18

Other Audio Applications

A number of audio and speech techniques have not yet been discussed in the book, and this chapter covers some of them. Four areas of application are briefly discussed in separate sections: *virtual reality*, *sonic interaction design*, *computational auditory scene analysis*, and *music information retrieval*. In addition, the last section of this chapter merely lists some techniques that are not covered at all in this book.

Virtual reality environments reproduce sound to the user using some of the sound reproduction techniques described in Chapter 14. However, the techniques required for virtual reality differ from those for basic sound reproduction because the interaction between the user and the virtual reality has a strong effect on the sound, in contrast to traditional sound reproduction where the listener does not affect the sound content. The audio engine synthesizes and reproduces a meaningful representation of the sound scene in the virtual world on-the-fly, depending on the actions of the user in the world. Another field involving interaction between humans and computers in the context of audio techniques is *sonic interaction design*, where new methods for human–computer interaction by means of sound are sought.

The speech recognition techniques described in Section 16.3 aim to analyse speech signals using complex pattern recognition techniques to obtain performance on a par with, or even better than human listeners are capable of. The last two techniques described in this chapter share some similarities with speech recognition, although the goal here is to recognize or analyse music and/or other auditory scenes in a natural world for different applications. *Computational auditory scene analysis* (CASA) refers to techniques that try to reveal all information from the ear canal signals accessible to a human listener. *Music information retrieval*, which can be seen partly as a subtopic of CASA, covers different techniques on how computers could recognize musical structures in audio tracks in the same manner as humans do and use this information to operate intelligently on music databases.

18.1 Virtual Reality and Game Audio Engines

In *virtual reality* techniques, the perception of physical presence in locations elsewhere in the real world or in imaginary worlds is created for a subject (Sherman and Craig, 2003).

Communication Acoustics: An Introduction to Speech, Audio, and Psychoacoustics, First Edition.
Ville Pulkki and Matti Karjalainen.

Optimally, all information received by the subject via his or her senses would mimic the conditions simulated as if they were real. The different modalities of perception in virtual reality applications have to be covered with different devices. The visuals are displayed either on a computer screen or through special stereoscopic displays, and audio may be delivered over loudspeakers or headphones.

The simulated virtual environment can be similar to the real world in order to create a lifelike experience – for example, for training purposes or for virtual tourism. Alternatively, it can differ significantly from reality, as in virtual reality games. A number of subjects may share the same virtual reality, with the ability to perceive each other's *avatars* (or *characters*), which are virtual representations of the users, and potentially to communicate with each other.

In the scope of this book, consideration of the usage of audio in virtual worlds is of interest (Savioja *et al.*, 1999; Svensson and Kristiansen, 2002; Laitinen *et al.*, 2012; Tsingos *et al.*, 2004; Vorländer, 2007). Virtual worlds include a number of *virtual objects*, which may or may not generate sound, as shown in Figure 18.1. *Virtual reality displays* reproduce the surroundings of the avatar to the user. The visual display is based on what the avatar sees from a particular point, which is restricted to certain directions, called the field of vision. Similarly, the audio engine is used to render all sounds that the avatar would hear in the location. In addition to the virtual objects, the virtual worlds may also contain enclosed spaces and acoustically reflective objects, the effect of which on the sound should be taken into account.

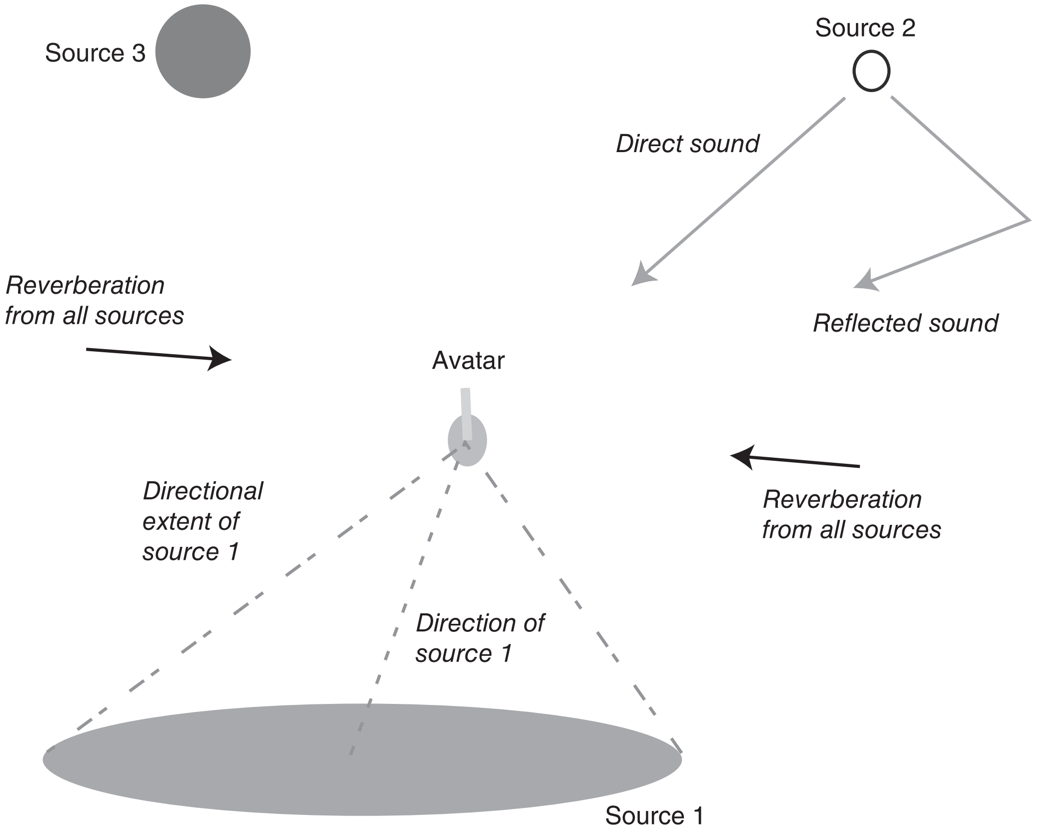
Source 3
Source 2
Direct sound
Reverberation from all sources
Reflected sound
Avatar
Directional extent of source 1
Reverberation from all sources
Direction of source 1
Source 1

Figure 18.1 A virtual world with three virtual objects producing sound. The user should perceive the sources in the directions relative to the avatar. The spatial extents of the sources should also be perceivable. Furthermore, reflections from nearby surfaces should be rendered, as should the diffuse reverberant field generated by all sources.

Audio content production for virtual realities and especially for computer games differs drastically from production for the music and cinema industries. Both the audio content and the engine rendering it have to be composed in production. Often, background music is also used, and this also changes dynamically. This needs some advanced methods for content production and extensive testing of the system. This is called *dynamic audio* or *adaptive audio*, and the traditional audio can then be referred to as *linear audio*.

The tasks performed by the engine are listed below, not all of which are necessary for plausible rendering of virtual reality audio, since the audio may be implemented with varying levels of detail and accuracy, depending on resources and application.

- *Reproduction or synthesis of source signals*. The sounds emanated by the virtual objects into the virtual space have to be generated somehow. In many cases, sound signals which were produced earlier and stored in the memory of a computer are used. A piece of audio to be rendered in virtual reality is also called an *audio asset*. In some cases, the sounds of the virtual objects may be synthesized using physical models of sound sources, as discussed by Cook (2007) and Farnell (2010).
- *Synthesis of source directivity*. If a realistic acoustic virtual environment is sought, the frequency-dependent directivity of the source should be taken into account. The virtual object then emanates different sound in different directions (Savioja *et al.*, 1999). Quite often this effect is ignored, as the plausibility of the virtual world is often not affected much if it is not implemented. The implementation would also be difficult for computational reasons, as the directional patterns of real sources are complex and frequency-dependent.
- *Simulation of the direct sound path*. The transfer function of the virtual transmission channel between the source and the listener may also be modelled with appropriate delaying, amplification, and filtering, as well as some filtering effects of the atmosphere (Savioja *et al.*, 1999). High-speed movement of the source relative to the listener in the physical world results in the Doppler effect, where the perceived pitch of the sound changes with the change in distance between the listener and the source. This effect can be modelled simply in the reproduction of recorded source signals with dynamic sample rate conversion. Alternatively, the effect can be implemented by updating the sound propagation delay correctly. The delay is computed dynamically from the distance between the avatar and the source.
- *Virtual source positioning*. The virtual source positioning techniques discussed in Section 14.5 can be used to reproduce the direct sound(s) arriving at the listener. The virtual source directions are defined by the positions of the virtual objects relative to the avatar.
- *Spatial extent of virtual sources*. A sound source with a considerable volume, such as a group of talking people, should be perceived as spatially broad when the avatar is near the source, and correspondingly it should be perceived to be spatially narrow when the avatar is far away. If the avatar is inside the group, the subject should perceive the virtual source as surrounding him or her. Some techniques to synthesize the extent of sources were discussed in Section 14.5.8.
- *Room effect simulation*. The image–source-method, discussed in Section 2.4.5, is, in principle, a feasible method to simulate the room effect in virtual reality. The positions of image sources can be computed, and they can be reproduced by applying the appropriate direction in virtual source positioning, as above. The delay and gain for sounds of image sources are specified by their geometric location in the virtual world. However, such accurate modelling often leads to computationally complex solutions. Plausible room effects can also be obtained with digital filter structures that simulate the effect of

room reverberation, as discussed in Section 14.8, with considerably lower computational complexity.

- *Distance rendering*. The sensation of *distance* is largely provided by simulating both the direct sound path and the room effect. When they are reproduced correctly, the distance cues (discussed in Section 12.7) based on loudness, direct-to-reverberant (D/R) ratio, early reflections, and spectral content are correct. In loudspeaker-based audio rendering, the virtual sources can be easily positioned at distances as far as or farther than the loudspeakers are. It is hard to bring the sound source closer than the distance of the loudspeakers without special techniques; only some wave-field-synthesis methods with a massive number of loudspeakers are capable of bringing the virtual source into the listening area (see Section 14.5.6). If accurately head-tracked and well-equalized headphone listening is available (see Section 14.6.2), the virtual sources can be rendered close to and far from the listener, since the distance cues based on range dependence of binaural cues at short distances can also be exploited, and since the listening room has no effect on the sound entering the ear canals, making the task easier. In fact, bringing the virtual sources close with headphones is trivial, since the virtual sources are localized inside the head without special treatment. The correct rendering of distances for virtual sources outside the head, on the other hand, demands special techniques.

18.2 Sonic Interaction Design

A field connected strongly to product sound quality is *sonic interaction design*. In sonic interaction design, methods that use sound to convey information, meaning, and aesthetic or emotional qualities in interactive contexts are exploited (Franinovic and Serafin, 2013), thus extending the field of communication acoustics in the direction of interaction design, where human–computer interfaces are studied. Digital devices communicate with the user using non-speech sounds, which is analogous to product sound quality, where sounds made by devices convey information on the state of the device. In this field, the sounds are specially designed and digitally synthesized to communicate this information.

The main research areas are:

- *Perceptual, cognitive, and emotional study of sonic interactions*. This is an area of the field that addresses human perception of sound in interactive conditions in general (Aglioti and Pazzaglia, 2010; De Lucia *et al.*, 2009).
- *Product sound design*. This area concerns the practices and principles of designing sounds of products to maximize positive effects and minimize negative effects in interaction (Franinovic *et al.*, 2007; Hug, 2008).
- *Sonification*. The goal of sonification is to represent data provided by any process in the form of sound, so that the user can perceive or interpret the conveyed information just by listening (Hermann *et al.*, 2011). In a practical example, the angling of a drilling machine is sonified in order to help the user to drill holes more accurately (Großhauser and Hermann, 2010).

Sonic interaction design is thus approached using knowledge from a number of fields, such as interactive arts, electronic music, cultural studies, psychology, cognitive sciences, and communication acoustics. It shifts the focus from reception-based psychoacoustic studies to studies of perception of sound in active, embodied, and emotionally engaging situations. Multimodality, especially the tight connection between audition, haptics, and gestures, is examined in a unified framework.

18.3 Computational Auditory Scene Analysis, CASA

Auditory scene analysis (ASA) was presented in Section 11.7.2. *Computational auditory scene analysis* (CASA) is defined by Wang and Brown (2006) as 'the field of computational study that aims to achieve human performance in ASA by using one or two microphone recordings of the acoustic scene.' Note that the definition does not say that the human auditory system should be imitated in the processing, only that the performance of the system should match that of a human. Human performance seems difficult to achieve, as we are very good at analysing complex acoustic scenes containing sound from multiple sources through reverberant rooms. The current CASA methods hardly achieve human performance levels in such tasks.

Some of the audio and speech technologies discussed in this book are included within, or at least related to, CASA. For example, speech recognition (Section 16.3), hearing aids and cochlear implants (Sections 19.5 and 19.6), and music information retrieval (Section 18.4) aim to analyse sound scenes as well as humans do. Recent trends in CASA are reviewed by Wang and Brown (2006), where the following topics are discussed:

- *Multiple-f_0 estimation techniques* for revealing the fundamental frequencies of simultaneous harmonic sources.
- *Segregation of monaural signals*, where a mixture of simultaneous speech or other signal sources and interfering noise sources is presented to a system, and different features are analysed from the summed signal in the time-frequency domain, targeting the segregation of individual source signals.
- *Binaural sound localization and segregation*, where the sources are localized and their signals are segregated.
- *Analysis of musical audio signals*, where musical structures are analysed from monophonic signals.
- *Robust speech recognition*, which aims to recognize speech as well as humans do in natural conditions, such as free-form discussions between individuals or speech in noisy conditions with unlimited vocabularies, to name but two.

18.4 Music Information Retrieval

Consider a track of musical content stored as a file containing the signals to be reproduced over each loudspeaker. When humans listen to the track, they perceive musical structures, such as tempo, beat, melodies, harmonies, and genre. The presentation of the music as loudspeaker signals does not *per se* imply anything about the musical content in the track, which is, of course, not a problem when reproducing the sound with audio devices. However, as the databases of music nowadays contain millions of tracks, it would be useful to be able to make automatic queries on the database based on some higher-level specifications. For example, if only audio signals are available, it is not possible to make the request 'I want to hear some sturdy bebop with no trumpet, a female vocalist singing in French, and with electric guitar'. A musically trained listener may be able to pick out such tracks using his or her memory and by listening to available tracks. A computer-based solution would be appealing, since such a task could potentially be automated and computers are much faster at such tasks than humans.

Music information retrieval (MIR) is the area where the strategies for enabling access to music collections, both new and historical, are developed in order to keep up with the expectations of search and browse functionality. These techniques are interesting to end users of music, since they would enable someone to find and use music in a personalized way. Also, different professionals in music, such as music performers, teachers, musicologists, copyright lawyers, and music producers would benefit from the ability to make such structured queries of music databases (Casey *et al.*, 2008).

A widely used application of MIR is content-based music description, where the system identifies what the user is seeking even when he does not know specifically what he or she is looking for. Many commercial applications can identify the original recording from a short and noisy sample taken with a handheld device, for example, in a public space where music is played from recordings. The application identifies the artist, album, and track title by finding the best match to the sample from a large music database.

The driver of MIR systems is metadata. Music tracks are analysed by different systems, and metadata that describes the tracks from different aspects is collected. The simplest type of metadata is *factual metadata*, which describes the objective truths about a track, such as the performer, the name of the piece, and so on. The factual metadata can also be accompanied by *cultural metadata*, which describes properties like mood, emotion, genre, style, and so forth. Most often, these attributes are specified to the database by expert listeners, implying that each track added to a database should first be listened to. The fast pace of music creation motivates the use of automatic systems for these tasks.

MIR systems would thus need automatic methods that analyse musical structures from audio tracks. There are several approaches to this, as summarized by Casey *et al.* (2008). The methods resemble speech recognition techniques, where the input signal is divided into short time frames and complex machine-learning techniques are utilized to decode the content of speech. The task of deciphering the musical structures from a single audio signal is a slightly different task. Music typically has multiple sources that are concurrently active, making the recognition of melody and harmony a complex task. Music also has prominent structures organized in both time and frequency.

Some subtopics that have been researched in MIR are:

- *Beat tracking*. The automatic estimation of the temporal structure of music, such as musical beat, tempo, rhythm, and meter is called beat tracking. Many approaches exist to this end. See, for example, Dannenberg (2005).
- *Melody and bass estimation*. The automatic estimation of melody and bass lines is important because the melody forms the core of Western and many other music genres, and it is also a strong indicator of the identity of a piece of music. Often, the bass line also helps in revealing the harmonic progression. Although the measurement of the lines has, in the past, been problematic, state-of-the-art technologies for automatic melody/bass estimation have reached a state whereby they are able to deal with polyphonic music recordings (Poliner *et al.*, 2007; Ryynänen and Klapuri, 2006).
- *Chord and key recognition*. The information on musical chords and keys is an important part of Western music, and it can be used to understand musical structures. Chord- and key-recognition systems based on the use of HMMs have been found to perform well in the task; these unify recognition and smoothing into a single probabilistic framework (Harte *et al.*, 2006; Lee and Slaney, 2008).

- *Music structure*. Music often includes nested repetitive structures, such as the drum pattern that often repeats itself within each bar and the melodic line that is often repetitive in longer units. The analysis of structures such as verse–verse–chorus–verse is useful for music database applications. Furthermore, such automatic structure extraction can be used in other applications, such as to facilitate the editing of audio in recording workflows (Fazekas and Sandler, 2007).

18.5 Miscellaneous Applications

A number of techniques in audio and speech processing that have not been covered elsewhere in this book are briefly mentioned in the list below, with the intention of communicating to the reader fields related to or within communication acoustics which fall beyond the scope of this book.

- *Beamforming* or *spatial filtering* is a signal processing technique used with arrays of microphones for capturing or enhancing sound sensitive to direction. It processes and combines the acoustic signals of the separate microphones to form a desired beam pattern. Traditional beamforming techniques, such as delay-and-sum or filter-and-sum, combine the signals of the microphones in such a way that signals at particular angles experience constructive interference while others experience destructive interference (Bitzer and Simmer, 2001). More advanced beamforming algorithms offer the capability of focusing in specific directions of a target sound while attenuating interferers and noise originating from other directions, with the most popular algorithm being the *linear constrained minimum variance* (LCMV) or the Frost algorithm (Benesty *et al.*, 2008). Perceptually motivated methods for beamforming have also been proposed to reduce perceivable processing artefacts by humans (Delikaris-Manias and Pulkki, 2013; Faller *et al.*, 2010). Due to the reciprocity between acoustic sources and receivers, loudspeakers can also be used for beamforming.
- *Blind source separation* – the goal here is to separate the components making up a combined signal. The task is relatively simple if the number of sources is at most equal to the number of microphones (Cardoso, 1998). Unfortunately, if the number of sources is higher, and if reflections and reverberation exist in the mixed signal, this task is very difficult. One of the main techniques in this field is *independent component analysis* (ICA) (Hyvärinen *et al.*, 2004).
- *Dereverberation* attempts to remove partially or completely the physical or perceptual effects generated by reverberation from audio or speech signals (Hatziantoniou and Mourjopoulos, 2004).
- *Watermarking* is the process of embedding information into an audio signal so that it is difficult to remove and impossible to perceive by listening. If the signal is copied, this information is also carried in the copy. Watermarking has become increasingly important to enable copyright protection and ownership verification (Bliem *et al.*, 2013; Cvejic and Seppanen, 2008). Watermarking is also used to detect tampering, for copy control, for broadcast monitoring, for transmission of metadata, and in some cases even for echo cancelling (Szwoch *et al.*, 2009).
- *Audio forensics* concerns the acquisition, analysis, and evaluation of sound recordings that may ultimately be presented as admissible evidence in a court of law or some other official venue (Maher, 2009). Its primary characteristics are 1) establishing the authenticity of audio

evidence, 2) enhancing audio recordings to improve speech intelligibility and audibility of low-level sounds, and 3) interpreting and documenting sonic evidence, such as identifying speakers, transcribing dialogue, and reconstructing crime or accident scenes and timelines.

- *Auditory displays* are strongly related to sonic interaction design, though the focus is more on how some information can be delivered from a computer to a human (Kramer, 1994) and less on interaction. A subfield is *sonification*, defined in Section 18.2, though in relation to auditory displays the focus is more on the sonification of complex and multidimensional data sets, as discussed by (Hermann and Ritter, 2004). Another area within auditory displays is *earcons* or *auditory icons*, which are brief, distinctive sounds used to represent a specific event or to convey other information, just like icons in visual displays (McGookin and Brewster, 2004).
- *Semantic audio* – semantic technology involves some kind of understanding of the meaning of the information it deals with. Semantic audio is the field of audio technology covering the development of applications which utilize semantic analysis methods on the audio content and thereby enable new features and functionalities. Examples of such applications are automated mixing of audio, source separation, upmixing, content retrieval, and intelligent audio effects. The definition of semantic audio thus overlaps with music information retrieval, sound reproduction, and audio effects. An overview of the topics in semantic audio can be seen in the programmes of AES conferences containing the title ‘Semantic Audio’, for example `www.aes.org/publications/conferences/?confNum=53`. The semantic analysis of audio is discussed by Lerch (2012) and Schuller (2013).

Summary

This chapter overviewed techniques for virtual reality audio, sonic interaction design, and music information retrieval. Furthermore, some other techniques, interesting in the scope of this book, were briefly reviewed.

Further Reading

The book by (Vorländer, 2007) might be interesting to readers who want to learn more about the physics behind acoustic virtual reality. More information on music information retrieval and the analysis of music signals is given by (Gold *et al.*, 2011; Muller *et al.*, 2011).

References

Aglioti, S.M. and Pazzaglia, M. (2010) Representing actions through their sound. *Exper. Brain Res.*, **206**(2), 141–151.

Benesty, J., Chen, J., and Huang, Y. (2008) *Microphone Array Signal Processing*. Springer.

Bitzer, J. and Simmer, K.U. (2001) Superdirective microphone arrays. In Brandstein, M. and Ward, D. (eds) *Microphone Arrays*. Springer, pp. 19–38.

Bliem, T., Galdo, G.D., Borsum, J., Craciun, A., and Zitzmann, R. (2013) A robust audio watermarking system for acoustic channels. *J. Audio Eng. Soc.*, **61**(11), 878–888.

Cardoso, J.F. (1998) Blind signal separation: statistical principles. *Proc. IEEE*, **86**(10), 2009–2025.

Casey, M.A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., and Slaney, M. (2008) Content-based music information retrieval: Current directions and future challenges. *Proc. IEEE*, **96**(4), 668–696.

Cook, P.R. (2007) *Real Sound Synthesis for Interactive Applications*. AK Peters.

Cvejic, N. and Seppanen, T. (eds) (2008) *Difital Audio Watermarking Techniques and Technologies: Applications and Benchmarks*. InormationScience Reference.

Dannenberg, R.B. (2005) Toward automated holistic beat tracking, music analysis and understanding. *Int. Conf. Music Inform. Retrieval*, pp. 366–373.

De Lucia, M., Camen, C., Clarke, S., and Murray, M.M. (2009) The role of actions in auditory object discrimination. *Neuroimage*, **48**(2), 475–485.

Delikaris-Manias, S. and Pulkki, V. (2013) Cross pattern coherence algorithm for spatial filtering applications utilizing microphone arrays. *IEEE Trans. Audio, Speech, and Language Proc.*, **21**(11), 2356 – 2367.

Faller, C., Favrot, A., Langen, C., Tournery, C., and Wittek, H. (2010) Digitally enhanced shotgun microphone with increased directivity. *Audio Eng. Soc. Convention 129* AES.

Farnell, A. (2010) *Designing Sound*. MIT Press.

Fazekas, G. and Sandler, M. (2007) Intelligent editing of studio recordings with the help of automatic music structure extraction. *Audio Eng. Soc. Convention 122* AES.

Franinovic, K. and Serafin, S. (eds) (2013) *Sonic Interaction Design*. MIT Press.

Franinovic, K., Hug, D., and Visell, Y. (2007) Sound embodied: Explorations of sonic interaction design for everyday objects in a workshop setting. *Proc. of the Intl. Conf. on Auditory Display* ICAD.

Gold, B., Morgan, N., and Ellis, D. (2011) *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. John Wiley & Sons.

Großhauser, T. and Hermann, T. (2010) Multimodal closed-loop human machine interaction. *Human Interaction with Auditory Displays – Proc. Interactive Sonification Workshop*, pp. 59–63.

Harte, C., Sandler, M., and Gasser, M. (2006) Detecting harmonic change in musical audio. *Proc. 1st ACM workshop on Audio and music computing multimedia*, pp. 21–26 ACM.

Hatziantoniou, P.D. and Mourjopoulos, J.N. (2004) Errors in real-time room acoustics dereverberation. *J. Audio Eng. Soc.*, **52**(9), 883–899.

Hermann, T. and Ritter, H. (2004) Sound and meaning in auditory data display. *Proc. IEEE*, **92**(4), 730–741.

Hermann, T., Hunt, A., and Neuhoff, J.G. (2011) *The Sonification Handbook*. Logos Verlag.

Hug, D. (2008) Genie in a bottle: Object-sound reconfigurations for interactive commodities. *Proc. of Audio Mostly Conf.*, Pitea, Sweden, pp. 56–63.

Hyvärinen, A., Karhunen, J., and Oja, E. (2004) *Independent Component Analysis*. John Wiley & Sons.

Kramer, G. (1994) *Auditory Display: Sonification, Audification, and Auditory Interfaces*. Addison-Wesley.

Laitinen, M-V., Pihlajamäki, T., Erkut, C., and Pulkki, V. (2012) Parametric time-frequency representation of spatial sound in virtual worlds. *ACM Trans. Appl. Percep. (TAP)*, **9**(2), 8–20.

Lee, K. and Slaney, M. (2008) Acoustic chord transcription and key extraction from audio using key-dependent HMMs trained on synthesized audio. *IEEE Trans. Audio, Speech, and Language Proc.*, **16**(2), 291–301.

Lerch, A. (2012) *An Introduction to Audio Content Analysis: Applications in signal processing and music informatics*. John Wiley & Sons.

Maher, R. (2009) Audio forensic examination. *Signal Proc. Mag., IEEE*, **26**(2), 84–94.

McGookin, D.K. and Brewster, S.A. (2004) Understanding concurrent earcons: Applying auditory scene analysis principles to concurrent earcon recognition. *ACM Trans. Appl. Percep. (TAP)*, **1**(2), 130–155.

Muller, M., Ellis, D.P., Klapuri, A., and Richard, G. (2011) Signal processing for music analysis. *IEEE J. Select. Topics in Signal Proc.*, **5**(6), 1088–1110.

Poliner, G.E., Ellis, D.P., Ehmann, A.F., Gómez, E., Streich, S., and Ong, B. (2007) Melody transcription from music audio: Approaches and evaluation. *IEEE Trans. Audio, Speech, and Language Proc.*, **15**(4), 1247–1256.

Ryynänen, M. and Klapuri, A. (2006) Transcription of the singing melody in polyphonic music. *Int. Conf. Music Inform. Retrieval*, pp. 222–227 ISMIR.

Savioja, L., Huopaniemi, J., Lokki, T., and Väänänen, R. (1999) Creating interactive virtual acoustic environments. *J. Audio Eng. Soc.*, **47**(9), 675–705.

Schuller, B.W. (2013) *Intelligent Audio Analysis*. Springer.

Sherman, W. and Craig, A. (2003) Understanding virtual reality: interface, application, and design. *The Morgan Kaufmann series in computer graphics and geometric modeling*.

Svensson, P. and Kristiansen, U.R. (2002) Computational modelling and simulation of acoustic spaces. *22nd Int. Audio Eng. Soc. Conf.: Virtual, Synthetic, and Entertainment Audio* AES.

Szwoch, G., Czyzewski, A., and Ciarkowski, A. (2009) A double-talk detector using audio watermarking. *J. Audio Eng. Soc.*, **57**(11), 916–926.

Tsingos, N., Gallo, E., and Drettakis, G. (2004) Perceptual audio rendering of complex virtual environments. *ACM Trans. Graphics*, **23**(3), 249–258.

Vorländer, M. (2007) *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*. Springer.

Wang, D. and Brown, G.J. (2006) *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press.