Toolbox for acoustic scene creation and rendering (TASCAR): Render methods and research applications

Giso GRIMM[a,b*]and **Joanna LUBERADZKA**[a] and
Tobias HERZKE[b] and **Volker HOHMANN**[a,b]
[a] Medizinische Physik and Cluster of Excellence Hearing4all
Universität Oldenburg, D-26111 Oldenburg, Germany
[b] HörTech gGmbH
Marie-Curie-Str. 2, D-26129 Oldenburg, Germany

Abstract

TASCAR is a toolbox for creation and rendering of dynamic acoustic scenes that allows direct user interaction and was developed for application in hearing aid research. This paper describes the simulation methods and shows two research applications in combination with motion tracking as an example. The first study investigated to what extent individual head movement strategies can be found in different listening tasks. The second study investigated the effect of presentation of dynamic acoustic cues on the postural stability of the listeners.

Keywords

Spatial audio, hearing research, motion tracking

1 Introduction

Hearing aids are evolving from simple amplifiers to complex processing devices. Algorithms in hearing devices, e.g., directional microphones, direction of arrival estimators, or binaural noise reduction, depend on the spatial properties of the surrounding acoustic environment [Hamacher et al., 2005]. Several studies show a large performance gap between laboratory measurements and real life experience, attributed to a changed user behavior [Smeds et al., 2006] as well as oversimplification of the test environment [Cord et al., 2004; Bentler, 2005]. To bridge this gap, a reproduction of complex listening environments in the laboratory is desired. To allow for a systematic evaluation of hearing device performance, these virtual acoustic environments need to be scalable and reproducible. There are several requirements for a virtual acoustic environment to make it suitable for hearing research. For human listening a high plausibility of the environments and a reproduction of the relevant perceptual cues is required. For machine listening and processing in multi-microphone hearing devices, a correct reproduction of relevant physical properties is needed. For an ecologically valid evaluation of hearing devices, the virtual acoustic environments need to reflect relevant every-day scenarios. Additionally, to assess limitations of hearing devices, realistic but challenging environments are required. In both cases, the reproduction need to allow for listener movements in the environment and may contain moving sources.

Existing virtual acoustic environment engines often target authentic simulations for room acoustics (e.g., EASE, ODEON), resulting in a large complexity. They typically render impulse responses for off-line analysis or auralization. Other tools, e.g., the SoundScapeRenderer [Ahrens et al., 2008], do not provide all features required here, such as room simulation and diffuse source handling. Therefore, a toolbox for acoustic scene creation and rendering (TASCAR) was developed as a Linux audio application. The aim of TASCAR is to interactively reproduce time varying complex listening environments via loudspeakers or headphones. For a seamless integration into existing measurement tools of psycho-acoustics and audiology, low-delay real-time processing of external audio streams in the time domain is applied, and interactive modification of the geometry is possible. TASCAR consists of a standalone application for the acoustic simulation, and a set of command line programs and Octave/Matlab scripts for recording from and playing to jack ports, and measuring impulse responses.

The simulation methods and implementation are described in section 2. Two research applications of TASCAR in combination with motion tracking are shown as an example. The first study (section 3.1) investigates to what extent individual head movement strategies can be found in different listening tasks. Results indicate that individual strategies exist in natural listening tasks, but task specific behavior can be found in tasks which include localization. The second study (section 3.2) investigates the effect

[*] g.grimm@uni-oldenburg.de

of presentation of dynamic acoustic cues on the postural stability of the listeners. Test subjects performed a stepping test while imposed with stationary or spatially dynamic sounds. Results show that in the absence of visual cues the spatial dynamics of acoustic stimuli have a significant effect on postural stability.

2 TASCAR: Methods and implementation

The implementation of TASCAR utilizes the jack audio connection kit [Davis and Hohn, 2003]. Audio content is exchanged between different components of TASCAR via jack ports. The jack time line is used as a base of all time-varying features. Audio signals are processed block-wise in the time domain. A rough signal and data flow chart of TASCAR is shown in Figure 1.

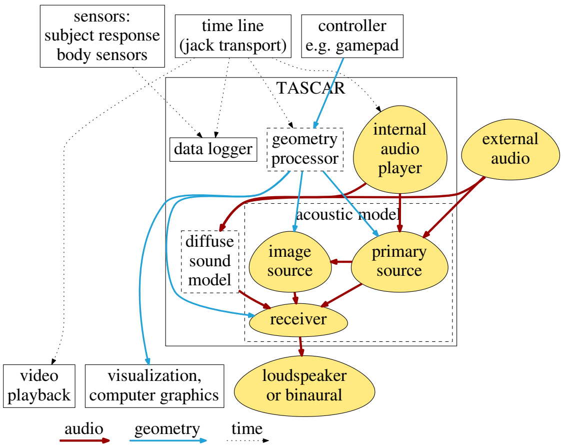

Figure 1: Schematic audio and control signal flow chart of TASCAR in a typical hearing research subjective test application.

The structure of TASCAR can be divided into three major components: Audio content is delivered by an audio player module. It provides a non-blocking method of accessing sound file portions. Audio content can also be delivered by external sources, e.g., from physical sources, audio workstations, or any other jack client. The second major block is the geometry processing of the virtual acoustic environment. The last block is the acoustic model, i.e., the combination of audio content and geometry information into an acoustic environment in a given render format.

2.1 Geometry processing

An acoustic scene in TASCAR consists of objects of several types: Sound sources, receivers, reflectors, and dedicated sound portals for coupled room simulations [Grimm et al., 2014]. All object types have trajectories defined by location in Cartesian coordinates and orientation on ZYX-Euler-coordinates. These trajectories are linearly interpolated between sparse time samples; the location is interpolated either in Cartesian coordinates, or in spherical coordinates relative to the origin. The orientation is interpolated in Euler coordinates. The geometry is updated once in each processing cycle.

Sound source objects can consist of multiple “sound vertices”, either as vertices of a rigid body, i.e., following the orientation of the object, or as a chain, i.e., at a given distance on the trajectory. Each “sound vertex” is a primary source.

$\mathbf{p}_{src}$ is the primary source position, $\mathbf{p}_{rec}$ is the receiver position, and O_{rec} is the rotation matrix of the receiver. Then $\mathbf{p}_{rel} = \mathrm{O}_{rec}^{-1}\left(\mathbf{p}_{src} - \mathbf{p}_{rec}\right)$ is the position of the sound source relative to the receiver, and $r = ||\mathbf{p}_{rel}||$ is the distance between source and receiver.

Reflectors can consist of polygon meshes with one or more faces. For each mesh, reflection properties can be defined. For a first order image source model, each pair of primary source and reflector face creates an image source. For higher order image source models, also the image sources of lower orders are taken into account. A schematic sketch of the image model geometry is shown in Figure 2. The image source position $\mathbf{p}_{img}$ is determined by the closest point on the (infinite) reflector plane $\mathbf{p}_{cut}$ to the source $\mathbf{p}_{src}$: $\mathbf{p}_{img} = 2\mathbf{p}_{cut} - \mathbf{p}_{src}$.

The image source position is independent of the receiver position. However, the visibility of an image source depends on the receiver position and the reflector dimension. If the intersection point of the connection from the image source to the receiver with the reflector plane $\mathbf{p}_{is}$ is within the reflector boundaries, the image source is visible, and a specular reflection is applied. If $\mathbf{p}_{is}$ is not within the reflector boundaries, the effective image source position is shifted into the direction of the closest point on the boundary to $\mathbf{p}_{is}$, and an “edge reflection” is applied. The differences between these two reflection types in terms of audio processing are described in section 2.2.2.

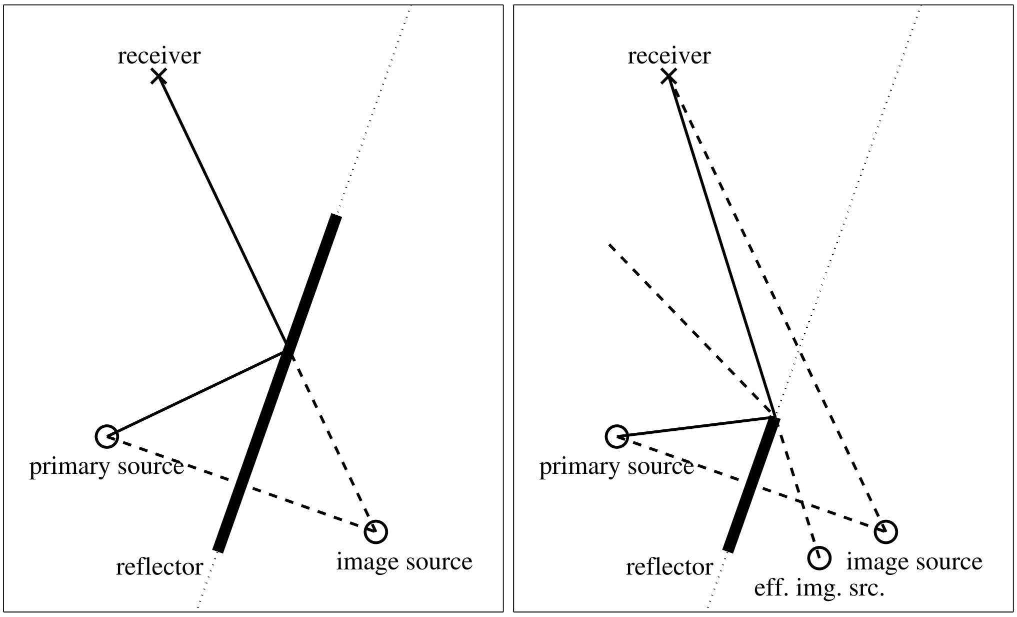

Figure 2: Schematic sketch of the image model geometry. Left panel: “specular” reflection, i.e., the image source is visible within the reflector; right panel: edge reflection.

2.2 Acoustic model

For each pair of receiver and sound source – primary or image source – an acoustic model is calculated. The acoustic model can be split into the transmission model, which depends only on the distance between source and receiver, an image source model, which depends on the reflection properties of the reflecting surfaces as well as on the “visibility” of the reflected image source, and a receiver model, which encodes the direction of the sound source relative to the receiver into the render output format.

2.2.1 Transmission model

The transmission model consists of air absorption, and a time-varying delay line for a simulation of Doppler-shift and time-varying comb-filter effects.

Point sources follow a $1/r$ sound pressure law, i.e., doubling the distance r results in half of the sound pressure. Air absorption is approximated by a simple first order low-pass filter model with the filter coefficients controlled by the distance:

$$\begin{aligned} y_k &= a_1 y_{k-1} + (1-a_1)x_k \qquad (1) \\ a_1 &= e^{-\frac{r f_s}{c\alpha}}, \qquad (2) \end{aligned}$$

where c is the speed of sound, x_k is the source signal at the sample k, and y_k is the filtered signal. The empiric constant $\alpha = 7782$ was manually adjusted to provide sensible values for distances below 50 meters. This approach is very similar to that of [Huopaniemi et al., 1997] who used a FIR filter to model the frequency response at certain distances. However, in this approach the distance parameter r can be varied dynamically.

The time varying delay line uses nearest neighbor interpolation[1].

2.2.2 Image source model

Early reflections are modeled using an image source model. In opposite to most commonly used models (e.g., [Allen and Berkley, 1979]) which calculate impulse responses for a rectangular enclosure (“shoebox model”), reflections are simulated for each reflecting polygon-shaped surface.

With finite reflectors, it is distinguished between a “specular” reflection, when the image source is visible from the receiver position within the reflector, and an “edge” reflection, when the image source would not be “visible”. In both cases, the source signal is filtered with a first order low pass filter[2] determined by a reflectivity coefficient ρ, and a damping coefficient δ:

$$y_k = \delta y_{k-1} + \rho x_k \qquad (3)$$

For “edge” reflections, the effective image source is shifted that it appears from the di-

[1] Other interpolation methods are planned.

[2] In later versions of TASCAR the reflection filter will be controlled by frequency-dependent absorption coefficients to avoid the sample rate dependency.

rection of a point on the reflector edge which is closest to $\mathbf{p}_{is}$. If receiver or sound source are behind the reflector, the image source is not rendered.

2.2.3 Receiver model

A receiver encodes the output of the transmission model of each sound source into the output format, based on the relative position between sound source and receiver, $\mathbf{p}_{k,rel}$. Each receiver owns one jack output port for each output channel n; the number of channels depends on the receiver type and configuration. The receiver output signal $z_k(n)$ for the output channel n and sound source k is

$$z_k(n) = w(\mathbf{p}_{k,rel}, n) y_k \tag{4}$$

with the transmission model output signal y_k. $w(\mathbf{p}_{rel}, n)$ are the driving weights for each output channel. The mixed output signal of the whole virtual acoustic environment is the sum of $z_k(n)$ across all sources k, plus the diffuse sound signals decoded for the respective receiver type (see section 2.3 for more details).

Several receiver types are implemented: Virtual omni-directional microphones simply return the output without directional processing, $w = 1$. Simple virtual cardioid microphones apply a gain g depending on the angle θ between source and receiver:

$$w = \frac{1}{2} (\cos(\theta) + 1) \tag{5}$$

For reproduction via multichannel loudspeaker arrays, receiver types with one output channel for each loudspeaker can be used. A "nearest speaker" receiver is a set of virtual loudspeakers at given positions in space (typically matched with the physical loudspeaker setup). The driving weights for each virtual loudspeaker are 1 for the least angular distance between the virtual loudspeaker and $\mathbf{p}_{rel}$, and 0 for all other channels. Other receiver types are horizontal and full-periphonic 3rd order Ambisonics [Daniel, 2001], VBAP [Pulkki, 1997], and "basic" as well as "in-phase" ambisonic panning [Neukom, 2007].

Since the geometry is updated only once in each processing block, all receiver types interpolate their driving weights so that the processed geometry is matched at the end of each block. For some receiver types, e.g., 3rd order Ambisonics, this may lead to a spatial blurring of the sources if the angular movement within one processing block is large compared to the spatial resolution of the receiver type.

2.3 Diffuse sources and reverberation

Diffuse sources, e.g., background signals, or diffuse reverberation [Wendt et al., 2014], are added in first order ambisonics (FOA) format. No distance law is applied to diffuse sound sources; instead, they have a rectangular spatial range box, i.e., they are only rendered if the receiver is within their range box, with a von-Hann ramp at the boundaries of the range box. Position and orientation of the range box can vary with time. The diffuse source signal is rotated by the difference between receiver orientation and box orientation. Each receiver type provides also a method to render FOA signals to the receiver-specific output format.

2.4 Further components

Besides the open source core of TASCAR in form of a command line application[3], a set of extension modules is commercially developed by HörTech gGmbH. These components include a graphical user interface, a time aligned data logging system for open sound control (OSC) messages, interfaces for motion trackers and electro-oculography, and specialized content controllers.

3 Example research applications

In this section, two studies related to hearing aid research which are based on TASCAR are briefly described, to illustrate possible applications.

3.1 Individualized head motion strategies

The hypothesis of this study was that task-specific head movement strategies can be measured on an individual basis. Head movements in a natural listening environment were assessed. A panel discussion with four talkers in a simulated room with early reflections was played back via an eight-channel loudspeaker array, using 3rd order Ambisonics. Head movements were recorded with the time aligned data logger using a wireless inertial measurement unit and a converter to OSC messages.

Figure 3 shows five individual head orientation trajectories. Systematic differences can be observed: Whereas one subject (green line) performs a searching motion, i.e., modulation

[3] https://github.com/gisogrimm/tascar

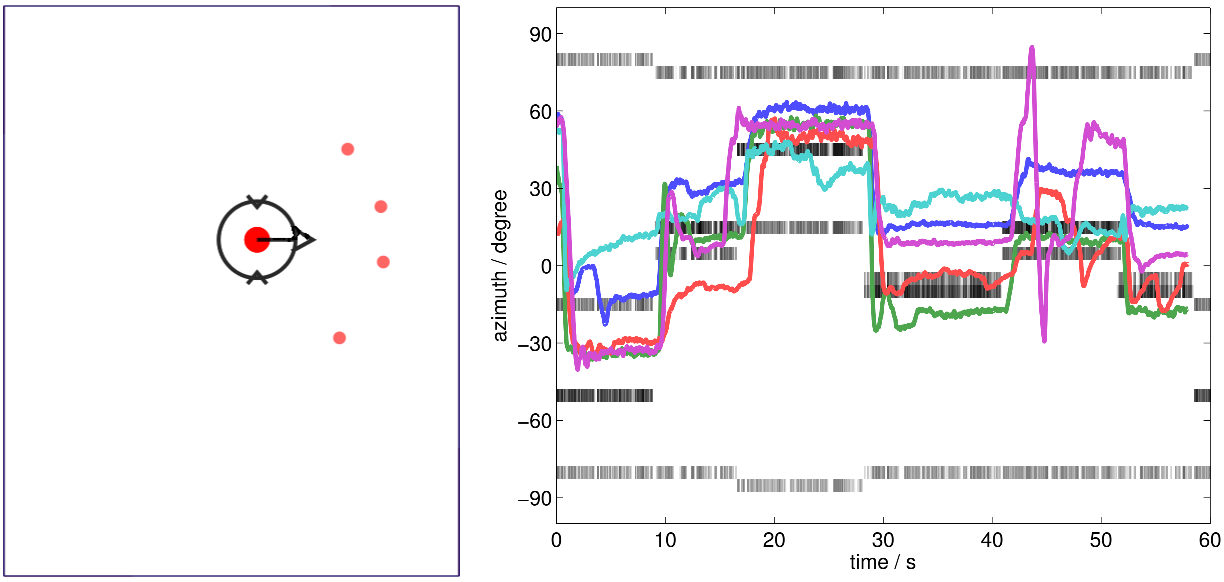

Figure 3: Intensity of a panel discussion in a room as a function of time and azimuth (shades of gray) with five individual head orientations.

around the final position, at each change of talkers, other subjects adapt slower to source position changes. One subject (blue line) shows a constant offset, possibly indicating a better-ear listening strategy.

3.2 Postural stability

Some hearing aid users feel disturbed by fast-acting automatics of hearing aids and the potentially resulting quickly changing binaural cues. To prepare the ground for further investigations of this effect, the second study assessed the effect of spatially dynamic acoustic cues on the postural stability [Büsing et al., 2015]. It is based on an experiment in which it was shown that the presence of a stationary sound can improve the postural stability in the absence of visual cues [Zhong and Yost, 2013]. A Fukuda stepping test was performed, in which the subjects were asked to step 100 steps in a fixed position. The subject drift was taken as a measure of postural stability.

In this study with 10 young participants with normal vision and hearing, the factors *vision* (open or closed eyes), *stimulus* (static or moving) and *spatial complexity* (two sources or many sources) on postural stability were analyzed. The stimuli were rendered with TASCAR; the factors *stimulus* and *spatial complexity* were realized by alternative virtual environments. The environment with low complexity was a kitchen scene with a frying pan and a clock, either rendered statically or with a sinusoidal rotate around the listener. The complex environment was a virtual amusement park, either from a carousel perspective or from a static position. The subjects were tracked with the microsoft kinect skeleton tracking library. The positions of the modeled nodes were send from the windows PC via OSC to the TASCAR data logger. The body rotation was measured as the rotation of the shoulder skeleton nodes. The results are shown in Figure 4. Vision has the largest effect on the body rotation; with open eyes the average body rotation during the test is small, independent from the stimulus and complexity condition. However, without visual cues, the spatially dynamic complex scene leads to a significantly higher body rotation than the corresponding complex static scene.

4 Conclusions

To bridge the gap between laboratory results and real-life experience in the domain of hearing research and hearing device evaluation, a tool for acoustic scene creation and rendering (TASCAR) was developed. The tool focuses on a reproduction of perceptual cues and physical properties of the sound field which are relevant for typical applications in hearing device research. Simplifications allow for computational efficiency. The implementation utilizes the jack audio connection kit, resulting in a large flexibility.

To compute the sound at a given position of the receiver, the signal coming from each source

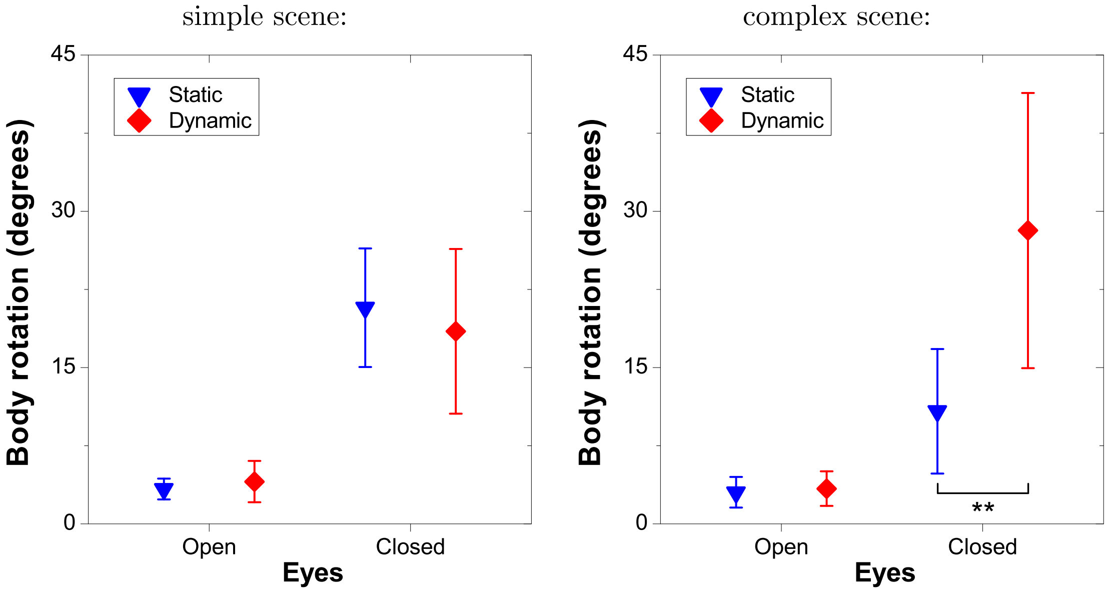

Figure 4: Body rotation in a Fukuda stepping test in a simple scene (left panel) and a complex scene (right panel). In the absence of visual cues, the dynamic cues (red diamonds) have a significant effect on the body rotation in the complex scene.

– primary or image source – is computed based on the transmission model, i.e., depending on the distance. The receiver output signal is computed depending on the type of the receiver and the angle between source and receiver. The receiver signals from all sources are added up and combined with diffuse sounds, resulting in the sound of a virtual acoustic environment in a given point.

Two studies based on the spatial audio reproduction of TASCAR demonstrate its applicability as a research tool for reproduction of spatially dynamic acoustic environments.

5 Acknowledgments

This study was funded by the research group DFG for1732. We thank Isabel Büsing and Tobias Neher from the group “Individualized Hearing Devices” for providing the data on postural stability and the good collaboration.

References

J. Ahrens, M. Geier, and S. Spors. 2008. The soundscape renderer: A unified spatial audio reproduction framework for arbitrary rendering methods. In *Audio Engineering Society Convention 124*. Audio Engineering Society.

J. B. Allen and D. A. Berkley. 1979. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65:943.

R. A. Bentler. 2005. Effectiveness of directional microphones and noise reduction schemes in hearing aids: A systematic review of the evidence. *Journal of the American Academy of Audiology*, 16(7):473–484.

I. Büsing, G. Grimm, and T. Neher. 2015. Einfluss von räumlich-dynamischen Schallinformationen auf das Gleichgewichtsvermögen (influence of spatially dynamic acoustic cues on postural stability). In *18. Jahrestagung der Deutschen Gesellschaft für Audiologie*, Bochum, Germany.

M. T. Cord, R. K. Surr, B. E. Walden, and O. Dyrlund. 2004. Relationship between laboratory measures of directional advantage and everyday success with directional microphone hearing aids. *Journal of the American Academy of Audiology*, 15(5):353–364.

J. Daniel. 2001. *Représentation de champs acoustiques, application à la transmission et à la reproduction de scnes sonores complexes dans un contexte multimdia*. Ph.D. thesis, Université Pierre et Marie Curie (Paris VI), Paris.

P. Davis and T. Hohn. 2003. Jack audio connection kit. In *Proc. of the Linux Audio Developer Conference. ZKM Karlsruhe.*

G. Grimm, T. Wendt, V. Hohmann, and

S. Ewert. 2014. Implementation and perceptual evaluation of a simulation method for coupled rooms in higher order ambisonics. In *Proceedings of the EAA Joint Symposium on Auralization and Ambisonics*, Berlin.

V. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, and U. Rass. 2005. Signal processing in high-end hearing aids: state of the art, challenges, and future trends. *EURASIP Journal on Applied Signal Processing*, 2005:2915–2929.

J. Huopaniemi, L. Savioja, and M. Karjalainen. 1997. Modeling of reflections and air absorption in acoustical spaces a digital filter design approach. In *Applications of Signal Processing to Audio and Acoustics, 1997. 1997 IEEE ASSP Workshop on*, pages 4–pp. IEEE.

M. Neukom. 2007. Ambisonic panning. In *Audio Engineering Society Convention 123*, 10.

V. Pulkki. 1997. Virtual sound source positioning using vector base amplitude panning. *J. Audio Eng. Soc*, 45(6):456–466.

K. Smeds, G. Keidser, J. Zakis, H. Dillon, A. Leijon, F. Grant, E. Convery, and C. Brew. 2006. Preferred overall loudness. ii: Listening through hearing aids in field and laboratory tests. *International Journal of Audiology*, 45(1):12–25.

T. Wendt, S. van der Par, and S. Ewert. 2014. Perceptual and room acoustical evaluation of a computational efficient binaural room impulse response simulation method. In *Proceedings of the EAA Joint Symposium on Auralization and Ambisonics*, Berlin.

X. Zhong and W. A. Yost. 2013. Relationship between postural stability and spatial hearing. *Journal of the American Academy of Audiology*, 24(9):782–788.