56], [57] enabled design of more sophisticated directivity patterns than those achievable by commonly used first-order microphones. This allowed the implementation of different panning laws and psychoacoustical panning functions in the multichannel microphone array design process.

D. Enlarging the optimal listening area

When a listener moves away from the center of the sweet spot, the auditory event shifts in the direction of the closest loudspeaker. This is due to the fact that the signal from the closest loudspeaker arrives earlier when compared to what is observed at the sweet spot. Position-independent (PI) stereo [58], [59] aims to alleviate this problem by designing loudspeaker directivity patterns in a manner that compensates for the incongruent time delay via appropriate intensity differences.

The design method proposed for this purpose by Rodeñas et. al. [60] consists of two separate optimization procedures. The first procedure involves finding a common directivity pattern for the left and right loudspeakers of a
stereophonic setup in order to provide level differences needed to compensate incongruent time differences over a desired listening area. Such directivity patterns can be obtained by beamforming using an array of loudspeakers. The other optimization procedure involves finding the filter coefficients to be used for beamforming.

Fig. 9 illustrates the problem of off-center listening and loudspeaker directivity patterns implemented to counteract this problem using a loudspeaker array with two drivers. It may be observed that when the listener moves to the left of the ideal listening position, the signal from the left loudspeaker will arrive earlier and at a higher level than the right loudspeaker, shifting the perceived direction of the virtual source towards the left. This problem can be compensated for by adjusting the right loudspeaker to have a higher level than the left loudspeaker at the corresponding direction, effectively shifting the virtual source back. The loudspeaker directivity patterns shown in the figure were designed to achieve this and thus to allow the listening area to be enlarged. Rodenas et al. [60] report results of an informal listening test with a system realized using loudspeaker arrays with two tweeter and two mid-range drivers each and state that the proposed approach widens the sweet spot for standard stereophonic material.

While designing loudspeaker directivity patterns for robust stereophony is a promising idea, technical difficulties such as the equalization of drivers, compensation
of diffraction from the edges of the loudspeaker cabinets, and the required number of loudspeakers used in the design may limit its practical use. These techniques, along with the generalization of the design approach to multichannel systems, and combination with other technologies such as PSR are interesting directions for future research.

V. PERCEPTUALLY MOTIVATED ROOM AURALIZATION

In cases where an acoustic scene actually exists, like a live concert or a tennis match, the scene is recorded and reproduced by the techniques reviewed in the above. These techniques also capture acoustics of the environments in which these recordings are made. There also are applications where such scenes exist only virtually, for example in computer games or virtual reality (VR) applications. In such cases, acoustics of the environment which contain the scene to be rendered need to be synthesized. The process of making the acoustics of a real or virtual environment such as a room or a concert hall audible is referred to as auralization [8].

Rooms are multipath environments where the recording of a source by a microphone will include not only the direct path but also early reflections, reverberation tail, and diffraction components. Many different models have been proposed in the past fifty years to simulate room acoustics. A recent review article provide summary of research on room acoustics modeling [61] and divides algorithms into three classes: i) convolutional algorithms, ii) delay networks, and iii) computational acoustics models. Convolutional algorithms involve measuring the impulse response of an actual room and convolving it with a desired input signal. Delay networks, which will be discussed in more detail in Sec. V-B, are algorithms where the input is filtered and fed back along a number of delay paths designed according to desired reverberation characteristics. Computational acoustics models aim at simulating the propagation of sound waves in the modeled space.

Among computational acoustics models there are geometric models, which use geometric arguments to calculate the room impulse response. These include the image-source method (ISM) [62], [63], ray tracing [64] or beam tracing [65] and its variants [66]. Other computational acoustics models such as finite-difference methods [67], digital waveguide mesh (DWM) [68], finite element methods (FEM) [69] and boundary element methods (BEM) [70] are based on the time and space discretized solutions of the wave equation, hence individual reflections are not rendered explicitly but their effects are merged into the overall simulated wave fields. Computational acoustics models are capable of providing very accurate results (at least for certain frequency ranges), and are therefore used in architectural acoustics. However, their physical accuracy comes at a very high computational cost. While some computation can be carried out offline, auralization will typically require real-time operation at interactive rates, for instance in order to allow a user to explore a virtual environment. The main computational bottleneck that this entails is associated with the different filtering operations involved in calculating and synthesizing reflections and edge diffraction components for each source.

Despite their high computational complexity, highly accurate room auralization will always be in demand for applications such as architectural acoustics. However, they are not suitable for applications such as immersive games and virtual reality, where a low computational cost is paramount. Such applications warrant the simplification of the model to the lowest possible number of components and sources to be rendered, which is typically achieved by removing perceptually irrelevant content.

A. Simplification of room acoustics models

The lack of a comprehensive mathematical model of the precedence effect, analogous to models of monaural
masking, has made it difficult for a long time to predict whether an individual reflection would be audible in the presence of the direct sound and other reflections. This is mainly due to the fact that the audibility of a reflection depends on many parameters.

One of the first models that aimed to parameterize the audibility of reflections, named Reflection Masked Threshold (RMT), was proposed by Buchholz et al. [71]. The RMT is the lowest level at which a reflection will be audible, and it is a function of the directions of the reflection and of the corresponding direct sound, the level delay of reflection with respect to the direct sound, the frequency spectra between the direct sound and the reflection, the effect of other reflections and reverberation, and the signal content. RMT can be used for simplifying room acoustic models via culling inaudible reflections.

A simpler decision rule for culling inaudible early reflections was proposed by Begault et al. [72]–[74] based on the relative level of the reflection. In the absence of reverberation, the audibility threshold of a reflection is 21 dB below the level of direct sound for a delay of 3 ms. The presence of diffuse reverberation has the effect of increasing this threshold by 11 dB. This threshold is also known to decrease with the angle between the direct sound and the early reflection.

Properties of binaural hearing, such as the precedence effect, may also make some reflections inaudible. The exclusion of those reflections from audio rendering pipeline can further reduce the associated computational cost. To that end, a model of the precedence effect was proposed in [75] according to which perceived directions of acoustic events are modeled as normally distributed variables. If the direct path and a reflection are present, then the distribution of the perceived direction is a mixture of two Gaussians. The audibility of the reflection was then shown to be related to the number of modes in the mixture: if the mixture is unimodal the reflection is masked and if it is bimodal it is audible. The derivation of the model parameters was made via subjective localization experiments. This model was applied for the culling of reflections in binaural room auralization [76]. More specifically, the image source method (ISM) was used to obtain a number of secondary sources and these were clustered according to their distance from the listener position and their azimuth angle. A single reflection masker was obtained for each cluster using the precedence effect model and the rest of the secondary sources in the same cluster are excluded from the rendering pipeline, thereby reducing the computational cost. Subjective evaluations were carried out using different audio material, different room geometries and different listening positions to compare the room auralizations using full room response, level-based reflection selection, and perceptually-motivated selection based on the precedence effect model. These experiments showed that reflection culling based on the precedence effect is capable of reducing the number of early reflections by over 60%, without any significant degradation on subjective localization, spaciousness, presence and envelopment experiences.

Another approach to perceptually-motivated simplification of auralsation based on absolute threshold of hearing was recently proposed [77]. According to this model, the duration of ray tracing for calculating the room impulse responses for a given source depends on a temporal cutoff point determined by the last audible ray. It was shown that this approach resulted in noticeable improvements in computation time of impulse responses without significantly degrading the auditory experience.

B. Perceptually-motivated artificial reverberation

Room impulse responses can be divided in two parts—early reflections, where reflections are separated in time and have strong directional characteristics, and the reverberation tail, where higher-order reflections begin to overlap in time and the sound field becomes diffuse. The human auditory system is sensitive to the direction of the direct wave front and the early reflections, while it cannot discern the directions of individual reflections within the reverberation tail [78]. Level and directions of lateral early reflections are related directly to the perception of the width of a sound source and the spatial impression of an enclosure [79].

As the density of reflections increases, the statistical properties like reflection density and decay slope become more important than the fine temporal structure. In real enclosures, sound energy decays exponentially, and the point at which the total energy of the room impulse response drops 60 dB below its initial value is called the reverberation time [80]. The reverberation time has a strong influence on how spacious an enclosure is perceived [78]. Other quantities that have a strong influence on the perceived quality of reverberation include, the density of the individual reflections in the late reverberation tail, called the reflection density [80], the time dependent profile of reflection density, called the echo density profile [81], and the number of damped resonant frequencies per Hz, called the mode density [82]. The typical objective of perceptually-motivated artificial reverberators is to render accurately the properties of reverberation described above.

Since the early part and the reverberation tail are perceived differently, a common approach is to model
and render them separately in a typical room auralization algorithm. For the reverberation tail, a statistically compatible model is usually acceptable, due to the fact that the human auditory system is not sensitive to its fine structure. Fig. 10 shows the diagram of a typical binaural auralization system. Here, one module simulates and renders binaurally the direct path and a number of early reflections, while an artificial reverberator unit renders the reverberation tail. In this context, we refer to an artificial reverberator as a room acoustic model (typically a delay network) that only aims at reconstructing important perceptual features of room reverberation with little regard to its physical accuracy. By only targeting the perceptual aspects of room reverberation, vast reductions in computational complexity are possible.

Various room auralization systems have been developed in the past twenty years [10], [83], [84]. The DIVA system [10], one of the first parametric interactive room auralization systems, simulates all the first- and second-order reflections and synthesizes them binaurally or for rendition over loudspeakers. It is capable of simulating the absorption characteristics of different wall materials, air absorption and source directivity. Late reverberation is provided via an artificial reverberator consisting of a recursive structure using comb and allpass filters.

The choice of artificial reverberation in a room auralization system is dictated not only by perceptual considerations but also by computational cost, and the holy grail in artificial reverberator design is an algorithm which can achieve good perceptual quality at a reasonable computational cost. The earliest digital artificial reverberators were proposed by Schroeder in the 1960s and consisted of comb filters connected in parallel to simulate the frequency modes of a room and all-pass filters to simulate a dense reverberation tail [85]. Original designs by Schroeder sometimes produced metallic sounding reverberation and various improvements were subsequently proposed [9], [86]. These improvements, however, did not provide means to explicitly or easily control the characteristics of the synthesized reverberation.

Feedback delay networks (FDNs) were developed as
a multichannel extension of the Schroeder reverberator [88], [89]. FDN is a recursive delay network which can generate reverberation for a number of input channels such as individual audio channels of a 4 channel (i.e. quadrophonic) system. Each of the input channels are delayed, fed back recursively through a feedback loop, attenuated, and mixed with the incoming channels. The delay lines are designed to have incommensurate lengths and the feedback loop consists of multiplication with a unitary matrix.

Jot and Chaigne extended the FDN design and proposed a simple and structured procedure to design good quality reverberators with a desired frequency-dependent reverberation time [87]. They also introduced the design principle that in order to avoid isolated ringing modes which tend to sound “metallic”, all the structure modes should decay at the same rate. A conceptual block diagram of Jot’s reverberator is shown in Fig. 11. Notice the absorption filters in the feedforward path that allow controlling the decay rate at different frequencies, and a tonal correction filter that is used to equalize the reverberator frequency response so that the generated reverberation sounds more natural. The original design uses a Householder matrix for the feedback path, however other unitary matrices can also be used [90]. These matrices can also be time-varying, resulting in improved perceptual characteristics [91].

Equivalent to a wide class of FDNs are the digital waveguide networks (DWN) [92]. A DWN consists of a number of digital waveguides (the digital equivalent of analog propagation lines, formed of two opposite delay lines with equal length) connected at lossless scattering junctions (see Fig. 12). Each scattering junction carries out a simple matrix multiplication to scatter the incoming signals on digital waveguides from each of its neighbors to generate outgoing signals to be distributed back to the same digital waveguides in the opposite direction. A signal reverberated using a DWN can be obtained by summing all of the outgoing signals of one of the scattering nodes. DWNs have appealing stability properties and have significant design flexibility owing to the different possible network graphs, types of lossless scattering, lengths of the digital waveguides. While both Jot’s reverberator and DWNs are capable of producing responses with a high perceptual quality, the parameters of these models are not explicitly linked to the physical characteristics of a particular room.

Artificial reverberators that are more tightly linked to room acoustics also exist. One of the earlier designs proposed by Kendall et al. [93] was based on recirculating delay elements whose lengths are determined by using an image source model of a rectangular room. A similar approach was also used in [94]. Karjalainen et al. proposed a class of DWNs designed to simulate early reflections and axial modes of rectangular rooms [95]. A drawback of their algorithm is that many of the algorithm’s internal parameters still require hand tuning in order to achieve a satisfactory reverberation.

An artificial reverberator that inherits all its parameters from physical characteristics of the room it simulates was recently proposed [12]. This reverberator, termed scattering delay network (SDN), is a modified DWN where the length of the digital waveguides and the topology of the network, as illustrated conceptually in Fig. 13, are derived directly from the geometry of the simulated space. In particular, SDN is a minimal network connecting as many scattering nodes as there are walls in the room, and where each scattering node is positioned at the point where first-order reflections impinge on the wall.

This design ensures that first-order reflections are rendered exactly, while second and higher-order reflections are simulated with a gradually diminishing accuracy. Since first-order early reflections are perceptually more important than the higher-order reflections, the resulting reverberation is perceptually realistic and statistically very similar to that of an actual room. A by-product of this design is that SDN does not require separate modules for early reflections and late reverberation while still allowing precise and explicit control of the room geometry, source and receiver directivity patterns, and wall absorption characteristics. Furthermore, it enables a
Fig. 12. Conceptual depiction of a DWN. The figure shows a DWN with 5 nodes (indicated as circles) connected via bidirectional delay lines (curves with double arrows). In order to maximize the reflection density, the delay lengths are chosen to be co-prime numbers. Input signal can be fed and the output can be obtained from any node. The inset shows the connection between two nodes where incoming and outgoing signals and the individual delay elements are clearly visible.

Fig. 13. Conceptual depiction of the SDN reverberator. The figure shows a rectangular room as observed from above with the associated delay lines interconnected at scattering junctions on the wall. Other delay lines associated to the floor and ceiling are also present, but are not shown here for clarity.

Fig. 14. Conceptual depiction of the SDN reverberator. The figure shows a rectangular room as observed from above with the associated delay lines interconnected at scattering junctions on the wall. Other delay lines associated to the floor and ceiling are also present, but are not shown here for clarity.

C. Audio Source Culling

Complex virtual environments typically include many sound sources, which makes synthesizing their acoustics a challenging task in terms of the associated computational cost. This difficulty is especially pronounced when rendering such audio content over devices with limited computational power such as mobile phones. In a typical scenario involving many concurrent sources, it may be necessary to select and render only a few.

State-of-the-art game engines typically use volumetric culling of sound sources. Each sound source has an associated culling volume (cube, sphere or, cylinder) and when the listener is within this volume, the sound is rendered. This is a simple approach which does not incur any significant computational cost apart from the relatively simple collision detection operation between the bounding boxes of the listener and each of the sound sources. However, this approach does not take into account the relative levels of the sound sources. It also does not limit the number of sound sources that can be simultaneously active. This makes the available computational power the only determinant in whether or not a sound source will be rendered, completely disregarding its perceptual salience.

Sound sources are dynamically activated in response to user generated events in interactive applications such as games and VR applications. For scenes comprising multiple concurrent sound sources, many of these sources will be masked by the others. This makes it redundant to process these inaudible sources.

Tsingos et. al. [96], [97] provide a perceptually-based source culling approach. The approach is based on ranking the sources in the scene using their binaural loudness at different frequency bands as a measure of perceptual salience. Loudness values used to calculate a masking threshold from a time-frequency representation of the sound sources and stored for use during runtime.
As a new sound event occurs, the decision to render the new sound source is made at the audio frame level. Each frame is compared with the existing mix for evaluating whether the mix can mask it. If it can, the frame is culled. As a result of the frame-level temporal resolution several frames from a single sound source can be culled while others are rendered. This results in each sound source being only partially culled. A similar algorithm was proposed by Metan and Hacıhabibo˘glu [98]. The audibility calculation in this algorithm is slightly different from the algorithm of Tsingos et. al. As a new sound event is generated, a look-ahead algorithm checks for the audibility of each frame of a sound source given the current mix being played. The decision to render a sound source is based on the ratio of audible frames to the total number of frames in the audio signal to be rendered. This way, the whole source and not a portion of it is rendered or culled. The advantage of these methods is that they potentially allow the preprocessing stage of the source culling process to be integrated with existing perceptual audio coding algorithms such as MPEG-1 Layer I Audio.

VI. SUMMARY

The body of knowledge on spatial hearing and the mechanisms which govern it has been steadily growing. However, a comprehensive model which can account for all the different aspects of spatial hearing is yet to be developed. Still, the existing knowledge can be used to design audio systems and algorithms, which have lower computational and hardware costs but can provide a subjective performance as good as more complicated physically-motivated systems. While the developments in computer hardware could make it possible to overcome issues that are due to computational limitations, the physical limitations such as the size of electroacoustic transducers or data bandwidth will remain. Similarly the energy cost of carrying out simple operations such as multiplication or memory access is likely to diminish, but will never vanish and power efficiency of mobile devices will also continue to be relevant. These issues will make it even more desirable to design simpler audio systems and algorithms. The importance of using knowledge on auditory perception to that aim will thus remain high.

ACKNOWLEDGMENT

This work was supported by the Turkish Scientific and Technological Research Council (TÜBİTAK) under the Research Grant 113E513 Spatial Audio Reproduction Using Analysis-based Synthesis Methods and by the European Commission under Grant Agreement no. 316969 within the FP7-PEOPLE Marie Curie Initial Training Network “Dereverberation and Reverberation of Audio, Music, and Speech (DREAMS)”.

REFERENCES


[74] ——, “Early reflection thresholds for anechoic and reverberant stimuli within a 3-d sound display,” in Proc. 18th Int. Congress on Acoust. (ICA’04), Kyoto, Japan, 2004.


James Johnston (F’02) received his BSEE and MSEE from Carnegie-Mellon University. He has worked on audio and acoustic signal processing in various research labs and companies such as AT&T Bell Laboratories, AT&T Labs, Neural Audio and Microsoft among others. He was the primary researcher and inventor of the MPEG-2 AAC audio coding algorithm, and a principal contributor to the “MP3” algorithm. Mr. Johnston was awarded the IEEE James L. Flanagan Signal Processing Field Award (2006); elected Fellow, Audio Engineering Society (1997); received AT&T Technology Medal and AT&T Standards Award (1998); received a New Jersey Inventor of the Year Award (2001); elected IEEE Fellow (2002). He was a distinguished lecturer of the IEEE Signal Processing Society in 2011. His current research interests include acoustic scene modelling, loudspeaker design, loudspeaker pattern control, cochlear modelling, masking threshold models, stereo imaging models and stereo imaging sensitivity models, methods of reproducing sound fields, microphone and sound field capture techniques, both actively steered and time-invariant, and speech and audio coding methods in general.

Zoran Cvetković (SM’04) is Professor of Signal Processing at Kings College London. He received his Dipl. Ing. and Mag. degrees from the University of Belgrade, Yugoslavia, the M.Phil. from Columbia University, and the Ph.D. in electrical engineering from the University of California, Berkeley. He held research positions at EPFL, Lausanne, Switzerland (1996), and at Harvard University (2002-04). Between 1997 and 2002 he was a member of the technical staff of AT&T Shannon Laboratory. His research interests are in the broad area of signal processing, ranging from theoretical aspects of signal analysis to applications in audio and speech technology, and neuroscience. From 2005 to 2008 he served as an Associate Editor of IEEE Transactions on Signal Processing.

Hüseyin Hachabiboğlu (S’96-M’00-SM’12) received the B.Sc. (honors) degree from the Middle East Technical University (METU), Ankara, Turkey, in 2000, the M.Sc. degree from the University of Bristol, Bristol, U.K., in 2001, both in electrical and electronic engineering, and the Ph.D. degree in computer science from Queens University Belfast, Belfast, U.K., in 2004. He held research positions at University of Surrey, Guildford, U.K. (2004-2008) and Kings College London, London, U.K. (2008-2011). Currently, he is an Associate Professor of Signal Processing and the Head of Department of Modelling and Simulation at Graduate School of Informatics, Middle East Technical University (METU), Ankara, Turkey. He also coordinates the recently established Multimedia Informatics graduate program at the same department. His research interests include audio signal processing, room acoustics, multichannel audio systems, psychoacoustics of spatial hearing, microphone arrays, and game audio. He is a member of the IEEE Signal Processing Society, Audio Engineering Society (AES), Turkish Acoustics Society (TAD), and the European Acoustics Association (EAA).

Enzo De Sena (S’11-M’14) received the B.Sc. in 2007 and M.Sc. cum laude in 2009, both from the Universita degli Studi di Napoli Federico II in Telecommunication Engineering. In 2013, he received the Ph.D. degree in Electronic Engineering from Kings College London, where he was also a Teaching Fellow from 2012 to 2013. Between 2013 and 2016 he was a Postdoctoral Research Fellow at the Department of Electrical Engineering of KU Leuven. Since September 2016 he is a Lecturer (Assistant Professor) in Audio at the Institute of Sound Recording of the University of Surrey. He held visiting positions at the Center for Computer Research in Music and Acoustics at Stanford University (2013), at the Signal and Information Processing section at Aalborg University (2014-2015) and at the Speech and Audio Processing Group at Imperial College London (2016). He is a former Marie Curie fellow. His current research interests include room acoustics modelling, multichannel audio systems, microphone beamforming and binaural modelling.