TECNIACUSTICA Cáceres · 2011

42º CONGRESO ESPAÑOL DE ACÚSTICA
ENCUENTRO IBÉRICO DE ACÚSTICA
EUROPEAN SYMPOSIUM ON ENVIRONMENTAL ACOUSTICS
AND ON BUILDINGS ACOUSTICALLY SUSTAINABLE

SPATIAL AUDIO REPRODUCTION METHODS FOR VIRTUAL REALITY

REFERENCIA PACS: 43.38.Md

Bruno Sanches Masiero, Michael Vorländer
Institute of Technical Acoustics. RWTH-Aachen University. Germany
email: bma@akustik.rwth-aachen.de

ABSTRACT

The Institute of Technical Acoustics (ITA) in Aachen has recently established its new Virtual Reality Lab, aimed for research and educational purpose. This laboratory is equipped with four potent computers for use on real-time room simulation and auralization, a stereoscopic projection system, an electromagnetic tracking device and eight high quality monitor loudspeakers array for immersive spatial audio reproduction.

The hardware mentioned above allows the synthesis and reproduction of either binaural signals, by means of binaural synthesis and dynamic crosstalk cancellation technique, or stereophonic signals, using the approaches of vector-base amplitude panning (VBAP) and higher order ambisonics (HOA). In this paper we will outline the above mentioned methods and describe how they are integrated with the virtual reality system under development in ITA.

INTRODUCTION

In a virtual acoustic environment the main goal is the simulation (rendering) of acoustic environments and an appropriate audio reproduction [1]. While the computer modeling of sound propagation is quite advanced and fast algorithms are available, the reproduction is often treated with insufficient care so that headphone systems or loudspeaker 3D surround sound systems are not perfectly matched to individual listeners, or they are not described and equalized properly to enable a reproducibility of the sound signal presentation. The presentation of the stimulus, however, is of utmost importance and adequate reproduction systems have to be used to reproduce these stimuli. The reproduction technique directly influences the quality of the presented stimulus and, accordingly, the level of *immersion* achieved by the listener.

The user of a virtual reality environment should be able to move freely inside this space, so hardware that has to be fastened on the user's body should be avoided. On basis of this argumentation, audio reproduction via headphones is often undesirable, i.e., spatial audio should preferably be reproduced over loudspeakers.

There are several approaches to loudspeaker-based spatial sound reproduction. *Binaural technique* aims at exactly reproducing the sound pressure at the listeners' eardrums [2-5], other simpler methods focus only on delivering plausible interaural time and/or amplitude cues at the listener's position – such as *Ambisonics* [6-8] or *Vector Base Amplitude Panning* (VBAP) [9] – while more complex methods focus on synthesizing the desired sound field within the whole listening area delimited by the driving loudspeakers – such as *wave field synthesis* (WFS) or *higher order Ambisonics* (HOA) [10], [11]. The main spatial audio techniques to be named are:

- Binaural (also known as *Transaural*)
- Stereophony
- Vector Base Amplitude Panning (VBAP)
- Ambisonics and its extension to Higher-Order Ambisonics (HOA)
- Wave Field Synthesis (WFS)

In the binaural technique, all spatial information perceived by a listener is provided by the sound pressures present at the left and right eardrums, called a *binaural* signal (thus having two channels). These signals can be recorded with a dummy head or be synthesized with the use of a *head-related transfer function* (HRTF) database. So, if we are able to reproduce at the user's eardrums the same signal that would be heard if this person were in the virtual scene, then all correct spatial information will be available. A prerequisite for an accurate spatial impression is an exact reproduction of both channels at their respective ear – and only there. This is easily achieved with a headphone playback but, as already mentioned, extra hardware should be avoided in immersive VR applications.

Binaural signals can also be reproduced through loudspeakers. Nevertheless, since the signal reproduced by a loudspeaker will be heard by the listener's both ears (see crosstalk effect on figure 1), a set of filters has to be used to achieve the required high channel separation between ears, usually called *Crosstalk Cancellation* (CTC) filters.

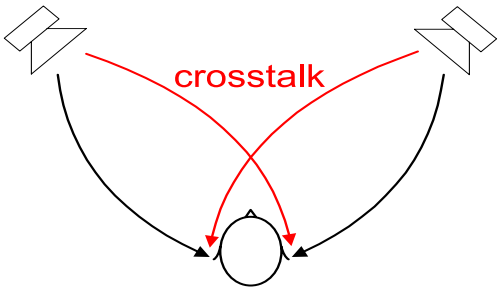

Figure 1: Example of the crosstalk's presence (red arrows).

Crosstalk is accepted in the co-called Stereophony which is at present the most popular commercial audio format, being in use by the phonographic industry since the 1940's. Stereophony does not aim at binaural reproduction but uses amplitude and time delay difference to create a phantom (virtual) sound source between two loudspeakers. VBAP may be seen as an extension of the classical intensity stereophony to more than two loudspeakers, allowing the virtual source to be placed anywhere inside the convex surface defined by the system's loudspeakers[1].

In its original formulation, Ambisonics is also an amplitude panning method, with the extension that it also allows phase shifts while VBAP uses only positive panning weights. On the one hand, VBAP is a technique mainly used for the reproduction of synthetic sounds. On the other hand, Ambisonics signals can be directly recorded with an adequate microphone array that is independent of the reproduction set-up, turning this technique especially interesting for the recording and broadcasting areas. Ambisonics also distinguishes itself from VBAP by the fact that all loudspeakers on the array may be active for any given virtual source position, delivering a more homogeneous sound field with the drawback of a broader virtual source.

The panning techniques mentioned so far rely on the psychoacoustic effect of "summing localization" to produce interaural cues that lead the listener to perceive a virtual source in space. Another category of spatial sound reproduction systems aims at faithfully reproducing the original/simulated sound field within a defined region (sweet spot). These methods are characterized by the requirement of many loudspeakers.

[1] The spatial audio systems used in movie theaters and home cinemas (e.g. surround 5.1 or 7.1) are also a variation of the amplitude panning technique.

The WFS technique is based on the Huygens-Fresnel principle, which states that each wave front can be described by an infinite numbers of elementary waves on the boundary surface. Consequently, any wave field can be generated inside a volume delimited by infinitely many sound sources. In practical implementation, to reduce the amount of loudspeakers installed, only two-dimensional arrays are used. And the use of a finite number of loudspeakers defines an upper frequency limit up to which wave field reconstruction is possible. HOA is the direct extension of the classical Ambisonics technique to a larger number of loudspeakers distributed on the surface of a sphere. Unlike WFS, where the sound pressure must be know on the border of the reproduction area, HOA is based on the spherical harmonic description of the sound field at single point in the center of the array. HOA arrays should be preferably spherical (3D) or circular (2D), while WFS arrays can be of the form of any concave polygon. HOA will have an upper frequency limit like the WFS technique.

Since these sound field reproduction techniques require a large number of loudspeakers, these techniques are not commonly used for virtual reality set-ups, as the positioning of the loudspeakers can severely interfere with the video reproduction aspects. For this reason we restrict the size of the loudspeaker array at our VR-Lab to eight transducers and we focus on binaural and panning techniques, which will be described in detail next.

BINAURAL LOUDSPEAKER TECHNOLOGY

The spatial impressions perceived by humans are extracted from cues imprinted on the signal arriving at the listener's ear. These cues are caused by alteration on the incoming wave front by reflection and diffraction on the listener's body, meaning that these cues are highly individual. The alterations cause to the wave front arriving from a given direction can be described by a set of filters, the so called head-related transfer functions (HRTF). With a HRTF database it is possible to synthesize virtual sources at any position in the room and even position sources very close to the listener's head by making use of near-field effects. With the other methods mentioned above, this is either not possible or only possible to a limited extend. If the directivity of the source is considered as well, a real and natural sounding spatial hearing sensation can be created.

The reproduction chain of a loudspeaker-based binaural reproduction consists of a binaural signal that is first filtered by crosstalk cancellation filters and then reproduced by loudspeakers after which the binaural signal is finally heard by the listener. Ideally, the CTC filters have to be constantly adjusted to the position and viewing direction of the listener. Combining a dynamic binaural synthesis with a dynamic crosstalk cancellation allows a realistic spatial reproduction with a few loudspeakers. The dynamic adjustment of crosstalk cancellation and synthesis filters in real time requires considerable system complexity, besides the need to constantly know the position and orientation of the listener inside the virtual reality environment. This is done with tracking devices based either on electromagnetic or optic input.

A pair of loudspeakers can deliver a sufficient channel separation – attenuation level of the crosstalk channel - only at a restricted range of head directions in relation to the speakers. To allow the user to freely move in the VR environment, it is possible to use at least three (in our case we use four) loudspeakers distributed in a plane and switch between the most adequate loudspeakers pair according to the head position [12].

CROSSTALK CANCELLATION FILTERS

The CTC filters are a set of filters applied to the binaural signal to reduce the crosstalk effect, i.e., to improve the channel separation between the signals arriving at the left and right ears. The aim is to achieve the best possible channel separation by using a suitable set of filters.

Figure 2 shows the iterative structure of the crosstalk cancellation. The impulse (a), that should only be audible by the left ear, is also reproduced – time shifted and with another amplitude - by the right loudspeaker (b), so that the crosstalk term a' is eliminated by destructive interference. This procedure has to be repeated to eliminate the compensation signal that arrives at the left ear as well, so the left loudspeaker emits the impulse (c) to cancel the second crosstalk term b'. Theoretically, this iteration has to be repeated indefinitely to achieve an infinitely high channel separation. In practice, however, this sequence decay below noise level after 5 to 7 iterations and the filter can then be truncated at that point.

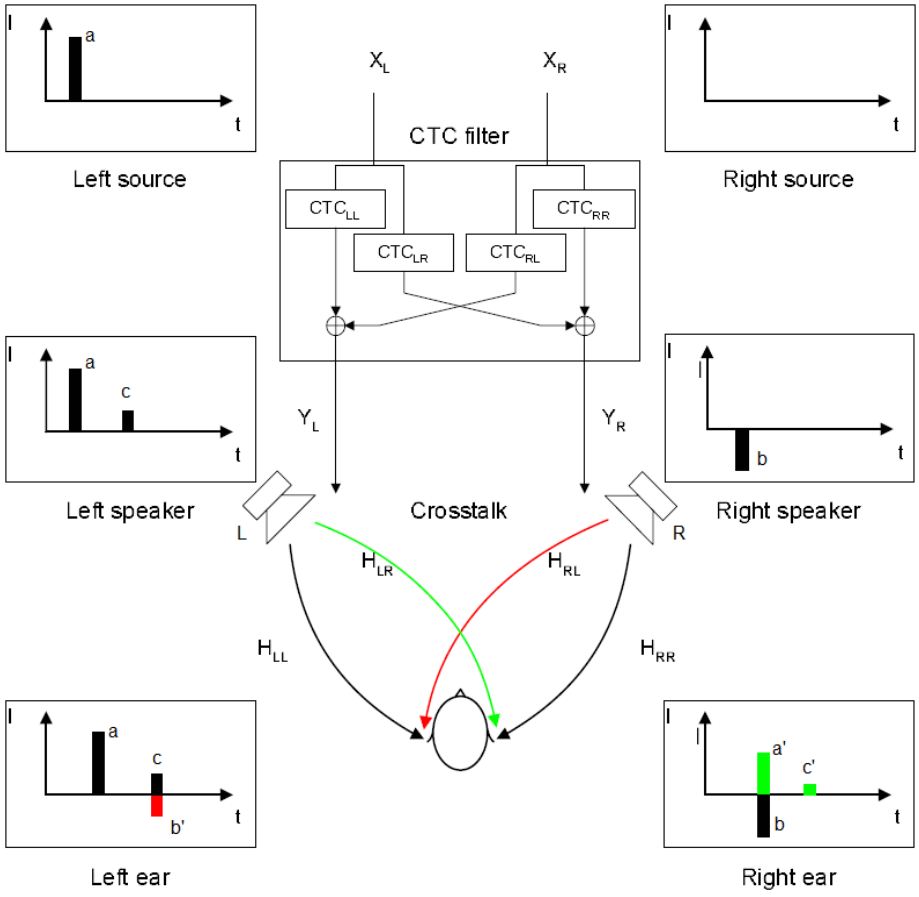

Figure 2: principle of crosstalk cancellation

For a mathematical analysis of the problem a closed-form solution is more suitable than the iterative method (that itself is more appropriate to understand the principles of CTC). Both solutions are, however, equivalent. The problem can be modeled in the following way:

$$Z_L = Y_L \cdot H_{LL} + Y_R \cdot H_{LR} \quad (1)$$

$$Z_R = Y_L \cdot H_{RL} + Y_R \cdot H_{RR} \quad (2)$$

These two equations can be rewritten in matrix form in the following way

$$\begin{bmatrix} Z_L \\ Z_R \end{bmatrix} = \begin{bmatrix} H_{LL} & H_{LR} \\ H_{RL} & H_{RR} \end{bmatrix} \begin{bmatrix} Y_L \\ Y_R \end{bmatrix} = \boldsymbol{Hy} = \boldsymbol{z}. \quad (3)$$

The filters for the crosstalk cancellation (CTC) are placed before the loudspeakers, so that $\boldsymbol{y} = \boldsymbol{C} \cdot \boldsymbol{x}$. The transfer function of the complete system is given in matrix form as

$$\boldsymbol{z} = \boldsymbol{H} \cdot \boldsymbol{C} \cdot \boldsymbol{x}. \quad (4)$$

For a binaural reproduction the output $\boldsymbol{z}$ should equal the input $\boldsymbol{x}$ apart of a time delay. Thus, the following equation has to be valid

$$\boldsymbol{H} \cdot \boldsymbol{C} = \boldsymbol{e}^{-j\Delta} \cdot \boldsymbol{I}, \quad (5)$$

where $\boldsymbol{I}$ is the identity matrix. The transfer matrix with the crosstalk cancellation filters $\boldsymbol{C}$ can be easily obtained by means of a pseudo-inverse of the transfer matrix $\boldsymbol{H}$, resulting in:

$$\boldsymbol{C} = \boldsymbol{e}^{-j\Delta} \cdot \boldsymbol{H}^{\dagger}, \quad (6)$$

The closed-form solution according to equation (2.6) is the exact solution for the entire crosstalk cancellation. It requires, however, infinitely long filters that are also prone to have stability problems. The later issue can be dealt with through a regularized matrix inversion approach. The regularization applies a constrain at the maximum gain allowed to the filters and can be expressed as follows

$$\boldsymbol{C} = \boldsymbol{e}^{-j\Delta} \cdot (\boldsymbol{H}^{H}\boldsymbol{H} + \beta(f)\boldsymbol{I})^{-1} \cdot \boldsymbol{H}^{H}. \quad (7)$$

This approach has no requirement on the matrix $\boldsymbol{H}$ to be square, meaning that more than two loudspeakers could be simultaneously used to achieve improved channel separation. That is a current topic of investigation.

STEREO DIPOLE

A special form of the binaural technique is found if the condition number of the matrix solution is investigated with regard to the loudspeaker placement. Then it is found that for high frequencies the setup should be given by a narrow triangle, whereas the low frequencies require more head shadow effect and correspondingly a wide-angle loudspeaker setup [13]. Accordingly, with standard stereo systems and crossover networks, the task is just to find an appropriate loudspeaker box shape which meets these requirements of optimum signal processing conditions. Unfortunately, the requirement for distributed loudspeakers is not practical in VR environments, for the same reasons as already commented for WFS and HOA. Commercial products based on this technology are known by the name “sound bars”.

STEREOPHONY

The term *stereophony* derives from the Greek words "στερεός" (stereos), meaning "firm, solid", plus the word "φωνή" (phone), meaning "sound, tone, voice". Even though the term *stereophonic* could be correctly applied to describe all "surround-sound" systems, it is *de facto* used only to describe the more common 2-channel systems – from now on the term *stereophony* will be used only to refer to the latter system.

In a natural hearing situation, the auditory event[2] produced by, e.g., a musical instrument stems from a single sound event[3]. Meanwhile, in stereo reproduction there are two sound sources or respectively two sound events that nevertheless produce a single auditory event as the sound recorded from the instrument and played back through both loudspeakers is perceived somewhere in between the loudspeakers. The perceived source is also called a *phantom source.*

In psychoacoustics this phenomenon is known as *summing localization* which states that a single auditory event may be provoked by many very similar sound events as they are superposed. Nevertheless, a theory to precisely explain why the human auditory system works this way is still missing, despite the large number of proposed models.

Stereophony will work best if the loudspeakers are placed in free field and the listener is located in a way that his head and the two loudspeakers form an equilateral triangle. Even though these conditions are rarely met in practice, as e.g. in a living room, stereophony still works with mild limitations.

Summing localization occurs when the signals played back by the loudspeakers:

a. are identical (monophony);
b. have the same time structure but different amplitude (intensity stereophony);
c. have the same time structure but are slightly delayed (time-of-arrival stereophony);
d. are a combination of a. and b. (mixed stereophony).

It is noteworthy that this effect will still occur, up to a given limit, if the time structure of the signals are not perfect identical. Loudspeakers playing the same signal will result in an auditory event in between both transducers. A level or time difference between the signals will move the phantom source over an arc, away from the center and in the direction of the louder/earlier loudspeaker. If the amplitude difference between the signals is too large, then the auditory event will coincide with the sound event generated by the louder transducer. Meanwhile, if the time difference between the signals becomes too large – greater than 1 ms to 5 ms, depending on the signal type – then the summing localization effect give away and two distinct sound sources are perceived, similar to echo.

A myriad of microphone configuration exist for live stereo recording, focusing either on level difference or on time-of-arrival difference or, as more commonly, delivering a mixture of both cues. For virtual reality applications, almost only intensity stereophony is used as it delivers a more robust phantom image. Early research already showed that in practical implementation intensity stereophony is superior then time-of-arrival stereophony. The dependency of the perceived direction with the head position and frequency is lower for phantom images generated with intensity stereophony then time-of-arrival stereophony. Also the localization error variance is lower for intensity stereophony, especially at high frequencies.

VECTOR BASE AMPLITUDE PANNING (VBAP)

Intensity stereophony relies on playing the same signal with different amplitude at two loudspeakers to place the phantom source in between the loudspeakers. The basic principle of stereophony can be expanded to allow the creation of phantom sources in any desired direction – also in the three-dimensional space. Such systems were already built in the 1950's and were

[2] Auditory Event is the description of the phenomenon of hearing on a perceptual level, i.e., as it is perceived by the human auditory system.

[3] Sound Event is the description of the phenomenon of hearing on a physical level, i.e., its physical aspects.

then called “Synthetic Soundfield”. Later, in the 1990’s this method was described in vectorial notation, hence the name “Vector Base Amplitude Panning”.

For a two-dimensional set-up, the two loudspeakers closest to the direction of the phantom source should be selected while for a three-dimensional arrangement the three loudspeakers closest to the direction of the desired phantom source should be selected, forming an active triangle. Source elevation cannot be as well perceived as the source’s azimuth. Regardless of 2D or 3D set-ups, VBAP will deliver a relatively small listening region (sweet spot) and not well defined phantom sources for lateral directions. One of the main drawbacks of VBAP is the variation on the apparent width of the virtual source. When placed in a central position between the loudspeakers, the virtual source is perceived as a large source. As the virtual source move towards one of the loudspeakers, the apparent width of the virtual source will decrease, up to the point that when the virtual source is placed on the same direction as one of the loudspeaker, only this loudspeaker will be active and the perceived width of the virtual source will match the width of the active loudspeaker.

It is important to emphasize that phantom sources can only be placed inside the convex polygon defined by the loudspeakers. Other than stereophony, reproduction of material specially recorded for this set-up is not practicable, so this method is used mainly for synthesized sound.

THREE-DIMENSIONAL SYNTHESIS

Let the loudspeakers be positioned on the surface of a sphere, equidistant from the listener. The three-dimensional unit-vectors $\boldsymbol{l_1}, \boldsymbol{l_2}$ and $\boldsymbol{l_3}$ define the directions of loudspeakers 1, 2 and 3 respectively. The unit-vector pointing at the virtual source defines the virtual source direction. We derive the normalized gain factors as follows

$$\widehat{\boldsymbol{p}} = g_1\widehat{\boldsymbol{l_1}} + g_2\widehat{\boldsymbol{l_2}} + g_3\widehat{\boldsymbol{l_3}}, \tag{8}$$

or, in matrix notation

$$\widehat{\boldsymbol{p}} = \boldsymbol{g^T L_{123}}, \tag{9}$$

where $\boldsymbol{L_{12}} = [\boldsymbol{l_1}\ \boldsymbol{l_2}\ \boldsymbol{l_3}]^T$ is a matrix containing the source position and $\boldsymbol{g} = [g_1\ g_2\ g_3]^T$ is a vector containing the gain factors. The values of $\boldsymbol{g}$ can be calculated by

$$\boldsymbol{g} = \widehat{\boldsymbol{p}}^T\boldsymbol{L_{12}^{-1}} = [p_1\ p_2\ p_3]\begin{bmatrix} l_{11} & l_{12} & l_{13} \\ l_{21} & l_{22} & l_{23} \\ l_{31} & l_{32} & l_{33} \end{bmatrix}^{-1}, \tag{10}$$

and finally energy normalization is required

$$\boldsymbol{g_{norm}} = \frac{\sqrt{C}\boldsymbol{g}}{|\boldsymbol{g}|_2}. \tag{11}$$

The three loudspeakers closest to the virtual source should be used to define the active triangles where the source is located. Note that each loudspeaker may belong to several bases. The active triangles should not be intersecting and they should be selected so that maximum localization accuracy in each direction is provided, i.e., the area of the active triangles should be kept as small as possible. In other words, when good localization accuracy on a large listening area is desired, the dimension of the active region must be decreased, or, equivalently, more loudspeakers are needed [9].

AMBISONICS

As already mentioned in the previous section, VBAP is primarily used for synthesized sound since a recording scheme independent of the playback setup is not available for VBAP. The Ambisonics technique was proposed mainly to overcome the lack of recording and broadcasting possibilities of multi-channel spatial audio systems. In its original formulation, the recording method of intensity stereophony with two perpendicularly superposed cardioid microphones, commonly known as X-Y arrangement, was upgraded by the use of an extra figure-of-eight microphone perpendicular to the other two microphones – note that the arrangement with two cardioid microphone can be substituted by two figure-of-eight and an omnidirectional microphone. The directivity pattern of these microphones corresponds to the form of the spherical harmonics of order 0 and 1, as depicted in Figure 3. Even though originally formulated only up to first order, Ambisonics can be expanded to use higher spherical harmonic order, the so called *Higher Order Ambisonics (HOA)*, that allows an ever more precise description and reproduction of the sound field at the cost of a more complex recording and reproduction system.

To construct an array made of three figure-of-eight microphones plus and omnidirectional microphone very close together is unpractical. But a set-up with four omnidirectional microphones distributed on the faces of a tetrahedron can be used instead and the signals can then be transformed into the equivalent omnidirectional and figure-of-eight patterns through a spherical harmonics transformation. The signals delivered by the microphones are called “A-Format” signals while the signals from the virtual microphones, i.e., the transformed signals, are called “B-Format” or portable signals. The B-Format signals can be independently saved or broadcasted and at the playback end of the chain these signals must be adequately decoded to drive the speaker set-up available for reproduction.

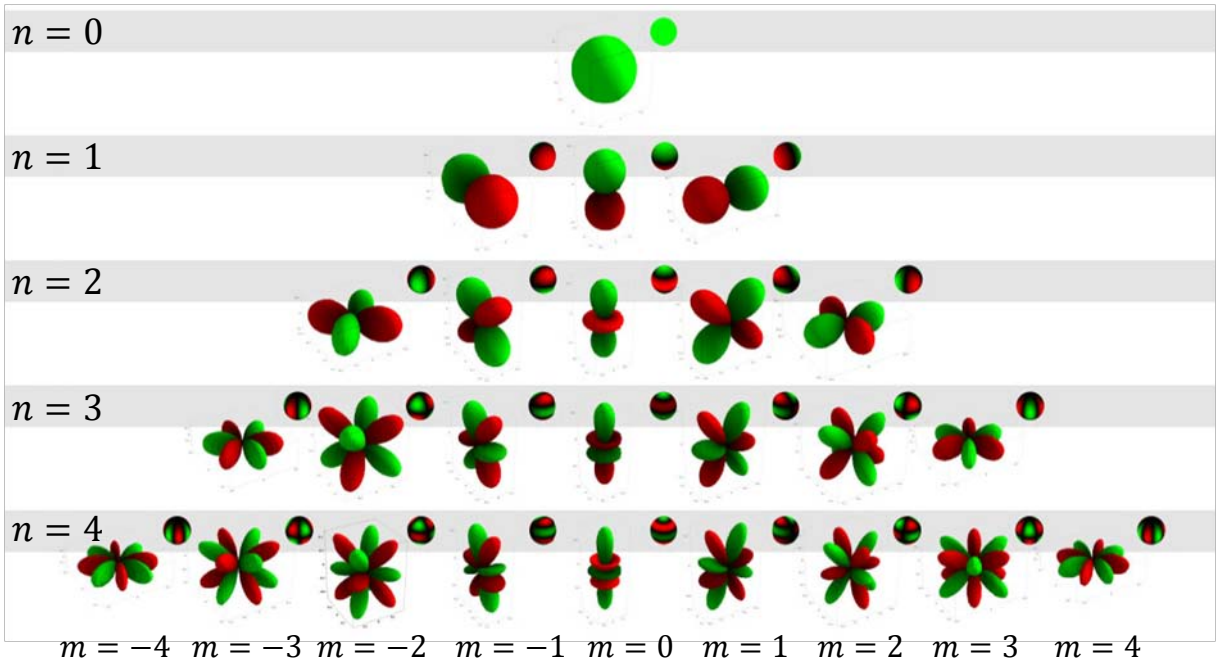

Figure 3: Example of spherical harmonics. Depicted are the first five orders of the real spherical harmonics.

In classical Ambisonics the decoding step is done assuming that the loudspeakers are plane wave sources that lead to real-valued and frequency independent panning weights. Different then in VBAP, in Ambisonics all loudspeakers may be simultaneously active, often with opposite phase. The superposition of the loudspeakers' signals in the center of the array results in a sound field as similar as possible to the original sound field described by the B-format signal. Even though the listener – usually located at the center of the array – acts as an obstacle that avoids perfect sound-field reconstruction, Ambisonics is still able to deliver correct spatial cues for sound localization since it works, like stereophony and VBAP, also based on amplitude panning and summing localization.

It is to be expected that the localization perception will improve with an increase in the number of loudspeakers and in the corresponding spherical harmonic order. The enhancement in spatial resolution will also lead to an extended sweet spot.

In Stereophony and VBAP distance coding is done through variation of gain and delay. This does not allow a virtual source to be placed closer then the distance of the loudspeakers. Under the assumption that the loudspeakers radiate plane waves Ambisonics suffers from the same limitation. But if the loudspeakers are modeled as point sources, then it becomes possible to synthesize point sources either outside the loudspeaker array or inside the array, the so-called focused sources. This does, however, require a frequency-dependent equalization, i.e., it cannot be achieved with simple panning weighting and a considerably dense array.

AMBISONICS SYNTHESIS

When used for playback of recorded material, Ambisonics is divided into encoding and decoding stages. For virtual reality application this two stages are merged into a single synthesis process, but the naming convention with encoding and decoding matrixes is kept unchanged.

When modeling the secondary sources as plane waves (loudspeakers are at a large distance from the listener) the loudspeakers weights can be calculated as the solution of a linear system of equations. In this case, the weights can be interpreted as panning functions (pure weighting). The aim is to synthesize an outgoing plane wave from arbitrary direction $\boldsymbol{n}_{pw} = -\left[1\ \theta_{pw}\ \varphi_{pw}\right]^T$ using plane wave sources radiating from directions $\boldsymbol{n}_l = -[1\ \theta_l\ \varphi_l]^T$ with amplitudes w_l. The notation with minus is used to define the direction vector pointing towards the origin, as is the case with incoming wave front as assumed in Ambisonics.

In spherical coordinates $\boldsymbol{x}_S = [r\ \theta\ \varphi]^T$ the solution to the wave equation may be written in terms of spherical Bessel functions and spherical harmonics. The spatial variation of the sound field at a radial frequency ω – wave number given by $k = {}^{\omega}/_{c}$, c being the speed of sound – may be expressed as

$$p_S(\boldsymbol{x_S}, \omega) = p(r, \theta, \varphi, \omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} C_n^m(\omega) j_n(kr) Y_n^m(\theta, \varphi), \tag{12}$$

where $j_n(kr)$ is the spherical Bessel function of the first kind, $C_n^m(\omega)$ so-called spherical harmonics expansion coefficients of the function $p_S(r, \theta, \varphi, w)$ and $Y_n^m(\theta, \varphi)$ are the spherical harmonics partially depicted in its real formulation in Figure 3 and defined in its complex formulation as

$$Y_n^m(\theta, \varphi) = \sqrt{\frac{(2n+1)}{4\pi} \frac{(n-m)!}{(n+m)!}} P_n^m(\cos\theta) e^{ik\varphi}, \tag{13}$$

where $P_n^m(\cdot)$ denotes the m-th order associated Legendre polynomial of n-th degree.

Real spherical harmonics bases are obtained from

$$\begin{cases} Y_n^m = \dfrac{Y_n^m + (-1)^m Y_n^{-m}}{\sqrt{2}}, \text{for } m > 0 \\ Y_n^m = \dfrac{Y_n^{-m} - (-1)^m Y_n^m}{i\sqrt{2}}, \text{for } m < 0 \end{cases}. \tag{14}$$

AMBISONICS DECODING: PLANE WAVE APPROACH

The spherical harmonic expansion of a plane wave arriving from direction $\boldsymbol{n}_{pw}$ is given as

$$\begin{aligned} S_{pw}(\boldsymbol{x}_S, \omega) &= \hat{S}_{pw}(\omega) e^{-ik(-\boldsymbol{n}_{pw} \cdot \boldsymbol{x}_S)} \\ &= \hat{S}_{pw}(\omega) \left(4\pi \sum_{n=0}^{\infty} i^n j_n(kr) \sum_{m=-n}^{n} Y_n^m(\theta, \varphi) Y_n^m(\theta_{pw}, \varphi_{pw})^* \right). \end{aligned} \tag{15}$$

The sound field generated by L plane waves at the center of the spherical array is given by

$$\tilde{P}_{pw}(r, \theta, \varphi, \omega) = 4\pi \sum_{n=0}^{\infty} i^n j_n(kr) \sum_{m=-n}^{n} Y_n^m(\theta, \varphi) \times \sum_{l=1}^{L} w_l \cdot Y_n^m(\theta_l, \varphi_l)^*. \tag{16}$$

The superposition of all arriving plane waves should approximate as well as possible the plane wave expansion of the desired plane wave arriving from direction $\boldsymbol{n}_{pw}$. The resulting mode-matching equation for each order *n* and degree *m* is

$$\sum_{l=1}^{L} w_l \cdot Y_n^m(\theta_l, \varphi_l)^* = Y_n^m(\theta_{pw}, \varphi_{pw})^*, \tag{17}$$

that can be expressed in matrix formulation as

$$\begin{pmatrix} Y_0^0(\theta_1, \varphi_1)^* & \cdots & Y_0^0(\theta_l, \varphi_l)^* \\ Y_1^{-1}(\theta_1, \varphi_1)^* & \cdots & Y_1^{-1}(\theta_l, \varphi_l)^* \\ Y_1^0(\theta_1, \varphi_1)^* & \cdots & Y_1^0(\theta_l, \varphi_l)^* \\ Y_1^1(\theta_1, \varphi_1)^* & \cdots & Y_1^1(\theta_l, \varphi_l)^* \\ \vdots & \ddots & \vdots \\ Y_n^m(\theta_1, \varphi_1)^* & \cdots & Y_n^m(\theta_l, \varphi_l)^* \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_l \end{pmatrix} = \begin{pmatrix} Y_0^0(\theta_{pw}, \varphi_{pw})^* \\ Y_1^{-1}(\theta_{pw}, \varphi_{pw})^* \\ Y_1^0(\theta_{pw}, \varphi_{pw})^* \\ Y_1^1(\theta_{pw}, \varphi_{pw})^* \\ \vdots \\ Y_n^m(\theta_{pw}, \varphi_{pw})^* \end{pmatrix}, \tag{18}$$

or in a more compact form as

$$\boldsymbol{C}\boldsymbol{w} = \boldsymbol{b}. \tag{19}$$

Equation (18) introduces a linear equation system to be solved for the L weights w_l for all spherical harmonics up to the truncation order N resulting in a linear equation system with $(N+1)^2$ equations. Best reproduction quality is achieved when $L \approx (N+1)^2$.

If $(N+1)^2 = L$, then the decoding matrix can be found from the inverse of the Matrix $\boldsymbol{C}$ (re-encoding matrix**)** whose elements are the encoding gains associated to the loudspeaker directions.

$$\boldsymbol{D} = \boldsymbol{C}^{-1}. \tag{20}$$

If the number of modes to be matched is less than the number of loudspeakers $(N+1)^2 < L$, then the linear equation system is under-determined and the decoding matrix is found from the pseudo-inverse of $\boldsymbol{C}$

$$\boldsymbol{D} = \boldsymbol{C}^{\dagger} = \boldsymbol{C}^{\boldsymbol{H}}(\boldsymbol{C} \cdot \boldsymbol{C}^{\boldsymbol{H}})^{-1}. \tag{21}$$

Finally, if the number of modes exceeds the number of loudspeakers $(N+1)^2 < L$, then there is no exact solution for the inversion problem. Typically, the solution with the minimum least-squared error, given by the pseudo-inverse, is applied

$$\boldsymbol{D} = \boldsymbol{C}^{\dagger} = (\boldsymbol{C} \cdot \boldsymbol{C}^{\boldsymbol{H}})^{-1}\boldsymbol{C}^{\boldsymbol{H}}. \tag{22}$$

CONCLUSIONS

In this paper we reviewed several spatial audio reproduction methods that were implemented at the VR Lab of the Institute of Technical Acoustics in the RWTH Aachen University. This is part of our research activities in Acoustics Virtual Reality where we compare the system performance for various kinds of applications. One example is the comparison of room acoustic impressions in real and virtual rooms, and another example is the psychological effort to perform the so-called attention switch in scenarios of verbal communication.

Also in education we use our VR Lab technologies. Students participating at the Acoustic Virtual Reality laboratory course offered by the institute have the opportunity to implement all this methods and subjectively compare their qualities and drawbacks.

With fixed electroacoustic installations in the laboratory and well-documented software in MATLAB we are aiming at a stable environment for education and future developments of 3D audio technology.

REFERENCES

[1] M. Vorländer, *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality (RWTHedition)*. Springer, 2007, p. 335.

[2] B. B. Bauer, “Stereophonic Earphones and Binaural Loudspeakers,” *Journal of the Audio Engineering Society*, vol. 9, no. 2, pp. 148-151, 1961.

[3] B. S. Atal, M. Hill, and M. R. Schroeder, “Apparent Sound Source Translator,” U.S. Patent 3,236,949.

[4] O. Kirkeby, P. A. Nelson, and H. Hamada, “The ‘stereo dipole’ a virtual source imaging system using two closely spaced loudspeakers,” *Journal of the Audio Engineering Society*, vol. 46, no. 5, pp. 387–395, 1998.

[5] T. Lentz, “Dynamic crosstalk cancellation for binaural synthesis in virtual reality environments,” *Journal of the Audio Engineering Society*, vol. 54, no. 4, pp. 283-294, 2006.

[6] M. Gerzon, “PERIPHONY: WITH-HEIGHT SOUND REPRODUCTION,” *J. Audio Eng. Soc.*, vol. 21, no. 1, pp. 2-10, 1973.

[7] J. Daniel, “Spatial Sound Encoding Including Near Field Effect: Introducing Distance Coding Filters and a Viable, New Ambisonic Format,” in *23rd International Conference: Signal Processing in Audio Recording and Reproduction*, 2003.

[8] F. Zotter, H. Pomberger, and M. Frank, “An Alternative Ambisonics Formulation: Modal Source Strength Matching and the Effect of Spatial Aliasing,” in *Proc. of the 126th AES Conv., Munich*, 2009, pp. 1-12.

[9] V. Pulkki, “Virtual sound source positioning using vector base amplitude panning,” *Journal of the Audio Engineering Society*, vol. 45, no. 6, pp. 456–466, 1997.

[10] J. Daniel, R. Nicol, and S. Moreau, “Further investigations of high order ambisonics and wavefield synthesis for holophonic sound imaging,” in *AES 114 th Convention*, 2003, vol. convention, no. 5, pp. 2764-2778.

[11] J. Ahrens and S. Spors, “An Analytical Approach to Sound Field Reproduction Using Circular and Spherical Loudspeaker Distributions,” *Acta Acustica united with Acustica*, vol. 94, no. 6, pp. 988-999, Nov. 2008.

[12] T. Lentz and G. K. Behler, “Dynamic Cross-Talk Cancellation for Binaural Synthesis in Virtual Reality Environments,” in *Audio Engineering Society Convention 117*, 2004.

[13] T. Takeuchi and P. A. Nelson, “Subjective and objective evaluation of the optimal source distribution for virtual acoustic imaging,” *JOURNAL-AUDIO ENGINEERING SOCIETY*, vol. 55, no. 11, p. 981, 2007.