Precise Near-to-Head Acoustics with Binaural Synthesis

Tobias Lentz[*], Ingo Assenmacher[†], Michael Vorländer[*], Torsten Kuhlen[†]

[*]Institute of Technical Acoustics
RWTH Aachen University
Neustrasse 50, 52066 Aachen, Germany
phone: +49 (0)241 80 97985 , email: tobias.lentz@akustik.rwth-aachen.de
www: www.akustik.rwth-aachen.de

[†]Virtual Reality Group
RWTH Aachen University
Seffenter Weg 23, 52074 Aachen, Germany
phone: +49 (0)241 80 22134, email: assenmacher@rz.rwth-aachen.de
www: www.rz.rwth-aachen.de/vr

Abstract

For enhanced immersion into a virtual scene more than just the visual sense should be addressed by a Virtual Reality system. Additional auditory stimulation appears to have much potential, as it realizes a multi-sensory system. This is especially useful when the user does not have to wear any additional hardware, e.g., headphones. Creating a virtual sound scene with spatially distributed sources requires a technique for adding spatial cues to audio signals and an appropriate reproduction. In this paper we present a real-time audio rendering system that combines dynamic crosstalk cancellation and multi-track binaural synthesis for virtual acoustical imaging. This provides the possibility of simulating spatially distributed sources and, in addition to that, near-to-head sources for a freely moving listener in room-mounted virtual environments without using any headphones. A special focus will be put on near-to-head acoustics, and requirements in respect of the head-related transfer function databases are discussed.

Keywords: Spatial acoustics, crosstalk cancellation, binaural synthesis, multi-modality, interactive Virtual Reality

First presented at the 2nd Workshop "Virtuelle und Erweiterte Realität der GI-Fachgruppe VR/AR", extended and revised for JVRB

1 Introduction

In Virtual Reality (VR) systems, in theory, all human sensory systems have to be stimulated in a natural way to enhance immersion into computer-generated environments with which users can interact freely and naturally. In practice only few sensory systems are involved in VR human computer interaction. With the presentation of stereoscopic images the visual sense is mainly addressed. With respect to immersion, the acoustic sense is a very important additional source of information for the user, as acoustical perception works precisely for close auditory stimuli and does provide cues from all directions. This enhances the liveliness and credibility of the virtual environment and improves user orientation. It is evident that modern VR environments use room-mounted displays like a Holobench or a CAVE-like display system. In such

environments, the listener should be free of any wearable active hardware, e.g., head-mounted displays or headphones.

From the acoustical point of view, the major drawback of room-mounted VR environments using large video walls is that it is extremely difficult to realize a proper loudspeaker setup. In this special field of application a complete binaural approach seems to be an appropriate solution for the problem. In contrast to wave field synthesis, which reproduces the whole sound field with a large number of loudspeakers, binaural synthesis in combination with dynamic crosstalk cancellation (CTC) is able to provide a spatial auditory representation by using few loudspeakers which are not preassigned for a horizontal position. Speakers can be mounted, e.g., above the screens but the placing of virtual sound sources is not restricted due to the binaural reproduction. Using loudspeakers for binaural synthesis introduces the problem of crosstalk, as sound waves determined for the right ear arrive at the left ear and vice versa. In order to realize a proper channel separation at any point of the listening area, CTC filters are calculated on-line when the listener moves his head. In the same manner the binaural synthesis is updated dynamically dependent on the head movement, in order to provide virtual sources which are located fixed to the room, not to the head.

Currently our system already has the ability of generating congruent visual and acoustical scenes providing the possibility of, e.g., research on the interaction between visual and auditory human perception systems. 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. It frees the user of any costly DSP technology or other custom hardware which additionally is hard to maintain. In addition to that it is to our knowledge the first approach to install a versatile and stable real-time binaural acoustics system with dynamic CTC in a CAVE-like environment.

The remainder of this paper is structured as follows. We will give a brief overview of related work in this area, but will mainly focus on VR software systems and comparable approaches. After that we will present some basic facts about human spatial hearing. Then, a short overview of approaches to the reproduction of three-dimensional sound fields is given, which concentrates on the binaural approach. A core component of our system is the dynamic CTC, which is presented later on. A special challenge is the rendering of near-to-head sound sources, which is discussed in more detail after the CTC. Our work is integrated in an existing VR system and some measurements of the system's performance and preliminary listening tests are presented in detail at the end of this paper.

2 Related Work

For the rendering of spatial audio there are already some commercially available solutions. Usually, they are realized by dedicated hardware that can be coupled via a network interface with any VR-application. Systems like these generally provide a plug-in mechanism to run different approaches of spatial audio rendering techniques on the same dedicated DSP hardware. An example for such a system can be found in the Lake HURON machine [Hur05] or more recent systems like [AuS06]. However, the binaural acoustics module of the HURON machine is limited to the usage of headphones in virtual environments, and the AuSIM approach seems to focus on multi listener environments using multiple tracked headphones, while we focus on an individual, headphoneless solution in CAVE-like environments. Hardware constraints also apply to approaches such as [Sto95] or [NSG02]. Recent systems of audio serving technology for various applications can be found in association to the DIVA project [Sav99], [SHLV99]. With regard to the software, [HGL+95] concentrate on data structures, description facilities and synchronization issues for sound and graphics in virtual environments, but do not describe any special sound rendering technique. Spatial acoustic sound layers, e.g., like OpenAL [Ope06], allow application programmers to easily create virtual acoustic worlds. These layers are specifications open for implementation by different vendors, and our approach can be partially implemented based on these interfaces. However, the specifications express their view on virtual sound source modeling which differs from our view in some points. An example can be found in the modeling of source directivity with a cone approach, whereas we model a frequency and direction dependent directivity. In addition to that, they do not take the physical (room-related) position of the user into account. This is obvious, as the specifications apply to the PC consumer market and desktop systems where amplitude panning systems are used and the users have a very limited interaction range, sitting in front of the monitor. Our system needs this information for a correct CTC calculation. But, as the specifi-

urn:nbn:de:0009-6-5890, ISSN 1860-2037

cations do present an extension layer, the information needed for our approach can possibly be implemented using the extension points as defined, e.g., by OpenAL 1.1. Some aspects have then to be explicitly known by the application programmer. This seems to be counter intuitive to the view of the specifications, being a transparent programming interface.

3 Sound Reproduction

For visual stimuli, the brain compares pictures from both eyes to determine the placing of objects in a scene. With this information it creates a three-dimensional cognitive representation that humans perceive as a three-dimensional image. In a straight forward analogy, stimuli that are present at the eardrums will be compared by the brain to determine the nature and the direction of a sound event [Bla96]. Depending on the horizontal angle of incidence, different time delays and levels will consequently arise between both ears. In addition, frequency characteristics dependent on the angle of incidence are influenced by the interference between the direct signal and the reflections of head, shoulder, auricle, and other parts of the human body. The interaction of these three factors permits humans to assign a direction to acoustic events [Møl89]. The characteristics of the sound pressure at the eardrum can be described in the time domain by the Head-Related Impulse Response (HRIR) and in the frequency domain by the Head-Related Transfer Function (HRTF). These transfer functions can be measured individually with small in-ear microphones or with an artificial head.

Figure 1 shows a measurement of the artificial head of the Institute of Technical Acoustics (ITA) of RWTH Aachen University under 120 degree relating to the frontal direction in the horizontal plane. The Interaural Time Difference (ITD) can be assessed in the time domain plot. The Interaural Level Difference (ILD) is shown in the frequency domain plot and clarifies the frequency dependent level increase at the ear turned toward the sound source and the decrease at the ear that is turned away from the sound source. The aspect of head movement will be discussed later in Section 3.2.

A popular technique for spatial acoustic imaging can be found in the vector-base amplitude panning approach. Here, sound source distance and position is modeled by modifying the amplitude on the output channels of a predefined static loudspeaker setup. As near-to-head imaging is impossible with this approach, 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 section will briefly introduce principles and problems of these technologies.

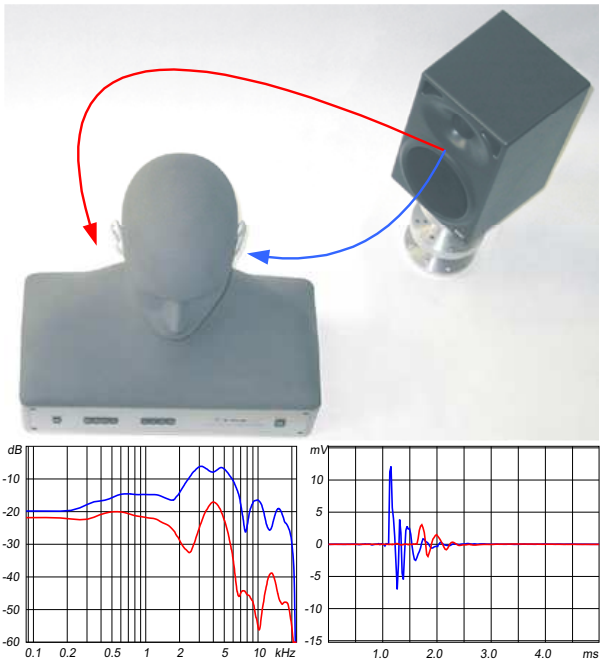


Figure 1: HRTF (smoothed with 1/6 octave bandwidth) and HRIR of a sound source under 120 degree measured using the artificial head of the Institute of Technical Acoustics of RWTH Aachen University.

3.1 Methods for Acoustical Imaging

The basic theory of the wave field synthesis is the *Huygens' principle*. An array of loudspeakers (ranging from just a few to some hundreds in number) is placed in the same position as a microphone array was placed at the time of recording the sound event in order to reproduce an entire real sound field. Figure 2 shows this principle of recording and reproduction [BVdV92, dVSV94]. In a VR environment the loudspeaker signal will then be calculated for a given position of one or more virtual sources.

By reproducing a wave field in the whole listening area, it is possible to walk around or turn the head while the spatial impression does not change. This is the big advantage of this principle. The main drawback, beyond the high effort of processing power, is the size of the loudspeaker array. Furthermore, mostly

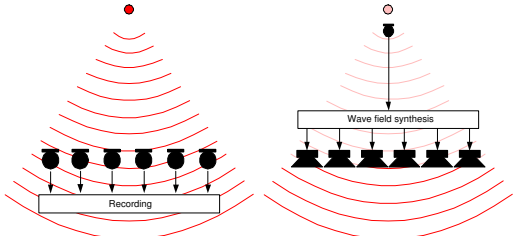


Figure 2: Setup for recording a sound field and reproducing it by wave field synthesis.

two-dimensional solutions have been presented so far. The placement of those arrays in video projection VR systems like a Holobench or a Powerwall is only just possible, but in display systems with four to six surfaces as in CAVE-like environments it is nearly impossible without any severe compromises.

As an illustration to this problem we give an example for a five sided projection CAVE-like environment, where the roof is open and five sides are passive stereo projection surfaces. The usage of polarized light for visual channel separation needs a special surface of the projection walls to retain the polarization. In addition to that, as more than one and large surfaces are used in this installation, only inflexible material for the walls, e.g., acrylic glass, can be considered. The user movements and interactions are tracked using an optical tracking approach, hence the construction uses a top mounted rack for the cameras to be placed. Figure 3 depicts this installation.

The only place left for the loudspeakers to be positioned is next to the cameras on the rack system. The application of a wave field synthesis approach is not feasible in this or a similar setup. First of all, speakers can only mounted high above the user's head. Second, tracking has a higher priority than the acoustic reproduction system, as it is essential for user centered projection and interaction. This results in the requirement that any loudspeaker setup has to take the placing of the cameras into consideration. Typically for devices that use back projection, there is no room for the speakers behind the projection surfaces to place loudspeakers, even when permeable or flexible wall material is used.

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 HRIRs 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. The major problem is the reproduction of a signal at these positions in such a way that the listener perceives it as natural. This is accomplished by the dynamic CTC system which will be described in detail in section 4.

3.2 Dynamic Binaural Synthesis

The 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, 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 software 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

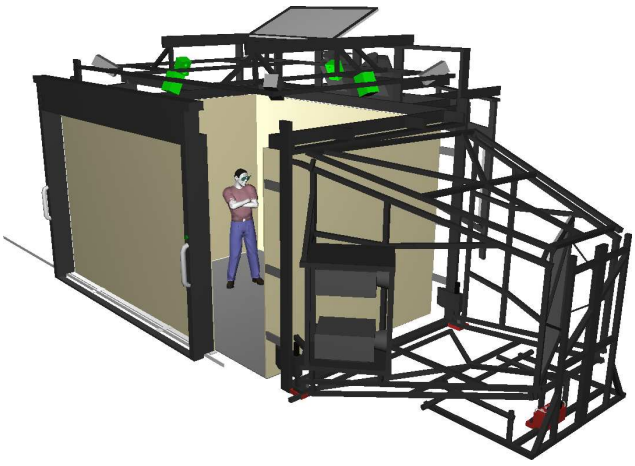


Figure 3: Five sided projection CAVE-like virtual environment using optical tracking. Possible loudspeakers are colored in green and mounted on the rack on top of the device.

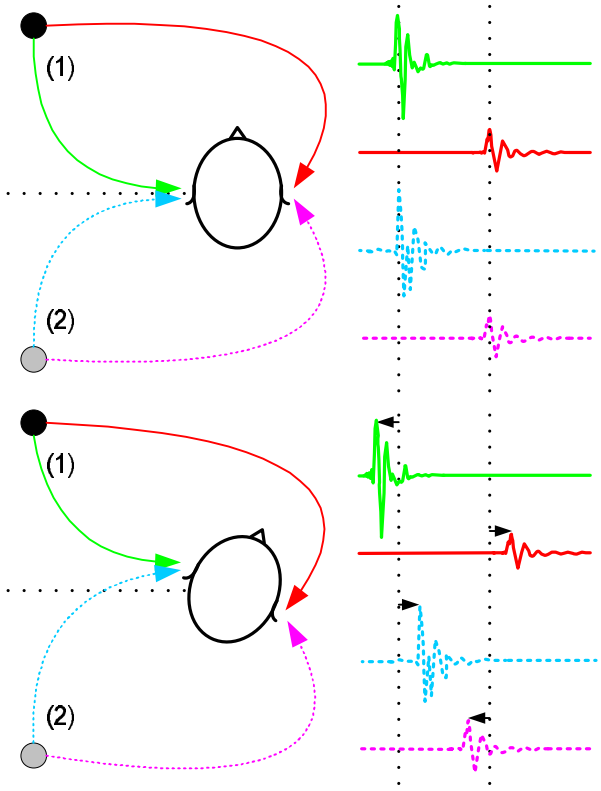


Figure 4: Variance of ITD depending on the relative head orientation of the user.

can be chosen from a database. It is also possible to synthesize many different sources and to create a complex three-dimensional acoustic scenario.

The main advantage of a dynamic synthesis is the almost complete elimination of front back confusion which often appears when using the static binaural synthesis with non-individualized HRTFs as, e.g., reported in [WAKW93]. Figure 4 shows the arrival time at each ear in relation to the listener's orientation. In this example the interaural time delay is almost equal regardless whether the sound is reproduced at position 1 or 2. Although the frequency dependent interaural level difference is still different for the two source positions, this is often not a sufficient cue. For that reason the signal could be perceived by the user as if coming from a non-existent mirror source due to the congruent differences in the ITD. Using a dynamic synthesis the ITD increases when the listeners median plane (see Figure 5) turns away from the source (1) and decreases when the median plane turns toward the source (2). Due to this fact a source is well defined in its position due to the ancillary information, the relative movement of the listener.

3.3 Near-to-Head Sources

Another advantage of binaural synthesis is the ability of near-to-head source imaging. In contrast to panning systems where the virtual sources are always on or behind the line spanned by the speakers, binaural synthesis can realize a source at any distance to the head by using an appropriate HRTF. Especially at closer distances the ILD is higher due to the shading of the head.

In addition to that, the near-to-head source imaging matches with the use case of interacting in room-mounted virtual environments. In common room-mounted visual VR displays, the area for the user movement is not much larger than the grasping range of a normal user, e.g., as it is limited by the projection walls of a CAVE-like system or by the cables needed for a tracking device. With respect to the more logical layer of interaction, the observation that direct interaction metaphors (such as by-hand-placements or manipulations) are more precise than indirect interaction metaphors (such as ray-based indirect transformations) holds. Furthermore, objects that are placed in the area close to the projection plane are perceived as most stereoscopic and provide the most insight into complex visualizations, e.g., data from CFD simulations. As a consequence, most objects reside rather near to the user's head and within the grasping range during a typical VR session, see Figure 6. This reveals why the binaural approach is a very suitable method for acoustic rendering in this field of application and display setup.

As mentioned above, it is possible to realize every virtual distance, in particular, one close to the listener's head, by choosing the appropriate HRTF. In practice, it is neither feasible nor necessary to store HRTFs for every possible distance. In order to reduce the amount of HRTFs that have to be present in the system it is useful to examine where a special HRTF is needed

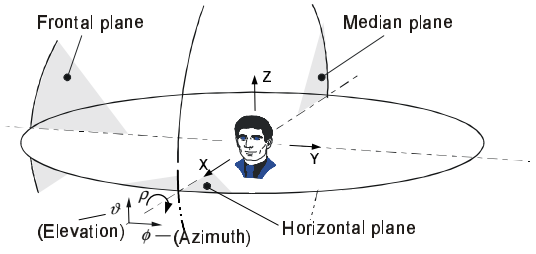


Figure 5: Specified frame of reference for head movements and rotations.

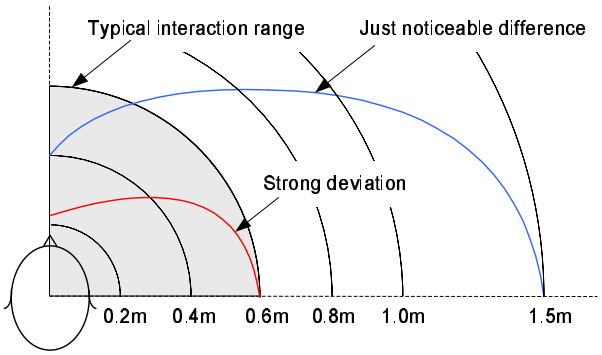


Figure 6: Limits of noticeable differences between near-field and far-field HRTFs. The gray area marks the typical interaction range in a room-mounted VR CAVE-like display.

and where interpolation or a simple level adaption is sufficient. Beyond a certain distance from the head the sound pressure level decreases in the same range at both ears. The ILD is nearly constant and is independent of the distance. This is the area where the waves emitted from the source have approximately spherical characteristics. The level decreases according to the spherical wave attenuation for the specific distance. So all dependency on the distance can be modified at runtime of the program. Therefore, HRTFs must be stored for different angles but all in the same distance. That fact reduces significantly the required RAM-space needed for the storage of HRTFs.

A simple listening test was performed in the hemianechoic chamber of the ITA, examining the ranges where different near-field HRTFs have to be applied. The listeners were asked to compare signals from simulated HRTFs with those from correspondingly measured HRTFs on two criteria, namely the perceived location of the source and any coloration of the signals. The simulated HRTFs were prepared from far-field HRTFs (measured in a distance of two meters) with a simple level correction applied likewise to both channels. From the comments of all 9 listeners, for each distance and azimuth, a coloration would earn 5 points, while a change in perceived location earned 10 points. No perceived difference earned 0 points. The total points were then added up. Therefore, the higher the total score, the greater the perceived audible difference between the original and simulated HRTFs for that point.

Figure 6 shows the result. The nearer boundary line defines where huge audible differences and a change in the perceived location of the sound were detected. The boundary line further away from the head is the boundary where any audible difference was detected. These results were taken as basis for the HRTF database design.

For the synthesis database, the HRTFs of the ITA head were measured in distances of 0.2m, 0.3m, 0.4m, 0.5 m, 0.75 m, 1.0 m, 1.5 m and 2.0 m. The spatial resolution is 1 degree for the azimuth angle and 5 degree for the elevation angle. Virtual sources between two measured distances will be realized by linear interpolation of two group delay compensated HRTFs. The higher amount of measurements at closer distances to the head have to be applied to minimize comb-filter effects. Consequently, our generated HRTF database covers all requirements for a near-to-head source imaging system. It should be noted that our system uses the HRTFs of the full sphere because the ITA head has asymmetrical pinnae and head geometry. Non-symmetrical pinnae cause positive effects on the perceived externalization of the generated virtual sources [BT05].

4 Crosstalk Cancellation

As stated above, binaural audio signals contain all information being necessary for the localization of a sound source and can have a complex three-dimensional character. One problem which still has to be solved is the correct reproduction of these signals. 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. But equalization of headphones and reproducibility when headphones are removed and replaced is not a trivial task. 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. For this reason, a reproduction by loudspeaker can be considered to be associated with less problems with regard to the naturalness of the presented sound. The problem with loudspeaker reproduction is the crosstalk between the channels that destroys the three-dimensional cues of the binaural signal. 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 CTC filter which is based on the transfer functions

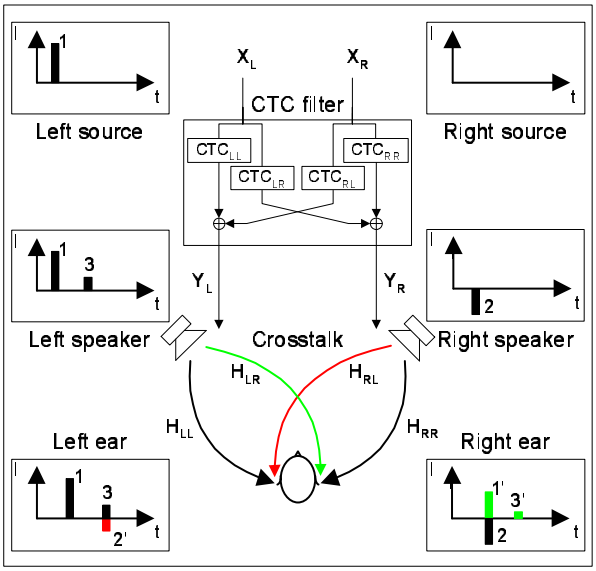


Figure 7: Principle of static crosstalk cancellation. Presenting an impulse addressed to the left ear (1) and the first steps of compensation (2, 3).

from each loudspeaker to each ear, see Section 3. 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, see Figure 7. A detailed description of static crosstalk cancellation systems can be found in [BC96].

For a moving listener it is necessary to adapt the CTC filters dependent on the current position and orientation of the listener [Gar97]. The system described in this paper calculates the CTC filters on-line using a HRTF database and the position and orientation of the tracking system. The CTC database contains HRTFs measured in a distance of 2 m in a resolution of 1 degree for azimuth and elevation. The distance has to be adjusted before the filter calculation by modifying group delay and the level according to the spherical wave attenuation for the actual distance of the head to the loudspeaker. The filter set is recalculated when the weighted sum of the listener's movement in all degrees of freedom is above 1 (see Equation 1). The threshold can be parameterized in six degrees of freedom, positional values $(\Delta x, \Delta y, \Delta z)$ and rotational values $(\Delta\phi, \Delta\vartheta, \Delta\rho)$ (see Figure 5). Movements and head rotation in the horizontal plane are most critical so Δx, Δy, and $\Delta\phi$ are chosen very low to dominate the filter update (see Table 1). The threshold always refers to the value where the last exceeding occurred in such a way as the resulting hysteresis prevents a permanent switching between two filters. This may occur when a fixed spacing determines the boundaries between two filters and the tracking data jitter a little bit.

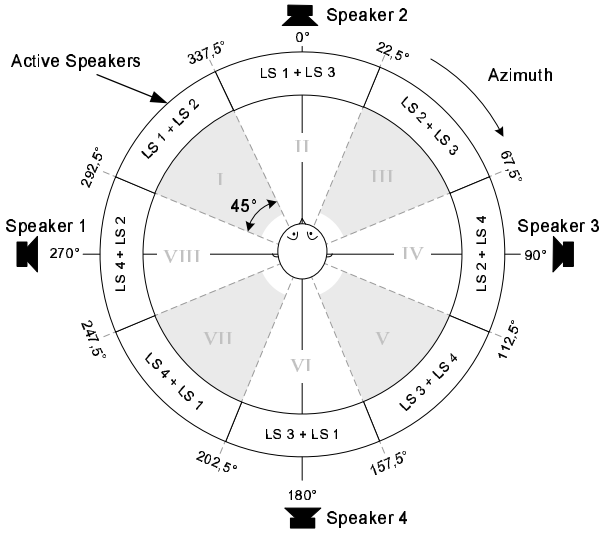
$$\begin{aligned} s &= \frac{|x_{new}-x_{old}|}{\Delta x} + \frac{|y_{new}-y_{old}|}{\Delta y} \\ &+ \frac{|z_{new}-z_{old}|}{\Delta z} + \frac{|\phi_{new}-\phi_{old}|}{\Delta\phi} \\ &+ \frac{|\vartheta_{new}-\vartheta_{old}|}{\Delta\vartheta} + \frac{|\rho_{new}-\rho_{old}|}{\Delta\rho} \\ &\geq 1 \end{aligned} \tag{1}$$

But a 2-channel dynamic CTC is only possible in the angle spanned by the loudspeakers. When the ITD in one HRTF decreases toward zero, as happens when the listener faces one speaker, an adequate cancellation is impossible [LS02]. To provide a full 360 degrees rotation for the listener we extended the well-known 2-speaker CTC solution to a 4-speaker setup. With four loudspeakers eight combinations of a normal CTC system are possible, yet, the validity areas of the 2-channel CTC systems will, however, overlap.

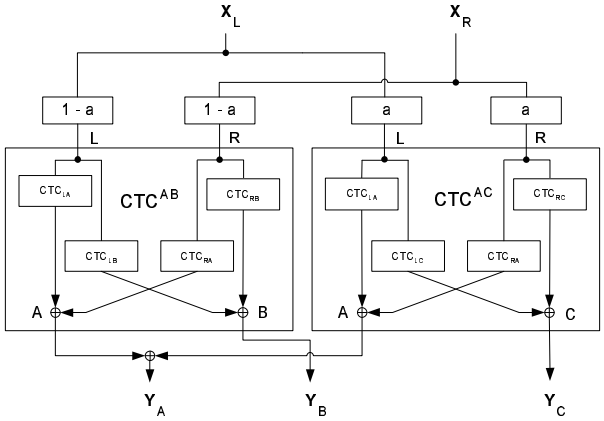
Measurements and listening tests showed that the cancellation achieves good results inside the different areas, but the switching between areas is still audible as a "click". Comparing CTC filters one step before and one step after switching the filters reveals some differences, in particular, at high frequencies. Inaccuracies in the speaker's placement and the determination of the head position by the head tracker can cause a mismatch in the time alignment so that a sufficient consistent cancellation is not possible. To reduce the interfering "clicks", a smoother changeover from one sector to the next is needed. Due to the fact that the CTC filter structure is a linear time-invariant system, a linear superposition of two classical 2-speaker systems is possible. Speakers in the active area are labeled with A and B, speakers in the destination area are labeled with A and C. For example, fading from sector I to II, speakers 1 and 2 (A, B) are active, and after the fading is complete, speakers 1 and 3 (A, C) are active, see Figure 8(a). Figure 8(b) shows the complete crosstalk cancellation filter structure of a three

Δx	Δy	Δz	$\Delta\phi$	$\Delta\vartheta$	$\Delta\rho$
1 cm	1 cm	3 cm	1 °	3 °	3 °

Table 1: Maximal deviations for filter recalculation in all degrees of freedom.



(a)



(b)

Figure 8: Partitioning into sectors and dynamic 3-channel CTC filter structure with cross-fade.

speaker solution. To provide full rotation, the ± 45 degree and ± 90 degree configurations alternate, as they do when applying the switching method. Only in between two sectors both configurations are active at the same time. A more detailed description of the four channel CTC algorithm can be found in [Len06]. The superposition of two CTC systems imply a high precision of the loudspeaker position and the determination of the current head position (tracking system). Inaccuracies could cause comb-filter effects if the two signals don't match perfectly. Listening tests using this fading method showed that "clicks" are apparently not audible anymore. Furthermore, sound colorations due to comb-filter effects were indiscernible, but a combined angle and time dependent fading method will be tested in the future to make the system even more robust.

The crosstalk cancellation achieves a channel separation at an average of 20 dB in an area of about $2\text{ m} \times 3\text{ m}$ (which is also the area of valid tracking in our system) in a frequency range of $100\text{ Hz} - 12\text{ kHz}$ [LAS05].

With this system an efficient CTC can be established in the full space around the user, and it is especially suited to enhance VR systems in CAVE-like environments with spatial audio to combine visual and acoustical stimuli in an excellent way. Our work is integrated as a module into a complete VR software system, called ViSTA, which is developed by the VR-Group at RWTH Aachen University.

5 Integration

The general architecture and interdependence of the subparts depicted above are explained as follows. Figure 9 shows the schematic implementation of combining binaural synthesis and CTC. The figure presents a box "ViSTA-Audio Manager" which is a placeholder for an arbitrary VR software system that is capable of delivering spatial information of different sound sources and the listener's head in real-time over a TCP/IP-based communication channel.

The connection between the two systems consists of two bidirectional (TCP) and one unidirectional (UDP) communication channel. The first TCP channel establishes the connection to the audio server and allows to control the sound system. The second TCP channel is automatically created by the system and is used for server sided events, errors and exception messages to the VR client application. The UDP channel exists for the fast rate transmission of spatial updates of the listener and various sound sources in the virtual environ-

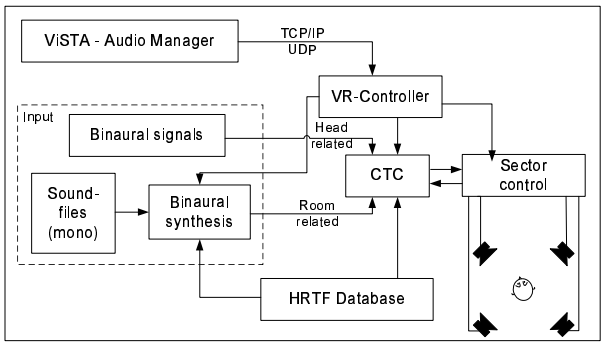


Figure 9: Complete system for virtual acoustical imaging.

ment, encoded as a table of positions and orientations. The spatial updates are used to change the filters online and have to be delivered with an update rate of about 100 Hz. In contrast to this, the TCP channels are expected to be used at a low frequency. A more complete overview of the system setup of the our system implementation is described in [AKLV04].

Two of the main audio to video synchronization issues will be discussed in the following section.

The first issue is the latency that the system has in starting, pausing, stopping and altering of attributes of virtual sound sources. This issue is important for the matching of a suddenly appearing sound to a specific object or event. [VdPK00] have shown that it is an advantage if a sound that indicates a specific situation (e.g., a sound that is emitted from a hammer that hits a steel plate) is optimally presented 35 ms after the situation has been visually perceived by the user.

The other synchronization issue to deal with is the latency for updates that are necessary when the is user moving. The latency discussion on that issue has to consider the head-tracking technology that is used in the system. This point deals with the lag in updates if the user moves his head, possibly at a very fast rate. [BSM+04] have stated that an update lag of 70 ms at maximum remains unnoticed for the user. This includes the runtime of the sound waves from the loudspeakers to the user's ears.

The following section will give details about the measurements that were taken using the following setup. The tracking PC comprises of a Pentium-4 with 2.4 GHz CPU speed and 256 MB of RAM in conjunction with a four camera set-up. In the current system we employ an optical tracking system, the A.R.T. tracking system which delivers spatial measures with a fixed update rate of 60 Hz [ART04]. As a client visual VR machine, a Linux dual Pentium-4 machine with 3 GHz CPU speed and 2 GB of RAM was used, in conjunction with an NVidia GeForce FX 5950 Ultra as graphics output device. The host for the auditory VR subsystem was a Pentium-4 machine with 3 GHz CPU speed and 1 GB of RAM. As audio hardware an RME Hammerfall system is used. The RME hardware allows the sound output streaming with a scalable buffer size and therefore a minimum latency of 1.5 ms. The network interconnection between the machines was a standard Gigabit Ethernet this is used for normal network communication in our laboratory.

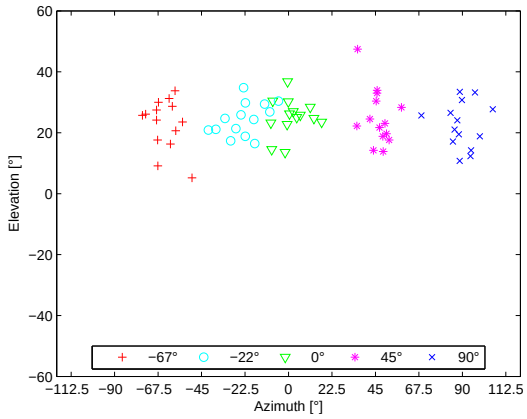
The latency of the audio system is the time elapsed between the incoming of a new position and orientation for either a source or the listener, and the point in time the output signal is generated with the updated filter functions. The output block length of the convolution is 256 taps as well as the chosen buffer size of the sound output device, resulting in a time between two buffer switches of 5.8 ms at 44.1 kHz sampling rate for the rendering of a single block. The calculation of a new CTC filter set (1024 taps) takes 3.5 ms on a 3 GHz PC and 0.1 ms to process a new binaural filter (512 taps) for each sound source. In a worst case scenario the filter calculation just finishes after the sound output device fetched the next block, so it takes the time playing this block until the updated filter becomes active at the output. That would cause a latency of one block. In such a case the overall latency accumulates to 11.8 ms.

Cost Item		Time (ms)
Tracking (60Hz)		18.20
UDP transport	+	0.70
VR app. transform	+	0.10
Serializing, deserializing	+	0.22
UDP spatial update	+	0.70
Filter processing	+	11.80
Sum	=	31.72

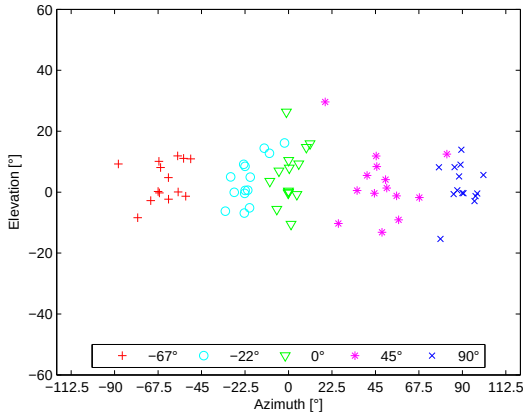
Table 2: Total costs for a end-to-end trip of a spatial update from a tracking sample to the output of the loudspeakers.

Table 2 shows the costs for an end-to-end trip starting with the tracking and ending with the sounds waves reaching the user's ear. The table shows worst case measurements which still meet the 70 ms update criteria for head movements. A more detailed discussion of synchronization and latency can be found in [AKL05].

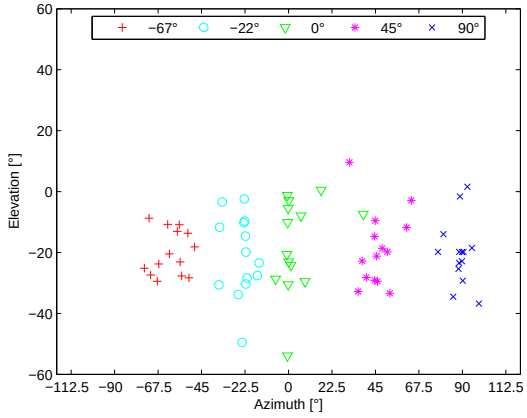
A preliminary listening test with 14 participants gives first a first impression of the quality of the auditory representation. To evaluate the spatial precision of the virtual sources a presentation of auditory stimuli together with visual stimuli was chosen. It is tested if the auditory perception matches with the visual image. The visual anchor stimulus is a white ball in a distance of 1.2 m to the center of the CAVE-like environment. The control feedback, a red ball, can be adjusted by the user using a FlyStick, a 6 degree of freedom input device of the A.R.T. tracking system. In this test the user is asked to determine the deviation of the visual and the auditory stimuli. The participants are instructed that the acoustic stimulus is in the region close



(a) Localization results for sources at 30 degree elevation



(b) Localization results for sources at 0 degree elevation



(c) Localization results for sources at -30 degree elevation

Figure 10: Results of the listening test with simultaneously presented auditory and visual stimuli.

to the visual anchor point (white ball), but does not have to match the exact its position (although possible). The perceived position of the sound is to be indicated by moving the control feedback (red ball) to the position assuming the sound is coming from. The user is fully free in walking around and turning his head. To determine the deviation the position of the control feedback and of the visual stimulus itself is stored in a logging file after confirming the choice. The stimuli were presented in random order at $-67, -22, 0, 45$ and 90 degree azimuth related to the frontal view. The reason for choosing these special azimuth angles for the test based on the speaker setup and the resulting different sectors of the crosstalk cancellation (see Section 4). Figure 10 shows the results for all tests with a stimuli presentation at different angles of elevation related to the frontal view. One reason for the scattering of the perceived angles could be the individual differences between the HRTFs of the listener and the HRTFs of the artificial head used for the synthesis and the crosstalk cancellation. It can also be noted, that sources with an elevation angle of -30 degree (below the listener) tend to be perceived higher. These two points will be subjects of further investigation.

6 Summary and Outlook

Within the scope of this paper, a system for reproduction of binaural signals using loudspeakers for a moving listener is presented. In order to realize this idea, it is necessary to update the CTC filter depending on the listener's position. The required filters are calculated at runtime of the program. In addition to that, it is possible to perform a binaural synthesis to transform mono sound sources into binaural sources. The number of sound sources is only limited by the computational resources being available in the system hardware. The approach is software-based and not restricted to the presence of specialized DSP hardware, in contrast to most existing approaches. It provides a flexible and scalable approach for spatialized acoustical rendering including the possibility of realizing near-to-head sources.

A future extension of the system will be the integration of a room acoustical simulation algorithm. With an effective low latency convolution algorithm it will be possible to auralize a room and present the signal by a crosstalk canceled loudspeaker system together with the visual image of the room. The local next step is an extensive perception-based evaluation of the localization performance by means of user studies.

7 Acknowledgment

The work of this paper is part of the DFG project KU1132/3 and VO600/13. The authors kindly thank the DFG for their support and funding.

References

[AKL05] Ingo Assenmacher, Torsten Kuhlen, and Tobias Lentz, *Binaural Acoustics for CAVE-like Environments Without Headphones*, Eurographics Symposium on Virtual Environments 2005, Eurographics ACM SIGGRAPH Symposium Proceedings, 2005.

[AKLV04] Ingo Assenmacher, Torsten Kuhlen, Tobias Lentz, and Michael Vorländer, *Integrating Real-time Binaural Acoustics into VR Applications*, egve04: Proceedings of the 10th Eurographics Symposium on Virtual Environments, Eurographics Association, June 2004, ISBN 3-905673-10-x, pp. 129–136.

[ART04] ART, *ART tracking systems, Manual*, ART GmbH, 2004, Last visited: 31.07.2006.

[AuS06] AuSim3D, *www.ausim3d.com*, 2006, Last visited: 31.07.2006.

[BC96] Jerry Bauck and Duane H. Cooper, *Generalization Transaural Stereo and Applications*, Journal of the Audio Engineering Society **44** (1996), no. 9, 683–705.

[Bla96] Jens Blauert, *Spatial Hearing: The psychophysics of Human Sound Localization*, revised edition ed., MIT Press, Cambridge MA, 1996, ISBN 0-262-02413-6.

[BSM+04] Douglas S. Brungart, Drian D. Simpson, Richard L. McKinley, Alexander J. Kordik, Ronald C. Dallman, and David A. Ovenshire, *The Interaction Between Head-Tracker Latency, Source Duration, and Response Time in the Localization of Virtual Sound Sources*, Proceedings of ICAD 04 - Tenth Meeting of the International Conference on Auditory Display, Sidney, Australia, July 2004.

[BT05] Tim Brookes and Chris Treble, *The Effect of Non-Symmetrical Left/Right Recording Pinnae on the Perceived Externalisation of Binaural Recordings*, Proceedings of the 118th Audio Engineering Society Convention Barcelona, Spain, no. Preprint 6439, May 2005.

[BVdV92] A.J. Berkhout, P. Vogel, and D. de Vries, *Use of Wave Field Synthesis for Natural Reinforced Sound*, Proceedings of the Audio Engineering Society Convention 92, no. Preprint 3299, 1992.

[dVSV94] Diemer de Vries, Evert W. Start, and Vincent G. Valstar, *The Wave-Field Synthesis Concept Applied to Generation of Reflections and Reverberation*, Proceedings of the 96th Audio Engineering Society Convention Amsterdam, The Netherlands, no. Preprint 3812, March 1994.

[Gar97] William G. Gardner, *3-D audio using loudspeakers*, Ph.D. thesis, MIT Media Lab, Massachusetts Institute of Technology, 1997.

[HGL+95] James K. Hahn, Joe Geigel, Jong Won Lee, Larry Gritz, Tapio Takala, and Suneil Mishra, *An Integrated Approach to Motion and Sound*, Journal of Visualization and Computer Animation **6** (1995), no. 2, 109–123.

[Hur05] Lake Huron, *http://www.lake.com.au*, 2005, Last visited: 31.07.2006.

[LAS05] Tobias Lentz, Ingo Assenmacher, and Jan Sokoll, *Performance of Spatial Audio Using Dynamic Cross-Talk Cancellation*, Proceedings of the 119th Audio Engineering Society Convention New York, USA, no. Preprint 6541, 2005.

[Len06] Tobias Lentz, *Dynamic Crosstalk Cancellation for Binaural Synthesis in Virtual Reality Environments*, Journal of the Audio Engineering Society **54** (2006), no. 4, 283–294.

[LS02] Tobias Lentz and Oliver Schmitz, *Realisation of an adaptive cross-talk cancellation system for a moving listener*, 21st Audio Engineering Society Conference, St. Petersburg, June 2002.

[Møl89] Henrik Møller, *Reproduction of artificial head recordings through loudspeakers*, Journal of the Audio Engineering Society. **37** (1989), no. 1/2, 30–33.

[NSG02] Martin Naef, Oliver Staadt, and Markus Gross, *Spatialized Audio Rendering for Immersive Virtual Environments*, Proceedings of the ACM symposium on Virtual reality software and technology, Hong Kong, China, 2002, pp. 65 – 72.

[Ope06] OpenAL, *www.openal.org*, 2006, Last visited: 31.07.2006.

[Sav99] Lauri Savioja, *Modeling Techniques for Virtual Acoustics*, Ph.D. thesis, Helsinki University of Technology, December 1999.

[SHLV99] Lauri Savioja, Jyri Huopaniemi, Tapio Lokki, and Riitta Väänänen, *Creating Interactive Virtual Acoustic Environments*, Journal of the Audio Engineering Society **49** (1999), no. 9, 675–705.

[Sto95] Russel L. Storms, *Npsnet-3d Sound Server: An Effective Use Of The Auditory Channel*, Master's thesis, Naval Postgraduate School, Monterey CA, 1995.

[VdPK00] Steven Van de Par and Armin Kohlrausch, *Sensitivity to Auditory-Visual Asynchrony and to Jitter in Auditory-Visual Timing*, Human Vision and Electronic Imaging V, Proceedings of the SPIE (Bernice E. Rogowitz and Thrasyvoulos N. Pappas, eds.), vol. 3959, 2000, pp. 234–242.

[WAKW93] Elisabeth Wenzel, Marianne Arruda, Doris J. Kistler, and Frederic Wightman, *Localisation using nonindividualized head-related transfer functions*, Journal of the Acoustical Society of America **94 (1)** (1993), 111–123.

Citation

Tobias Lentz, Ingo Assenmacher, Michael Vorländer, Torsten Kuhlen
Precise Near-To-Head Acoustics With Binaural Synthesis
Journal of Virtual Reality and Broadcasting, urn:nbn:de:0009-6-5890, ISSN 1860-2037.