Second Edition

HANDBOOK OF VIRTUAL ENVIRONMENTS

Design, Implementation, and Applications

Edited by

Kelly S. Hale
Kay M. Stanney

CRC Press
Taylor & Francis Group

Second Edition

HANDBOOK OF VIRTUAL ENVIRONMENTS

Design, Implementation, and Applications

Edited by
Kelly S. Hale
Kay M. Stanney

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper
Version Date: 20140619

International Standard Book Number-13: 978-1-4665-1184-2 (Hardback)

Library of Congress Cataloging-in-Publication Data

Handbook of virtual environments : design, implementation, and applications / editors, Kelly S. Hale,
Kay M. Stanney. -- Second edition.
pages cm -- (Human factors and ergonomics)
Includes bibliographical references and index.
ISBN 978-1-4665-1184-2 (hardback)
1. Computer simulation--Handbooks, manuals, etc. 2. Human-computer interaction--Handbooks, manuals, etc. 3. Virtual reality--Handbooks, manuals, etc. I. Hale, Kelly S. II. Stanney, Kay M.

QA76.9.C65H349 2014
004.01'9--dc23 2014023644

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Perspective by Jaron Lanier xiii
Preface xix
Editors xxi
Contributors xxiii

SECTION I Introduction

Chapter 1 Virtual Environments in the Twenty-First Century 3
Kay M. Stanney, Kelly S. Hale, and Michael Zyda

Chapter 2 Virtual Environments Standards and Terminology 23
Richard A. Blade and Mary Lou Padgett

SECTION II System Requirements: Hardware

Chapter 3 Vision and Virtual Environments 39
David R. Badcock, Stephen Palmisano, and James G. May

Chapter 4 Virtual Auditory Displays 87
Michael Vorländer and Barbara Shinn-Cunningham

Chapter 5 Dynamic Haptic Interaction with Video 115
Nuray Dindar, A. Murat Tekalp, and Cagatay Basdogan

Chapter 6 Olfactory Interfaces 131
David L. Jones, Sara Dechmerowski, Razia Oden, Valerie Lugo, Jingjing Wang-Costello, and William Pike

Chapter 7 Perception of Body Motion 163
Ben D. Lawson and Bernhard E. Riecke

Chapter 8 Eye Tracking in Virtual Environments 197
Xuezhong Wang and Brent Winslow

4 Virtual Auditory Displays

Michael Vorländer and Barbara Shinn-Cunningham

CONTENTS

4.1 Introduction 88
4.1.1 Why Are Virtual Auditory Interfaces Important? 88
4.1.1.1 Environmental Realism and Ambience 88
4.1.1.2 Presence/Immersion and Perceived Simulation Quality 89
4.1.1.3 Selective Auditory Attention 89
4.1.1.4 Spatial Auditory Displays 89
4.1.1.5 Cross-Modal Interactions 90
4.2 Physical Acoustics 90
4.2.1 Properties of Sound 90
4.2.1.1 Frequency 90
4.2.1.2 Strength 91
4.3 Psychophysics 91
4.3.1 Frequency Analysis in the Auditory System 91
4.3.2 Intensity Perception 92
4.3.3 Masking Effects 92
4.3.4 Pitch and Timbre 93
4.3.5 Temporal Resolution 93
4.3.6 Spatial Hearing 94
4.3.6.1 Head-Related Transfer Functions 94
4.3.6.2 Binaural Cues 95
4.3.6.3 Spectral Cues 97
4.3.6.4 Anechoic Distance Cues 97
4.3.6.5 Reverberation 97
4.3.6.6 Dynamic Cues 98
4.3.6.7 Effects of Stimulus Characteristics on Spatial Perception 98
4.3.6.8 Top-Down Processes in Spatial Perception 98
4.3.6.9 Benefits of Binaural Hearing 99
4.3.6.10 Adaptation to Distorted Spatial Cues 99
4.3.6.11 Intersensory Integration of Spatial Information 99
4.3.7 Auditory Scene Analysis 100
4.3.8 Speech Perception 100
4.4 Spatial Simulation 101
4.4.1 Room Modeling 101
4.4.2 Headphone Simulation 103
4.4.2.1 Diotic Displays 103
4.4.2.2 Dichotic Displays 103
4.4.2.3 Spatialized Audio 104
4.4.2.4 Practical Limitations on Spatialized Audio 104
4.4.3 Simulation Using Speakers 105
4.4.3.1 Nonspatial Display 105
4.4.3.2 Stereo Display 105

4.4.3.3 Multichannel Loudspeaker Systems ... 106
4.4.3.4 Cross-Talk Cancellation and Transaural Simulations ... 107
4.4.3.5 Lessons from the Entertainment Industry ... 107
4.5 Design Considerations ... 108
4.5.1 Defining the Auditory Environment ... 108
4.5.2 How Much Realism Is Necessary? ... 109
References ... 110

4.1 INTRODUCTION

Auditory processing is often given little attention when designing virtual environments (VEs) or simulations. This lack of attention is unfortunate because auditory cues play a crucial role in everyday life. Auditory cues increase awareness of surroundings, cue visual attention, and convey a variety of complex information without taxing the visual system. The entertainment industry has long recognized the importance of sound to create ambience and emotion, aspects that are often lacking in VEs. Placing someone in a virtual world with an improperly designed auditory interface is equivalent to creating a *virtual* hearing impairment for the user, making them less aware of their surroundings, and contributing to feelings of isolation.

Auditory perception, especially localization, is a complex phenomenon affected by physiology, expectation, and even the visual interface. This chapter will consider different methods for creating auditory interfaces. As will be discussed, spatialized audio using headphones or transaural systems is necessary to create compelling sound, and spatialized sound offers the sound engineer the greatest amount of control over the auditory experience of the listener. Multichannel audio systems can produce virtual sound events in 3D to some extent, but they require complex equipment and signal processing. For many applications, especially using projection screens, standard stereo speaker systems may be simpler to implement and provide benefits not available to headphone systems. Properly designed speaker systems, especially using subwoofers, may contribute to emotional context. The positives and negatives associated with each option will be discussed.

It is impossible to include everything that needs to be known about designing auditory interfaces in a single chapter. The current aim is to provide a starting point, laying out the essential theory behind implementing sound in a VE without overwhelming the novice designer. Instead of trying to review all perceptual and technical issues related to creating virtual auditory displays, this chapter focuses on fundamental aspects of spatial auditory perception and the generation of spatial auditory cues in VEs. Specifically, the chapter begins by introducing basic properties of sound and discussing the perception of these sound properties (psychoacoustics), with a special emphasis on spatial hearing. General techniques for producing auditory stimuli, both with and without spatial cues, are then considered (see Letowski et al., 2001, for a lexicon for understanding auditory displays). Unlike the visual channel, very little effort has been put into formulating theories concerning the creation of synthetic sound sources in VEs; the question of how to generate realistic sounds (rather than using sources from some precomputed, stored library of source sounds) is beyond the scope of this chapter. In addition, the technology involved in producing spatialized audio is rapidly changing, with new products introduced all the time as others disappear, so that any specific recommendations would quickly be dated. However, an overview of current technology and solutions is presented at the conclusion of the chapter.

4.1.1 Why Are Virtual Auditory Interfaces Important?

4.1.1.1 Environmental Realism and Ambience

If it does nothing else, an auditory interface should convey basic information about the VE to the user. For instance, in the real world, pedestrians walking through a shopping area are aware of everything from the sound of their own footsteps to the sounds of other shoppers to the mechanical sounds from cash registers, scanners, escalators, and other machines. In *control room* situations such as nuclear power plants, air traffic control centers, or the bridge of a ship, sounds such as alarms, switches being toggled, and verbal communications with other people in the room (including sounds of stress or uncertainty) provide vital information for participants. The location of these voices, switches, and alarms also provides information concerning their function and importance. In the absence of these basic auditory cues, situational awareness is severely degraded. The same is true in VEs.

The entertainment industry has recognized that sound is a vital aspect of creating ambience and emotion for films. George Lucas, best recognized by the public for stunning visual effects in his movies, has stated that sound is 50% of the movie experience (THX Certified Training Program, 2000). In VEs, the argument is often erroneously made that sound is secondary, since the visual image of a police car chasing down a city street can be compelling on its own. However, without appropriate sounds (squealing tires, a police siren, the tortured breathing of the driver, etc.), the emotional impact of a simulation is muted. The sound quality of footsteps depends on whether you are in grass, on pavement, or in a hallway. Likewise, the sound of one's own voice differs depending on whether you are inside a room or in an open field. These are the types of things that create ambience and feeling in film; the same is true in VEs.

4.1.1.2 Presence/Immersion and Perceived Simulation Quality

Presence (Chertoff & Schatz, 2013; Chapter 34, this book) can be defined as the "sense of being immersed in a simulation or virtual environment." Such a nebulous concept is difficult to quantify. Although definitive evidence is lacking, it is generally believed that the sense of presence is dependent on auditory, visual, and tactile fidelity (Sheridan, 1996). Referring back to the previous section, it can be inferred that as environmental realism increases, the sense of presence increases. However, although realism probably contributes to the sense of presence, it is not necessarily true that an increased sense of presence results in a greater sense of realism. Specifically, although virtual or spatial audio does not necessarily increase the perceived realism of a VE, it does increase the sense of presence (Hendrix & Barfield, 1996). Thus, if implemented properly, appropriately designed audio increases the overall sense of presence in a VE or simulation. Indeed, using medium- and high-quality auditory displays can enhance the perceived quality of visual displays. Inversely, using low-quality auditory displays reduces the perceived quality of visual displays (Storms, 1998).

4.1.1.3 Selective Auditory Attention

In a multisource sound environment, it is easier to segregate, attend to, and comprehend sound sources if they are separated in space, something known as the *cocktail party effect* (Cherry, 1953; Shinn-Cunningham, 2008; Yost, 2006). This ability to direct selective auditory attention, which enables a listener to process whatever sound source is most important at a given moment, is critical in many common situations such as teleconferencing (Begault, 1999) or multichannel radio communications (Begault, 1993; Begault & Wenzel, 1992; Haas, Gainer, Wightman, Couch, & Shilling, 1997). Even when spatial sound cues are imperfect (which can degrade sound localization accuracy), they can improve communication in multichannel situations (Drullman & Bronkhorst, 2000; Shinn-Cunningham, Ihlefeld, & Satyavarta, 2005).

4.1.1.4 Spatial Auditory Displays

While graphical displays are an obvious choice for displaying spatial information to a human operator (particularly after considering the spatial acuity of the visual channel), the visual channel is often overloaded, with operators monitoring a myriad of dials, gauges, and graphic displays. In these cases, spatial auditory cues can provide invaluable information to an operator, particularly when the visual channel is saturated (Begault, 1993; Bronkhorst, Veltman, & van Breda, 1996; Shilling & Letowski, 2000). Spatial auditory displays are also being developed for use in applications for which visual information provides no benefit, for instance, in limited field-of-view (FOV) applications or when presenting information to the blind. In command/control applications, the primary goal is to convey unambiguous information to the human operator. In such situations, realism, per se, is not useful, except to the extent that it makes the operator's task easier (i.e., reduces the workload); however, spatial resolution is critical. In these applications, signal-processing schemes that could enhance the amount of information transferred to the human operator may be useful, even if the result is *unnatural*, as long as the user is able to extract this information (e.g., see Durlach, Shinn-Cunningham, & Held, 1993). It should be noted that when designing spatialized auditory displays for noisy environments such as cockpits, electronic noise cancellation technology should be employed and user's hearing loss taken into account to make certain that the displayed information is perceptible to the user (Begault, 1996). Also, for high-g environments, more work needs to be conducted to discover the contribution of g-forces to displacements in sound localization (e.g., Clark & Graybiel, 1949; DiZio, Held, Lackner, Shinn-Cunningham, & Durlach, 2001).

4.1.1.5 Cross-Modal Interactions

The importance of multimodal interactions involving the auditory system cannot be ignored (Popescu et al., 2014, Chapter 17; Simpson Cowgill, Gilkey, & Weisenberger, 2014, Chapter 13). A whole range of studies show that judgments about one sensory modality are influenced by information in other sensory modalities. For instance, localized auditory cues reduce response times to visual targets (Frens, van Opstal, & van der Willigen, 1995; Perrott, Saberi, Brown, & Strybel, 1990; Perrott, Sadralodabai, Saberi, & Strybel, 1991). Similarly, the number of auditory events affects the perceived number of visual events occurring at that time (e.g., see Shams, Kamitani, & Shimojo, 2004). Even noninformative auditory cues can improve accuracy of perception of visual motion (Kim, Peters, & Sham, 2012), demonstrating the power of cross-modal perceptual effects. Auditory cues also augment or even substitute for tactile and/or visual information about events that are difficult to perceive in these other modalities, such as visual information outside a limited FOV. Through such cross-modal interactions, auditory cues can play an important role in conveying information that may, superficially, seem to be more naturally communicated through some other sensory channel.

4.2 PHYSICAL ACOUSTICS

4.2.1 Properties of Sound

Sound is a pressure wave produced when an object vibrates rapidly back and forth. The diaphragm of a speaker produces sound by pushing against molecules of air, thus creating an area of high pressure (*condensation*). As the speaker's diaphragm returns to its resting position, it creates an area of low pressure (*rarefaction*). This localized disturbance travels through the air as a wave of alternating low pressure and high pressure at approximately 344 m/s or 1128 ft/s (at 70°F), depending on temperature and humidity.

4.2.1.1 Frequency

If the musical note "A" is played as a pure sinusoid, there will be 440 condensations and rarefactions per second. The distance between two adjacent condensations or rarefactions, typically represented by the symbol λ, equals the wavelength of the sound wave. The velocity at which the sound wave is traveling is denoted as c. The time one full oscillatory cycle (condensation through rarefaction) takes is called the frequency (f) and is expressed in Hertz or cycles per second. The relationship between frequency, velocity, and wavelength is given by $f = c/\lambda$.

From a modeling standpoint, this relationship is important when considering the Doppler shift. As a sound source is moving toward a listener, the perceived frequency increases because the wavelength is compressed as a function of the velocity (v) of the moving source. This compression can be explained by the equation $\lambda = (c - v)/f$. For negative velocities (i.e., for sources moving away), this expression describes a relative increase in the wavelength (and a concomitant decrease in frequency).

4.2.1.2 Strength

The amplitude of the waveform determines the intensity of a sound stimulus. It should not be confused, however, with the sound intensity defined in physics as the sound energy propagating per second through a reference area. The meaning of intensity in the context of this chapter is simply the meaning of strength. It is measured in decibels (dB). Decibels give the level of sound (on a logarithmic scale) relative to some reference level. One common reference level is 2×10^{-5} N/m². Decibels referenced to this value are commonly used to describe sound intensity expressed in units of dB sound pressure level (SPL). The sound level in dB SPL can be computed by the following equation:

$$\text{dB SPL} = 20 \log_{10}\left(\frac{\text{RMS sound pressure}}{20 \times 10^{-6}\ \text{N/m}^2}\right)$$

The threshold of hearing is in the range of 0–10 dB SPL for most sounds, although the actual threshold depends on the spectral content of the sound. When measuring sound strength in the *real world*, it is measured with a sound pressure meter. Most sound pressure meters allow one to collect sound-level information using different scales that weight energy in different frequencies differently in order to approximate the sensitivity of the human auditory system to sound at low-, moderate-, or high-intensity levels. These scales are known as A, B, and C weighted scales, respectively. The B scale is rarely used; however, the C scale (dBC) is useful for evaluating noise levels in high-intensity environments such as traffic noise and ambient cockpit noise. The frequency response of the dBC measurement is closer to an unfiltered (flat) response than dBA. In fact, when conducting *sound surveys* in a complex noise environment, it is prudent to measure sound level in both dBA and flat response (or dBC) to make an accurate assessment of the audio environment.

Frequency, intensity, and complexity are physical properties of an acoustic waveform. The perceptual analogs for frequency, intensity, and complexity are pitch, loudness, and timbre, respectively. Although the distinction between physical and perceptual measures of sound properties is an important one, both physical and perceptual descriptions are important when designing auditory displays.

4.3 PSYCHOPHYSICS

The basic sensitivity of the auditory system is reviewed in detail in a number of textbooks (e.g., see Gelfand, 1998; Moore, 1997; Yost, 2006; Zwicker & Fastl, 2007). This section provides a brief overview of some aspects of human auditory sensitivity that are important to consider when designing auditory VEs.

4.3.1 Frequency Analysis in the Auditory System

In the cochlea, acoustic signals are broken down into constituent frequency components by a mechanical Fourier-like analysis. Along the length of the cochlea, the frequency to which that section of the cochlea responds varies systematically from high to low frequencies. The strength of neural signals carried by the auditory nerve fibers arrayed along the length of the cochlea varies with the mechanical displacement of the corresponding section of the cochlea. As a result, each nerve fiber can be thought of as a frequency channel that conveys information about the energy and timing of the input signal within a restricted frequency region. At all stages of the auditory system, these multiple frequency channels are evident.

Although the bandwidth changes with the level of the input signal and with input frequency, to a crude first-order approximation, one can think of the frequency selectivity of the auditory system as constant on a log-frequency basis (approximately one-third octave wide). Thus, a particular auditory nerve responds to acoustic energy at and near a particular frequency.

Humans are sensitive to acoustic energy at frequencies between about 20 and 22,000 Hz. Absolute sensitivity varies with frequency. Humans are most sensitive to energy at frequencies around 2000 Hz and are less sensitive for frequencies below and above this range.

The fact that input waveforms are deconstructed into constituent frequencies affects all aspects of auditory perception. Many behavioral results are best understood by considering the activity of the auditory nerve fibers, each of which responds to energy within about a third of an octave of its particular *best frequency*. For instance, the ability to detect a sinusoidal signal in a noise background degrades dramatically when the noise spectrum is within a third octave of the sinusoid frequency. When a noise is spectrally remote from a sinusoidal target, it causes much less interference with the detection of the sinusoid. These factors are important when one considers the spectral content of different sounds that are to be used in an auditory VE. For instance, if one must monitor multiple kinds of alerting sounds, choosing signals that are spectrally remote from one another will improve a listener's ability to detect and respond to different signals.

4.3.2 Intensity Perception

Listeners are sensitive to sound intensity on a logarithmic scale. For instance, doubling the level of a sound source causes roughly the same perceived change in the loudness independent of the reference level. This logarithmic sensitivity to sound intensity gives the auditory system a large dynamic range. For instance, the range between just detectable sound levels and sounds that are so loud that they cause pain is roughly 110–120 dB (i.e., an increase in sound pressure by a factor of a million). The majority of the sounds encountered in everyday experience span a dynamic intensity range of 80–90 dB. Typical sound reproduction systems use 16 bits to represent the pressure of the acoustic signal (providing a useful dynamic range of about 90 dB), which is sufficient for most simulations.

While sound intensity (a physical measure) affects the loudness of a sound (a perceptual measure), loudness does not grow linearly with intensity. In addition, the same decibel increase in sound intensity can result in different increments in loudness, depending on the frequency content of the sound. Thus, intensity and loudness, while closely related, are not equivalent descriptions of sound.

4.3.3 Masking Effects

As mentioned earlier, when multiple sources are presented to a listener simultaneously or in rapid succession, the sources interfere with one another in various ways. For instance, a tone that is audible when played in isolation may be inaudible when a loud noise is presented simultaneously. Such effects (known as *masking* effects) arise from a variety of mechanisms, from physical interactions of the separate acoustic waves impinging on the ear to high-level, cognitive factors. For a more complete description of these effects than is given in the following, see Yost (2006, pp. 153–167) or Moore (1997, pp. 111–120).

Simultaneous masking occurs when two sources are played concurrently. However, signals do not have to be played simultaneously for them to interfere with one another perceptually. For instance, both *forward* masking (in which a leading sound interferes with perception of a trailing sound) and *backward* masking (in which a lagging sound interferes with perception of a leading sound) occur. Generally speaking, many simultaneous and forward-masking effects are thought to arise from peripheral interactions that occur at or before the level of the auditory nerve. For instance, the mechanical vibrations of the basilar membrane are nonlinear, so that the response of the membrane

to two separate sounds may be less than the sum of the response to the individual sounds. These nonlinear interactions can suppress the response to what would (in isolation) be an audible event.

Other, more central factors influence masking as well. For instance, backward masking may reflect higher-order processing that limits the amount of information extracted from an initial sound in the presence of a second sound. The term *informational masking* refers to all masking that cannot be explained by peripheral interactions in the transduction of sound by the auditory periphery. Most such effects can be traced to problems with segregating a source of interest from other, competing sources, problems with identifying which source in a sound mixture is the most important (*target*) source, or some combination thereof (Shinn-Cunningham, 2008). These failures of selective auditory attention can have significant impact on perception, even for sounds that are clearly audible. For instance, perceptual sensitivity in discrimination and detection tasks is often degraded when there is uncertainty about the characteristics of a target source (e.g., see Yost, 2006, pp. 219–220).

4.3.4 Pitch and Timbre

Just as sound intensity is the physical correlate of the percept of loudness, source frequency is most closely related to the percept of pitch. For sound waves that are periodic (including pure sinusoids, for instance), the perceived *pitch* of a sound is directly related to the inverse of the period of the sound signal. Thus, sounds with low pitch have relative long periods and sounds with high pitch generally have short periods. Many real-world sounds are not strictly periodic in that they have a temporal pattern that repeats over time but has fluctuations from one cycle to the next. Examples of such pseudo-periodic signals include the sound produced by a flute or a vowel sound spoken aloud. The perceived pitch of such sounds is well predicted by the average period of the cyclical variations in the stimulus.

The percept of pitch is not uniformly strong for all sound sources. In fact, nonperiodic sources such as noise do not have a salient pitch associated with them. For relatively narrow sources that are aperiodic, or for band-limited noise, a weak percept of pitch can arise that depends on the center frequency or the cutoff frequency of the spectral energy of the signal, respectively. In fact, perceived pitch is affected by a wide variety of stimulus attributes, including temporal structure, frequency content, harmonicity, and even loudness. Although the pitch of a pure sinusoid is directly related to its frequency, there is no single physical parameter that can predict perceived pitch for more complex sounds. Nonetheless, for many sounds, pitch is a very salient and robust perceptual feature that can be used to convey information to a listener. For instance, in music, pitch conveys melody. In speech, pitch conveys a variety of information (ranging from the gender of a speaker to paralinguistic, emotional content of a speech). Pitch is also a very important cue for segregating competing sound sources and allowing a listener to focus selective auditory attention on a target source (e.g., see Carlyon, 2004).

The percept of timbre is the sound property that enables a listener to distinguish an oboe from a trumpet. Like pitch, the percept of timbre depends on a number of physical parameters of sound, including spectral content and temporal envelope (such as the abruptness of the onset and offset of sound). Like pitch, timbre is an important property for enabling listeners to identify a target source and thus can be used to convey information through an auditory display (e.g., see Brewster, Wright, & Edwards, 1993). However, sounds with different timbres have different perceptual weight, a factor that should be considered in designing discrete sounds for an auditory display (e.g., see Chon & McAdams, 2012). As with pitch, timbre is a feature that allows listeners to direct selective auditory attention to a desired target amidst competing sounds, enabling them to extract desired information from that source despite the presence of interfering information (e.g., Maddox & Shinn-Cunningham, 2011).

4.3.5 Temporal Resolution

The auditory channel is much more sensitive to temporal fluctuations in sensory inputs than either visual (Badcock, Palmisano, & May, 2013, Chapter 3) or proprioceptive (Dindar, Tekalp, & Basdogan, 2013, Chapter 5; Lawson & Riecke, 2014, Chapter 7) channels. For instance, the auditory system can detect amplitude fluctuations in input signals up to 50 Hz (i.e., a duty cycle of 20 ms) very easily (e.g., see Yost, 2006, pp. 146–149). Sensitivity degrades slowly with increasing modulation rate, so that some sensitivity remains even as the rate approaches 1000 Hz (i.e., temporal fluctuations at a rate of 1 per ms). The system is also sensitive to small fluctuations in the spectral content of an input signal for roughly the same modulation speeds. Listeners not only can detect rapid fluctuations in an input stimulus, but they can react quickly to auditory stimuli. For instance, reaction times to auditory stimuli are faster than visual reaction times by 30–40 ms (an improvement of roughly 20%; e.g., see Welch & Warren, 1986).

4.3.6 Spatial Hearing

Spatial acuity of the auditory system is far worse than that of the visual (Badcock et al., 2013, Chapter 3) or proprioceptive (Dindar et al., 2013, Chapter 5; Lawson & Riecke, 2013, Chapter 7) systems (for a review, see Middlebrooks & Green, 1991). For a listener to detect an angular displacement of a source from the median plane, the source must be displaced laterally by about a degree. For a source directly to the side, the listener does not always detect a lateral displacement of 10°. Auditory spatial acuity is even worse in other spatial dimensions. A source in the median plane must be displaced by as much as 15° for the listener to perceive the directional change accurately. While listeners can judge relative changes in source distance, absolute distance judgments are often surprisingly inaccurate even under the best of conditions (e.g., see Zahorik, Brungart, & Bronkhorst, 2005).

Functionally, spatial auditory perception is distinctly different from that of the other *spatial* senses of vision and proprioception. For the other spatial senses, position is neurally encoded at the most peripheral part of the sensory system. For instance, the photoreceptors of the retina are organized topographically so that a source at a particular position (relative to the direction of gaze) excites a distinct set of receptors (Badcock et al., 2013, Chapter 3). In contrast, spatial information in the auditory signals reaching the left and right ears of a listener must be computed from the peripheral neural representations. The way in which spatial information is carried by the acoustic signals reaching the eardrums of a listener has been the subject of much research. This section provides a brief review of how acoustic attributes convey spatial information to a listener and how the perceived position of a sound source is computed in the brain (for more complete reviews, see Blauert, 1997; Middlebrooks & Green, 1991; Mills, 1972; Wightman & Kistler, 1993).

4.3.6.1 Head-Related Transfer Functions

The pairs of spatial filters that describe how sound is transformed as it travels through space to impinge on the left and right ears of a listener are known as head-related transfer functions (HRTFs). HRTFs describe how to simulate the direct sound reaching the listener from a particular position but do not generally include any reverberant energy. Empirically measured HRTFs vary mainly with the direction from head to source but also vary with source distance (particularly for sources within reach of the listener). For sources beyond about a meter away, the main effect of distance is just to change the overall gain of the HRTFs. In the time domain, the HRTF pair for a particular source location provides the pressure waveforms that would arise at the ears if a perfect impulse were presented from the spatial location in question. Often, HRTFs are represented in the frequency domain by taking the Fourier transform of time-domain impulse responses.

HRTFs contain most of the spatial information present in real-world listening situations. In particular, binaural cues are embodied in the relative phase and magnitude (respectively) of the linear filters for the left and right ears. Spectral cues and source intensity are present in the absolute frequency-dependent magnitudes of the two filters.

Figure 4.1 shows two HRTF pairs from a human subject in the time domain (left side of figure) and in the frequency domain (magnitude only, right side of figure). All panels are for a source at azimuth 90 and elevation 0. The top two panels show the HRTF for a source very close to the head (15 cm from the center of the head). The bottom two panels show the HRTF for a source 1 m from the head. In the time domain, it is easy to see the interaural differences in time and intensity, while the frequency-domain representation shows the spectral notches that occur in HRTFs, as well as the frequency-dependent nature of the interaural level difference (ILD). The ILDs are larger in all frequencies for the nearer source (top panels) as expected. In the time domain, the 1 m source must traverse a greater distance to reach the ears than the near source, resulting in additional time delay before the energy reaches the ears (note time-onset differences in the impulse responses in the left top and left bottom panels).

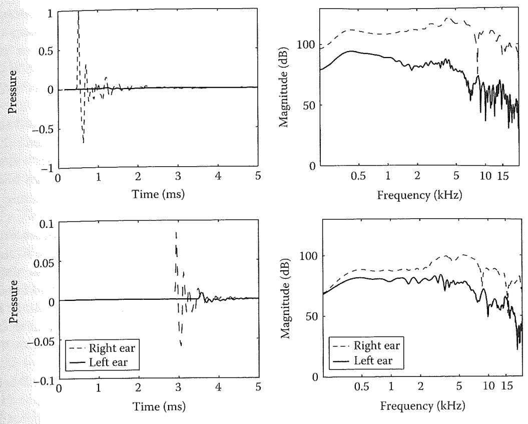

FIGURE 4.1 Time-domain (left panels) and magnitude spectra (right panels) representations of anechoic HRTFs for a human subject. All panels show a source at 90 azimuth and 0 elevation. Top panels are for a source at 15 cm, and bottom panels for a source at 1 m.

4.3.6.2 Binaural Cues

The most robust cues for source position depend on differences between the signals reaching the left and right ears. Such *interaural* or *binaural* cues are robust specifically because they can be computed by comparing the signals reaching each ear. As a result, binaural cues allow a listener to factor out those acoustic attributes that arise from source content from those attributes that arise from source position.

Depending on the angle between the interaural axis and a sound source, one ear may receive a sound earlier than the other. The resulting interaural time differences (ITDs) are the main cue indicating the laterality (left/right location) of the direct sound. The ITD grows with the angle of the source from the median plane; for instance, a source directly to the right of a listener results in an ITD of 600–800 μs favoring the right ear. ITDs are most salient for sound frequencies below about 2 kHz but occur at all frequencies in a sound. At higher frequencies, listeners use ITDs in signal *envelopes* to help determine source laterality but are insensitive to differences in the interaural phase of the signal.

Listeners can reliably detect ITDs of 10–100 μs (depending on the individual listener), which grossly correspond to ITDs that would result from a source positioned 1°–10° from the median plane. Sensitivity to changes in the ITD deteriorates as the reference ITD gets larger. For instance, the smallest detectable change in ITD around a reference source with an ITD of 600–800 μs (corresponding to the ITD of a source far to the side of the head) can be more than a factor of 2 larger than for a reference with zero ITD.

At the high end of the audible frequency range, the head of the listener reflects and diffracts signals so that less acoustic energy reaches the far side of the head (causing an *acoustic head shadow*). Due to the acoustic head shadow, the relative intensity of a sound at the two ears varies with the lateral location of the source. The resulting ILDs generally increase with source frequency and angle between the source and median plane. ILDs are perceptually important for determining source laterality for frequencies above about 2 kHz.

When a sound source is within reach of a listener, extralarge ILDs (at all frequencies) arise due to differences in the relative distances from source to left and right ears (e.g., see Brungart & Rabinowitz, 1999; Duda & Martens, 1998). These additional ILDs are due to differences in the relative distances from source to left and right ears and help to convey information about the relative distance and direction of the source from the listener (Shinn-Cunningham, Santarelli, & Kopco, 2000). Other low-frequency ILDs that may arise from the torso appear to help determine the elevation of a source (Algazi, Avendano, & Duda, 2001). Most listeners are able to detect ILDs of 0.5–1 dB, independent of source frequency.

The perceived location of a sound source usually is consistent with the ITD and ILD information available. However, there are multiple source locations that cause roughly the same ITD and ILD cues. For sources more than a meter from the head, the locus of such points is approximately a hyperbolic surface of rotation symmetric about the interaural axis that is known as the *cone of confusion* (see left side of Figure 4.2). When a sound is within reach of the listener, extralarge ILDs provide additional robust, binaural information about the source location. For a simple spherical head model, low-frequency ILDs are constant on spheres centered on the interaural axis (see right side of Figure 4.2). The rate at which the extralarge ILD changes with spatial location decreases as sources move far from the head or near the median plane. In fact, once a source is more than a meter or so from the head, the contribution of this *near-field* ILD is perceptually insignificant (Shinn-Cunningham et al., 2000). In general, positions that give rise to the same binaural cues (i.e., the intersection of constant ITD and ILD contours) form a circle centered on the interaural axis (Shinn-Cunningham et al.). Since ITD and ILD sensitivity is imperfect, the locus of positions that cannot be resolved from binaural cues may be more accurately described as a *torus of confusion* centered on the interaural axis. Such tori of confusion degenerate to the more familiar cones of confusion for sources more than about a meter from a listener.

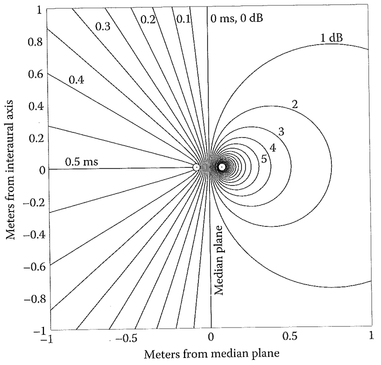

FIGURE 4.2 Iso-ITD (left side of figure) and iso-ILD (right side of figure) contours for sources near the head. On the left, sources at each location along a contour give rise to nearly the same ITD. On the right, sources at each location along a contour give rise to nearly the same unique near-field ILD component.

4.3.6.3 Spectral Cues

The main cue to resolve source location on the torus of confusion is the spectral content of signals reaching the ears. These spectral cues arise due to interactions of the outer ear (pinna) with the impinging sound wave that depend on the relative position of sound source and listener's head (Batteau, 1967). Spectral cues only occur at relatively high frequencies, generally above 6 kHz. Unlike interaural cues for source location, spectral cues can be confused with changes in the spectrum of the source itself. Perhaps because of this ambiguity, listeners are more likely to make localization errors in which responses fall near the correct torus of confusion but are not in the right direction. Individual differences in spectral filtering of the pinnae are large and are important when judging source direction (e.g., see Wenzel, Arruda, Kistler, & Wightman, 1993).

4.3.6.4 Anechoic Distance Cues

In general, the intensity of the direct sound reaching a listener (i.e., sound that does not come off of reflective surfaces like walls) decreases with the distance of the source. In addition, the atmosphere absorbs energy in high, audible frequencies as a sound propagates, causing small changes in the spectrum of the received signal with changes in source distance. If a source is unfamiliar, the intensity and spectrum of the direct sound are not robust cues for distance because they can be confounded with changes in the intensity or spectral content (respectively) of the signal emitted from the source. However, even for unfamiliar sources, overall level and spectral content provide relative distance information (Mershon, 1997).

4.3.6.5 Reverberation

Reverberation (acoustic energy reaching a listener from indirect paths, via walls, floors, etc.) generally has little affect on or degrades the perception of source direction (e.g., see Begault, 1993; Hartmann, 1983; Shinn-Cunningham, 2000b). However, it actually aids in the perception of source distance (e.g., see Mershon, Ballenger, Little, McMurty, & Buchanan, 1989; Shinn-Cunningham, Kopco, & Santarelli, 1999). At least grossly, the intensity of reflected energy received at the ears is independent of the position of the source relative to the listener (although it can vary dramatically from one room to another). As a result, the ratio of direct to reverberant energy provides an absolute measure of source distance for a given listening environment.

Reverberation not only provides a robust cue for source distance, but it also provides information about the size and configuration of a listening environment. For instance, information about the size and *spaciousness* of a room can be extracted from the pattern of reverberation in the signals reaching the ears. While many psychophysical studies of sound localization are performed in anechoic (or simulated anechoic) environments, reverberation is present (in varying degrees) in virtually all everyday listening conditions. Anechoic environments (such as those used in many simulations and experiments) seem subjectively unnatural and strange to naive listeners. Conversely, adding reverberation to a simulation causes all sources to seem more realistic and provides robust information about relative source distance (e.g., see Begault, Wenzel, Lee, & Anderson, 2012; Brungart & D'Angelo, 1995). While reverberation may improve distance perception and improve the realism of a display, it can decrease accuracy in directional perception, albeit slightly, and may interfere with the ability to extract information in a source signal (e.g., degrade speech) and to attend to multiple sources (e.g., see Section 4.3.6.8).

4.3.6.6 Dynamic Cues

In addition to *static* acoustic cues like ITD and ILD, changes in spatial cues with source or listener movement also influence perception of source position and help to resolve torus-of-confusion ambiguities (e.g., see Wallach, 1940; Wightman & Kistler, 1989). For instance, a source either directly in front or directly behind a listener would cause near-zero ITDs and ILDs; however, a leftward rotation of the head results in ITDs and ILDs favoring either the right ear (for a source in front) or the left ear (for a source behind).

While the auditory system generally has good temporal resolution, the temporal resolution of the binaural hearing system is much poorer. For instance, investigations into the perception of moving sound sources imply that binaural information averaged over a time window lasting 100–200 ms results in what has been termed *binaural sluggishness* (e.g., see Grantham, 1997; Kollmeier & Gilkey, 1990).

4.3.6.7 Effects of Stimulus Characteristics on Spatial Perception

Characteristics of a source itself affect auditory spatial perception in a number of ways. For instance, the bandwidth of a stimulus can have a large impact on both the accuracy and precision of sound localization. As a result, one must consider how nonspatial attributes of a source in a VE will impact spatial perception of a signal. In cases where one can design the acoustic signal (i.e., if the signal is a warning signal or some other arbitrary waveform), these factors should be taken into consideration when one selects the source signal.

For instance, the spectral filtering of the pinnae cannot be determined if the sound source does not have sufficient bandwidth. This makes it difficult to unambiguously determine the location of a source on the torus of confusion for a narrowband signal. Similarly, if a source signal does not have energy above about 5 kHz, spectral cues will not be represented in the signals reaching the ears and errors along the torus of confusion are more common (e.g., Gilkey & Anderson, 1995).

Ambiguity in narrowband source locations arises in other situations as well. For instance, narrowband, low-frequency signals in which ITD is the main cue can have ambiguity in their heard location because the auditory system is only sensitive to interaural phase. Thus, a low-frequency sinusoid with an ITD of half cycle favoring the right ear may also be heard far to the left side of the head. However, binaural information is integrated across frequency so that ambiguity in lateral location is resolved when interaural information is available across a range of frequencies (Brainard, Knudsen, & Esterly, 1992; Stern & Trahiotis, 1997; Trahiotis & Stern, 1989). When narrowband sources are presented, the heard location is strongly influenced by the center frequency of the source (Middlebrooks, 1997).

While spectral bandwidth is important, temporal structure of a source signal is also important. In particular, onsets and offsets in a signal make source localization more accurate, particularly when reverberation and echoes are present. A gated or modulated broadband noise will generally be more accurately localized in a reverberant room (or simulation) than a slowly gated broadband noise (e.g., see Rakerd & Hartmann, 1985, 1986).

4.3.6.8 Top-Down Processes in Spatial Perception

Experience with or knowledge of the acoustics of a particular environment also affects auditory localization, and implicit learning and experience affects performance (e.g., see Clifton, Freyman, & Litovsky, 1993; Shinn-Cunningham, 2000b). In other words, spatial auditory perception is not wholly determined by stimulus parameters but also by the state of the listener. Although such effects are not due to conscious decision, they can measurably alter auditory localization and spatial perception. For instance, when localizing a sound followed by a later *echo* of that sound, the influence of the later sound diminishes with repetition of the sound pairing, as if the listener learns to discount the lagging echo (Freyman, Clifton, & Litovsky, 1991).

4.3.6.9 Benefits of Binaural Hearing

Listeners benefit from receiving different signals at the two ears in a number of ways. As discussed earlier, ITD and ILD cues allow listeners to determine the location of sound sources. However, in addition to allowing listeners to locate sound sources in the environment, binaural cues allow the listener to selectively attend to sources coming from a particular direction. This ability is extremely important when there are multiple competing sources in the environment (e.g., see Shinn-Cunningham, 2008).

Imagine a situation in which there is both a speaker (whom the listener is trying to attend) and a competing source (that is interfering with the speaker). If the speaker and competitor are both directly in front of the listener, the competitor degrades speech reception much more than if the competitor is off to one side, spatially separated from the speaker. This *binaural advantage* arises in part because when the competitor is off to one side of the head, the energy from the competitor is attenuated at the far ear. As a result, the signal-to-noise ratio at the far ear is larger than when the competitor is in front. In other words, the listener has access to a cleaner signal in which the speaker is more prominent when the speaker and noise are spatially separated. However, the advantage of the spatial separation is even larger than can be predicted on the basis of energy.

A homologous benefit can be seen under headphones. In particular, if one varies the level of a signal until it is just detectable in the presence of a masker, the necessary signal level is much lower when the ITD of the signal and masker are different than when they are the same. The difference between these thresholds, referred to as the masking level difference (MLD), can be as large as 10–15 dB for some signals (e.g., see Durlach & Colburn, 1978; Zurek, 1993).

The binaural advantage affects both signal detection (e.g., see Gilkey & Good, 1995) and speech reception (e.g., see Bronkhorst & Plomp, 1988). It is one of the main factors contributing to the ability of listeners to monitor and attend multiple sources in complex listening environments (i.e., the *cocktail party effect*; see, e.g., Shinn-Cunningham, 2008; Yost, 1997). Thus, the binaural advantage is important for almost any auditory signal of interest. In order to get these benefits of binaural hearing, signals reaching a listener must have appropriate ITDs and/or ILDs.

4.3.6.10 Adaptation to Distorted Spatial Cues

While a naive listener responds to ITD, ILD, and spectral cues based on their everyday experience, listeners can learn to interpret cues that are not exactly like those that occur naturally. For instance, listeners can learn to adapt to unnatural spectral cues when given sufficient long-term exposure (Hofman, Van Riswick, & Van Opstal, 1998). Short-term training allows listeners to learn how to map responses to spatial cues to different spatial locations than normal (Shinn-Cunningham, Durlach, & Held, 1998). These studies imply that for applications in which listeners can be trained, *perfect* simulations of spatial cues may not be necessary. However, there are limits to the kinds of distortions of spatial cues to which a listener can adapt (e.g., see Shinn-Cunningham, 2000a).

4.3.6.11 Intersensory Integration of Spatial Information

Acoustic spatial information is integrated with spatial information from other sensory channels (particularly vision) to form spatial percepts (e.g., see Welch & Warren, 1986). In particular, auditory spatial information is combined with visual (and/or proprioceptive) spatial information to form the percept of a single, multisensory event, especially when the inputs to the different modalities are correlated in time (e.g., see Popescu et al., 2013; Warren, Welch, & McCarthy, 1981; Chapter 17). When this occurs, visual spatial information is much more potent than that of auditory information, so the perceived location of the event is dominated by the visual spatial information (although auditory information does affect the percept to a lesser degree, e.g., see Pick, Warren, & Hay, 1969;

Welch & Warren, 1980). *Visual capture* refers to the perceptual dominance of visual spatial information, describing how the perceived location of an auditory source is captured by visual cues.

Summarizing these results, it appears that the spatial auditory system computes source location by combining all available acoustic spatial information. Perhaps even more importantly, a priori knowledge and information from other sensory channels can have a pronounced effect on spatial perception of auditory and multisensory events.

4.3.7 Auditory Scene Analysis

Listeners in real-world environments are faced with the difficult problem of listening to many competing sound sources that overlap in both time and/or frequency. The process of separating out the contributions of different sources to the total acoustic signals reaching the ears is known as *auditory scene analysis* (e.g., see Bregman, 1990; Carlyon, 2004).

In general, the problem of grouping sound energy across time and frequency to reconstruct each sound source is governed by a number of basic (often intuitive) principles. For instance, naturally occurring sources are often broadband, but changes in the amplitude or frequency of the various components of a single source are generally correlated over time. Thus, comodulation of sound energy in different frequency bands tends to group these signal elements together and cause them to fuse into a single perceived source. Similarly, temporal and spectral proximity both tend to promote grouping so that signals close in time or frequency are grouped into a single perceptual source (sometimes referred to as a stream). Spatial location can also influence auditory scene analysis such that signals from the same or similar locations are grouped into a single stream. Other factors affecting streaming include (but are not limited to) harmonicity, timbre, and frequency or amplitude modulation (Bregman, 1990; Darwin, 1997).

For the development of auditory displays, these grouping and streaming phenomena are very important because they can directly impact the ability to detect, process, and react to a sound. For instance, if a masker sound is played simultaneously with a target sound, the ability to process the target is significantly worse if the target is heard as just one component of a single sound source comprised of the masker plus the target; when the target is heard as a distinct sound source, a listener is much better at both detecting the target's presence and extracting meaning from the target. Such *grouping* effects cannot be explained solely by peripheral mechanisms, since many times, the target sound is represented faithfully in activity on the auditory nerve. Instead, such effects arise from *central* limitations (e.g., see Shinn-Cunningham, 2008).

4.3.8 Speech Perception

Arguably the most important acoustic signal is that of speech. The amount of information transmitted via speech is larger than any other acoustic signal. For many applications, accurate transmission of speech information is the most critical component of an auditory display.

Speech perception is affected by many of the low-level perceptual issues discussed in previous sections. For instance, speech can be masked by other signals, reducing a listener's ability to determine the content of the speech signal. Speech reception in noisy environments improves if the speaker is located at a different position than the noise source(s), particularly if the speaker and masker are at locations giving rise to different ILDs. Speech reception is also affected by factors that affect the formation of auditory streams, such as comodulation, harmonic structure, and related features. However, speech perception is governed by many high-level, cognitive factors that do not apply to other acoustic signals. For instance, the ability to perceive a spoken word improves dramatically if it is heard within a meaningful sentence rather than in isolation. Speech information is primarily conveyed by sound energy between 200 and 5000 Hz. For systems in which speech communication is critical, it is important to reduce the amount of interference in this range of frequencies or it will impede speech reception.

4.4 SPATIAL SIMULATION

Spatial auditory cues can be simulated using headphone displays or loudspeakers. Headphone displays generally allow more precise control of the spatial cues presented to a listener, both because the signals reaching the two ears can be controlled independently and because there is no indirect sound reaching the listener (i.e., no echoes or reverberation). However, headphone displays are generally more expensive than loudspeaker displays and may be impractical for applications in which the listener does not want to wear a device on the head. While it is more difficult to control the spatial information reaching a listener in a loudspeaker simulation, loudspeaker-based simulations are relatively simple and inexpensive to implement and do not physically interfere with the listener.

Simulations using either headphones or speakers can vary in complexity from providing no spatial information to providing nearly all naturally occurring spatial cues. This section reviews both headphone and speaker approaches to creating spatial auditory cues.

4.4.1 Room Modeling

HRTFs generally do not include reverberation or echoes, although it is possible to measure binaural transfer functions (known as binaural room transfer functions) that incorporate the acoustic effects of a room. While possible, such approaches are generally not practical because such filters vary with listener and source position in the room as well as the relative position of listener and source to produce a combinatorially large number of transfer functions. In addition, such filters can be an order of magnitude longer than traditional HRTFs, increasing both computational and storage requirements of the system.

There has been substantial effort devoted to developing computational models for room reverberation, including high-quality commercial software packages (e.g., see www.odeon.dk or www.catt.se). The required computations are quite intensive; in order to simulate each individual reflection, one must calculate the distance the sound wave has traveled, how the waveform was transformed by every surface on which it impinged, and the direction from which it is arriving at the head. The resulting waveform must then be filtered by the appropriate anechoic HRTF based on the direction of incidence with the head (Vorländer, 2008, pp. 141–146).

If one looks at the resulting reflections as a function of time from the initial sound, the number of reflections in any given time slice increases quadratically with time. At the same time, the level of each individual reflection decreases rapidly, both due to energy absorption at each reflecting surface and increased path length from source to ear. Moreover, those reflections lose their coherence due to surface scattering and edge diffraction. Although second- or third-order reflections may be individually resolvable, higher-order reflections occur so densely in time that the distinct specular content of each *echo* becomes practically irrelevant. Instead, from a certain transition time (which depends on the size of the room and on the surface corrugations), the reflections are smeared in time and heavily overlap to the point that they are well approximated as a so-called *diffuse* sound field. Therefore, many simulations only *spatialize* a relatively small number of the loudest, earliest-arriving reflections (e.g., up to second or third order) and then add random noise that dies off exponentially in time (uncorrelated at the two ears) to simulate later-arriving reflections that are dense in time and arriving from essentially random directions. Even with such simplifications, the computations necessary to generate *realistic* reverberation (particularly in a system that tries to account for movement of a listener) are a challenge; however, with today's computational power, such simulations are feasible and produce plausible results.

Figure 4.3 shows the room impulse response at the right ear for a source located at 45 azimuth, 0 elevation, and distance of 1 m. This impulse response was measured in a moderate-size classroom in which significant reverberant energy persists for as long as 450 ms. The initial few milliseconds of the response are shown in the inset. In the inset, the initial response is that caused by the sound

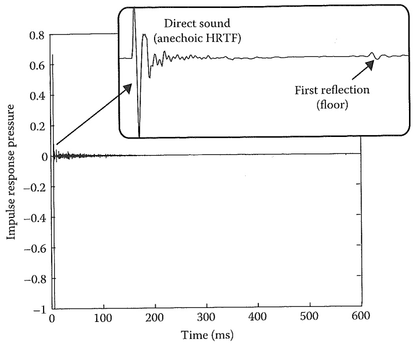

FIGURE 4.3 Impulse response at the right ear for a source at 45 azimuth, 0 elevation, and distance 1 m in a standard classroom. Inset shows first 10 ms of total impulse response.

wave that travels directly from the source to the ear. The first reflection is also evident at the end of the inset, at a much reduced amplitude. In the main figure, the decay of the reverberant energy can be seen with time.

The development of tractable reverberation algorithms for real-time systems has achieved some promising results but is still an ongoing area of research. Most algorithms are based on geometrical acoustics. They are known as hybrid specular and diffuse reflection models and combination of early (mainly specular) and late (mainly diffuse) components (e.g., see Vorländer, 2008, pp. 216–226).

Wave models are also in use in order to compute wave effects such as eigenmodes or diffraction (Aretz, Maier, & Vorländer, 2010; Botteldooren, 1995; Savioja, 2010). Usually this is implemented as a second level into the hybrid models so that the low-frequency range and the mid- and high-frequency range are treated by wave models and geometric models, respectively.

The rendering process using geometrical acoustics is based on a 3D room model consisting of polygons defining the room boundaries. These polygons represent the reflecting surfaces, and accordingly, they are tagged with acoustic absorption and scattering coefficients. For the simulation of reflections, the amplitude and the time of arrival must be calculated. For this, geometrical construction methods are used, which result in identifying the reflection paths between the reflection boundaries of sound propagation in the room.

As soon as the reflection components of the room impulse response are found, they must be convolved with the directional-dependent HRTF and with the sound source stimulus—the so-called *dry* signal.

One crucial part of real-time systems is a rapid calculation of the geometric reflection paths using the polygon model of the room. This task can be highly accelerated by using tree structures in the polygon database (Schröder & Lentz, 2006) or other search algorithms such as spatial hashing (e.g., Schröder, Ryba, & Vorländer, 2010) or frustum tracing (Chandak, Lauterbach, Taylor, Ren, & Manocha, 2008).

Another crucial part of the sound rendering process is real-time convolution of the binaural impulse response with the sound source. Room impulse responses are typically of some seconds in length. The corresponding binaural room impulse response is a finite impulse response (FIR) filter.

For achieving appropriate immersion, the latency of the audio output cannot exceed more than a few milliseconds. Discrete convolution in the time domain and synchronous audio output may be straightforward conceptually but is inefficient in the number of operations required to compute the output samples one by one. Fast Fourier transform (FFT)-based convolution is more efficient, computationally, but requires block processing; this, in turn, leads to block-size-dependent latency in the output. One approach to solving these problems is to use nonuniform segmented block convolution algorithms (Garcia, 2002; Wefers & Vorländer, 2012) or infinite impulse response (IIR) filters.

The binaural filters can be interpreted in the time domain as a temporal series of reflections. Alternatively, they can be interpreted in the frequency domain as a binaural transfer function. The extent to which listeners are sensitive to interaural (temporal) and spectral details in reverberant energy is also not well understood and requires additional research. Nonetheless, there is clear evidence that the inclusion of reverberation can have a dramatic impact on the subjective realism of a virtual auditory display and can aid in perception of source distance (e.g., see Brungart & D'Angelo, 1995).

4.4.2 Headphone Simulation

In order to simulate any source somewhere in space or in a room over headphones, one must simply play a stereo headphone signal that recreates at the eardrums the exact acoustic waveforms that would actually arise from a source at the desired location. This is generally accomplished by empirically measuring the HRTF that describes how an acoustic signal at a particular location in space is transformed as it travels to and impinges on the head and ears of a listener. Then, in order to simulate an arbitrary sound source at a particular location, the appropriate transfer functions are used to filter the desired (known) source signal. The resulting stereo signal is then corrected to compensate for transfer characteristics of the display system (for instance, to remove any spectral shaping of the headphones) and presented to the listener. This holds for a single source at a particular direction or for numerous sources or room reflections, provided they have been filtered with each particular directional HRTF.

4.4.2.1 Diotic Displays

The simplest headphone displays present identical signals to both ears (*diotic* signals). With a diotic display, all sources are perceived as inside the head (not *externalized*), at midline. This internal sense of location is known as *lateralization* not *localization* (Plenge, 1974). While a diotic display requires no spatial auditory processing, it also provides no spatial information to a listener. Such displays may be useful if the location of an auditory object is not known or if spatial auditory information is unimportant. However, diotic displays are the least realistic headphone display. In addition, as discussed in Section 4.3.6.9, benefits of spatial hearing can be extremely useful for detection and recognition of auditory information. For instance, when listeners are required to monitor multiple sound sources, spatialized auditory displays are clearly superior to diotic displays (Haas et al., 1997).

4.4.2.2 Dichotic Displays

While normal interaural cues vary with frequency in complex ways, simple frequency-independent ITDs and ILDs affect the perceived lateral position of a sound source (e.g., see Durlach & Colburn, 1978). Stereo signals that only contain a frequency-independent ITD and/or ILD are herein referred to as *dichotic* signals (although the term is sometimes used to refer to any stereo signal in which left and right ears are different).

Generation of a constant ITD or ILD is very simple over headphones since it only requires that the source signal be delayed or scaled (respectively) at one ear. Just as with diotic signals, dichotic signals result in sources that appear to be located on an imaginary line inside the head, connecting the two ears. Varying the ITD or ILD causes the lateral position of the perceived source to move toward the ear receiving the louder and/or earlier-arriving signal. For this reason, such sources are usually referred to as *lateralized* rather than *localized*.

Dichotic headphone displays are simple to implement but are only useful for indicating whether a sound source is located to the left or right of a listener. On the other hand, when multiple sources are lateralized at different locations (using different ITD and/or ILD values), some binaural unmasking can be obtained (see Section 4.3.6.9).

4.4.2.3 Spatialized Audio

Using signal-processing techniques, it is possible to generate stereo signals that contain most of the normal spatial cues available in the real world. In fact, if properly rendered, spatialized audio can be practically indistinguishable from free-field presentation (Langendijk & Bronkhorst, 2000). When coupled with a head-tracking device, spatialized audio provides a true virtual auditory interface. Using a spatialized auditory display, a variety of sound sources can be presented simultaneously at different directions and distances. One of the early criticisms of spatialized audio was that it was expensive to implement; however, as hardware and software solutions have proliferated, it has become feasible to include spatialized audio in most systems. Spatialized audio solutions can be fit into any budget, depending on the desired resolution and number of sound sources required. Most virtual reality (VR) systems are currently outfitted with headphones of sufficient quality to reproduce spatialized audio, making it relatively easy to incorporate spatialized audio in an immersive VE system.

4.4.2.4 Practical Limitations on Spatialized Audio

While in theory, HRTF simulation should yield stimuli that are perceptually indistinguishable from natural experience, a number of practical considerations limit the realism of stimuli simulated using HRTFs. Measurement of HRTFs is a difficult, time-consuming process. In addition, storage requirements for HRTFs can be prohibitive. As a result, HRTFs are typically measured only at a single distance, relatively far from the listener, and at a relatively sparse spatial sampling. Changes in source distance are simulated simply by scaling the overall signal intensity. Because the HRTFs are only measured for a finite number of source directions at this single source distance, HRTFs are interpolated to simulate locations for which HRTFs are not measured. While this approach is probably adequate for sources relatively far from the listener and when some inaccuracy can be tolerated, the resulting simulation cannot perfectly recreate spatial cues for all possible source locations (Wenzel & Foster, 1993). Individual differences in HRTFs are very important for some aspects of sound source localization (particularly for distinguishing front/back and up/down). However, most systems employ a standard set of HRTFs that are not matched to the individual listener. Using these *nonindividualized* HRTFs reduces the accuracy and externalization of auditory images but still results in useful performance increases (Begault & Wenzel, 1993). Researchers have explored a variety of HRTF compression schemes in which individual differences are encoded in a small number of parameters that can be quickly or automatically fit to an individual (e.g., see Kistler & Wightman, 1991; Middlebrooks & Green, 1992). Nonetheless, many *typical* systems cannot simulate source position along a cone of confusion because they do not use individualized HRTFs.

The most sophisticated spatialized audio systems use trackers to measure the movement of a listener and update the HRTFs in real time to produce appropriate dynamic spatial cues. The use of head tracking dramatically increases the accuracy of azimuthal localization (Moldrzyk, Ahnert, Feistel, Lentz, & Weinzierl, 2004; Sorkin, Kistler, & Elvers, 1989). However, time lag in such systems (from measuring listener movement, choosing the new HRTF, and filtering the ongoing source signal) can be greater than 30 ms. While the binaural system is sluggish, the resulting delay can be perceptible. Real-time systems are also too complex and costly for some applications. Instead, systems may compute signals off-line and either ignore or limit the movement of the listener; however, observers may hear sources at locations inside or tethered to the head (i.e., moving with the head) with such systems.

Many simulations do not include any echoes or reverberation in the generated signals. Although reverberation has little impact (or degrades) perception of source direction, it is important for distance perception. In addition, anechoic simulations sound subjectively artificial and less realistic than do simulations with reverberation.

4.4.3 Simulation Using Speakers

The total acoustic signal reaching each ear is simply the sum of the signals reaching that ear from each source in an environment. Using this property, it is possible to vary spatial auditory cues (e.g., ITD, ILD, and spectrum) by controlling the signals played from multiple speakers arrayed around a listener. In contrast with headphone simulations, the signals at the two ears cannot be independently manipulated; that is, changing the signal from any of the speakers changes the signals reaching both ears. As a result, it is difficult to precisely control the interaural differences and spectral cues of the binaural signal reaching the listener to mimic the signals that would occur for a real-world source. However, various methods for specifying the signals played from each loudspeaker exist to simulate spatial auditory cues using loudspeakers.

To reduce the variability of audio signals reaching the ears, careful attention should be given to speaker placement and room acoustics. If speaker systems are not properly placed and installed in a room, even the best sound systems will sound inferior. Improperly placed speakers can reduce speech comprehension, destroy the sense of immersion, and dramatically reduce bass response (Holman, 2000). This is especially true when dealing with small rooms. If the system is installed properly, there will be a uniform (flat) frequency response at the listening area.

One example of a four-speaker system (two front and two surround) is described in an International Telecommunications Union (ITU) recommendation and places speakers at ±30° in front of the listener and at ±110° behind the listener (ITU-R BS. 775-1). It is further recommended that the signals emanating from the two surround channels be decorrelated to increase the sense of spaciousness. Correlated mono signals may give a sense of lateralization rather than localization. If a subwoofer is used, it is usually placed in front of the room. Placing the subwoofer too close to a corner may increase bass response but may result in a muddier sound. The subwoofer should be moved to achieve the best response in the listening area. Unfortunately, speaker placement will vary depending on the dimensions and shape of the room, as well as the number of speakers employed. If the system is mobile, the sound system will have to be readjusted for every new location, unless the simulation incorporates its own enclosure. If the simulation will be housed in different sized rooms, the audio system (amplifiers and speakers) must have enough headroom (power) to accommodate both large enclosures as well as small. When possible, acoustical tile and diffusers should be employed where appropriate to reduce reverberation and echoes.

4.4.3.1 Nonspatial Display

Many systems use free-field speakers in which each speaker presents an identical signal. Such systems are analogous to diotic headphone systems; although simple to develop, these displays (like diotic headphone displays) provide no spatial information to the listener. Such systems can be used when spatial auditory information is unimportant and when segregation of simultaneous auditory signals is not critical. For instance, if the only objects of interest are within the visual field and interference between objects is not a concern, this kind of simplistic display may be adequate.

4.4.3.2 Stereo Display

The analog of the dichotic headphone display presents signals from two speakers simultaneously in order to control the perceived laterality of a *phantom* source. For instance, simply varying the level of otherwise identical signals played from a pair of speakers can alter the perceived laterality of a phantom source. Most commercial stereo recordings are based on variations of this approach.

Imagine a listener sitting equidistant from two loudspeakers positioned symmetrically in front of the listener. When the left speaker is played alone, the listener hears a source in the direction of the left speaker (and ITD and ILD cues are consistent with a source in that leftward direction). When the right speaker is played alone, the listener hears a source in the direction of the right speaker. When identical signals at identical levels are played from both speakers, each ear receives two direct signals, one from each of the symmetrically placed speakers. To the extent that the listener's head is left–right symmetric, the total direct sound in each ear is identical, and the resulting percept will be of a single source at a location that gives rise to zero ITD and zero ILD (e.g., in the listener's median plane). Varying the relative intensity of otherwise identical signals played from the two speakers causes the gross ITD and ILD cues to vary systematically, producing a phantom source whose location between the two speakers varies systematically with the relative speaker levels (e.g., see Bauer, 1961).

This simple *panning* technique produces a robust perception of a source at different lateral locations; however, it is nearly impossible to precisely control the exact location of the phantom image. In particular, the way in which the perceived direction changes with relative speaker level depends upon the location of the listener with respect to the two loudspeakers. As the listener moves outside a restricted area (the *sweet spot*), the simulation degrades rather dramatically. In addition, reverberation can distort the interaural cues, causing biases in the resulting simulation. Nonetheless, such systems provide some information about source laterality and can be very effective when one wishes to simulate sounds from angular positions falling between the loudspeaker positions.

4.4.3.3 Multichannel Loudspeaker Systems

Two-channel stereo can be extended to multichannel panning techniques, so-called vector base amplitude panning (VBAP). The technique can be used to place virtual source in 3D space if the loudspeaker arrays are surrounding the listener (Pulkki, 1997). Loudspeaker triplets are used with particular amplitudes in order to create a sound image in the corresponding triangle. Larger areas around the listener are covered by several of those triplets.

Another technique, actually one of the most popular techniques of spatial audio, is *Ambisonics* (Gerzon, 1976). The mathematical basis is the set of spherical harmonics (SH), a set of orthogonal functions in spherical coordinates. With these, sound fields can be decomposed into their directional components (SH coefficients); these exact coefficients are used to filter the speakers of the reproduction array. The speaker arrays can be freely designed in 3D space, but simple solutions correspond to Platonic bodies such as cubes, dodecahedrons, or icosahedrons. The SH coefficients can be derived from simulation or from recordings with spherical microphone arrays. The first-order approach as defined by Gerzon requires a setup of an omnidirectional and three figure-eight microphones. The spatial resolution and the corresponding details of directional sounds are limited by the low-order SH representation, which creates a kind of spatial smoothing. Higher-order Ambisonics (HOA) allow reproduction of more spatial detail; for this, higher-order microphone arrays must be used (Meyer & Elko, 2002). Wave field synthesis (WFS) technology is another theoretical approach to wave field reconstruction (Berkhout, 1988). In WFS, sound waves are decomposed into plane waves sampled on linear or circular microphone arrays, typically in 2D. If the discrete spatial sampling of the array is sufficiently high, any wave field can be reconstructed with a corresponding large number of loudspeakers in a dense distribution. This goes back to the Huygens principle that explains wave propagation by arrays of numerous point sources, each radiating elementary waves, which interfere to form wave fronts (e.g., see De Vries, 2012). To use WFS, the process of sound recording and mixing is different from usual techniques applied in audio engineering. The spatial decoding of virtual sources is integrated in a flexible way so that position, orientation, and movements of the listener are not restricted in dynamic scenes.

All of the multichannel techniques listed suffer from the fact that a large number of loudspeakers must be used with accordingly a large amount of signal processing, control, and amplifier units. For VR installations with surrounding displays, a practical problem often arises from the conflicting demands of trying to place loudspeaker arrays along with video screens with an undisturbed video image and free line of sight. Acoustically transparent video screens would solve the problem, but the image resolution of such screens is significantly less than for high-quality hard projection screens.

4.4.3.4 Cross-Talk Cancellation and Transaural Simulations

More complex signal-processing schemes can also be used to control spatial cues using a small setup of loudspeakers. In such approaches, the total signal reaching each ear is computed as the sum of the signals reaching that ear from each of the speakers employed. By considering the timing and content of each of these signals, one can try to reproduce the exact signal desired at each ear.

The earliest such approach attempted to recreate the sound field that a listener would have received in a particular setting from stereo recordings taken from spatially separated microphones. In the playback system, two speakers were positioned at the same relative locations as the original microphones. The goal of the playback system was to play signals from the two speakers such that the total signal at each ear was equal to the recorded signal from the nearer microphone. To the extent that the signal from the far speaker was acoustically canceled, the reproduction would be accurate. Relatively simple schemes involving approximations of the acoustic alterations of the signals as they impinged on the head were used to try to accomplish this *cross-talk cancellation*.

As signal-processing approaches have been refined and knowledge of the acoustic properties of HRTFs improves, more sophisticated algorithms have been developed. In particular, it is possible to calculate the contribution of each speaker to the total signal at each ear by considering the HRTF corresponding to the location of the speaker. The total signal at each ear is then the sum of the HRTF-filtered signals coming from each speaker. If one also knows the location and source of the signal that is to be simulated, one can write equations for the desired signals reaching each ear as HRTF-filtered versions of the desired source. Combining these equations yields two frequency-dependent, complex-valued equations that relate the signals played from each speaker to the desired signals at the ears. To the extent that one can find and implement solutions to these equations, it is possible (at least in theory) to recreate the desired binaural signal by appropriate choice of the signals played from each speaker. The problem with such approaches is that the simulation depends critically on the relative location of the speakers and the listener. In particular, if the listener moves his head outside of the sweet spot, the simulation degrades rapidly. Head tracking and dynamic cross-talk filtering is one solution to this problem, which also allows head movements explicitly so that the listener can benefit from a more natural behavior in the virtual scene (Lentz & Behler, 2004). Head trackers can be used in conjunction with multispeaker simulations in order to improve the simulation. However, this requires that computations be performed in real time and significantly increases the cost and complexity of the resulting system. It can be difficult to compute the required loudspeaker signals and the computations are not particularly stable numerically (Masiero & Vorländer, 2012). The technique called *stereo dipole* or *sound bar* is actually more stable. It is an optimized distributed source approach where the high frequencies are radiated from the frontal region while the loudspeakers for the low frequencies span a larger angle (Takeushi, Teschl, & Nelson, 2001). To the extent that reverberation in the listening space further distorts the signals reaching the ears, the derived solutions are less robust than those for headphones. In all cases of binaural reproduction using headphones or cross-talk cancellation, the quality is best if the HRTFs used in the equations are matched to the listener.

4.4.3.5 Lessons from the Entertainment Industry

The ability to generate an accurate spatial simulation using loudspeakers increases dramatically as the number of speakers used in the display increases. With an infinite number of speakers around the listener, one would simply play the desired signal from the speaker at the desired location of the source to achieve a *perfect* reproduction. Surround sound technologies, which are prevalent in the entertainment industry, are implemented via a variety of formats. Surround sound systems find their genesis in a three-channel system created for the movie *Fantasia* in 1939.

Fantasound speakers were located in front of the listener at left, middle, and right. The *surround* speakers consisted of approximately 54 speakers surrounding the audience and carried a mixture of sound from the left and right front speakers (i.e., it was not true stereo; Garity & Hawkins, 1941).

Currently, the most common surround sound format is the 5.1 speaker system, in which speakers are located at the left, middle, and right in front of the listener and left and right behind the listener. The so-called 0.1 speaker is a subwoofer. The middle speaker, located in front of the listener, reproduces most of the speech information to the listener. Typical 5.1 surround formats are Dolby Digital Surround and Digital Theater Systems (DTS).

More complex surround sound formats include Dolby Digital Surround EX, which is a 6.1 speaker system (adding a center speaker behind the listener); a 10.2 system has also been developed (Holman, 2000). In the future, even greater numbers of speakers and more complex processing are likely to become standard. However, it is important to note that adding additional speakers may be detrimental to producing a sense of immersion, especially in a small room. As the number of speakers increases, explicit care must be taken to assure that the sound field in the room is diffuse enough that the speakers themselves are not obtrusive. If the user notices the speakers in the room, the illusion of reality will be destroyed. To negate this possibility, extreme care must be taken into account for room acoustics, speaker design and placement, and the location of the listener in the room (Holman).

In general, it is possible to generate relatively realistic phantom sources using multiple loudspeakers, but such techniques often only simulate sound in the left–right (azimuthal) angle direction of the desired source precisely. More complex simulations in which spectral cues accurately simulate the direction of the desired sound source on the cone of confusion are rarely undertaken as they require much more complex signal processing that is tailored to the individual user (and his or her HRTFs). However, oftentimes, robust left–right angular simulations are sufficient, depending on the goals of an auditory display.

4.5 DESIGN CONSIDERATIONS

4.5.1 Defining the Auditory Environment

When creating the auditory portion of a VE, careful attention should be placed on what is absolutely essential for the task. The adage of motion picture sound designers, *see it, hear it* (Holman 1997; Yewdall, 1999), is also valid for designing audio for VEs. The sense of immersion experienced in a movie theater is a carefully orchestrated combination of expertly designed sound effects and skillfully applied auditory ambiences. It is also interesting to note that *realistically* rendered sound is often perceived as emotionally flat in motion pictures. Sound effects are often designed as exaggerated versions of reality to convey emotion or to satisfy the viewers' expectations of reality (Holman). Sound design in VE needs to balance the need for accurate reproduction with the need to make the user emotionally involved in a synthetic environment.

On the hardware side, a simple sound card solution may be adequate if spatialized audio and fidelity are not paramount. However, if there are multiple sound sources and/or a multiuser interface, a special-purpose audio server may be necessary. Just as photos and films of the visual environment are taken during the development process, it is a good idea to make audio recordings, including sound-level measurements, when developing an auditory interface. In addition to cataloguing the different sounds in a real environment, it is also important to systematically measure the intensity of sounds being experienced by the listener. In this manner, the VE developer has a detailed reference with which to compare the real-world auditory environment with the virtual auditory environment. Given the wide dynamic range involved with recording sounds ranging from concussive events to footsteps and the necessity that recordings be absolutely clean and accurate, the best solution may be to rely on professionals for making appropriate recordings that contain the requisite content and

ambience. Different combinations of microphones and recording equipment produce vastly differing sound quality. Choosing the appropriate combination for a particular application is still more of an art than a science. In addition, there are many high-quality commercially available sound libraries for obtaining a wide variety of sound effects and ambiences (Holman, 1997; Yewdall, 1999). Given these facts, the use of prerecorded, high-quality sound content is the de facto standard for many current auditory displays.

4.5.2 How Much Realism Is Necessary?

Much of the research devoted to developing and verifying virtual display technology emphasizes the subjective realism of the display (Chertoff & Schatz, 2013; Chapter 34); however, this is not the most important consideration for all applications. In some cases, signal processing that improves realism actually interferes with the amount of information a listener can extract. For instance, inclusion of echoes and reverberation can significantly increase the perceived realism of a display and improve distance perception. However, echoes can degrade perception of source direction. For applications in which information about the direction of a source is more important than the realism of the display or perception of source distance, including echoes and reverberation may be ill advised.

In headphone-based systems, realism is enhanced with the use of individualized HRTFs, particularly in the perception of up/down and front/back positions. Ideally, HRTFs should also be sampled in both distance and direction at a spatial density dictated by human sensitivity. Thus, while the most *realistic* system would use individualized HRTFs that are sampled densely in both direction and distance, most systems use generic HRTFs sampled coarsely in direction and at only one distance.

If a particular application requires a listener to extract 3D spatial information from the auditory display, HRTFs may have to be tailored to the listener to preserve directional information and reverberation may have to be included to encode source distance. On the other hand, if a particular application only makes use of one spatial dimension (for instance, to indicate the direction that a blind user must turn), coarse simulation of ITD and ILD cues (even without detailed HRTF simulation) is probably adequate.

If information transfer is of primary importance, it may be useful to present acoustic spatial cues that are intentionally distorted so that they are perceptually more salient than more *realistic* cues. For instance, it may be useful to exaggerate spatial auditory cues to improve auditory resolution; however, such an approach requires that listeners are appropriately trained with the distorted cues. With such training, listeners can respond accurately with altered cues but also respond accurately to *normal* localization cues with little or no negative aftereffects of training, suggesting that both the new and the old mappings from spatial cues to exocentric location can be maintained simultaneously (e.g., see Hofman et al., 1998; Shinn-Cunningham, 2000a).

The processing power needed to simulate the most realistic virtual auditory environment possible is not always cost-effective. For instance, the amount of computation needed to create realistic reverberation in a VE may not be justifiable when source distance perception and subjective realism are not important. Other acoustic effects are often ignored in order to reduce computational complexity of an acoustic simulation, including the nonuniform radiation pattern of a realistic sound source, spectral changes in a sound due to atmospheric effects, and the Doppler shift of the received spectrum of moving sources. The perceptual significance of many of these effects is not well understood; further work must be done to examine how these factors affect the realism of a display, as well as what perceptual information such cues may convey.

In command-and-control applications, the goal is to maximize information transfer to the human operator; subjective impression (i.e., realism) is unimportant. In these applications, both technological and perceptual issues must be considered to achieve this goal. If nonverbal warnings or alerts are created, the stimuli must be wideband enough to be localizable. In addition, stimuli should be significantly intense to be at least 15 dB above background noise level. Stimulus onset should be fairly gradual so as not to be excessively startling to the user. In many instances, one may want the stimuli to be aesthetically pleasing to the user. As can be imagined, creating acceptable spatialized auditory displays is no trivial chore and should involve formal evaluations to ensure perceptual accuracy and system usability. For applications in which speech is the main signal of interest, basic interaural cues are important for preserving speech intelligibility, particularly in noisy, multisource environments. On the other hand, there is probably little benefit gained from including the detailed frequency dependence of normal HRTFs. In entertainment applications (Greenwood-Ericksen, & Stafford, 2013; Chapter 49), cost is the most important factor; the precision of the display is unimportant as long as the simulation is subjectively satisfactory. For scientific research (Polys, 2013; Chapter 49), high-end systems are necessary in order to allow careful examination of normal spatial auditory cues. In clinical applications, the auditory display must only be able to deliver stimuli that can distinguish listeners with normal spatial hearing from those with impaired spatial hearing. Such systems must be inexpensive and easy to use, but there is no need for a perfect simulation.

REFERENCES

Algazi, V. R., Avendano, C., & Duda, R. O. (2001). Elevation localization and head-related transfer function analysis at low frequencies. *Journal of the Acoustical Society of America*, *109*(3), 1110–1122.

Aretz, M., Maier, P., & Vorländer, M. (November 2010). Simulation based auralization of the acoustics in a studio room using a combined wave and ray based approach. VDT Tonmeistertagung, Leipzig, Germany. (8 pages)

Badcock, D. R., Palmisano, S., & May, J. G. (2014). Vision and virtual environments. In K. S. Hale & K. M. Stanney (Eds.), *Handbook of virtual environments* (2nd ed., pp. 39–86). Boca Raton, FL: CRC Press.

Batteau, D. W. (1967). The role of the pinna in human localization. *Proceedings of the Royal Society of London*, *B168*, 158–180.

Bauer, B. B. (1961). Phasor analysis of some stereophonic phenomena. *Journal of the Acoustical Society of America*, *33*(11), 1536–1539.

Begault, D., Wenzel, E., Lee, A., & Anderson, M. (2012, July). *Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source*. In the 108th Audio Engineering Society (AES) Convention, Vol. 5134, Paris, France.

Begault, D. R. (1993a). Head-up auditory displays for traffic collision avoidance system advisories: A preliminary investigation. *Human Factors*, *35*(4), 707–717.

Begault, D. R. (1996, November). *Virtual acoustics, aeronautics, and communications*. Presented at the 101st convention of the Audio Engineering Society, Los Angeles, CA.

Begault, D. R. (1999). Virtual acoustic displays for teleconferencing: Intelligibility advantage for "telephone-grade" audio. *Journal of the Audio Engineering Society, 47*(10), 824–828.

Begault, D. R., & Wenzel, E. M. (1992). Techniques and applications for binaural sound manipulation in human–machine interfaces. *The International Journal of Aviation Psychology*, 2(1), 1–22.

Begault, D. R., & Wenzel, E. M. (1993). Headphone localization of speech. *Human Factors*, *35*(2), 361–376.

Berkhout, A. J. (1988). A holographic approach to acoustic control. *Journal of the Audio Engineering Society*, *36*, 977–995.

Blauert, J. (1997). *Spatial hearing* (2nd ed.). Cambridge, MA: MIT Press.

Botteldooren, D. (1995). Finite-difference time-domain simulation of low-frequency room acoustic problems. *Journal of the Acoustical Society of America*, *98*(6), 3302–3308.

Brainard, M. S., Knudsen, E. I., & Esterly, S. D. (1992). Neural derivation of sound source location: Resolution of ambiguities in binaural cues. *Journal of the Acoustical Society of America*, *91*, 1015–1027.

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound.* Cambridge, MA: MIT Press.

Brewster, S. A., Wright, P. C., & Edwards, A. D. N. (1993). An evaluation of earcons for use in auditory human-computer interfaces. In *Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems* (pp. 222–227). New York, NY: ACM.

Bronkhorst, A. W., & Plomp, R. (1988). The effect of head-induced interaural time and level differences on speech intelligibility in noise. *Journal of the Acoustical Society of America*, *83*, 1508–1516.

Bronkhorst, A. W., Veltman, J. A., & van Breda, L. (1996). Application of a three-dimensional auditory display in a flight task. *Human Factors*, *38*(1), 23–33.

Brungart, D. S., & D'Angelo, W. (1995). Effects of reverberation cues on distance identification in virtual audio displays. *The Journal of the Acoustical Society of America*, *97*, 3279.

Brungart, D. S., & Rabinowitz, W. M. (1999). Auditory localization of nearby sources: Near-field head-related transfer functions. *Journal of the Acoustical Society of America*, *106*, 1465–1479.

Carlyon, R. P. (2004). How the brain separates sounds. *Trends in Cognitive Science*, *8*, 465–471.

Chandak, A., Lauterbach, C., Taylor, M., Ren, Z., & Manocha, D. (2008, October 19–24). AD-frustum: Adaptive frustum tracing for interactive sound propagation. In *Proceedings of the IEEE Visualization*, Columbus, OH.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, *25*, 975–979.

Chertoff, D., & Schatz, S. (2014). Beyond presence. In K. S. Hale & K. M. Stanney (Eds.), *Handbook of Virtual Environments* (2nd ed., pp. 855–870). Boca Raton, FL: CRC Press.

Chon, S. H., & McAdams, S. (2012). Investigation of timbre saliency, the attention-capturing quality of timbre. *Journal of the Acoustical Society of America*, *13*, 3433–3433.

Clark, B., & Graybiel, A. (1949). The effect of angular acceleration on sound localization: The audiogyral illusion. *The Journal of Psychology*, *28*, 235–244.

Clifton, R. K., Freyman, R. L., & Litovsky, R. L. (1993). Listener expectations about echoes can raise or lower echo threshold. *Journal of the Acoustical Society of America*, *95*, 1525–1533.

Darwin, C. J. (1997). Auditory grouping. *Trends in Cognitive Sciences*, *1*, 327–333.

De Vries, D. (2012). Sound reinforcement by wavefield synthesis: Adaptation of the synthesis operator to the loudspeaker directivity characteristics. In *Proceedings of the Audio Engineering Society Convention* (Paper No. 8661), San Francisco, CA.

Dindar, N., Tekalp, A. M., & Basdogan, C. (2014). Dynamic haptic interaction with video. In K. S. Hale & K. M. Stanney (Eds.), *Handbook of virtual environments* (2nd ed., pp. 117–132). Boca Raton, FL: CRC Press.

DiZio, P., Held, R., Lackner, J. R., Shinn-Cunningham, B. G., & Durlach, N. I. (2001). Gravitoinertial force magnitude and direction influence head-centric auditory localization. *Journal of Neurophysiology*, *85*, 2455–2560.

Drullman, R., & Bronkhorst, A. W. (2000). Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation. *Journal of the Acoustical Society of America*, *107*(4), 2224–2235.

Duda, R. O., & Martens, W. L. (1998). Range dependence of the response of a spherical head model. *Journal of the Acoustical Society of America*, *104*(5), 3048–3058.

Durlach, N. I., & Colburn, H. S. (1978). Binaural phenomena. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (Vol. 4, pp. 365–466). New York, NY: Academic Press.

Durlach, N. I., Shinn-Cunningham, B. G., & Held, R. M. (1993). Supernormal auditory localization. I. General background. *Presence*, *2*(2), 89–103.

Frens, M. A., van Opstal, A. J., & van der Willigen, R. F. (1995). Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. *Perception & Psychophysics*, *57*, 802–816.

Freyman, R. L., Clifton, R. K., & Litovsky, R. Y. (1991). Dynamic processes in the precedence effect. *Journal of the Acoustical Society of America*, *90*, 874–884.

García, G. (2002). Optimal filter partition for efficient convolution with short input/output delay. In *Proceedings of Audio Engineering Society Convention* (Paper No. 5660), Los Angeles, CA.

Garity, W. E., & Hawkins, J. A. (1941, August). Fantasound. *Journal of the Society of Motion Picture Engineers* *37*, 127–146.

Gelfand, S. A. (1998). *Hearing: An introduction to psychological and physiological acoustics*. New York, NY: Marcel Dekker, Inc.

Gerzon, M. (1976). Multidirectional sound reproduction systems. UK-Patent 3,997,725.

Gilkey, R. H., & Anderson, T. R. (1995). The accuracy of absolute localization judgments for speech stimuli. *Journal of Vestibular Research*, *5*(6), 487–497.

Gilkey, R. H., & Good, M. D. (1995). Effects of frequency on free-field masking. *Human Factors*, *37*(4), 835–843.

Grantham, D. W. (1997). Auditory motion perception: Snapshots revisited. In R. Gilkey & T. Anderson (Eds.), *Binaural and spatial hearing in real and virtual environments* (pp. 295–314). New York, NY: Lawrence Erlbaum Associates.

Greenwood-Ericksen, A., Kennedy, R. C., & Stafford, S. (2014). Entertainment applications of virtual environments. In K. S. Hale & K. M. Stanney (Eds.), *Handbook of virtual environments* (2nd ed., pp. 1291–1316). Boca Raton, FL: CRC Press.

Haas, E. C., Gainer, C., Wightman, D., Couch, M., & Shilling, R. D. (1997, September). Enhancing system safety with 3-D audio displays. In *Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting* (pp. 868–872), Albuquerque, NM.

Hartmann, W. M. (1983). Localization of sound in rooms. *Journal of the Acoustical Society of America*, *74*, 1380–1391.

Hendrix, C., & Barfield, W. (1996). The sense of presence in auditory virtual environments. *Presence*, *5*(3), 290–301.

Hofman, P. M., Van Riswick, J. G. A., & Van Opstal, A. J. (1998). Relearning sound localization with new ears. *Nature Neuroscience*, *1*(5), 417–421.

Holman, T. (1997). *Sound for film and television*. Boston, MA: Focal Press.

Holman, T. (2000). *5.1 Surround sound: Up and running*. Boston, MA: Focal Press.

Kim, R., Peters, M. A., & Shams, L. (2012). 0+1 > 1: How adding noninformative sound improves performance on a visual task. *Psychological Science*, *23*(1), 6–12.

Kistler, D. J., & Wightman, F. L. (1991). A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *Journal of the Acoustical Society of America*, *91*, 1637–1647.

Kollmeier, B., & Gilkey, R. H. (1990). Binaural forward and backward masking: Evidence for sluggishness in binaural detection. *Journal of the Acoustical Society of America*, *87*(4), 1709–1719.

Langendijk, E. H., & Bronkhorst, A. W. (2000). Fidelity of three-dimensional-sound reproduction using a virtual auditory display. *Journal of the Acoustical Society of America*, *107*(1), 528–537.

Lawson, B. D., & Riecke, B. E. (2014). Perception of body motion. In K. S. Hale & K. M. Stanney (Eds.), *Handbook of virtual environments* (2nd ed., pp. 115–130). New York, NY: CRC Press.

Lawson, B. D., & Riecke, B. E. (2014). Perception of body motion. In K. S Hale & K. M. Stanney (Eds.), *Handbook of virtual environments* (2nd ed., pp. 163–196). Boca Raton, FL: CRC Press.

Lentz, T., & Behler, G. K. (2004). Dynamic crosstalk cancellation for binaural synthesis in virtual environments. In *Proceedings of the 117th Convention of the Audio Engineering Society* (p. 6315), San Francisco, CA, October 2004.

Letowski, T., Karsh, R., Vause, N., Shilling, R., Ballas, J., Brungart, D., & McKinley, R. (2001). Human factors military lexicon: Auditory displays. ARL Technical Report, ARL-TR-2526; APG (MD).

Maddox, R., & Shinn-Cunningham, B. B. (2011). Influence of task-relevant and task-irrelevant feature continuity on selective auditory attention. *Journal of the Association for Research in Otolaryngology*, *13*, 119–129.

Masiero, B., & Vorländer, M. (2012). A framework for the calculation of dynamic crosstalk cancellation filters. *IEEE Transactions on Audio, Speech and Language Processing* (under review).

McMahan, R. P., Kopper, R., & Bowman, D. A. (2014). Principles for designing effective 3D interaction techniques. In K. S Hale & K. M. Stanney (Eds.), *Handbook of virtual environments* (2nd ed., pp. 285–312). Boca Raton, FL: CRC Press.

Mershon, D. H. (1997). Phenomenal geometry and the measurement of perceived auditory distance. In R. Gilkey & T. Anderson (Eds.), *Binaural and spatial hearing in real and virtual environments* (pp. 251–214). New York, NY: Lawrence Erlbaum Associates.

Mershon, D. H., Ballenger, W. L., Little, A. D., McMurty, P. L., & Buchanan, J. L. (1989). Effects of room reflectance and background noise on perceived auditory distance. *Perception*, *18*, 403–416.

Meyer, J., & Elko, G. W. (2002, May 13–17). A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield. In *Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing* (Vol. II, pp. 1781–1784), Orlando, FL.

Middlebrooks, J. C. (1997). Spectral shape cues for sound localization. In R. Gilkey & T. Anderson (Eds.), *Binaural and spatial hearing in real and virtual environments* (pp. 77–98). New York, NY: Lawrence Erlbaum Associates.

Middlebrooks, J. C., & Green, D. M. (1991). Sound localization by human listeners. *Annual Review of Psychology*, *42*, 135–159.

Middlebrooks, J. C., & Green, D. M. (1992). Observations on a principal components analysis of head-related transfer functions. *Journal of the Acoustical Society of America*, *92*, 597–599.

Mills, A. W. (1972). Auditory localization. In J. V. Tobias (Ed.), *Foundations of modern auditory theory* (pp. 303–348). New York, NY: Academic Press.

Moldrzyk, C., Ahnert, W., Feistel, S., Lentz, T., & Weinzierl, S. (2004). Head-tracked Auralization of acoustical simulation. In *Proceedings of the 117th Convention of the Audio Engineering Society*, San Francisco, CA, October 2004, p. 6275.

Moore, B. C. J. (1997). *An introduction to the psychology of hearing* (4th ed.). San Diego, CA: Academic Press.

Perrott, D. R., Saberi, K., Brown, K., & Strybel, T. (1990). Auditory psychomotor coordination and visual search behavior. *Perception & Psychophysics*, *48*, 214–226.

Perrott, D. R., Sadralodabai, T., Saberi, K., & Strybel, T. (1991). Aurally aided visual search in the central visual field: Effects of visual load and visual enhancement of the target. *Human Factors*, *33*, 389–400.

Pick, H. L., Warren, D. H., & Hay, J. C. (1969). Sensory conflict in judgements of spatial direction. *Perception & Psychophysics*, *6*, 203–205.

Plenge, G. (1974). On the differences between localization and lateralization. *Journal of the Acoustical Society of America*, *56*(3), 944–951.

Polys, N. (2014). Information visualization in virtual environments. In K. S. Hale & K. M. Stanney (Eds.), *Handbook of virtual environments* (2nd ed., pp. 1267–1296). New York, NY: CRC Press.

Popescu, G. V., Trefftz, H., & Burdea, G. C. (2014). Multimodal interaction modeling. In K. S. Hale & K. M. Stanney (Eds.), *Handbook of virtual environments* (2nd ed., pp. 411–434). Boca Raton, FL: CRC Press.

Pulkki, V. (1997). Virtual sound source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society*, *45*(6), 456–466.

Rakerd, B., & Hartmann, W. M. (1985). Localization of sound in rooms. II. The effects of a single reflecting surface. *Journal of the Acoustical Society of America*, *78*, 524–533.

Rakerd, B., & Hartmann, W. M. (1986). Localization of sound in rooms. III. Onset and duration effects. *Journal of the Acoustical Society of America*, *80*, 1695–1706.

Savioja, L. (2010). Real-time 3D finite-difference time-domain simulation of mid-frequency room acoustics. In *Proceedings of the 13th International Conference on Digital Audio Effects, DAFx* (p. 43), Graz, Austria.

Schröder, D., & Lentz, T. (2006). Real-time processing of image sources using binary space partitioning. *Journal of the Audio Engineering Society*, *54*, 604–619.

Schröder, D., Ryba, A., & Vorländer, M. (2010). Spatial data structures for dynamic acoustic virtual reality. *Proceedings of the International Congress on Acoustics*, Sydney, Australia, August 2010.

Shams, L., Kamitani, Y., & Shimojo, S. (2004). Modulations of visual perception by sound. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *Handbook of multisensory processes* (pp. 27–34). Cambridge, MA: MIT Press.

Sheridan, T. B. (1996). Further musings on the psychophysics of presence. *Presence*, *5*(2), 241–246.

Shilling, R. D., & Letowski, T. (2000). *Using spatial audio displays to enhance virtual environments and cockpit performance*. NATO Research and Technology Agency Workshop entitled What is essential for Virtual Reality to Meet Military Human Performance Goals, The Hague, The Netherlands.

Shinn-Cunningham, B. G. (2000a). Adapting to remapped auditory localization cues: A decision-theory model. *Perception & Psychophysics*, *62*(1), 33–47.

Shinn-Cunningham, B. G. (2000b, April 2–5). Learning reverberation: Implications for spatial auditory displays. In *Proceedings of the International Conference on Auditory Displays* (pp. 126–134), Atlanta, GA.

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences*, *12*, 182–186.

Shinn-Cunningham, B. G., Durlach, N. I., & Held, R. (1998). Adapting to supernormal auditory localization cues. 1: Bias and resolution. *Journal of the Acoustical Society of America*, *103*(6), 3656–3666.

Shinn-Cunningham, B. G., Ihlefeld, A., Satyavarta, & Larson, E. (2005). Bottom-up and top-down influences on spatial unmasking. *Acta Acustica united with Acustica*, *91*, 967–979.

Shinn-Cunningham, B. G., Kopco, N., & Santarelli, S. G. (1999). *Computation of acoustic source position in near-field listening*. Paper presented at the Third International Conference on Cognitive and Neural Systems, Boston, MA.

Shinn-Cunningham, B. G., Santarelli, S., & Kopco, N. (2000). Tori of confusion: Binaural localization cues for sources within reach of a listener. *Journal of the Acoustical Society of America*, *107*(3), 1627–1636.

Sorkin, R. D., Kistler, D. S., & Elvers, G. C. (1989). An exploratory study of the use of movement-correlated cues in an auditory head-up display. *Human Factors*, *31*(2), 161–166.

Stern, R. M., & Trahiotis, C. (1997). Binaural mechanisms that emphasize consistent interaural timing information over frequency. In A. R. Palmer, A. Rees, A. Q. Summerfield, & R. Meddis (Eds.), *Psychophysical and physiological advances in hearing, Proceedings of the XI international symposium on hearing*, August 1997, Grantham, London, U.K.: Whurr Publishers, 1998.

Storms, R. L. (1998). *Auditory-visual cross-modal perception phenomena* (Doctoral dissertation). Naval Postgraduate School, Monterey, CA.

Takeushi, T., Teschl, M., & Nelson, P. A. (2001, June). Subjective evaluation of the optimal source distribution system for virtual acoustic imaging. *Proceedings of the 19th Audio Engineering Society International Conference* (pp. 373–385), Schloss Elmau, Germany, June 21–24, 2001.

THX Certified Training Program. (2000, June). *Presentation materials*. San Rafael, CA.
Trahiotis, C., & Stern, R. M. (1989). Lateralization of bands of noise: Effects of bandwidth and differences of interaural time and phase. *Journal of the Acoustical Society of America*, *86*(4), 1285–1293.
Vorländer, M. (2008). *Auralization—Fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality*. Berlin, Germany: Springer.
Wallach, H. (1940). The role of head movements and vestibular and visual cues in sound localization. *Journal of Experimental Psychology*, *27*, 339–368.
Warren, D. H., Welch, R. B., & McCarthy, T. J. (1981). The role of visual-auditory "compellingness" in the ventriloquism effect: Implications for transitivity among the spatial senses. *Perception & Psychophysics*, *30*, 557–564.
Wefers, F., & Vorländer, M. (2012, October 26–29). Optimal filter partitions for non-uniformly partitioned convolution. In *Proceedings of the 133rd Audio Engineering Society Convention* (pp. 6–4) San Francisco, CA, October 2012.
Welch, R., & Warren, D. H. (1986). Intersensory interactions. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance* (Vol. 2, pp. 25.1–25.36). New York: John Wiley & Sons.
Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin*, *88*, 638–667.
Wenzel, E. M., Arruda, M., Kistler, D. J., & Wightman, F. L. (1993). Localization using nonindividualized head-related transfer functions. *Journal of the Acoustical Society of America* (pp. 102–105), *94*, 111–123.
Wenzel, E. M., & Foster, S. H. (1993, October). Perceptual consequences of interpolating head-related transfer functions during spatial synthesis. In *Proceedings of the 1993 Workshop on the Applications of Signal Processing to Audio and Acoustics*, New York.
Wightman, F. L., & Kistler, D. J. (1989). Headphone simulation of free-field listening. II. Psychophysical validation. *Journal of the Acoustical Society of America*, *85*, 868–878.
Wightman, F. L., & Kistler, D. J. (1993). Sound localization. In W. A. Yost, A. N. Popper, & R. R. Fay (Eds.), *Human psychophysics* (pp. 155–192), New York, NY: Springer Verlag.
Yewdall, D. L. (1999). *Practical art of motion picture sound*. Boston, MA: Focal Press.
Yost, W. A. (1997). The cocktail party problem: Forty years later. In R. Gilkey & T. Anderson (Eds.), *Binaural and spatial hearing in real and virtual environments* (pp. 329–346). New York, NY: Lawrence Erlbaum Associates.
Yost, W. A. (2006). *Fundamentals of hearing: An introduction* (5th ed.). San Diego, CA: Academic Press.
Zahorik, P., Brungart, D. S., & Bronkhorst, A. W. (2005). Auditory distance perception in humans: A summary of past and present research. *Acta Acustica united with Acustica*, *91*, 409–420.
Zurek, P. M. (1993). Binaural advantages and directional effects in speech intelligibility. In G. Studebaker & I. Hochberg (Eds.), *Acoustical factors affecting hearing aid performance* (pp. 255–276). Boston, MA: College-Hill Press.
Zwicker, E., & Fastl, H. (2007). *Psychoacoustics—Facts and models* (3rd ed.). Berlin, Germany: Springer.

5 Dynamic Haptic Interaction with Video

Nuray Dindar, A. Murat Tekalp, and Cagatay Basdogan

CONTENTS

5.1 Introduction 116
5.2 Computation of Force Experienced by the User 117
5.2.1 Static Force 118
5.2.2 Dynamic Force 118
5.2.3 Implementation of Depth Image-Based Haptic Rendering 119
5.2.3.1 Mapping between Image Coordinates and Haptic Coordinates 119
5.2.3.2 Interpolation of Depth Map 120
5.2.3.3 Selection of Mass for Video Object and Haptic Interface Point 121
5.2.3.4 Calculation of Accelerations and Forces 121
5.2.3.5 Temporal Interpolation of Forces 121
5.3 Video Processing for Haptic Rendering 121
5.3.1 Enhancement of Depth Images 122
5.3.2 Motion Estimation 122
5.4 Case Studies and Results 124
5.4.1 Simulation Results 124
5.4.2 Results with a Real Video Clip 127
5.5 Conclusion 127
Acknowledgment 129
References 129

This chapter introduces the notion of passive dynamic haptic interaction with video and describes the computation of force due to relative motion between an object in a video and the haptic interface point (HIP) of a user, given associated pixel-based depth data. While the concept of haptic video, that is, haptic rendering of forces due to geometry and texture of objects in a video from the associated depth data, has already been proposed, passive dynamic haptic interaction with video has not been studied before. It is proposed that in passive dynamic interaction, a user experiences motion of a video object and dynamic forces due to its movement, even though the content of the video shall not be altered by this interaction. To this effect, the acceleration of a video object is estimated using video motion estimation techniques while the acceleration of the HIP is estimated from the HIP position acquired by the encoders of the haptic device. Mass values are assigned to the video object and HIP such that user interaction shall not alter the motion of the video object according to the laws of physics. Then, the dynamic force is computed by using Newton's second law. Finally, it is scaled and displayed to the user through the haptic device in addition to the static forces due to the geometry and texture of the object. Experimental results are provided to demonstrate the difference in rendered forces with and without including the dynamics.