A True Spatial Sound System for CAVE-Like Displays Using Four Loudspeakers

Torsten Kuhlen[1], Ingo Assenmacher[1], and Tobias Lentz[2]

[1] Virtual Reality Group, RWTH Aachen University,
Seffenter Weg 23, 52074 Aachen, Germany
[2] Institute of Technical Acoustics, RWTH Aachen University,
Neustraße 50, 52066 Aachen, Germany
{kuhlen,assenmacher}@rz.rwth-aachen.de,
tobias.lentz@akustik.rwth-aachen.de

Abstract. The paper introduces an audio rendering system based on the binaural approach, which allows a real-time simulation of spatially distributed sound sources and, in addition to that, near-to-head sources in room-mounted virtual environments. We have been developing a dynamic crosstalk cancellation, allowing the listener to freely move around without wearing any headphones. The paper gives a comprehensive description of the system, concentrating on the dual dynamic crosstalk cancellation and aspects of the integration of a real-time room acoustical simulation. Finally, two applications are described to show the wide applicability of the system.

Keywords: Virtual Reality, 3D Audio, Spatial Acoustics, Binaural Synthesis.

1 Introduction

In contrast to the human visual system, the human auditory system can perceive input from all directions and has no limited field of view. As such, it provides valuable cues for navigation and orientation in virtual environments. With respect to immersion, the acoustic sense is a very important additional source of information for the user, as acoustical perception works precisely especially for close auditory stimuli. This enhances the liveliness and credibility of the virtual environment. The ultimate goal thus would be the ability to place virtual sounds in any three dimensions and distance around the user in real-time. However, although multimodality is a crucial feature in Virtual Reality, the acoustical component is often neglected. While with head-mounted displays it might be quite acceptable to wear headphones, in CAVE-like environments loudspeakers are favored. Therefore, we have been developing a high-quality sound system based on the binaural approach. This approach is a powerful method for generating spatially distributed sounds, as a binaural signal represents the sound pressure at the eardrum of the listener. In contrast to other loudspeaker-based systems, the main benefit is the ability to reproduce near-to-head sources realistically. The binaural approach relies on crosstalk compensation when loudspeakers are used instead of headphones, which in existing systems only works when the user's head is

R. Shumaker (Ed.): Virtual Reality, HCII 2007, LNCS 4563, pp. 270–279, 2007.

positioned in a “sweet spot”. Since in a virtual environment a user is typically walking around however, we have developed a dynamic crosstalk suppression which even works properly for a moving user. In particular, we are using a setup of four loudspeakers, where only two of them are active at a time in a way that a full 360 degrees rotation for the listener is provided. In addition to that, we coupled the acoustical rendering system with a visual VR system in order to realize a multi-modal dynamic virtual environment where the user and all sound sources can freely move without constraints, including near-to-head sources. With our approach, a versatile and stable real-time binaural sound system based on loudspeakers with dynamic crosstalk cancellation is available, generating congruent visual and acoustical scenes. One of the major contributions of this comprehensive system is the realization as a software-only solution that makes it possible to use this technology on a standard PC basis.

The remainder of this paper is structured as follows. First, we will give a brief overview of different reproduction approaches. We will, however, concentrate on the dynamic binaural synthesis approach with crosstalk cancellation and present this in more detail. After that, we briefly discuss aspects of the integration of a real-time room-acoustical simulation in our system. Room acoustics is an important aspect of a virtual environment, as sound is never correct without reference to the room it is present in. Two showcase applications will be presented in section 4 to show the width of applications that are possible to realize with a system like this. After that, a short discussion will end the paper.

2 A Brief Taxonomy of Audio Rendering in VR Systems

2.1 Panning

A popular technique for spatial acoustic imaging can be found in the intensity panning approach. Here, sound source distance and position is modeled by modifying the amplitude on the output channels of a predefined static loudspeaker setup. Such a multi channel audio is often used in home cinema systems to surround the listener with sound and also work quite well for VR applications which do not require a very exact placing of virtual sound sources. However, intensity panning is not able to provide authentic virtual sound scenes. In particular, it is impossible to crate virtual near-to-head sound sources, although these sources are of high importance in direct interaction metaphors for virtual environments, such as by-hand-placements or manipulations. Therefore, we will not discuss it in more detail.

There are mainly two different techniques reproducing a sound event with true spatial relation, i.e. wave field synthesis and the binaural approach. The following sections will briefly introduce principles and problems of these technologies.

2.2 Wave Field Synthesis

The basic theory of the wave field synthesis (WFS) is the *Huygens' principle*. An array of loudspeakers (ranging from just a few to some hundreds in number) used to reproduce a complete wave field. The simulation reproduces the sound field by virtually placing an array of microphones in the field that would be used to record a

sound event at these points [3]. Placing loudspeakers at these points can reproduce an entire wave field for this area. Hence, this approach is especially adequate for multi-user virtual environments, where even non-tracked moving users get the real spatial sound impression.

The main drawback, beyond the high effort of processing power, is the size of the loudspeaker array. Furthermore, mostly solutions have been presented so far, where the sound field is only reproduced in one horizontal plane by a line of loudspeakers instead of a two-dimensional array. The placement of the loudspeakers in immersive projection-based VR displays with four to six surfaces as in CAVE-like environments is nearly impossible without any severe compromises. However, for semi-immersive displays like an L-shaped bench or a PowerWall, wave field synthesis is an excellent option when the focus is on multi-user virtual environments [12].

2.3 Binaural Synthesis

In contrast to wave field synthesis, a binaural approach does not deal with the complete reproduction of the sound field. It is convenient and sufficient to reproduce the sound field only at two points, the ears of the listener. In this case only two signals have to be calculated for a complete three-dimensional sound scene. The procedure of convolving a mono sound source with an appropriate pair of Head-Related Transfer Functions (HRTFs) in order to obtain a synthetic binaural signal is called *binaural synthesis.* The synthesized signals contain the directional information of the source, which is provided by the information in the HRTFs.

A *static binaural synthesis* transforms a sound source without any position information to a virtual source being related to the listener′s head. This already works for a non-moving head, as the applied transfer function is related to the head and not to the room. For a moving head, this implies that the virtual source moves with the listener. For the realization of a room-related virtual source, a *dynamic binaural synthesis* has to be applied, where the HRTF must be changed when the listener turns or moves his head. In a VR system, the listener′s position is always known and can also be used to realize a synthetic sound source with a fixed position corresponding to the room coordinate system. The system calculates the relative position and orientation of the listener′s head to the imaginary point where the source should be localized. By knowing the relative position and orientation, the appropriate HRTF can be chosen from a database. It is also possible to synthesize many different sources and to create a complex three-dimensional acoustic scenario.

Headphones versus Loudspeakers. From a technical point of view, the presentation of binaural signals by headphones is the easiest way since the acoustical separation between both channels is perfectly solved. However, unsatisfying results are often obtained in the subjective sense of listeners. Furthermore, when the ears are covered by headphones, the impression of a source located at a certain point and distance to the listener often does not match the impression of a real sound field.

Another point is that while wearing a head-mounted display (HMD) in combination with headphones may fit quite well, in projection-based VR displays the focus is on non-intrusive components, i.e., the user should be as free as possible of

any wearable hardware. This leads to the need for loudspeaker based reproduction techniques not only for the WFS approach, but also for the binaural synthesis.

The problem with loudspeaker reproduction of binaural signals is the crosstalk between the channels that destroys the three-dimensional cues. The requirement for a correct binaural presentation is that the right channel of the signal is audible only in the right ear and the left one is audible only in the left ear. This problem can be solved by a crosstalk cancellation (CTC) filter which is based on the transfer functions from each loudspeaker to each ear. For a static CTC system the four transfer functions from the speakers to the ears are used to calculate the filters for the elimination of crosstalk at one specific point. A detailed description of static CTC systems can be found in [2]. A static CTC system may only be quite acceptable for desktop VR systems with a user sitting in front of a monitor. In all other cases, it is necessary to adapt the CTC filters in real-time dependent on the current position and orientation of the listener [5]. While static CTC systems are state of the art, a newly developed, purely software-based approach of such a dynamic CTC – allowing a user to freely move within immersive virtual environments without wearing headphones – will be described in the next section.

Table 1 summarizes the benefits and drawbacks of the particular audio rendering approaches for different VR displays. For a high-quality integration of spatial audio into a virtual environment, only WFS and dynamic binaural synthesis with loudspeakers come into question. While WFS is rather costly and can only be installed in

Table 1. Assessment of available audio rendering approaches for different VR display technologies (LS: Loudspeaker, HP: Headphones, BS: Binaural Synthesis, WFS: Wave Field Synthesis). The benefits and drawbacks of the particular approaches are described in section 2. The last row depicts our approach introduced in this paper.

	Desktop VR	HMD	Non-immersive Projection	Immersive Projection	Remarks
Panning (LS)	😐	😐	😐	😐	No near-to-head sources
WFS (LS)	☺	☺	☺	☹	High quality, high costs, multi-user
Static BS, HP	😐	☹	☹	☹	Sound moves with user
Dyn. BS, HP	😐	☺	😐	😐	Bad naturalness of presented sounds
Static BS, LS	😐	☹	☹	☹	Sweet Spot: user may not move
Dyn. BS, LS	☺	☺	☺	☺	**Flexible, low costs, user-centered**

non- or semi-immersive environments without compromises, binaural synthesis promises to cope with few loudspeakers and flexible setup. Additionally, in particular near-to-head sound sources can be very precisely reproduced by the binaural approach. Especially for direct interaction such as by-hand placements or manipulations), which are preferred interaction metaphors in virtual environments, most objects reside within the grasping range and thus rather near to the user's head. Although no systematic studies have been carried out so far, there are strong hints that for such near-to-head sources, the binaural approach produces better results than WFS. A major drawback of the binaural synthesis is that it cannot be used in multi-user environments, however.

Dynamic 360 degrees Crosstalk Cancellation. To reproduce the binaural signal at the ears with a sufficient channel separation without using headphones, a CTC system is needed [9, 6]. Getting the CTC work in an environment where the user should be able to walk around and turn his head, a dynamic CTC system which is able to adapt during the listener's movements [5, 7] is required. The dynamic solution overrides the sweet spot limitation of a normal static crosstalk cancellation.

Figure 1 (left) shows the four transfer paths from the loudspeakers to the ears of the listener (H_{1L}) illustrates the transfer function from loudspeaker 1 to the left ear). A correct binaural reproduction means that the complete transfer function from the left input to the left ear, including the transfer function H_{1L}, is meant to become a flat spectrum. The same is intended for the right transfer path, accordingly. The crosstalk indicated by H_{1R} and H_{2L} has to be canceled by the system.

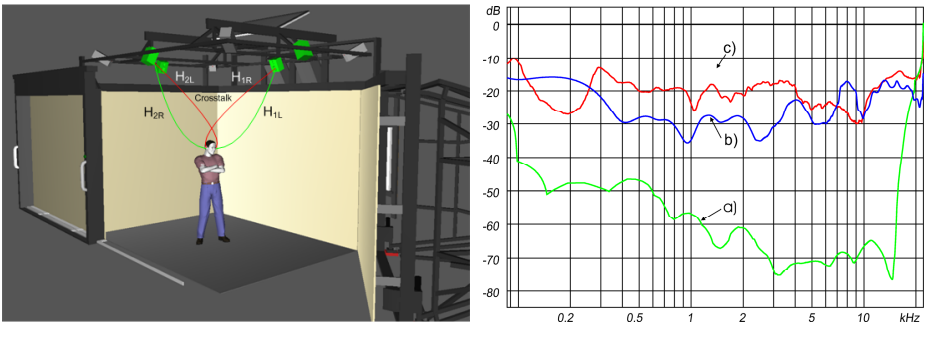

Fig. 1. *Left*: The CAVE-like environment at RWTH Aachen University. Four loudspeakers are mounted on the top rack of the system. The door, shown on the left, and a moveable wall, shown on the right, can be closed to allow a 360 degree view with no roof projection. *Right*: Measurement of the accessible channel separation using a filter length of 1,024 taps. a) calculated, b) static solution, c) dynamic system.

Since the user of a virtual environment is already tracked to generate the correct stereoscopic video images, it is possible to calculate the CTC filter on-line for the current position and orientation of the user. The calculation at runtime enhances the flexibility of the acoustic system regarding the validity area and the flexibility of

the loudspeaker setup which can hardly be achieved with preprocessed filters. As a consequence, a database containing 'all' possible HRTFs is required. For our system, we use a database with a spatial resolution of one degree for both azimuth and elevation. The HRTFs were measured at a frequency range of 100Hz – 20kHz, allowing a cancellation in the same frequency range.

To provide a full head rotation of the user, a two loudspeaker setup will not be sufficient as the dynamic cancellation will only work in between the angle spanned by the loudspeakers. Thus, a dual CTC algorithm with a four speaker setup has been developed, which is further described in [8]. With four loudspeakers eight combinations of a normal two-channel CTC system are possible and a proper cancellation can be achieved for every orientation of the listener. An angle dependent fading is used to change the active speakers in between the overlapping validity areas of two configurations. Each time the head-tracker information is updated in the system, the deviation of the head to the position and orientation compared to the information given that caused the preceding filter change is calculated.

To classify the performance that could be reached theoretically by the dynamic system, measurements of a static system were made to have a realistic reference for the achieved channel separation. Under absolute ideal circumstances, the HRTFs used to calculate the crosstalk cancellation filters are the same as during reproduction (individual HRTFs of the listener). In a first test, the crosstalk cancellation filters were processed with HRTFs of an artificial head in a fixed position. The calculated channel separation using this filter set is plotted in Figure 1 (right) a). Thereafter, the achieved channel separation was measured at the ears of the artificial head, which had not been moved since the HRTF measurement (Figure 1 (right) curve b)). In comparison to the ideal reference cases, Figure 1 (right) curve c) shows the achieved channel separation of the dynamic CTC system. The main difference between the static and the dynamic system is the set of HRTFs used for filter calculation. The dynamic system has to choose the appropriate HRTF from a database and has to adjust the delay and the level depending on the position data. All these adjustments cause minor deviations from the ideal HRTF measured directly at this point. For this reason, the channel separation of the dynamic system is not as high as one that can be achieved by a system with direct HRTF measurement.

3 Dynamic Real-Time Room-Acoustics

For VR applications, it is not only important to have a flexible and precise reproduction system, but also the synchronization between the presented aural cues and the visual scene (or other modalities) is very important. The synchronization aspect does not only cover the correct timing of the presentation, but also the validity of all the cues with respect to the presented scenery. In that sense, for a correct auditory rendering, the virtual room has to be taken into account.

For that purpose, we coupled our reproduction system, which is controlled by a VR application, with the room-acoustics simulation software RAVEN [10], which enables us to simulate the dynamic room impulse responses for moving sources in real-time. RAVEN is basically an upgrade and enhancement of the hybrid room acoustical simulation method by [14]. Image sources are used for determining early reflections in order to provide a most accurate localization of primary sound sources (precedence effect [4]) during the simulation. Scattering and reverberation are estimated on-line by means of an improved stochastic ray tracing method. This aspect is an innovation in real-time virtual acoustics, which is to be considered as an important extension of the perceptive dimension. Since this contribution focuses on the implementation and applications of the complete system, no further details are presented here. A detailed description of the implementation and test results can be found in [11]. However, our complete system of binaural reproduction with dynamic crosstalk cancellation and room-acoustical simulation is currently able to simulate up to 10 moving sources with 2 seconds reverberation in real-time on today's standard PCs.

4 Applications

4.1 Interactive Concert Hall

As a case study application which features all aspects of complex acoustical sceneries, we developed a simulation of an existing convention center, the Eurogress located in Aachen, which is occasionally used as a concert hall. The concert hall has a volume of approximately 15.000m³. In this application, users can freely move within the scenery and study, for example, different listening positions and source configurations. In addition to that, they are able to move really close to the sound sources that are located in the scenery, grab them and move them around, while at all the time, the simulation produces a correct binaural representation at the ears of the listener. As sound sources, we modeled five instruments, two violins, a viola, a cello and a double bass.

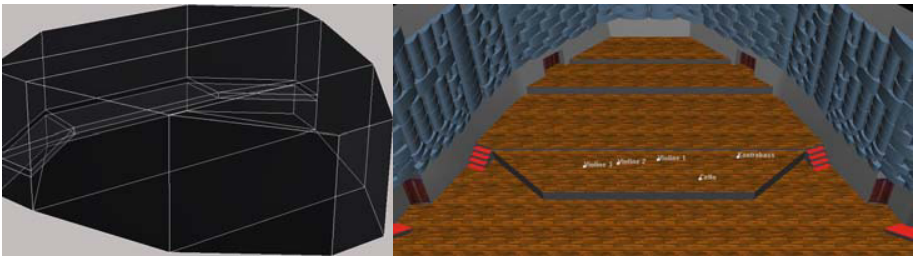

Fig. 2. *Left*: The room model used for the room acoustical simulation. *Right*: View from above the balcony on to the stage on the visual model of the concert hall. The labels indicate the positions of the sound sources in the scene at the beginning of the application. During the simulation, sound sources as well as listener positions may vary without constraints.

A source material we used midi tracks for each instrument, as it is difficult to get musical tracks in the field of classic music as anechoic and truly channel separated material. The five instruments are placed on stage of the concert hall, see Figure 2 (right).

The basis for the room-acoustical simulation is a reduced model of the visual model, see Figure 2 (left), and consists of 105 polygons and 74 planes respectively. This model has to be created before the simulation, as currently it is not possible for us to dynamically change the room-geometry during the simulation run. Although it is kept quite simple, the model contains all indoor elements of the room which are acoustically of interest, e.g., the stage, the skew wall elements, and the balustrade. Details of small elements are neglected and represented by equivalent scattering. Surface properties and coefficients are defined through standardized material data (ISO17497-1, ISO354:2003). For demonstration purposes, we implemented that users can switch between free-field simulation and room-acoustical simulation during the application run, allowing to pair-compare the difference in the sound-field with and without room-acoustics.

Although no formal studies in this environment were carried out, the responses to this dense and interactive scenario by users are very positive. The localization with enabled room-acoustics seems to be less precise, as there are more cues in the signal due to the simulated reflection of the walls, but the overall immersion is much higher, as the sound becomes more vivid and realistic. In addition to that, the presentation of a room inside a 360 degree CAVE-like environment in addition to a 360 degrees dynamic sound field with near-to-head sources seems to compensate the imperfection of the installation, e.g., as we do not compensate reflections from the real projection walls of the projection device in the binaural signal.

4.2 A Visuo-Acoustic VR Study in Neuro-Psychology

The audio rendering system has recently been integrated into *NeuroMan*, a toolkit developed at RWTH Aachen University which allows for VR-based behavioural studies in the neuro-psychological field [13]. Currently, a study is being carried out which aims at developing a three-dimensional cube-shaped paradigm to assess visual, auditory as well as cross-modal cueing effects on spatial orientation of attention. Up to now, cueing effects have only been assessed in two-dimensional space. In the three-dimensional space it remains unclear if cues need to have exactly the same spatial position as the target stimulus in order to provoke an optimal validity effect.

The cube is constructed with cues and targets being presented at exactly the same spatial positions in both the visual and the auditory modalities at an L-shaped workbench. The auditory condition is realised by the audio rendering system presented in the previous sections. The spatial positions in the cube are marked by virtual loudspeakers placed at all eight corners of the cube (see Figure 3 (left)). Spatial cueing is provided by changing the colour of one of these loudspeakers, whereas the loudspeaker itself topples down to serve as a target. There are valid as well as invalid cues (in an 80:20 ratio). The response times depending on the spatial position of these cues in relation to the target position are measured in milliseconds. A preliminary test with

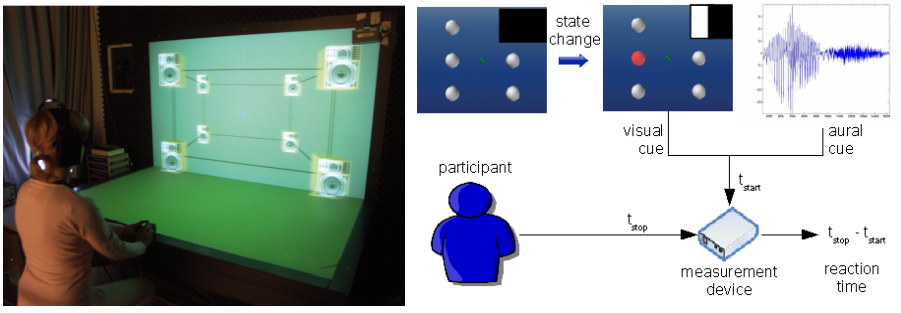

Fig. 3. *Left*: A VR-based neuro-psychological study about cross-modal cueing, performed on an L-bench and making use of the audio rendering system. *Right*: The experiment's setup and schematic view of the reaction time measurement.

9 volunteers has shown that the dynamic binaural approach based on loudspeakers is particularly suitable to perform such experiments and by far preferable to a headphones solution.

5 Discussion

We developed and integrated techniques to realize a fully dynamic acoustic-visual VR environment. We put emphasis on the interactivity aspect of the system, realizing a physically plausible real-time simulation of sound sources located in a room, where both are possible: the listener as well as the sources can be moved independently without constraints. The 360 degrees CTC system enables the use of a flexible, yet low cost, sound installation in CAVE-like environments without the need for headphones. This realizes a non-intrusive technology that can be used directly without further calibration. The complete system is the basis for extensive research on the interplay between acoustic and visual modalities on different levels. While one aspect covers the very basic interactions on a perceptive level, another focus will be research on a subjective level, where immersion or presence aspects will be focused on. Another great challenge will be the work on dynamic room simulation, where users will be able to change parameters of the room-acoustical simulation on-line, e.g., changing the room geometry or wall materials and inspect the differences in the sound-field. This, of course has the precondition of a simulation that has, as an on-line simulation comparable results (in terms of correctness and quality) to off-line simulations usually employed in room acoustical simulation.

Acknowledgements. The work in this paper is part of the DFG project KU 1132/3 and VO600/13. The authors kindly thank the German Research Foundation (DFG) for their support and funding. In addition, we would like to thank Solveyg Anders and Walter Sturm, University Hospital Aachen, for their collaboration in the study about cross-modal cueing, funded by the Interdisciplinary Center for Clinical Research (IZKF BIOMAT).

References

1. Steinberg Media Technologies GmbH: ASIO2.0 Audio Streaming Input Output Development Kit, (2007) last visited: 12.02.2007 URL: http//www.steinberg.net
2. Bauck, J., Cooper, D.H.: Generalization Transaural Stereo and Applications. In: Journal of the AES 44, 683–705 (1996)
3. Berkhout, A.J., Vogel, P., de Vries, D.: Use of Wave Field Synthesis for Natural Reinforced Sound. In: Proceedings of the Audio Engineering Society Convention 92 (1992)
4. Cremer L., Müller, H. A.: Die wissenschaftlichen Grundlagen der Raumakustik. S. Hirzel Verlag (1978)
5. Gardner, W.G.: 3-D audio using loudspeakers; PhD Thesis, Massachusetts Institute of Technology (1997)
6. Kirkeby, O., Nelson, P.A., Hamada, H.: Local sound field reproduction using two loudspeakers. Journal of the Acoustical Society of Amerika 104, 1973–1981 (1998)
7. Lentz, T., Schmitz, O.: Realisation of an adaptive cross-talk cancellation system for a moving listener. In: Proceedings of the 21st Audio Engineering Society Conference, St. Petersburg (2002)
8. Lentz, T., Behler, G.: Dynamic Cross-Talk Cancellation for Binaural Synthesis in Virtual Reality Environments. In: Proceedings of the 117th Audio Engineering Society Convention, San Francisco, USA (2004)
9. Møller, H.: Reproduction of artificial head recordings through loudspeakers, vol. 37 (1989)
10. Schröder, D., Lentz, T.: Real-Time Processing of Image Sources Using Binary Space Partitioning. Journal of the Audio Engineering Society 54,Nr.: 7/8, 604–619 (2006)
11. Schröder, D., Dross, P., Vorländer, M.: A Fast Reverberation Estimator for Virtual Environments. In: Audio Engineering Society, 30th international Conference, Saariselka, Finnland (2007)
12. Springer, J.P., Sladeczek, C., Scheffler, M., Jochstrate, J., Melchior, F., Fröhlich, B.: Combining Wave Field Synthesis and Multi-Viewer Stereo Displays. In: Proc.of the IEEE Virtual Reality 2006 Conference, pp. 237–240 (2006)
13. Valvoda, J.T., Kuhlen, T., Wolter, M., Armbrüster, C., Spijkers, W., Vohn, R., Sturm, W., Fimm, B.: NeuroMan: A Comprehensive Software System for Neuropsychological Experiments. CyberPsychology & Behaviour 8(4), 366–367 (2005)
14. Vorländer, M.: Ein Strahlverfolgungsverfahren zur Berechnung von Schallfeldern in Räumen. ACUSTICA 65, 138–148 (1988)