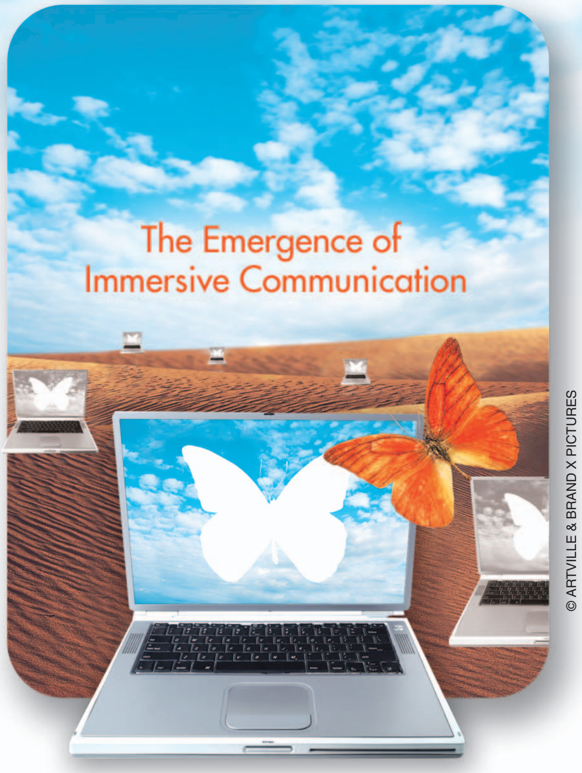
[V. Ralph Algazi and Richard O. Duda]

Headphone-Based Spatial Sound

[Exploiting head motion for immersive communication]



© ARTVILLE & BRAND X PICTURES

With its power to transport the listener to a distant real or virtual world, realistic spatial audio has a significant role to play for immersive communications. Headphone-based rendering is particularly attractive for mobile communications systems. Augmented realism and versatility in applications can be achieved when the headphone signals respond dynamically to the motion of the listener. The timely development of miniature low-power motion sensors is making this technology possible. This article reviews the physical and psychoacoustic foundations, practical methods, and engineering challenges to the realization of motion-tracked sound over headphones. Some new applications that are enabled by this technology are outlined.

INTRODUCTION

Advances in communication infrastructure and technology from the cell phone to the Internet are placing us at the threshold of a new generation of mobile applications that will deliver immersive communication. Such developments will spread rapidly and will impact both the workplace and the general public.

This article is concerned with the generation and reproduction of spatial sound for mobile immersive communications. Properly reproduced over headphones, spatial sound can provide an astonishingly lifelike sense of being remotely immersed in the presence of people, musical instruments, and environmental sounds whose origins are either far distant, virtual, or a mixture of local, distant, and virtual. For voice communication, spatial sound can go beyond increased realism to enhancing intelligibility and can provide the natural binaural cues needed for spatial discrimination. Mobile voice communications applications that use these new capabilities include audio teleconferencing or telepresence in a meeting. For music, immersive sound can go beyond reproduction that places the listener in the performance venue (or perhaps positioned on stage among the performers) to enabling the creation of entirely new audio effects. For environmental monitoring or games, it can provide unparalleled awareness of both the sound-generating objects and the surrounding acoustic space. Spatial sound will also be used in conjunction with video in remote monitoring to provide rapid sonic detection and orientation of events for subsequent detailed analysis by video.

Spatial sound technology has a long history [1]. The familiar stereo and multichannel surround-sound systems were designed for loudspeaker reproduction [2], [3]. By contrast, in this article, we focus on mobile systems, where the low power, light weight, high fidelity, low cost, and simple convenience of

Digital Object Identifier 10.1109/MSP.2010.938756
Date of publication: 17 December 2010

1053-5888/11/$26.00©2011IEEE

headphones make them the obvious choice. Thus, this article focuses on the generation and reproduction of headphone-based spatial sound.

CHALLENGES

The delivery of a high-quality spatial sound experience over headphones requires reproduction of the complex dynamic signals encountered in natural hearing. This goes well beyond current commercial practice. When sound is heard over only one earphone—as is typical for contemporary cell phones—the listening experience is severely limited. A pair of earphones enables binaural reproduction, which provides a major improvement. However, if a single voice channel is used to feed both earphones, most listeners will hear the voice internalized in or near the center of their heads. Relevant auditory cues can be produced by changing the balance and/or by introducing interaural time delays. These changes can shift the apparent location to a different point on a line between the ears, but the sound remains inside the head and unnatural.

PROPERLY REPRODUCED OVER HEADPHONES, SPATIAL SOUND CAN PROVIDE AN ASTONISHINGLY LIFELIKE SENSE OF BEING REMOTELY IMMERSED.

Binaural recordings made with two microphones embedded in a dummy head introduce such basic cues as the proper interaural time and level differences and add the important acoustic cues of room reflections and reverberation. They can produce a compellingly realistic listening experience. However, because of the lack of response to head motion, there are still major problems with conventional binaural technology: a) front/back confusion (and the related failure of binaural pickup to produce externalized sound for sources that are directly in front or in back), and b) significant sensitivity to the size and shape of the listener's head and outer ears. Further, the common experience of focusing attention by turning towards the source of the sound is not possible.

As we shall explain, there are basically two different ways to exploit dynamic cues to solve these problems. One approach uses so-called head-related transfer functions (HRTFs) to filter the signals from the source in a way that accounts for the propagation of sound from the source to the listener's two ears. This approach requires having HRTFs and isolated signals for every source and uses HRTF interpolation to account for head motion. The other approach, motion-tracked binaural (MTB), is based on sampling the sound field sparsely in the space around a real or virtual dummy head. MTB requires knowing the signals at multiple points around the head and uses interpolation of the signals from these microphones to account for head motion. For both methods, the essential dynamic cues that are generated by head motion can now be achieved by low-cost, low-power, small-size head trackers based on microelectromechanical systems (MEMS) technology. Thus, the development of new signal processing methods that respond to the dynamics of human motion promises a new era in immersive binaural audio applications for mobile communications.

Understanding any binaural technology requires knowledge of both the physics of sound propagation and the psychophysics of auditory perception. We begin with a brief review of the psychoacoustic cues for sound localization and then review their physical basis.

SOUND LOCALIZATION CUES

There is a large body of literature on the psychoacoustics of sound localization which can only be summarized briefly here. Blauert's book [4] is the classic reference for the psychoacoustics of spatial sound. Chapters 2 and 3 of Begault's book [5] provide an excellent overview for engineers. Begault also surveys the effects of visual and other nonauditory cues on spatial sound perception [6]. The primary auditory cues used by people include

1) the interaural time difference (ITD)
2) the interaural level difference (ILD)
3) monaural spectral cues that depend on the shape of the outer ear or pinna
4) cues from torso reflection and diffraction
5) the ratio of direct to reverberant energy
6) cue changes induced by voluntary head motion
7) familiarity with the sound source.

Except for source familiarity, all of these cues stem from the physics of sound propagation and vary with azimuth, elevation, range, and frequency. Although some of these cues are stronger than others, for optimum sound reproduction all of them should be present and consistent. When a strong cue conflicts with a weak one, the strong cue will often dominate. However, if the conflicts are too great, the listener will become bewildered, and the apparent location of the sound source will either be in error or be indeterminate.

The ITD and ILD are the primary cues for estimating the so-called lateral angle, the angle between the vertical median plane and a ray from the center of the head to the sound source. These cues have the important property of being largely independent of the source spectrum. According to Lord Rayleigh's well-known duplex theory, the ITD prevails at low frequencies, where head shadowing is weak, and the ILD prevails at high frequencies, where interaural phase difference is ambiguous [7]. The crossover frequency is around 1.5 kHz, where the wavelength of sound becomes less than the distance between the ears. Subsequent research has shown that the interaural envelope delay (IED) provides a temporal localization cue at high frequencies [8]. However, the low-frequency ITD is a particularly strong cue, and can override other, weaker localization cues [9].

The cues for elevation are not as robust as those for the lateral angle. It is generally accepted that the monaural spectral changes introduced by the outer ears or pinnae provide the primary static cues for elevation [10], although they can be overridden by head motion cues [11]. These spectral changes occur above 3 kHz, where the wavelength of sound becomes smaller than the size of the pinna. The reflection and refraction of sound by the torso provides even weaker elevation cues, although they appear at lower frequencies and can be important for sources that have little high-frequency content [12]. Monaural pinna cues present a special problem for sound reproduction because they vary so much from person to person, and they may not be faithfully reproduced by uncompensated headphones [13], [14].

The three primary cues for range are the absolute loudness level combined with familiarity with the source [15], the low-frequency ILD for close sources [16], and the ratio of direct to reverberant energy for distant sources [17]. In particular, reverberant energy decorrelates the signals reaching the two ears [18], and the differences between the timbre of direct and reverberant energy provides another localization cue, one that might be important for front/back discrimination as well. All of these cues contribute to externalization—the sense that the origin of the sound is outside of the head. Achieving convincing externalization with headphone-based sound reproduction has proved to be a difficult challenge, particularly for sources directly in front of or directly behind the listener.

[THE ITD AND ILD ARE THE PRIMARY CUES FOR ESTIMATING THE SO-CALLED LATERAL ANGLE.]

All of the cues mentioned so far are static. However, it has long been recognized that people also use dynamic cues from head motion to help localize sounds. Over 60 years ago, Wallach demonstrated that motion cues dominate pinna cues in resolving front/back confusion [11]. Although the pinna also provides important front/back cues, and although head motion is not effective for localizing very brief sounds, subsequent research studies have confirmed the importance of dynamic cues for resolving front/back ambiguities, improving localization accuracy, and enhancing externalization [19].

This summary of a large body of literature is necessarily brief, and a word of caution is needed. In particular, as is commonly done in the psychoacoustic literature, we have described the localization cues in the frequency domain, as if the ear were a Fourier spectrum analyzer. Because the auditory system performs an unusual kind of nonlinear, adaptive, short-time spectral analysis, classical spectral arguments require caution. The Franssen effect, for example, cannot be explained by a simple spectral analysis (see [4], p. 280). The fact that multiple sound sources are almost always present further complicates spectral arguments. In saying that the ITD and ILD are largely independent of the source spectrum, for example, we are tacitly assuming that the source spectrum is not changing rapidly and that there are time periods when the signal-to-noise ratio is high across the spectrum. Despite these limitations, spectral arguments provide insight into how humans localize sounds.

THE HRTF, HRIR, AND BRIR

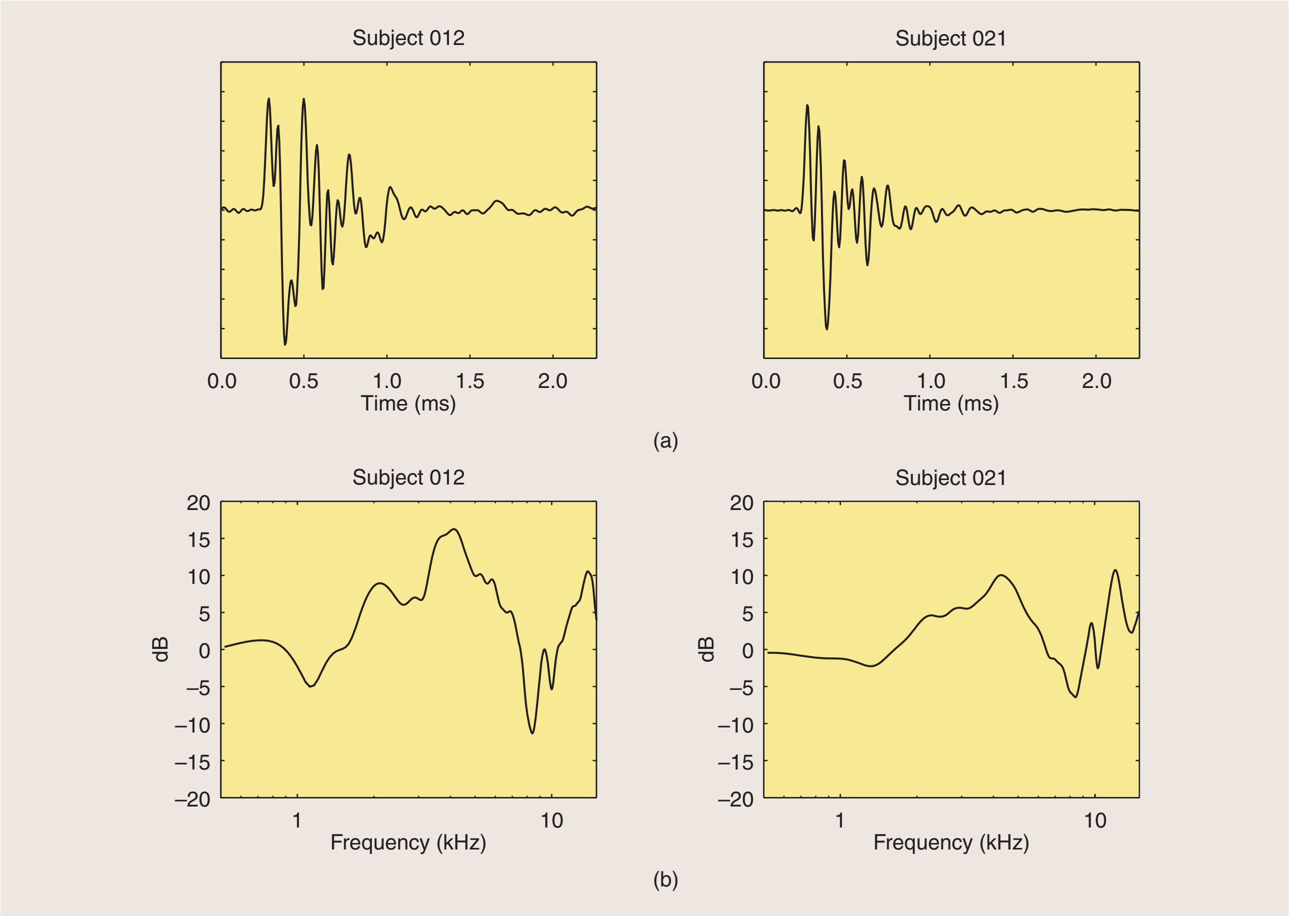
The acoustic cues for sound localization are a consequence of the physical processes of sound generation, propagation, diffraction, and scattering by objects in the environment, including the listener's own body. In principle, these processes can be analyzed by solving the wave equation subject to the appropriate boundary conditions. In practice, the irregularities of the boundary surfaces produce extremely complex phenomena, and measuring the boundary surfaces (particularly, the pinnae) with sufficient accuracy can be challenging. Analytical solutions are available only for very simple geometries. Standard numerical methods are limited by the need to have at least two spatial samples for the shortest wavelength of interest, and by execution times that grow as the cube of the number of sample points. Thus, most of what is known about the acoustic cues has come from acoustic measurements.

Fortunately, at typical sound pressure levels and object velocities, the physical processes are essentially linear and time invariant, and linear systems theory applies. The effects of the listener's own body on sounds coming from an isotropic point source in an anechoic environment are captured by the so-called HRTF [20], [21]. The HRTF is defined as the ratio of the Fourier transform of the sound pressure developed at the ear to the Fourier transform of the sound pressure developed at the location of the center of the listener's head with the listener absent. This frequency-domain definition has the advantage that the resulting HRTF is essentially independent of range when the source is in the far field. Most HRTF measurements are made under these conditions. The far-field range dependence is easily obtained merely by adding the propagation delay and the inverse range dependence.

The inverse Fourier transform of the HRTF is the head-related impulse response (HRIR). If $h(t)$ is the head-related impulse response for a distant source and c is the speed of sound, then the anechoic pressure response to an impulsive velocity source at a distance r is proportional to $h(t - r/c)/r$. The situation is more complicated when the source has a complicated radiation pattern or is distributed or is close to the head [16], and we limit our discussion to an isotropic point source in the far-field.

The temporal structure (especially multipath effects) is most easily seen in the HRIR, whereas the spectral structure is best revealed by the HRTF magnitude. Figure 1 shows experimentally measured HRIRs and HRTFs for two different subjects for a sound source located directly ahead. The complex behavior seen above 3 kHz is due primarily to the pinna, and the subject-to-subject differences are primarily due to differences in the sizes and shapes of the subjects' pinnae. The results shown in Figures 1–3 were taken from the CIPIC HRTF databse. The complete database and its documentation can be downloaded from http://interface.ece.ucdavis.edu/CIL_html/CIL_HRTF_database.htm.

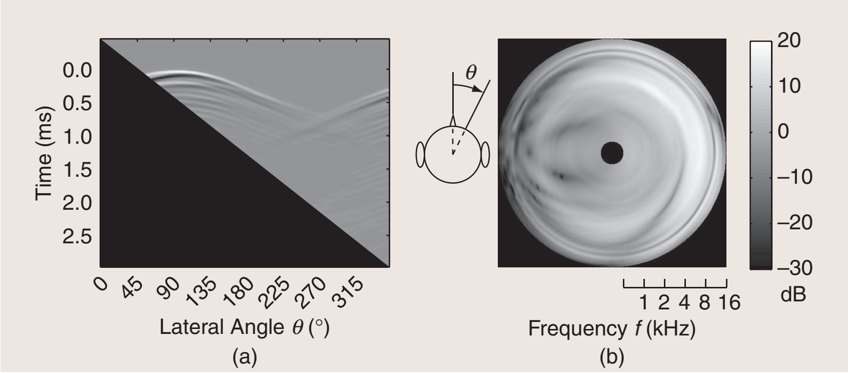
The directional dependence of the response for Subject 021 is illustrated in the images shown in Figures 2 and 3. Figure 2 shows how the right-ear HRIR changes when the source circles



[FIG1] Part (a) shows the HRIRs for Subject 012 and Subject 021 in the CIPIC HRTF database. Part (b) shows the magnitudes of the HRTFs.

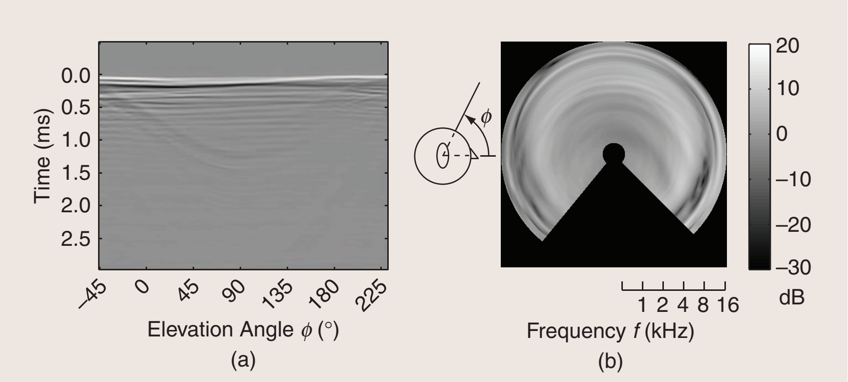
around the subject in the horizontal plane. The impulse response is strongest and begins soonest when the lateral angle θ is close to 100^c and the source is radiating directly into the ear. The HRTF reveals that the magnitude response is essentially constant in all directions at low frequencies, but above 3 kHz the response on the ipsilateral side $(0° < \theta < 180°)$ is clearly greater than the response on the contralateral side $(180° < \theta < 360°)$. To a first approximation, the response of the left ear can be found by changing the sign of θ. From the plots, we see that the time of arrival and the magnitude of signals, and thus the ITD and the ILD, also vary systematically with θ, and it is not surprising that the ITD and the ILD are strong cues for θ.

The variation of the HRTF with the elevation angle ϕ is more subtle. Figure 3 shows results in the median plane, where interaural differences are usually negligible. The HRIR reveals various pinna resonances and faint torso reflections. The HRTF shows that the strengths of the resonances and the frequencies and depths of various interference notches do change systematically with elevation. These spectral changes provide the monaural cues for elevation. The spectral profile varies significantly from person to person, and individualized HRTFs are required for accurate static elevation perception [22].



[FIG2] (a) Horizontal-plane variation of the right-ear HRIR and (b) the HRTF magnitude for Subject 021. In these images, the response is indicated by the brightness level. For the HRIR in (a), each vertical line corresponds to the impulse response at a particular lateral angle θ (see the diagram of the head). For the HRTF in (b), each radial line corresponds to the magnitude response (in decibels) at the corresponding lateral angle. Thus, the frequency response for the straight-ahead direction $\theta = 0$ is revealed by the brightness along a line from the center to the top of the plot. Frequencies range from 500 Hz near the center to 15 kHz at the periphery.

In these plots, the HRTFs and HRIRs are presented as continuous functions of lateral angle and elevation. In practice, they are always sampled at discrete angles. When results are needed at intermediate angles, interpolation is required. This raises the question of how densely the HRTFs need to be sampled to achieve accurate reconstruction. The answer depends on the tolerable reconstruction error, and is ultimately a psychoacoustic question [23]. In practice, the sampling density that is typically used is on the order of five degrees, which has received support from theoretical analysis [24].



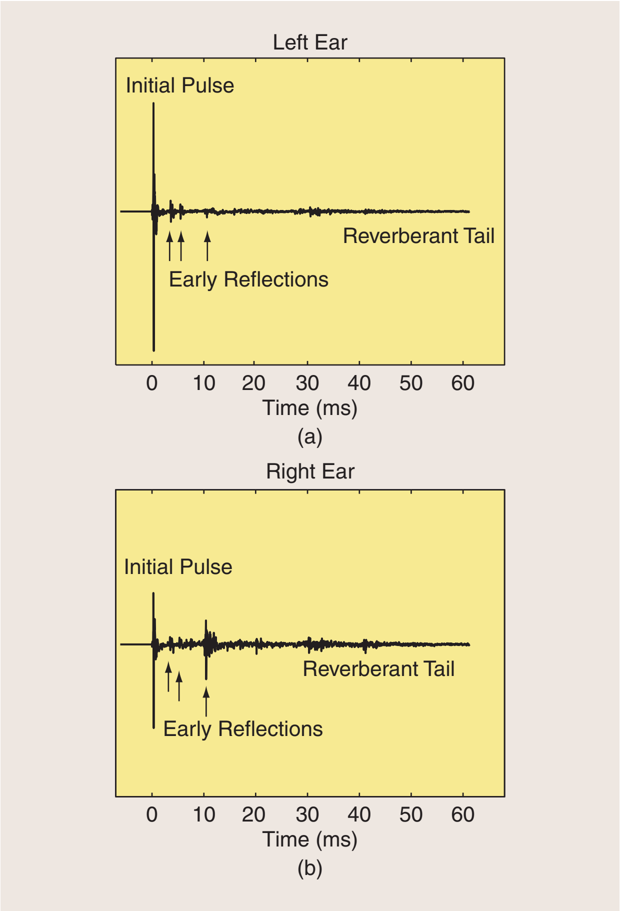
[FIG3] **Median-plane variation of the (a) HRIR and the (b) HRTF magnitude with elevation angle ϕ. The sector at the bottom of the HRTF image is blank because the low-elevation area was physically inaccessible.**

For practical applications as well as theoretical understanding, it is often useful to be able to replace an experimentally measured HRTF by a mathematical model. By including only a small number of terms or a small number of coefficients, these models can often be simplified or smoothed to provide HRTF approximations. Many models have been proposed, including principal components models [25], spherical-harmonic models [26], neural network models [27], pole-zero models [28], and structural models [29]. Unfortunately, the literature is too large to be reviewed here, and the references cited only provide representative examples.

Listening to a sound signal filtered by individualized HRTFs produces the auditory experience of hearing that sound in an anechoic chamber. However, anechoic chambers are very unusual and unpleasant listening environments. Although we are usually not aware of our acoustic surroundings, reflections of sound energy from objects in the environment have a profound effect on the nature and quality of the sound that we hear. In particular, for a distant source in a normal setting, the acoustic energy coming directly from the source can be significantly less than the subsequent energy arriving from multiple reflections. When the reflected sounds are missing, the perception is that the source must be very close.

It is unfortunate for the developers of spatial sound systems that most people believe that they are much better at judging the distance to a sound source than they actually are. Without visual cues, people usually greatly underestimate the distance to a source from its sound alone. Interestingly, we do best when the source is a person speaking, where familiarity with the source allows us to estimate range from the loudness level [30]. In general, proper gain settings, which listeners ordinarily want to control, are important for accurate distance judgments, and this is particularly important in the case of speech.

A natural way to accommodate the effects of the environment is to measure the impulse response in a room, thereby including all of the early reflections and subsequent reverberation caused by multiple reflections. When separate measurements are made for each ear, this is called the binaural room impulse response (BRIR). As Figure 4 illustrates, BRIRs are much longer than HRIRs. Thus, in filtering a



[FIG4] **(a) An example BRIR for a small room with the sound source on the left. The response at the left ear is shown in (a) and the response of the right ear is shown in (b). The initial pulse is the HRIR. Early reflections from the floor, ceiling, and walls are clearly visible. The multiple reflections that constitute the reverberant tail decay exponentially and last beyond the 60-ms time segment shown. Reverberation times in concert halls can extend to several seconds.**

sound signal with BRIRs, the issues of latency and computation time must be addressed.

THE EXPLOITATION OF HEAD MOTION DOES AMELIORATE THE LIMITATIONS INTRODUCED BY HAVING TO USE A DUMMY HEAD.

RENDERING SPATIAL SOUND OVER HEADPHONES

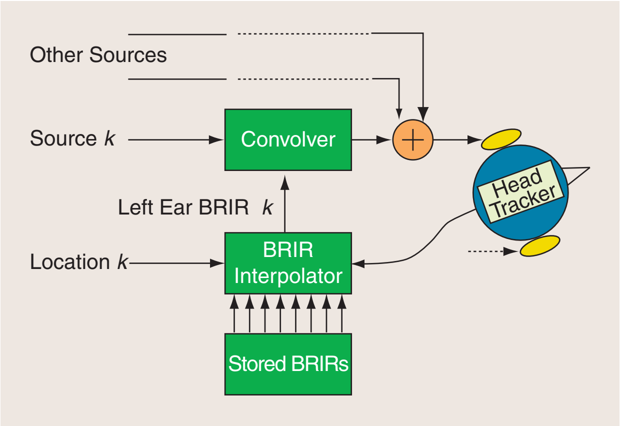
As we mentioned earlier, there are basically two different ways to render spatial sound over headphones: 1) through the use of HRTFs and 2) through the process of sampling and reconstructing the sound field. Both methods employ interpolation, but in quite different ways, and we consider each in turn.

HRTF-BASED RENDERING OF VIRTUAL SOUND FIELDS

The HRTF approach has been widely used to provide spatial sound over headphones, particularly for the virtual acoustic environments encountered in computer games and military training systems [5]. Here separate signals are available for each source, the spatial locations of the sources are all known, and a head tracker is used to determine the location and orientation of the listener in the room.

A conceptually simple example is the binaural room scanning (BRS) system illustrated in Figure 5 [31]. In a typical application, BRS is used to reproduce over headphones the experience of listening to a very high-quality surround-sound system. Here the source signals are the feeds sent to high-performance loudspeakers properly positioned in an acoustically optimized listening room. BRIRs are measured from the speakers to the microphones in a dummy head located at the ideal listening location, with separate BRIRs measured for every few degrees of head rotation. During playback, the signal from the head tracker is used to control an interpolator that, for each source, combines adjacent BRIRs to produce left-ear and right-ear BRIRs that vary continuously with head rotation. The results of convolving the source signals with their corresponding BRIRs are summed and fed to the headphones.

Properly implemented, BRS captures the room characteristics faithfully and produces very high-quality spatial sound. However, several difficult problems must be solved to realize these results [14]. The dummy head must adequately approximate the listener's head. A large number of long BRIRs must be measured. The error introduced by the interpolation algorithm must be unnoticeable. The combined process of head tracking, interpolation, and convolution cannot introduce detectable latency. And, as with all headphone-based systems, the headphones must be adequately compensated.



[FIG5] Elements of the BRS system.

The exploitation of head motion does ameliorate the limitations introduced by having to use a dummy head. Because the pinnae for the dummy head may differ greatly from the pinnae of the listener, listeners frequently report that the virtual loudspeakers appear to be elevated, particularly for the speaker that is directly in front.

For sources at the side, the large ITDs and ILDs that are generated are incompatible with an overhead location. These powerful cues dominate any confusion caused by conflicting pinna cues, and the source is perceived to be at a low elevation. For sources near the median plane, the pinna mismatch becomes more important. In the authors' experience, when head motion is tracked, after a short time listeners will adapt and experience reduced frontal elevation. Nevertheless, pinna mismatch is a troublesome problem for all headphone-based spatial sound systems.

IMPLEMENTATION ISSUES

The two major issues in the implementation of HRTF-based rendering are computational cost and latency. Computational requirements depend on the complexity of the auditory scene, the allowed motion of the listener, and the efficiency of the implementation of the algorithms. The approach to these issues depends on the application and on the decision as to what constitutes an acceptable auditory experience, as opposed to one that is indistinguishable from actually being present.

A simple analysis of the rendering of sound by direct convolution with a BRIR indicates the scope of the issues. Brute-force convolution of a sound signal with the 0.5-s impulse response of a small room (about 22,000 samples at 44.1 kHz) will require approximately one giga operations per second. Requirements are doubled for two ears and scale linearly with the number of sound sources. Thus the computational load can be very large. Further, motion of the listener will require a rapid change in the BRIRs. Unless fast, low-latency algorithms are used, this may result in an unacceptable delay in the response to head motion.

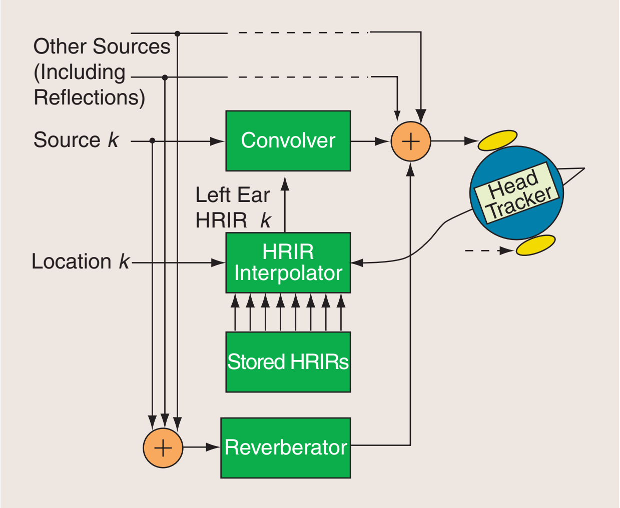
A variety of approximations have been introduced to address these problems. One illustrative example is sketched in Figure 6. Here the long BRIRs are replaced by short HRIRs combined with an approximate room model. Individualized HRIRs are used for high-performance systems, and generic HRIRs are used for consumer-grade products. Early reflections are represented by a small number of spatialized image sources, and the reverberant tail is approximated by filtering the sum of the source signals by an appropriate IIR filter. Many commercial systems are variations on this basic theme [5], [32].

MTB IS COMPUTATIONALLY SIMPLE. IT IS HIGHLY EFFECTIVE FOR LIVE SOUND AND FAITHFULLY CAPTURES THE ACOUSTICS OF THE RECORDING SPACE.

This architecture is particularly well suited to single-listener virtual environments, where the source signals are computer generated and their locations are under computer control. Like BRS, it can also be used to reproduce conventional stereo or surround sound recordings. It is not well suited to capturing natural sounds faithfully, for several reasons: a) it is difficult to obtain the separate source signals, b) it is difficult to determine the locations of the sources, and c) it is computationally very expensive to capture the complexity of the reflections and reverberation in natural listening spaces. In practice, one is forced to employ a two-stage process, using conventional recording practices to produce a surround sound mix, and then applying the HRIR-based procedure to the results. The standard recording practice is to make a virtue out of necessity, using post-production techniques to enhance an experience, e.g., by using spot microphones to highlight sounds that might otherwise not be heard. However, the results will not faithfully reproduce the original sonic landscape.

COMPUTING AND RENDERING NATURAL SOUND FIELDS

For many applications, we would like to be able to capture a natural sound field, with no prior knowledge of the number or locations of the sources, or the structure of the acoustic environment. Two basic methods have been developed for this purpose—Ambisonics and MTB. We consider each in turn.



[FIG6] An HRIR-based rendering system that employs a simple room model that uses image sources to account for early reflections and a single filter to simulate room reverberation.

AMBISONICS

The goal of Ambisonics is to recreate the acoustic waves that are incident on a listener's head [33]. The core idea is to use a coincident microphone array (called a sound field microphone) to capture pressure waves coming from different directions, and to reproduce those waves through loudspeakers positioned around the listener. The original method used four microphones, which produced a first-order approximation of the incident sound field. Higher-order Ambisonics uses additional microphones and the mathematics of spherical-harmonic expansions to achieve a more faithful approximation [34].

To use this approach for headphone reproduction, one can employ any of the HRTF-based methods described in the previous section to render the signals that would be sent to the loudspeakers [35]. This has the advantage that it eliminates the effects that the listening space has on loudspeaker reproduction. However, it inherits the limitations of HRTF-based rendering.

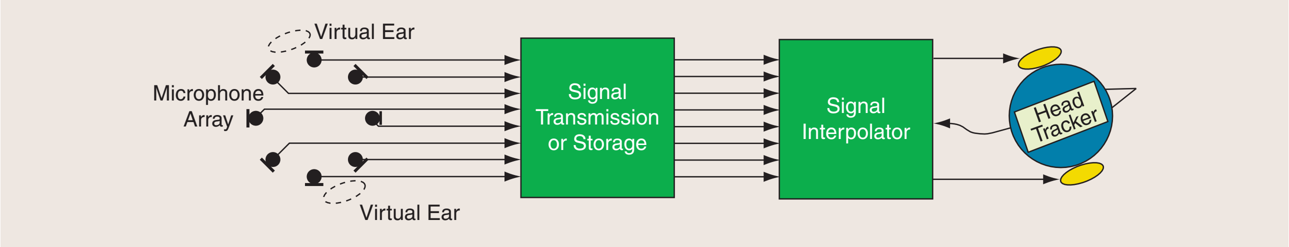
Another loudspeaker-based approach called wavefield synthesis employs hundreds of loudspeakers to recreate the sound field over a large area, such as an area occupied by an audience [36]. Although quite interesting, this approach is not relevant to headphone reproduction.

MOTION-TRACKED BINAURAL SOUND CAPTURE AND RENDERING

Binaural recording is particularly effective at capturing the acoustics of a natural listening space but was long thought to be unable to account for the important effects of head motion. However, once it was realized that a dummy-head microphone array is merely sampling the sound field at two points in space, it became clear that one could account for head motion by sampling at additional points and interpolating. The resulting generalization of binaural recording is called MTB [37], [38].

The basic components of an MTB system are shown in Figure 7. Sounds in the recording space are captured by microphones that are mounted around the diameter of a sphere or cylinder that is roughly the size of a human head. These signals can either be sent directly to the listener, or recorded for subsequent playback. The head tracker is used to control the interpolation between signals from the microphones that bridge the listener's ears.

Signal interpolation is much simpler than HRIR interpolation followed by convolution. However, for exact waveform reconstruction, Nyquist sampling theory requires the microphones to be no more than half a wavelength apart. If signals from adjacent microphones are directly interpolated, when the wavelength is shorter than half the intermicrophone distance, interference notches will appear in the spectrum. If a is the radius of the microphone array, N is the number of microphones, and c is the speed of sound, direct interpolation will



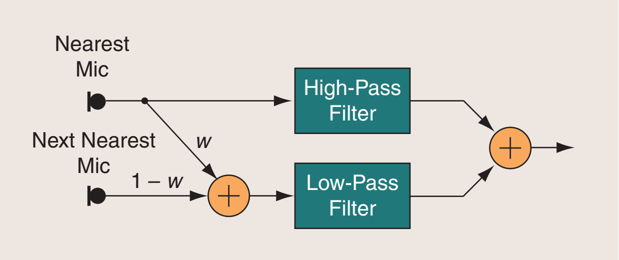
[FIG7] Basic components of an MTB system.

produce deep spectral notches at odd multiples of the frequency $f_{max} = Nc/4\pi a$ [37]. To cover the full 20-kHz bandwidth without suffering a significant spectral notch would require distributing about 128 microphones around a typical dummy head.

A NATURAL WAY TO ACCOMMODATE THE EFFECTS OF THE ENVIRONMENT IS TO MEASURE THE IMPULSE RESPONSE IN A ROOM.

Fortunately, exact waveform reconstruction is not necessary. The phase sensitivity needed for reconstruction is most important for the low-frequency ITD. In our experience, eight microphones produce results that are acceptable for speech, and 16 seem to be sufficient for music. To eliminate the flanging sounds associated with the spectral notches, the microphone signals are split into low-frequency components (below $0.5f_{max}$) and high-frequency components (above $0.5f_{max}$). The low-frequency components are interpolated, and then the high-frequency components are somehow restored.

Several methods have been investigated for restoring the high frequencies [38]. One of the simplest is illustrated in Figure 8. Here the interpolation weight w varies from $w = 1$ when the listener's ear is coincident with one of the microphones to $w = 0.5$ when the listener's ear is halfway between two microphones. The low-pass and high-pass filters are complementary, with a crossover frequency at $0.5f_{max}$. For $N = 16$ and $a = 8.75$ cm, the crossover frequency is 2.5 kHz. Because this method cannot provide exact waveform reconstruction, it generates artifacts, and controlled listening tests are needed to evaluate listening quality levels. Melick provides a systematic listing of the artifacts produced by the MTB procedure, together with suggestions for reducing them [39].



[FIG8] A simple method for low-frequency interpolation and high-frequency restoration, where the high-frequency components are always taken from the nearest microphone. The interpolation weight w varies between 0.5 and 1 depending on the azimuth angle of the listener.

By contrast to HRTF rendering, binaural sound captured with microphones in the ears of a dummy can be directly presented to a listener with no processing at all, and the processing demands for the signal interpolation used by the MTB method are small. However, the conversion of legacy stereo recordings through convolution leads to the same kinds of computational demands faced by HRTF-based methods, with the exception that the number of HRTFs required may be small. An alternative to real-time rendering is to perform the computations off-line for each sound source, and to store the resulting sound files for playback. A complex spatial soundscape is then created by a superposition of sounds files. Real-time computations are eliminated in exchange for an increase of the storage needed for sound files, and a new communication load for the remote access of these files [40].

The HRTF approach and the MTB approach have complementary strengths and weaknesses. MTB is computationally simple. It is highly effective for live sound and faithfully captures the acoustics of the recording space. It efficiently supports multiple simultaneously head-tracked listeners in broadcast or streamed applications, and to some extent it can be individualized to specific listeners [39]. It does not allow the listener to move around in the recording space, and it does not readily support conventional recording practices, such as the use of spot microphones.

EVALUATION OF THE QUALITY OF SPATIAL SOUND SYSTEMS

Spatial sound systems can be evaluated along various dimensions: accurate and stable sound source location, convincing externalization, faithful spectral quality, faithful reproduction of the acoustic environment, and freedom from audible artifacts. Psychoacoustic considerations influence all of these considerations. We cannot provide a systematic exposition of all of the approximation techniques, but we can list some examples.

Correct sound source localization in azimuth is provided by the HRTF and its principal cues, the ITD and ILD. Stabilization can be achieved by head tracking. Head tracking also reduces the need for personalization of the HRTF. In our experience, a simple head model without pinnae is often satisfactory. Room reflections and some reverberation are needed for externalization and distance perception. Discrete room reflections of 50 ms in total duration may be sufficient, and considerable effort has been devoted to developing room models with various degrees of tradeoff between auditory quality and computational complexity [32], [41], [42]. Reverberation decorrelates the signals at the two ears, which is particularly important for sources in the median plane, and contributes in a broad sense to externalization and the sense of distance [18]. Reverberation may have a long duration and is essentially random. Nonrandom recursive models are widely used and can approximate real reverberation very efficiently. While simple room models and artificial reverberation will not provide the sound quality of a good acoustic space such as a concert hall, computational efficiency at the cost of sound quality is an acceptable tradeoff for many applications.

TECHNICALLY, THESE OPPORTUNITIES STEM FROM THE INCREASINGLY WIDESPREAD USE OF SENSORS IN PORTABLE DEVICES.

DISCUSSION AND CONCLUSIONS

Research on spatial sound has yielded a spectrum of techniques for reproducing spatial sound. These techniques are particularly valuable for mobile communication, where—by contrast with the limitations of mobile visual displays—one can provide very high-quality immersive reproduction that creates a genuine experience of "being there" [43].

Although the main topic of this article has been the exploitation of head motion in the delivery of spatial audio, the opportunities for new ways to combine audio and video for immersive communication deserve comment. Technically, these opportunities stem from the increasingly widespread use of sensors in portable devices and the development of video technology such as virtual panoramic video. Psychologically, they stem from the fact that the auditory channel naturally provides the alerting and orienting cues to direct the attention of the visual channel. Here are a few of many possible applications.

- Internet-based services such as "street view" in Google or on-the-spot panoramic recording of news events such as CNN's Haiti: 360 can be augmented by simultaneously recorded spatial audio that increases the experience of presence. A similar technology can be employed for surveillance and remote monitoring. In general, any communications service that employs video broadcasting can include spatial audio broadcasting as well [40].
- Location-dependent information can be provided in audio form for services, tourism, and various kinds of guides, while affording hands-free and eyes-free operation. In particular, spatial audio can speed the access to information by providing alerting and orienting cues.
- The use of spatial audio in teleservices such as teleconferencing, telemedicine, and telerobotics can provide a remote specialist an enhanced presence.
- Finally, although not specifically relevant to mobile communication, training systems and various forms of entertainment (music, games, social networking) can all be enhanced by including spatial audio.

There are many obstacles to achieving a virtual experience that is indistinguishable from the real experience. The complexity of natural sound fields, the person-to-person variations in HRTFs, the limitations of transducers, and the usual costs of computation, bandwidth, storage, and hardware present the system designer with the need to compromise. As is the case with bandwidth compression, the key to finding effective solutions lies in exploiting psychoacoustics.

In the case of spatial sound, the most powerful psychoacoustic cues come from the interaural difference cues, room effects, and the dynamic cues produced by head motion. Although the importance of head motion on all aspects of the sound perception has been recognized for a number of years, the development of low-cost, low-power, miniature head trackers is a turning point in the use of motion tracking in spatial sound reproduction.

In this article, we have focused on two general methods for delivering spatial sound over headphones. Both of these methods provide the interaural cues, and both provide the head motion cues. The two general methods differ in the way that they account for the acoustic environment. HRTF-based methods can handle translation as well as rotation but require separate signals for every sound source and must employ room models to account for the complex reflection and reverberation patterns found in real acoustic spaces. MTB-based methods only handle rotation. By sampling and reconstructing the actual sound field in the vicinity of the head, they exchange a simulation problem for a sampling and reconstruction problem. Although either method is capable of handling both real and virtual environments, HRTF-based methods are more suitable for generating virtual auditory spaces, and MTB-based methods are more suitable for reproducing real auditory spaces.

A natural option is to combine the two, superimposing a limited number of artificial sound objects on a natural sound field and thus producing an augmented audio reality. The proper mix of recorded and synthetic sounds clearly depends on the application. However, we expect to see the emergence of hybrid systems that combine these approaches to provide the powerful immersive communication systems of the future.

ACKNOWLEDGMENTS

The authors would like to acknowledge the long-term support of their research on spatial sound by the National Science Foundation and by the University-Industry programs of the University of California. The contributions of Dennis Thompson and several University of California, Davis graduate students to the development of MTB have been invaluable. We are also grateful for the constructive suggestions of the reviewers.

AUTHORS

V. Ralph Algazi (vralgazi@ucdavis.edu) received the Ingenieur Radio degree from École Supérieure d'Électricité (ESE), Paris, France, and the M.S. and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology in 1952, 1955, and 1963, respectively. He has been on the faculty of the University of California, Davis, since 1965, where he was chair of the Department of Electrical and Computer Engineering from 1975 to 1986. He founded the Center for Image Processing and Integrated Computing (CIPIC). His research has focused principally on engineering applications concerned with human perception. He is a Life Senior Member of the IEEE and a member of the Audio Engineering Society.

Richard O. Duda (richard.o.duda@gmail.com) received the B.S. and M.S. degrees in engineering from the University of California, Los Angeles in 1958 and 1959, respectively, and the Ph.D. degree in electrical engineering from the Massachusetts Institute of Technology in 1962. During his career, he held appointments at SRI International, Fairchild Semiconductor, San Jose State University, and the University of California, Davis. His research interests are in the areas of pattern recognition and auditory perception. He is a coauthor of *Pattern Classification,* second edition. He is a member of the Audio Engineering Society, and is a Fellow of the IEEE and the American Association for Artificial Intelligence.

REFERENCES

[1] M. F. Davis, "History of spatial coding," *J. Audio Eng. Soc.*, vol. 51, no. 6, pp. 554–569, June 2003.

[2] C. Kyriakakis, P. Tsakalides, and T. Holman, "Surrounded by sound," *IEEE Signal Processing Mag.*, vol. 16, no. 1, pp. 55–66, Jan. 1999.

[3] F. Rumsey, *Spatial Audio*. Oxford, England: Focal Press, 2001.

[4] J. P. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization* (Revised Edition). Cambridge, MA: MIT Press, 1997.

[5] D. R. Begault. (1994). *3-D Sound for Virtual Reality and Multimedia*, Boston, MA: Academic [Online]. Available: http://human-factors.arc.nasa.gov/publications/Begault_2000_3d_Sound_Multimedia.pdf

[6] D. B. Begault, "Auditory and non-auditory factors that potentially influence virtual acoustic imagery," in *Proc. AES 16th Int. Conf. Spatial Sound Reproduction*, Rovaniemi, Finland, Apr. 1999.

[7] E. A. Macpherson and J. C. Middlebrooks, "Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited," *J. Acoust. Soc. Amer.*, vol. 111, pt. 1, no. 5, pp. 2219–2236, May 2002.

[8] J. C. Middlebrooks and D. M. Green, "Directional dependence on interaural envelope delays," *J. Acoust. Soc. Am.*, vol. 87, no. 5, pp. 2149–2162, May 1990.

[9] F. L. Wightman and D. J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *J. Acoust. Soc. Amer.*, vol. 91, no. 3, pp. 1648–1661, Mar. 1992.

[10] J. C. Middlebrooks, "Narrow-band sound localization related to external ear acoustics," *J. Acoust. Soc. Amer.*, vol. 92, no. 5, pp. 2607–2624, Nov. 1992.

[11] H. Wallach, "On sound localization," *J. Acoust. Soc. Amer.*, vol. 10, no. 4, pp. 270–274, 1939.

[12] V. R. Algazi, C. Avendano, and R. O. Duda, "Elevation localization and head-related transfer function analysis at low frequencies," *J. Acoust. Soc. Amer.*, vol. 109, no. 3, pp. 1110–1122, Mar. 2001.

[13] D. Pralong and S. Carlile, "The role of individualized headphone calibration for the generation of high fidelity virtual auditory space," *J. Acoust. Soc. Amer.*, vol. 100, no. 6, pp. 3785–3793, Dec. 1996.

[14] D. Griesinger, "Binaural techniques for music reproduction," in *Proc. Audio Engineering Society 8th Int. Conf.*, Washington, DC, 1990.

[15] M. B. Gardner, "Distance estimation of 0° or apparent 0°-oriented speech signals in anechoic space," *J. Acoust. Soc. Amer.*, vol. 45, no. 1, pp. 47–53, Jan. 1969.

[16] D. S. Brungart, "Auditory localization of nearby sources. III. Stimulus effects," *J. Acoust. Soc. Amer.*, vol. 106, no. 6, pp. 3589–3602, Dec. 1999.

[17] A. W. Bronkhorst and T. Houtgast, "Auditory distance perception in rooms," *Nature*, vol. 397, pp. 517–520, Feb. 1999.

[18] G. S. Kendall, "The decorrelation of audio signals and its impact on spatial imagery," *Comput. Music J.*, vol. 19, no. 4, pp. 71–87, 1995.

[19] F. L. Wightman and D. J. Kistler, "Resolution of front-back ambiguity in spatial hearing by listener and source movement," *J. Acoust. Soc. Amer.*, vol. 105, no. 5, pp. 2841–2853, May 1999.

[20] F. L. Wightman and D. J. Kistler, "Headphone simulation of free-field listening. I. Stimulus synthesis," *J. Acoust. Soc. Amer.*, vol. 85, no. 2, pp. 858–867, Feb. 1989.

[21] R. Nicol, *Binaural Technol.* New York: Audio Eng. Soc., 2010.

[22] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, "Binaural technique: Do we need individual recordings," *J. Audio Eng. Soc.*, vol. 44, no. 6, pp. 451–469, June 1996.

[23] A. Kulkarni and H. S. Colburn, "Role of spectral detail in sound-source localization," *Nature*, vol. 396, pp. 747–749, Dec. 1998.

[24] T. Ajdler, C. Faller, L. Sbaiz, and M. Vetterli, "Sound field analysis along a circle and its application to HRTF interpolation," *J. Audio Eng. Soc.*, vol. 56, no. 3, pp. 156–175, Mar. 2008.

[25] D. J. Kistler and F. L. Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," *J. Acoust. Soc. Am.*, vol. 91, pp. 1637–1647, Mar. 1992.

[26] R. Duraiswami, D. N. Zotkin, and N. A. Gumerov, "Interpolation and range extrapolation of HRTFs," in *Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing (ICASSP'04)*, May 2004, vol. 4, pp. iv–45–iv–48.

[27] R. L. Jenison and K. Fissell, "A spherical basis function neural network for modeling auditory space," *Neural Comput.*, vol. 8, no. 1, pp. 115–128, 1996.

[28] Y. Haneda, S. Makino, Y. Kaneda, and N. Kitawaki, "Common-acoustical-pole and zero modeling of head-related transfer functions," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 2, pp. 188–195, Mar. 1999.

[29] C. P. Brown and R. O. Duda, "A structural model for binaural sound synthesis," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 5, pp. 476–488, Sept. 1998.

[30] D. S. Brungart and K. R. Scott, "The effects of production and presentation level on the auditory distance perception of speech," *J. Acoust. Soc. Amer.*, vol. 110, no. 1, pp. 425–440, July 2001.

[31] P. Mackensen, U. Felderhoff, G. Theile, U. Horbach, and R. Pellegrini, "Binaural room scanning—A new tool for acoustic and psychoacoustic research," *J. Acoust. Soc. Amer.*, vol. 105, no. 2, pp. 1343–1344, Feb. 1999.

[32] J.-M. Jot, "Real-time spatial processing of sounds for music, multimedia and interactive human-computer interfaces," *Multimedia Syst.*, vol. 7, no. 1, pp. 55–69, 1999.

[33] M. A. Gerzon, "Ambisonics in multichannel broadcasting and video," *J. Audio Eng. Soc.*, vol. 33, no. 11, pp. 859–871, Nov. 1985.

[34] R. Duraiswami, D. N. Zotkin, Z. Li, E. Grassi, N. A. Gumerov, and L. S. Davis, "High order spatial audio capture and its binaural head-tracked playback over headphones with HRTF cues," in *Proc. 119th Convention of the Audio Engineering Society*, Preprint 6540, New York, NY, Oct. 2005.

[35] D. S. McGrath, "Methods and apparatus for processing spatial audio," U.S. Patent 6 259 795, July 2001.

[36] D. de Vries, *Wave Field Synthesis*. New York: Audio Eng. Soc., 2010.

[37] V. R. Algazi, R. O. Duda, and D. M. Thompson, "Motion-tracked binaural sound," *J. Audio Eng. Soc.*, vol. 52, no. 11, pp. 1142–1156, Nov. 2004.

[38] V. R. Algazi, R. O. Duda, and D. Thompson, "Dynamic binaural sound capture and reproduction," U.S. Patent 7 333 622, Feb. 2008.

[39] J. B. Melick, V. R. Algazi, R. O. Duda, and D. M. Thompson, "Customization for personalized rendering of motion-tracked binaural sound," in *Proc. 117th Convention of the Audio Engineering Society*, Preprint 6225, San Francisco, CA, Oct. 2004, p. 20.

[40] V. R. Algazi and R. O. Duda, "Immersive spatial sound for mobile multimedia," in *Proc. IEEE Int. Symp. Multimedia (ISM'05)*, Irvine, CA, Dec. 2005, pp. 739–746.

[41] M. Kleiner, B.-I. Dalenbäck, and P. Svensson, "Auralization—An overview," *J. Audio Eng. Soc.*, vol. 41, no. 11, pp. 861–875, Nov. 1993.

[42] U. Zölzer, *Digital Audio Signal Processing*. Chichester, England: Wiley, 1997.

[43] J. Huopaniemi, "Future of personal audio—Smart applications and immersive communication," in *Proc. AES 30th Int. Conf. Intelligent Audio Environments*, Saariselkä, Finland, Mar. 2007.