Fundamental and Technological Limitations of Immersive Audio Systems

CHRIS KYRIAKAKIS, MEMBER, IEEE

Numerous applications are currently envisioned for immersive audio systems. The principal function of such systems is to synthesize, manipulate, and render sound fields in real time. In this paper, we examine several fundamental and technological limitations that impede the development of seamless immersive audio systems. Such limitations stem from signal-processing requirements, acoustical considerations, human listening characteristics, and listener movement. We present a brief historical overview to outline the development of immersive audio technologies and discuss the performance and future research directions of immersive audio systems with respect to such limits. Last, we present a novel desktop audio system with integrated listener-tracking capability that circumvents several of the technological limitations faced by today's digital audio workstations.

Keywords*—*Acoustic signal processing, audio systems, auditory system, multimedia systems, signal processing.

I. INTRODUCTION

Emerging integrated media systems seamlessly combine digital video, digital audio, computer animation, text, and graphics into common displays that allow for mixed media creation, dissemination, and interactive access in *real time*. Immersive audio and video environments based on such systems can be envisioned for applications that include teleconferencing and telepresence; augmented and virtual reality for manufacturing and entertainment; air-traffic control, pilot warning, and guidance systems; displays for the visually or aurally impaired; home entertainment; distance learning; and professional sound and picture editing for television and film. The principal function of immersive systems is to synthesize multimodal perceptions that do not exist in the current physical environment, thus immersing users in a seamless blend of visual and aural information. Significant resources have been allocated over the past 20 years to promote research in the area of image and video processing, resulting in important advances in these fields. On the other hand, audio signal processing, and particularly immersive audio, have been largely neglected.

Accurate spatial reproduction of sound can significantly enhance the visualization of three-dimensional (3-D) information for applications in which it is important to achieve sound localization relative to visual images. The human ear–brain interface is uniquely capable of localizing and identifying sounds in a 3-D environment with remarkable accuracy. For example, human listeners can detect time-of-arrival differences of about 7 μs. Sound perception is based on a multiplicity of cues that include level and time differences and direction-dependent frequency-response effects caused by sound reflection in the outer ear, head, and torso, cumulatively referred to as the head-related transfer function (HRTF). In addition to such directional cues, human listeners use a multiplicity of other cues in the perception of timbre, frequency response, and dynamic range. Furthermore, there are numerous subjective sound qualities that vary from listener to listener but are equally important in achieving the "suspension of disbelief" desired in an immersive audio system. These include attributes such as the apparent source width, listener envelopment, clarity, and warmth [1], [2]. Vision also plays an important role in localization and can overwhelm the aural impression. In fact, a mismatch between the aurally perceived and visually observed positions of a particular sound causes a cognitive dissonance that can seriously limit the visualization enhancement provided by immersive sound. The amount of mismatch required to cause such a dissonance is subjective and can vary in both the level of perception and annoyance. For professional sound designers, a mere 4° offset in the horizontal plane between the visual and aural image is perceptible, whereas it takes a 15° offset before the average layperson will notice [3].

In this paper, we discuss several issues that pertain to immersive audio system requirements that arise from fundamental physical limitations as well as current technological drawbacks. We will address these issues from three complementary perspectives: identification of fundamental physical limitations that affect the performance of immersive audio systems, evaluation of the current status of immersive audio system development with respect to

Manuscript received September 8, 1997; revised December 4, 1997. The Guest Editor coordinating the review of this paper and approving it for publication was T. Chen.

The author is with the Integrated Media Systems Center, University of Southern California, Los Angeles, CA 90089-2564 USA (e-mail: ckyriak@imsc.usc.edu).

Publisher Item Identifier S 0018-9219(98)03283-6.

such fundamental limits, and delineation of technological considerations that affect present and future system design and development. In the final sections, we will present a novel sound-reproduction system that addresses several of the current technological limitations that currently affect the quality of audio at the desktop. This system incorporates a video-based tracking method that allows real-time processing of the audio signal in response to listener movement.

II. The Nature of Limitations in Immersive Audio Systems

There are two classes of limitations that impede the implementation of immersive audio systems. The first class encompasses fundamental limitations that arise from physical laws, and its understanding is essential for determining the feasibility of a particular technology with respect to the absolute physical limits. Many such fundamental limitations are not directly dependent on the choice of systems but instead pertain to the actual process of sound propagation and attenuation in irregularly shaped rooms. The physical properties of the acoustic environment are *encoded* in the sound field and must be *decoded* by an immersive audio system in order to accurately simulate the original environment. The influence of the local acoustic environment is reflected in the perception of spatial attributes such as direction and distance, as well as in the perception of room spaciousness and source size [1], [2]. The situation is further complicated by the fact that the decoding process must include the transformations associated with human hearing. These include the conversion of spatial sound cues into level and time differences and direction-dependent frequency-response effects caused by the pinna, head, and torso through a set of amplitude and phase transformations known as the HRTF's. The *seamless* incorporation of such cues in immersive audio systems is a very active area of research that, if successful, will give rise to systems that begin to approach performance near the fundamental limits.

The second class of limitations consists of constraints that arise purely from technological considerations. These are equally useful in understanding the potential applications of a given system and are imposed by the particular technology chosen for system implementation. For example, the process of encoding parameters associated with room acoustics into sound fields can be modeled using numerical methods. In theory, this would involve the solution of the wave equation for sound subject to the boundary conditions dictated by the complex (absorptive, reflective, and diffractive) room surfaces. The computational complexity of this problem is very high and involves the calculation of estimated 10^{10} normal modes that fall within the range of human hearing (20 Hz–20 kHz) for a large hall [4]. More recent methods have been developed for rendering sound fields through a process called auralization. Such methods utilize a combination of scaled models, digital filtering, and special-purpose hardware for real-time convolution to predict and render the sound field [5]. As the processing power of digital signal processing (DSP) hardware increases, the capability of auralization systems to render complex sound fields will increase proportionally.

III. Brief Historical Overview

A. Two-Channel Stereo

Although many of the principles of stereophonic sound were developed through research efforts in the early 1930's, there still remains a misconception as to the meaning of the word "stereo" itself. While it is generally associated with sound reproduction from two loudspeakers, the word originates from the Greek *stereos,* meaning solid or three-dimensional. The two-channel association came about in the 1950's because of technological limitations imposed by the phonograph record that had only two groove walls for encoding information.

Stereophony started with the work of Blumlein [6] in the United Kingdom, who recognized early on that it was possible to locate a sound within a range of azimuth angles by using an appropriate combination of delay and level differences. His work focused on accurate reproduction of the sound field at each ear of the listener and on the development of microphone techniques that would allow the recording of the amplitude and phase differences necessary for stereo reproduction. Fletcher, Steinberg, and Snow at Bell Laboratories in the United States [7]–[9] took a different approach. They considered a "wall of sound" in which an infinite number of microphones is used to reproduce a sound field through an infinite number of loudspeakers, similar to the Huygens principle of secondary wavelets. While this made for an interesting theoretical result, the Bell Labs researchers realized that practical implementations would require a significantly smaller number of channels. They showed that a three-channel system consisting of left, center, and right channels in the azimuth plane could represent the lateralization and depth of the desired sound field with acceptable accuracy [Fig. 1(a)]. The first such stereophonic three-channel system was demonstrated in 1934 with the Philadelphia Orchestra performing remotely for an audience in Washington, DC, over wide-band telephone lines.

B. Four-Channel Matrixed Quadraphonic System

While stereophonic methods can be a powerful tool in the reproduction of the spatial attributes of a sound field, they fall short of true three-dimensional reproduction. The quadraphonic system attempted to circumvent such limitations by capturing and transmitting information about the direct sound and the reverberant sound field [10], [11]. To deliver the four channels required by quadraphonic recordings over a two-channel medium (e.g., the phonograph record), it was necessary to develop an appropriate encoding and decoding scheme. Several such schemes were proposed based on $4:2:4$ matrix encoding/decoding that relied on phase manipulation of the original stereo signals [12]. Quadraphonic systems were capable of reproducing sound images fairly accurately in the front and rear sectors

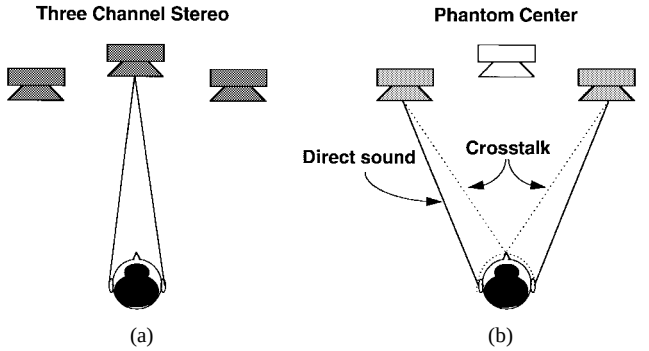

Fig. 1. (a) Stereo was originally invented based on three loudspeakers. Sound images in the center are rendered by a real loudspeaker that delivers the same direct sound to both ears. (b) In the two-loudspeaker stereo configuration with a phantom center, the cross-talk terms give rise to a less stable center image as well as a loss of clarity.

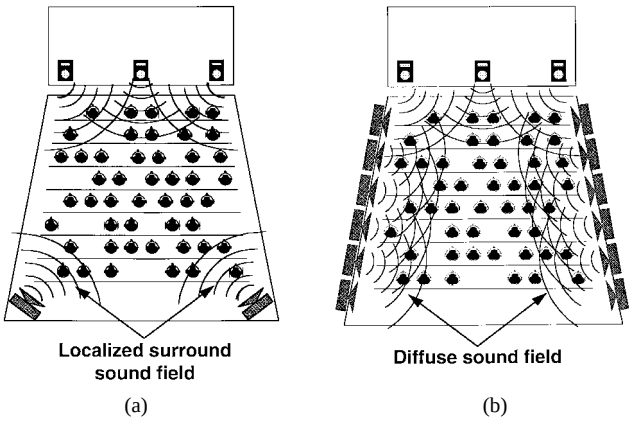

Fig. 2. (a) In early surround-sound systems with a mono surround channel, listeners seated off-center perceived the sound as if it originated from the effects loudspeaker that was closest to them, thus destroying the desired sense of envelopment. (b) Current systems use stereo surrounds reproduced over an array of loudspeakers along the sides of the theater to create a more diffuse sound field.

of the azimuthal plane, but they exhibited serious limitations when attempting to reproduce sound images to the side of the listener. Experiments showed [13], [14] that this was a limitation associated as much with sound-field synthesis using only four channels as with human psychoacoustic mechanisms. These technical limitations and the presence of two competing formats in the consumer marketplace contributed to the early demise of quadraphonic systems.

C. Multichannel Surround Sound

In the early 1950's, the first multichannel sound format was developed by 20th Century Fox. The combination of wide-screen formats such as CinemaScope (35 mm) and six-track Todd-AO (70 mm) with multichannel sound was the film industry's response to the growing threat of television. Stereophonic film sound was typically reproduced over three front loudspeakers, but these new formats included an additional monophonic channel that was reproduced over two loudspeakers behind the audience and was known as the effects channel. This channel increased the sense of space for the audience, but it also suffered from a serious technological limitation. Listeners seated on-center with respect to the rear loudspeakers perceived "inside-the-head" localization similar to the effect of stereo images reproduced over headphones. Listeners seated off-center localized the channel to the effects loudspeaker that was closest to them as dictated by the law of the first-arriving wavefront, thus destroying the sense of envelopment desired [14] [Fig. 2(a)]. The solution to these problems was found by introducing a second channel reproduced over an array of loudspeakers along the sides of the theater to create a more diffuse sound field [Fig. 2(b)].

In the mid-1970's, a new sound technology was introduced by Dolby Laboratories called Dolby Stereo. It was based on the optical technology that had been used for sound on film since the 1930's, and it circumvented the problems associated with magnetic multitrack recording. Dolby developed a matrix method for encoding four channels (left, center, right, and mono surround) into two channels using a technique derived from the matrix methods

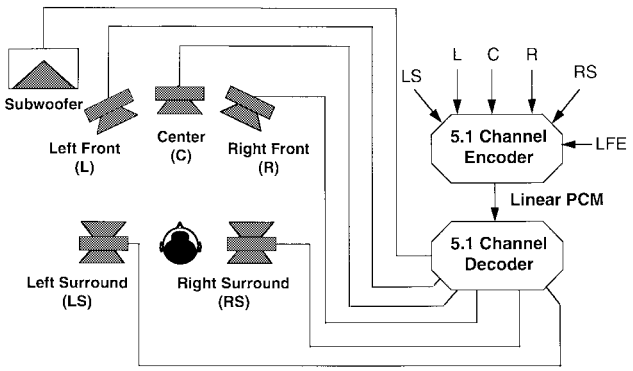

Fig. 3. Current commercial multichannel systems encode the LFE, three front, and two surround channels into a bit stream that is decoded at the user's end. With proper loudspeaker selection and placement, it is possible to simulate the experience of a movie theater. Dipole surround loudspeakers that do not radiate sound directly in the direction of the listener's ears produce the best envelopment.

used in quadraphonic systems but also ensured mono and stereo backward compatibility. In 1992, further enhancements by Dolby were introduced through a new format called Dolby Stereo Digital (SR•D). This format eliminated matrix-based encoding and decoding and provided five discrete channels (left, center, right, and independent left and right surround) in a configuration known as stereo surround. A sixth, low-frequency-enhancement (LFE) channel was introduced to add more head room and prevent the main speakers from overloading at low frequencies. The bandwidth of the LFE channel is limited between 0 and 120 Hz, a frequency regime that is outside the localization range for human listeners in a reverberant room, thus simplifying the placement requirements for the subwoofer used for LFE reproduction (Fig. 3).

Recent advances in digital audio compression and optical storage have made it possible to deliver up to six discrete audio channels in a consumer format centered around

the Dolby AC-3 compression scheme.[1] With exciting new formats such as an audio-only digital video disc just around the corner, the number of channels could easily increase to ten or more. While there are several million consumer systems capable of reproducing more than two channels, the majority of users (particularly those with desktop computer systems) would find the use of multiple loudspeakers impractical. In the sections that follow, we examine the requirements of systems that allow delivery of multiple channels over two loudspeakers using DSP to simulate certain characteristics of human listening.

IV. Spatial (3-D) Audio

A. Physiological Signal Processing

The human hearing process is based on the analysis of input signals to the two ears for differences in intensity, time of arrival, and directional filtering by the outer ear. Several theories were proposed as early as 1882 [15] that identified two basic mechanisms as being responsible for source localization: 1) interaural time differences (ITD's) and 2) interaural level differences (ILD's). A later theory by Lord Rayleigh [16] was based on a combination of ITD and ILD cues that operated in different wavelength regimes. For short wavelengths (corresponding to frequencies in the range of about 4–20 kHz), the listener's head casts an acoustical shadow giving rise to a lower sound level at the ear farthest from the sound source (ILD) [Fig. 4(b)]. At long wavelengths (corresponding to frequencies in the range of about 20 Hz–1 kHz), the head is very small compared to the wavelength, and localization is based on perceived differences in the time of arrival of sound at the two ears (ITD) [Fig. 4(a)]. The two mechanisms of interaural time and level differences formed the basis of what became known as the *duplex* theory of sound localization. In the frequency range between approximately 1 and 4 kHz, both of these mechanisms are active, which results in several conflicting cues that tend to cause localization errors.

While time or intensity differences provide source direction information in the horizontal (azimuthal) plane, in the median plane, time differences are constant and localization is based on *spectral filtering*. The reflection and diffraction of sound waves from the head, torso, shoulders, and pinnae, combined with resonances caused by the ear canal, form the physical basis for the HRTF. The outer ear can be modeled (in the static case) as a linear time-invariant system that is fully characterized by the HRTF in the frequency domain. As Blauert [17] describes it, the role of the outer ear is to superimpose angle- and distance-specific linear distortions on the incident sound signal. Spatial information is thus encoded onto the signals received by the eardrums through a combination of direction-dependent and direction-independent filters [18], [19]. The magnitude and phase of these head-related transfer functions vary significantly for each sound direction but also from person to person.

[1] See Dolby Laboratories at http://www.dolby.com.

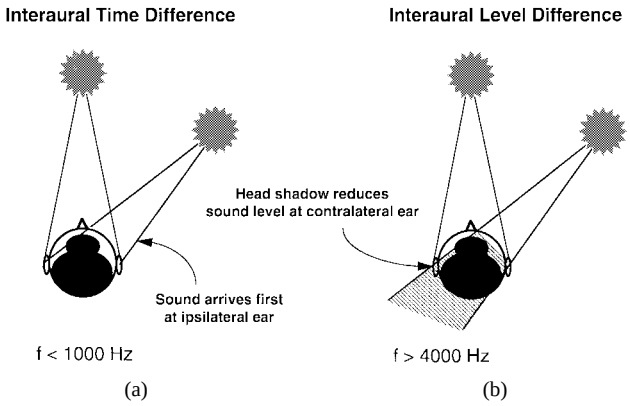

Fig. 4. (a) In the low-frequency regime, sound is localized based on differences in the time of arrival at each ear. (b) At higher frequencies, the wavelength of sound is short relative to the size of the head, and localization is based on perceived level differences caused by head shadowing. In the intermediate-frequency region, both mechanisms are active, and this can give rise to conflicting cues.

The emerging field of 3-D audio is based on digital implementations of such HRTF's. In principle, it is possible to achieve excellent reproduction of three-dimensional sound fields using such methods; however, it has been demonstrated that this requires precise measurement of each listener's individual HRTF's [20]. This seemingly fundamental requirement that derives from inherent physiological and cognitive characteristics of the human ear–brain interface has rendered such systems impractical for widespread use. Current research in this area is focused on achieving good localization performance while using synthetic (nonindividualized) HRTF's derived through averaging or modeling or based on the HRTF's of subjects that have been determined to be "good localizers" [21], [22]. In his review of the challenges in 3-D audio implementations, Begault [23] points out that there are currently three major barriers to successful implementation of such systems: 1) psychoacoustic errors such as front–back reversals typical in headphone-based systems, 2) large amounts of data required to represent measured HRTF's accurately, and 3) frequency- and phase-response errors that arise from mismatches between nonindividualized and measured HRTF's. It should be noted that front–back reversals can be reduced if the listener is allowed to move his head and that the lack of externalization experienced with headphone listening can be alleviated with appropriate use of reverberation.

A fourth challenge arises from technological limitations of current computing systems. One capability that we envision for immersive audio systems is the simulation of room acoustics and listener characteristics for interactive, virtual-, and augmented-reality applications. In addition to the computational requirements for photorealistic rendering of visual images, the synthesis of such acoustical environments requires computation of the binaural room response and subsequent convolution with the HRTF's of the listener *in real time* as the listener moves around the room. Typical impulse response duration is 3 s, which, when sampled at 48 kHz, requires a processor capable of operating at

more than 13 Gflops/channel [24]. This problem can be circumvented using special-purpose hardware or hybrid block fast Fourier transform/direct convolution methods [25].[2] The main goal is to reduce the number of operations for such computations, thus making them suitable for real-time interactive applications.

B. Spatial Audio Rendering

A critical issue in the implementation of immersive audio is the reproduction of 3-D sound fields that preserve the desired spatial location, frequency response, and dynamic range. There are two general methods for 3-D audio rendering that can be categorized as "head related" based on headphone reproduction and "nonhead related" based on loudspeaker reproduction [19]. A hybrid category, called transaural stereo, also exists that allows loudspeaker rendering of head-related signals. It should be noted that there are other methods for three-dimensional sound-field capture and synthesis such as ambisonics [26] and wave-field synthesis [27], [28], but these will not be examined in this paper.

Nonheadroom-related methods typically use multiple loudspeakers to reproduce multiple matrixed or discrete channels. Such systems can convey precisely localized sound images that are primarily confined to the horizontal plane and diffuse (ambient) sound to the sides of and behind the listener. In addition to the left and right loudspeakers, they make use of a center loudspeaker that helps create a solidly anchored center-stage sound image, as well as two loudspeakers for the ambient surround sound field. The most prevalent systems currently available to consumers are based on the formats developed by Dolby for film sound, including Pro Logic (four-channel matrixed encoded on two channels) and Dolby Digital (5.1-channel discrete based on the AC-3 compression scheme).[1] Other 5.1-channel schemes include DTS Digital Surround[3] and MPEG-2. Multichannel systems were designed primarily for authentic reproduction of sound associated with movies but have recently started to be used for music recordings and games on CD-ROM. The 5.1-channel Dolby Digital system was adopted in the U.S. standard of the upcoming advanced (high-definition) television system [29]. The design requirements for such loudspeaker-based systems include uniform audience coverage, accurate localization relative to visual images on the screen, diffuse rendering of ambient sounds, and capability for reproduction of the wide (up to 105 dB) dynamic range present in film soundtracks.

Head-related binaural recording, or dummy-head stereophony, methods attempt to accurately reproduce at each eardrum of the listener the sound pressure generated by a set of sources and their interactions with the acoustic environment [30]. Such recordings can be made with specially designed probe microphones that are inserted in the listener's ear canal or by using a dummy-head microphone system that is based on average human characteristics. Sound recorded using binaural methods is then reproduced through headphones that deliver the desired sound to each ear. It was concluded from early experiments that in order to achieve the desired degree of realism using binaural methods, the required frequency-response accuracy of the transfer function was ± 1 dB [31]. Other related work [32] compared direct listening and binaural recordings for the same subject and concluded that directional hearing was accurately preserved using binaural recording.

While there are several commercially available dummy-head systems, binaural recordings are not widely used primarily due to limitations that are associated with headphone listening [20], [33], [34]. These drawbacks can be summarized as follows.

1) Individualized HRTF information does not exist for each listener and the averaged HRTF's that are used make it impossible to match each individual's perception of sound.
2) There are large errors in sound position perception associated with headphones, especially for the most important visual direction, out in front.
3) Headphones are uncomfortable for extended periods of time.
4) It is very difficult to externalize sounds and avoid the "inside-the-head" sensation.

In many applications, however, such as in aircraft cockpits or multiuser environments, the use of headphones is required for practical reasons.

The use of loudspeakers for reproduction can circumvent the limitations associated with headphone reproduction of binaural recordings. To deliver the appropriate binaural sound field to each ear, however, it is necessary to eliminate the cross talk that is inherent in all loudspeaker-based systems. This is a technological limitation of *all* loudspeaker systems, and it arises from the fact that while each ear receives the desired sound from the same-side (ipsilateral) loudspeaker, it also receives undesired sound from the opposite-side (contralateral) loudspeaker.

Several schemes have been proposed to address cross-talk cancellation. The basic principle of such schemes relies on preconditioning the signal into each loudspeaker such that the output sound generates the desired binaural sound pressure at each ear. If we denote the sound pressures that must be delivered to each ear as P_L(ear) and P_R(ear) and the transfer functions from each loudspeaker to each ear as H_{LL}, H_{LR}, H_{RL}, and H_{RR}, then we can write (Fig. 5)

$$\begin{aligned} P_L(\text{speaker}) &= H_{LL}S_L + H_{RL}S_R \\ P_R(\text{speaker}) &= H_{LR}S_L + H_{RR}S_R \end{aligned} \tag{1}$$

in which we denote by S_L and S_R the input signals to each loudspeaker and $P_{L,R}$(speaker) the sound pressure delivered by each loudspeaker. To accurately reproduce the desired binaural signal at each ear, the input signals S_L and

[2] See Lake DSP at http://www.lakedsp.com.

[3] See http://www.dtstech.com/.

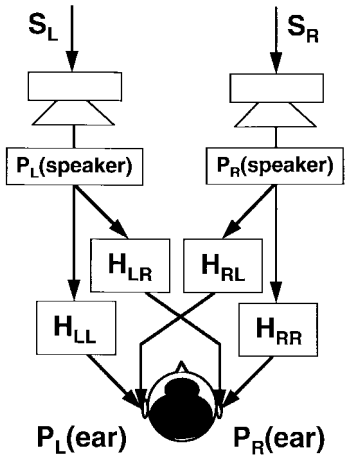

Fig. 5. Transfer functions associated with a loudspeaker sound-rendering system. To deliver the correct binaural sound, it is necessary to prefilter the signal to the loudspeakers so that the cross-talk terms H_{LR} and H_{RL} are cancelled during reproduction.

S_R must be chosen such that

$$P_L(\text{ear}) = P_L(\text{speaker})$$
$$P_R(\text{ear}) = P_R(\text{speaker}). \tag{2}$$

The desired loudspeaker input signals are then found from

$$S_L = \frac{H_{RR}P_L(\text{ear}) - H_{RL}P_R(\text{ear})}{H_{LL}H_{RR} - H_{LR}H_{RL}}$$
$$S_R = \frac{H_{LL}P_R(\text{ear}) - H_{LR}P_L(\text{ear})}{H_{LL}H_{RR} - H_{LR}H_{RL}}. \tag{3}$$

The only requirement is that S_L and S_R must be realizable filter responses. The first such cross-talk cancellation scheme was proposed by Bauer [35], later by Atal and Schroeder [4], and by Damaske and Mellert [36], [37] using a system called "True Reproduction of All Directional Information by Stereophony" (TRADIS). The main limitation of these early systems was the fact that any listener movement that exceeded 75–100 mm completely destroyed the spatial effect. Cooper and Bauck [38], [39] showed that under the assumption of left–right symmetry, a much simpler shuffler filter can be used to implement cross-talk cancellation as well as synthesize virtual loudspeakers in arbitrary positions. They went on to use results from Mehrgard and Mellert [40], who showed that the head-related transfer function is minimum phase to within a frequency-independent delay that is a function of the angle of incidence. This new "transaural" system significantly reduced the computational requirements by allowing implementations that use simple finite-duration impulse response filters.

The functionality and practical use of immersive audio systems based on such transaural cross-talk cancellation methods can be greatly enhanced by eliminating the requirement that the user remain stationary with a fixed head position. This increased functionality requires the capability to implement the requisite filters (and associated algorithms) in real time. A further requirement is precise information about the location of the listener's ears relative to the loudspeakers. This is achieved, with reasonable accuracy, in desktop-based audio systems in which the listener is seated at the keyboard at a fixed distance from the loudspeakers. Ultimately, the listener's head (and ear) location must be tracked in order to allow for head rotation and translation. Several issues related to both the desktop and the tracking implementations are discussed below.

V. Immersive Audio Rendering for Desktop Applications

For desktop applications, in addition to the user-imposed limitation of (typically) two loudspeakers, there exists an entirely different set of design requirements specific to applications such as professional sound editing for film and television, teleconferencing and telepresence, augmented and virtual reality, and home personal-computer (PC) entertainment. Such applications require *high-quality* audio for a single listener in a desktop environment. Issues that must be addressed include the optimization of the frequency response over a given frequency range, the dynamic range, and stereo imaging subject to constraints imposed by room acoustics and human listening characteristics. Several problems are particular to the desktop environment, including frequency-response anomalies that arise due to the local acoustical environment, the proximity of the listener to the loudspeakers, and the acoustics associated with small rooms.

A. Acoustical Limitations

In a typical desktop sound-monitoring environment, delivery of stereophonic sound is achieved through two loudspeakers that are placed on either side of a video or computer monitor. This environment, combined with the acoustical problems of small rooms, causes severe problems that contribute to audible distortion of the reproduced sound. Among these problems, one of the most important is the effect of discrete early reflections [41]–[43]. It has been shown [43] that these reflections are the dominant source of monitoring nonuniformities. These nonuniformities appear in the form of frequency-response anomalies in rooms where the difference between the direct and reflected sound level for the first 15 ms is less than 15 dB [44], [45] (Fig. 6). High levels of reflected sound cause comb filtering in the frequency domain, which in turn gives rise to severe changes in timbre. The perceived effects of such distortions were quantified with psychoacoustic experiments [41], [46] that demonstrated their importance.

A solution that has been proposed to alleviate the problems of early reflections is near-field monitoring. In theory, the direct sound is dominant when the listener is very close to the loudspeakers, thus reducing the room effects to below audibility. In practice, however, there are several issues that must be addressed in order to provide high-quality sound [47]. One such issue relates to the large reflecting surfaces that are typically present near the loudspeakers. Strong reflections from a console or a video/computer monitor act as baffle extensions for the loudspeaker, resulting in a boost of

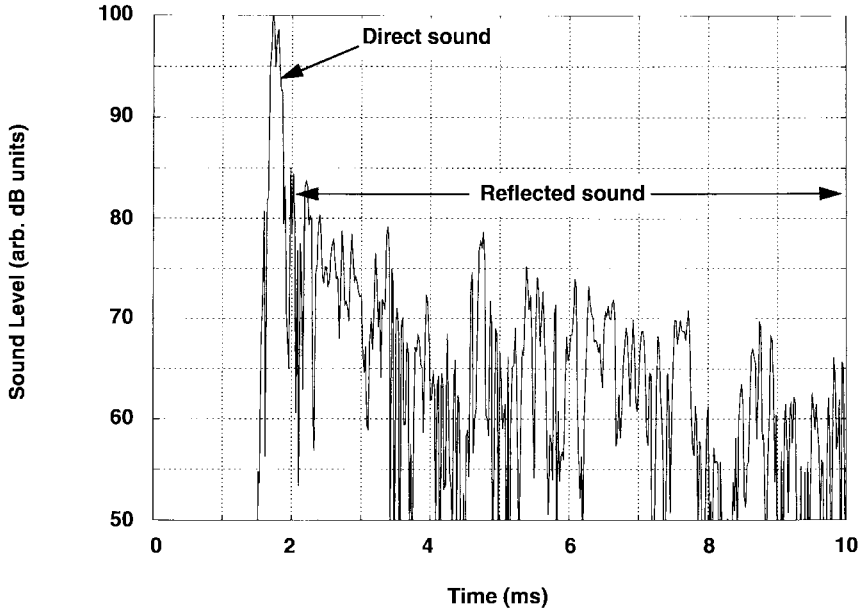

Fig. 6. The time-domain response of a loudspeaker system includes the direct sound as well as the sound due to multiple reflections from the local acoustical environment. Psychoacoustic evidence indicates that in order for these reflections not to be perceived, their spectrum level should be 15 dB below the level of the direct sound.

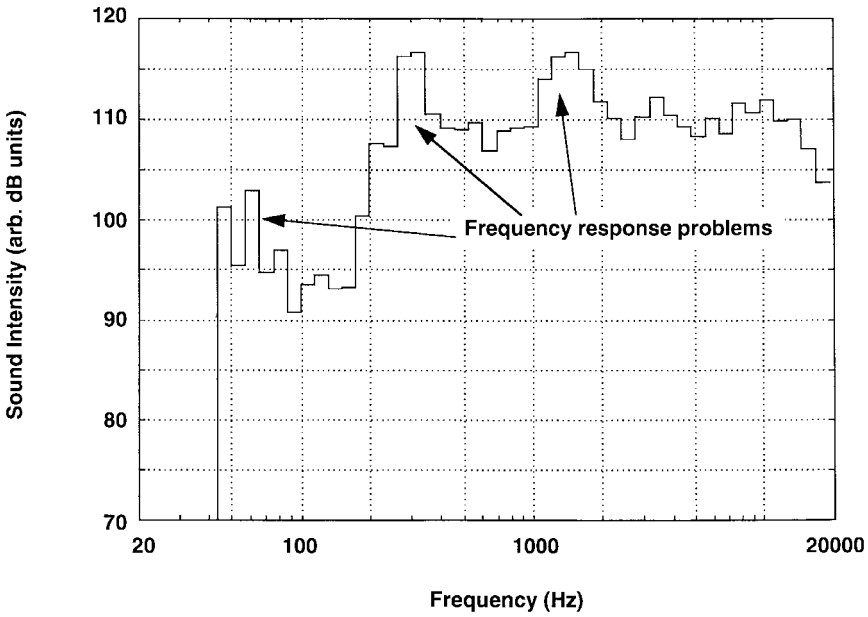

Fig. 7. Frequency-response problems that arise in the low frequencies due to standing-wave buildup in small rooms and in higher frequencies due to interactions with elements in the local acoustical environment (e.g., CRT screen, table top, untreated walls).

midbass frequencies. Furthermore, even if it were possible to place the loudspeakers far away from large reflecting surfaces, this would only solve the problem for middle and high frequencies. Low-frequency room modes do not depend on surfaces in the local acoustical environment but rather on the physical size of the room. These modes produce standing waves that give rise to large variations in frequency response (Fig. 7). Such amplitude and phase distortions can completely destroy carefully designed 3-D audio reproduction that relies on the transaural techniques described above.

B. Design Requirements

To circumvent these limitations, a set of solutions has been developed for single-listener desktop reproduction that delivers sound quality equivalent to a calibrated dubbing stage [43]. These solutions include direct-path dominant design and correct low-frequency response.

Based on our current understanding of psychoacoustic principles, it is possible to combine such cues to place the listener in a direct sound field that is dominant over the reflected and reverberant sound. The design considerations

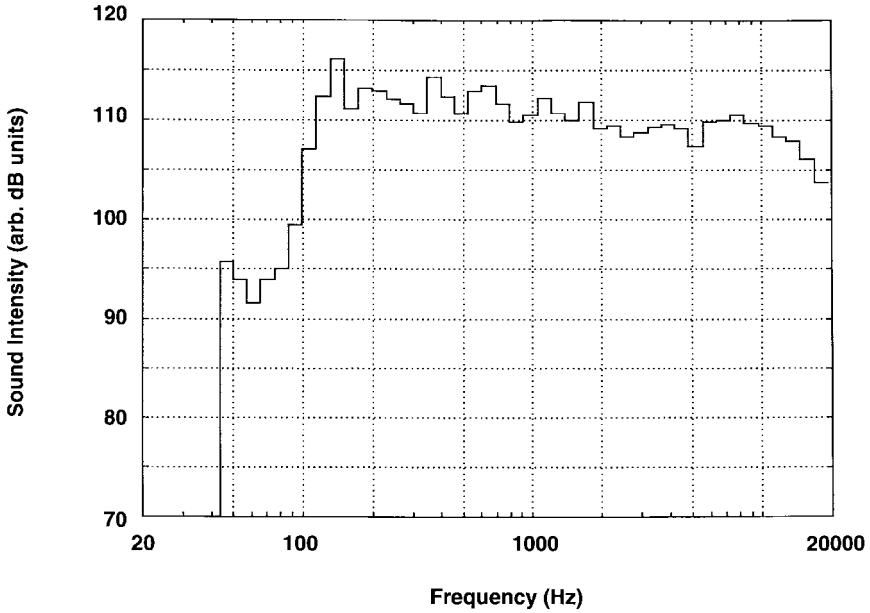

Fig. 8. A properly designed direct-path dominant system that compensates for frequency anomalies produces a much flatter frequency response. Frequencies below 100 Hz are reproduced with a separate subwoofer (response not shown) that is placed at a known distance from the listener to alleviate anomalies from standing waves.

for this direct-path dominant design include compensation of the physical (reflection and diffraction) effects of the video/computer monitor that extends the loudspeaker baffle as well as the large reflecting surface on which the computer keyboard typically rests. The distortions that arise from amplitude and phase anomalies are eliminated, and this results in a listening experience that is dramatically different than what is achievable through traditional near-field monitoring methods (Fig. 8).

Standing waves associated with the acoustics of small rooms give rise to fundamental limitations in the quality of reproduced sound, particularly in the uniformity of low-frequency response. Variations in this frequency regime can be as large as ± 15 dB for different listening locations in a typical room. The advantage of immersive audio rendering on desktop systems lies in the fact that the position of the loudspeakers and the listener are known *a priori*. It is therefore possible to use signal processing (equalization) to correct the low-frequency response. This smooth response, however, can only be achieved for a relatively small region around the listener. To correct over a larger region and compensate for listener movement, it is necessary to track the listener's position and use adaptive signal-processing methods that allow real-time correction of spatial as well as frequency-response attributes.

C. Listener-Location Considerations

In large rooms, multichannel sound systems are used to convey sound images that are primarily confined to the horizontal plane and are uniformly distributed over the audience area. Typical systems used for cinema reproduction use three front channels (left, center, right), two surround channels (left and right surround), and a separate low-frequency channel. Such 5.1-channel systems (a term coined by Holman to represent five full-spectrum channels and a low-frequency-only channel) are designed to provide accurate sound localization relative to visual images in front of the listener and diffuse (ambient) sound to the sides and behind the listener. The use of a center loudspeaker helps create a solid sound image between the left and right loudspeakers and anchors the sound to the center of the stage.

For desktop applications, in which a single user is located in front of a CRT display, we no longer have the luxury of a center loudspeaker because that position is occupied by the display. Size limitations prevent the front loudspeakers from being capable of reproducing the entire spectrum; thus, a separate subwoofer loudspeaker is used to reproduce the low frequencies. The two front loudspeakers can create a virtual (phantom) image that appears to originate from the exact center of the display provided that the listener is seated symmetrically with respect to the loudspeakers. With proper head and loudspeaker placement, it is possible to recreate a spatially accurate sound field with the correct frequency response in *one* exact position, the sweet spot. Even in this static case, however, the sound originating from each loudspeaker arrives at each ear at different times (about 200 μs apart), thereby giving rise to acoustic cross talk [Fig. 1(b)]. These time differences, combined with reflection and diffraction effects caused by the head, lead to frequency-response anomalies that are perceived as a lack of clarity [48].

This problem can be solved by adding a cross-talk cancellation filter (as described above in the description of transaural methods) to the signal of each loudspeaker. While this solution may be satisfactory for the static case, as soon as the listener moves even slightly, the conditions

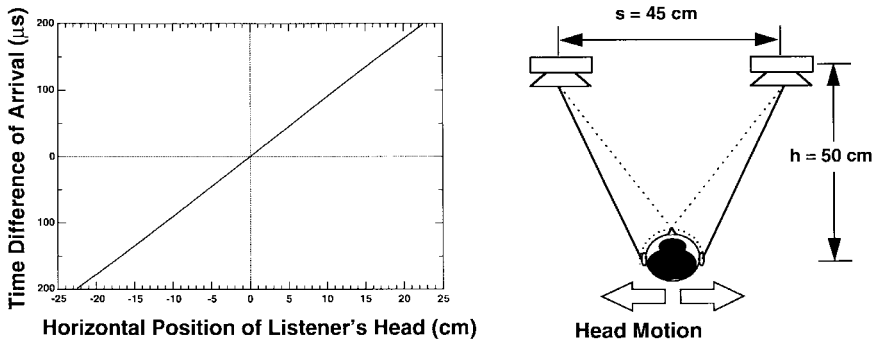

Fig. 9. Desktop sound system with vision-based head tracking. In this early prototype, the time difference of arrival at the two ears is adjusted in real time as the listener moves in the plane parallel to the loudspeakers. Current research is focused on tracking, pose estimation (for head rotations), and pinna shape recognition for real-time cross-talk cancellation and individualized HRTF synthesis.

for cancellation are no longer met, and the phantom image moves toward the closest loudspeaker because of the precedence effect. In order, therefore, to achieve the highest possible quality of sound for a nonstationary listener and preserve the spatial information in the original material, it is necessary to know the precise location of the listener relative to the loudspeakers [47], [49], [50]. In the section below, we describe an experimental system that incorporates a novel listener-tracking method in order to overcome the difficulties associated with two-ear listening as well as the technological limitations imposed by loudspeaker-based desktop audio systems.

VI. Future Research Directions

A. Vision-Based Methods for Listener Tracking

Computer vision has historically been considered problematic, particularly for tasks that require object recognition. Up to now, the complexity of vision-based approaches has prevented their being incorporated into desktop-based integrated media systems. Recently, however, von der Malsburg's Laboratory of Computational and Biological Vision at the University of Southern California (USC) has developed a vision architecture that is capable of recognizing the identity, spatial position (pose), facial expression, gesture identification, and movement of a human subject in *real time*.

This highly versatile architecture integrates a broad variety of visual cues in order to identify the location of a person's head within the image. Object recognition is achieved through pattern-based analysis that identifies convex regions with skin color that are usually associated with the human face and through a stereo algorithm that determines the disparities among pixels that have been moving [51]. This pattern-recognition approach is based on the elastic graph matching method that places graph nodes at appropriate fiducial points of the pattern [52]. A set of features is extracted at each graph node corresponding to the amplitudes of complex Gabor wavelets. The key advantage of this method is that a new pattern (face or ear) can be recognized on the basis of a small number of example images (10–100). For audio applications, in which the system must remember the last position of a listener that may have stopped moving, a hysteresis mechanism is used to estimate the current position and velocity of the head with a linear predictive filter.

While there are several alternative methods for tracking humans (e.g., magnetic, ultrasound, infrared, laser), they typically are based on tethered operations or require artificial fiducial markings (e.g., colored dots, earrings) to be worn by the user. Furthermore, these methods do not offer any additional functionality to match what can be achieved with vision-based methods (e.g., face and expression recognition, ear classification).

B. Desktop Audio System with Head Tracking

A novel multichannel desktop audio system that meets all the design requirements and acoustical considerations described above has been developed by Holman of TMH Corporation[4] in collaboration with the Immersive Audio Laboratory at USC's Integrated Media Systems Center (IMSC).[5] This system uses two loudspeakers that are positioned on the sides of a video monitor at a distance of 45 cm from each other and 50 cm from the listener's ears (Fig. 9). The seating position height is adjusted so that the listener's ears are at the tweeter level of the loudspeakers (117 cm from the floor), thus eliminating any colorations in the sound due to off-axis lobing. We have also incorporated the vision-based tracking algorithm described above using a standard video camera connected to an SGI Indy workstation. This tracking system provides us with the coordinates of the center of the listener's head relative to the loudspeakers and is currently capable of operating at 10 frames/s with a 3% accuracy.

In this single-camera system, it is possible to track listener movement that is confined in a plane parallel to loudspeakers and at a fixed distance from them. When the listener is located at the exact center position (the sweet spot), sound from each loudspeaker arrives at the corresponding ear at the exact same time (i.e., with zero ipsilateral time delay). At any other position of the listener

[4] See http://www.tmhlabs.com.

[5] See http://imsc.usc.edu.

in this plane, there is a relative time difference of arrival between the sound signals from each loudspeaker. To maintain proper stereophonic perspective, the ipsilateral time delay must be adjusted as the listener moves relative to the loudspeakers. The head coordinates provided from the tracking algorithm are used to determine the necessary time-delay adjustment. This information is processed by a 32-b DSP processor board (ADSP-2106x SHARC) resident in a Pentium-II PC. In this early version of our system, the DSP board is used to delay the sound from the loudspeaker that is closest to the listener so that sound arrives with the same time difference as if the listener were positioned in the exact center between the loudspeakers. In other words, we have demonstrated stereophonic reproduction with an adaptively optimized sweet spot.

We are currently in the process of identifying the bottlenecks of both the tracking and the audio signal-processing algorithms and integrating both into a single, PC-based platform for real-time operation. Furthermore, we are expanding the capability of the current single-camera system to include a second camera in a stereoscopic configuration that will provide distance (depth) information.

C. Pinna Classification for Enhanced Sound Localization

Immersive audio systems based on averaged HRTF's suffer from serious drawbacks. To map the entire three-dimensional auditory space requires a large number of tedious and time-consuming measurements, which is very difficult to do with human subjects. A further, and perhaps insurmountable, complication arises from the fact that this process must be repeated *for every* listener in order to produce accurate results. Last, discrete point measurements represent a quantization of 3-D space that is inherently continuous, thus requiring sophisticated interpolation algorithms that are computationally intensive and can give rise to errors [53]. Several methods have been proposed to overcome such limitations. Functional HRTF representations that make use of models to represent HRTF's have been proposed [54]–[56]; however, most are not suitable for real-time applications because they require significant computational resources.

There is significant evidence to suggest that the identification and incorporation of pinna physical characteristics may be a key factor limiting the development of seamless immersive audio systems. The human pinna is a rather complicated structure that for many years was considered to be a degenerate remain from past evolutionary forms. It was assumed to be a sound-collecting horn whose purpose was to direct sound into the ear canal. If this were true, then its physical dimensions would limit its role to a collector of high frequencies (short wavelengths). Experimental results, however, have shown that the pinna is a much more sophisticated instrument [17], [54], [57], [58]. The pinna folds act as miniature reflectors that create small time delays, which in turn give rise to comb-filtering effects in the frequency domain [59]. Also, the pinna is asymmetric relative to the opening of the ear canal. These ridges are arranged in such a way as to optimally translate a change in angle of the incident sound into a change in the pattern of reflections. It has been demonstrated [57] that the human ear–brain interface can detect delay differences as short as 7 μs. Furthermore, as the sound source is moved toward 180° in azimuth (directly behind the listener), the pinna also acts as a low-pass filter, thus providing additional localization cues.

To understand the fundamental limitations imposed by such pinna transformations, the IMSC Immersive Audio Laboratory, in collaboration with the USC Laboratory for Computational Vision, is developing a novel method for classification and cross comparison of pinna characteristics. We are currently in the process of implementing a data base of pinna images and associated measured directional characteristics (HRTF's). A picture of the pinna from every new listener allows us to select the HRTF from this data base that corresponds to the ear whose pinna shape is closest to the new ear. The algorithm that will be used for this identification is a modified version of the face-recognition algorithm described above. Initial results have shown successful matching of ears from unknown listeners to those already in our data base, including two artificial ears from the KEMAR dummy-head system. A planned extension of this matching method will select characteristics from several stored ears that best match the corresponding characteristics of the new pinna. An appropriate set of weighting factors will then be determined to form a synthetic HRTF that closely resembles that of the new listener. It is important to note that this method offers significant advantages over previous model-based attempts because it can be performed very fast and with minimum computational overhead.

ACKNOWLEDGMENT

The author wishes to thank Prof. T. Holman of the USC IMSC and TMH Corporation for his continued guidance and support, as well as Prof. C. von der Malsburg and Dr. H. Neven from the USC Laboratory for Computational Vision for the development and integration of the vision-based head-tracking and pinna classification algorithms.

REFERENCES

[1] Y. Ando, *Concert Hall Acoustics.* Berlin, Germany: Springer-Verlag, 1985.
[2] L. Beranek, *Concert and Opera Halls: How They Sound.* Woodbury, NY: Acoustical Society of America, 1996.
[3] S. Komiyama, "Subjective evaluation of angular displacement between picture and sound directions for HDTV sound systems," *J. Audio Eng. Soc.,* vol. 37, pp. 210–214, 1989.
[4] M. R. Schroeder and B. S. Atal, "Computer simulation of sound transmission in rooms," in *IEEE Int. Conv. Rec.*, 1963, vol. 7.
[5] M. Kleiner, B.-I. Dalenback, and P. Svensson, "Auralization—An overview," *J. Audio Eng. Soc.,* vol. 41, pp. 861–945, 1993.
[6] A. D. Blumlein, "Improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems," U.K. Patent 394 325, 1931.
[7] H. Fletcher, "Auditory perspective—Basic requirements," *Elect. Eng.,* vol. 53, pp. 9–11, 1934.
[8] W. B. Snow, "Basic principles of stereophonic sound," *SMPTE J.,* vol. 61, pp. 567–589, 1953.
[9] J. C. Steinberg and W. B. Snow, "Physical factors," *Bell Syst. Tech. J.,* vol. 13, pp. 245–258, 1934.

[10] P. Scheiber, "Quadrasonic sound system," U.S. Patent 3 632 886, 1973.
[11] ——, "Multidirectional sound system," U.S. Patent 3 746 792, 1973.
[12] D. H. Cooper and T. Shiga, "Discrete-matrix multichannel stereo," *J. Audio Eng. Soc.*, vol. 20, pp. 346–360, 1972.
[13] G. Theile and G. Plenge, "Localization of lateral phantom sources," *J. Audio Eng. Soc.*, vol. 25, pp. 196–199, 1977.
[14] T. Holman, "Channel crossing," *Studio Sound*, pp. 40–42, 1996.
[15] S. P. Thompson, "On the function of the two ears in the perception of space," *Philos. Mag.*, vol. 13, pp. 406–416, 1882.
[16] J. W. Strutt and L. Rayleigh, "On the perception of sound direction," *Philos. Mag.*, vol. 13, pp. 214–232, 1907.
[17] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, revised ed. Cambridge, MA: MIT Press, 1997.
[18] K. Genuit, "Ein modell zur beschreibung von außenohrübertragungseigenschaften," Ph.D. dissertation, RWTH Aachen, Germany, 1984.
[19] H. W. Gierlich, "The application of binaural technology," *Appl. Acoust.*, vol. 36, pp. 219–243, 1992.
[20] F. L. Wightman and D. J. Kistler, "Headphone simulation of free-field listening: Psychophysical validation," *J. Acoust. Soc. Amer.*, vol. 85, pp. 868–878, 1989.
[21] E. M. Wenzel, M. Arruda, and D. J. Kistler, "Localization using nonindividualized head-related transfer functions," *J. Acoust. Soc. Amer.*, vol. 94, pp. 111–123, 1993.
[22] D. R. Begault and E. M. Wenzel, "Headphone localization of speech," *Human Factors*, vol. 35, pp. 361–376, 1993.
[23] ——, "Challenges to the successful implementation of 3-D sound," *J. Audio Eng. Soc.*, vol. 39, pp. 864–870, 1991.
[24] H. Lehnert and J. Blauert, "Principles of binaural room simulation," *Appl. Acoust.*, vol. 36, pp. 259–291, 1992.
[25] W. G. Gardner, "Efficient convolution without input–output delay," *J. Audio Eng. Soc.*, vol. 43, pp. 127–136, 1995.
[26] M. A. Gerzon, "Ambisonics in multichannel broadcasting and video," *J. Audio Eng. Soc.*, vol. 33, pp. 859–871, 1985.
[27] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustics control by wavefield synthesis," *J. Acoust. Soc. Amer.*, vol. 93, p. 2764, 1993.
[28] M. M. Boone, E. N. G. Verheijen, and P. F. von Tol, "Spatial sound-field reproduction by wavefield synthesis," *J. Audio Eng. Soc.*, vol. 43, p. 1003, 1995.
[29] D. J. Meares, "Multichannel sound systems for HDTV," *Appl. Acoust.*, vol. 36, pp. 245–257, 1992.
[30] H. Moller, "Fundamentals of binaural technology," *Appl. Acoust.*, vol. 36, pp. 171–218, 1992.
[31] J. Blauert and P. Laws, "Group delay distortions in electroacoustical systems," *J. Acoust. Soc. Amer.*, vol. 63, pp. 1478–1483, 1978.
[32] H.-J. Platte, P. Laws, and H. v. Hövel, "Apparatus for the exact reproduction of ear input signals (in German)," *Fortschritte der Akustik*, vol. DAGA'75, pp. 361–364, 1975.
[33] F. L. Wightman and D. J. Kistler, "Headphone simulation of free-field listening: stimulus synthesis," *J. Acoust. Soc. Amer.*, vol. 85, pp. 858–867, 1989.
[34] H. Moller, C. B. Jensen, and D. Hammershoi, "Design criteria for headphones," *J. Audio Eng. Soc.*, vol. 43, pp. 218–232, 1995.
[35] B. B. Bauer, "Stereophonic earphones and binaural loudspeakers," *J. Audio Eng. Soc.*, vol. 9, pp. 148–151, 1961.
[36] P. Damaske and V. Mellert, "A procedure for generating directionally accurate sound images in the upper half-space using two loudspeakers," *Acustica*, vol. 22, pp. 154–162, 1969.
[37] P. Damaske, "Head related two channel stereophony with loudspeaker reproduction," *J. Acoust. Soc. Amer.*, vol. 50, pp. 1109–1115, 1971.
[38] D. H. Cooper and J. L. Bauck, "Prospects for transaural recording," *J. Audio Eng. Soc.*, vol. 37, pp. 3–19, 1989.
[39] J. Bauck and D. H. Cooper, "Generalized transaural stereo and applications," *J. Audio Eng. Soc.*, vol. 44, pp. 683–705, 1996.
[40] S. Mehrgard and V. Mellert, "Transformation characteristics of the external human ear," *J. Acoust. Soc. Amer.*, vol. 51, pp. 1567–1576, 1977.
[41] F. E. Toole, "Loudspeaker measurements and their relationship to listener preferences," *J. Audio Eng. Soc.*, vol. 34, pp. 227–235, 1986.
[42] S. Bech, "Perception of timbre of reproduced sound in small rooms: Influence of room and loudspeaker position," *J. Audio Eng. Soc.*, vol. 42, pp. 999–1007, 1994.
[43] T. Holman, "Monitoring sound in the one-person environment," *SMPTE J.*, vol. 106, pp. 673–680, 1997.
[44] R. Walker, "Room modes and low-frequency responses in small enclosures," *Audio Eng. Soc.*, preprint no. 4194, 1994.
[45] T. Holman, "Report on mixing studios sound quality," *J. Jpn. Audio Soc.*, 1994.
[46] F. E. Toole, "Subjective measurements of loudspeaker sound quality and listener performance," *J. Audio Eng. Soc.*, vol. 33, pp. 2–32, 1985.
[47] C. Kyriakakis and T. Holman, "High quality audio for the desktop," *J. Visual Commun. Image Representation*, to be published.
[48] T. Holman, "New factors in sound for cinema and television," *J. Audio Eng. Soc.*, vol. 39, pp. 529–539, 1991.
[49] T. Holman and C. Kyriakakis, "Acoustics and psychoacoustics of desktop sound systems," in *Proc. Ann. Meeting Acoustical Society of America*, San Diego, CA, 1997, p. 3092.
[50] W. G. Gardner, "Head-tracked 3-D audio using loudspeakers," in *Proc. WASPAA*, New Paltz, NY, 1997.
[51] O. Groetenherdt, "Video-based detection of heads using motion and stereo vision (in German)," in *Institute for Neuroinformatics*. Bochum, Germany: Univ. of Bochum, 1997.
[52] L. Wiskott, J. M. Fellous, N. Krueger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," Institute for Neuroinformatics, Tech. Rep. no. 8, 1996.
[53] F. L. Wightman, D. J. Kistler, and M. Arruda, "Perceptual consequences of engineering compromises in synthesis of virtual auditory objects," *J. Acoust. Soc. Amer.*, vol. 101, pp. 1050–1063, 1992.
[54] D. W. Batteau, "The role of the pinna in human localization," *Proc. Royal Soc. London*, vol. B168, pp. 158–180, 1967.
[55] A. J. Watkins, "The monaural perception of azimuth: A synthesis approach," in *Localization of Sound: Theory and Applications*, R. W. Gatehouse, Ed. Groton, CT: Amphora, 1979.
[56] J. Chen, B. D. Van Veen, and K. E. Hecox, "A spatial feature extraction and regularization model for the head-related transfer function," *J. Acoust. Soc. Amer.*, vol. 97, pp. 439–452, 1995.
[57] J. Hebrank and D. Wright, "Spectral cues used in the localization of sound sources in the median plane," *J. Acoust. Soc. Amer.*, vol. 56, pp. 1829–1834, 1974.
[58] S. S. Stevens and E. B. Newman, "The localization of actual sources of sound," *Amer. J. Physiology*, vol. 48, pp. 297–306, 1936.
[59] C. A. P. Rodgers, "Pinna transformations and sound reproduction," *J. Audio Eng. Soc.*, vol. 29, pp. 226–234, 1981.

Chris Kyriakakis (Member, IEEE) was born in Thessaloniki, Greece, in 1963. He received the B.S. degree in electrical engineering from the California Institute of Technology, Pasadena, in 1985 and the M.S. and Ph.D. degrees in electrical engineering from the University of Southern California (USC), Los Angeles, in 1987 and 1993, respectively.

He currently is an Assistant Professor in the Electrical Engineering Department, USC, and an Investigator in the Integrated Media Systems Center (IMSC), a National Science Foundation Engineering Research Center at USC. He is the head of the IMSC Immersive Audio Laboratory and is currently involved in the development of algorithms, systems, and architectures for immersive audio. These efforts include audio signal processing for accurate spatial rendering of sound at the desktop environment as well as experimental investigation of human listening characteristics. He is involved in several collaborative efforts within the IMSC, including the incorporation of a vision-based system to track a listener's head for precise, spatially accurate rendering of sound in a three-dimensional immersive audio environment as well as implementation of microphone arrays for robust speaker identification and tracking.