The Future of Audio Reproduction

Technology – Formats – Applications

Matthias Geier[1], Sascha Spors[1], and Stefan Weinzierl[2]

[1] Deutsche Telekom Laboratories, Quality and Usability Lab, TU Berlin,
Ernst-Reuter-Platz 7, 10587 Berlin, Germany
{Matthias.Geier,Sascha.Spors}@telekom.de
http://www.qu.tu-berlin.de
[2] Audio Communication Group, TU Berlin,
Einsteinufer 17, 10587 Berlin, Germany
Stefan.Weinzierl@tu-berlin.de
http://www.ak.tu-berlin.de

Abstract. The introduction of new techniques for audio reproduction such as binaural technology, *Wave Field Synthesis* and *Higher Order Ambisonics* is accompanied by a paradigm shift from *channel-based* to *object-based* transmission and storage of spatial audio. The separate coding of source signal and source location is not only more efficient considering the number of channels used for reproduction by large loudspeaker arrays, it will also open up new options for a user-controlled soundfield design. The paper describes the technological change from stereophonic to array-based audio reproduction techniques and introduces a new proposal for the coding of spatial properties related to auditory objects.

1 Introduction

Audio recording, transmission and reproduction has been a very active field of research and development in the past decades. Techniques for the reproduction of audio signals including a respresentation of the spatial configuration and the acoustical environment of the corresponding sound sources have been an important aspect of recent innovations such as multichannel audio for cinema and DVD and new techniques for audio reproduction, which are primarily used in a research context so far.

Stereophonic reproduction is currently the most widespread audio reproduction technique. However, the spatial cues of an auditory scene, which allow the listener to localize sound sources and to identify features of the acoustical environment, are only preserved to a limited degree. This has led to a variety of new techniques for audio reproduction such as binaural technology, *Wave Field Synthesis* (WFS) and *Higher Order Ambisonics* (HOA). The introduction of these techniques is accompanied by a paradigm shift from *channel-based* to *object-based* transmission and storage of spatial audio features. The separate coding of source signal and source location is not only mandatory with respect to the high number of sometimes several hundred reproduction channels used for large loudspeaker

M. Detyniecki, U. Leiner, and A. Nürnberger (Eds.): AMR 2008, LNCS 5811, pp. 1–17, 2010.

arrays for WFS or HOA, it will also be the basis for interactive installations in which the user has access to the spatial properties of the reproduced soundfield and is able to adapt it to his individual requirements or aesthetic preferences.

This contribution discusses the technological change from stereophonic to advanced audio reproduction techniques, highlights the need for an object-based description of audio scenes and discusses formats for the coding of spatial properties related to auditory objects.

2 Channel-Based Audio Reproduction

The *stereophonic* approach to transmission and storage of spatial audio implies a multichannel reproduction system, from traditional two-channel stereophony to modern configurations with five, seven or even more loudspeakers, as they are used for cinema, home theater or – more recently – also for pure audio content. Based on an auditory illusion, the so-called *phantom source*, stereophony spans a panorama between pairs of loudspeakers, on which sound sources can appear. These virtual locations are hard-coded as signal relations between the two channels feeding the respective pair of loudspeakers. It can thus be considered a channel-based approach to spatial coding and transmission, as opposed to more recent, object-based approaches, where spatial information and audio signals are transmitted independently.

2.1 Psychoacoustics of Stereophony

Phantom sources emerge when pairs of loudspeakers produce largely identical signals exhibiting only a small time lag or a level difference between them. The listener will then perceive a virtual sound source located on the loudspeakers basis. The perceived location is determined by the time lag, by the level difference, or by a combination of both effects, whereby the effect of a level difference of 1 dB is approximately equivalent to a time lag of 60 μs (Fig. 1). The intended location will, however, only appear for a listener on the symmetry axis of the loudspeaker pair. A configuration with loudspeakers and listener on the corners of an equilateral triangle, yielding an aperture angle of 60°, is generally considered as ideal. Any deviation of this equidistant listener location will introduce additional time differences and thus offset the correct source positions as they are indicated in Fig. 1 (right).

It should be noted that the emergence of a phantom source is an auditory illusion generated by an artificial soundfield. In natural acoustic environments, nearly identical sound signals arriving from different directions of incidence do not exist. In this artificial situation, the auditory system obviously tries to find a natural model matching the perceived sensory information as closely as possible, hence suggesting a source location which would yield the same interaural time and/or level differences, which are known to be the most important cues for sound localization. The ear signals of a frontal phantom source and a frontal real sound source are, however, significantly different. Considering the four transfer

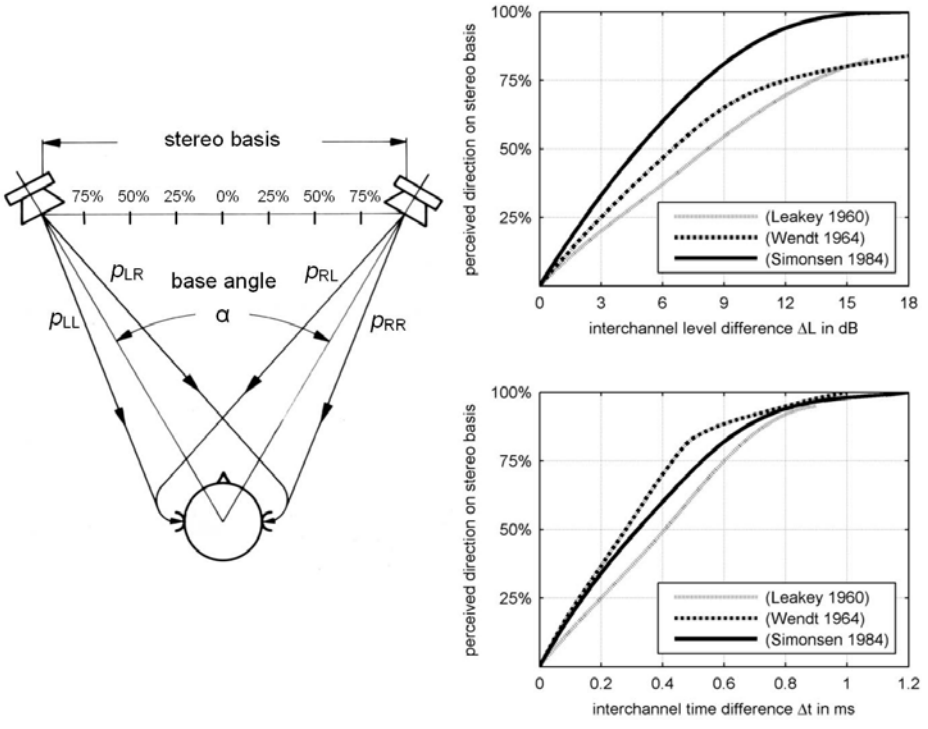

Fig. 1. *Left:* Loudspeaker configuration for two-channel stereophony with loudspeaker-ear transmission paths p_{XX}. The base angle α is usually chosen to be 60°. *Right:* Localisation of phantom sources on the loudspeaker basis against time and level differences between stereophonic signals. Values have been determined in listening tests with speech stimuli (Leakey [1], Simonsen, cited in [2]) and switched clicks (Wendt [3]).

paths from left/right loudspeaker to left/right ear ($p_{LL}, p_{LR}, p_{RL}, p_{RR}$, see Fig. 1, left) and the different source locations (±30° vs. 0° for a stereophonic vs. a real source), a considerable spectral difference can be expected, due to the comb filter caused by two consecutive signals at both ears (p_{LL} and p_{RL} for the left ear) and due to different effective *Head-Related Transfer Functions* (HRTFs) related to different angles of incidence. However, the perceived timbral distortion is much smaller than can be expected from a spectral analysis. A convincing explanation for this effect has still to be given. Theile has suggested that our auditory system might first determine a sorce location and then form a timbral impression after having compensated for the respective HRTF ([4], discussed by [5]), although there is no neurophysiological evidence for this hypothesis yet.

2.2 History and Formats

The formation of a fused auditory event generated by two signals of adjacent loudspeakers is the basis for spatial audio reproduction by all stereophonic

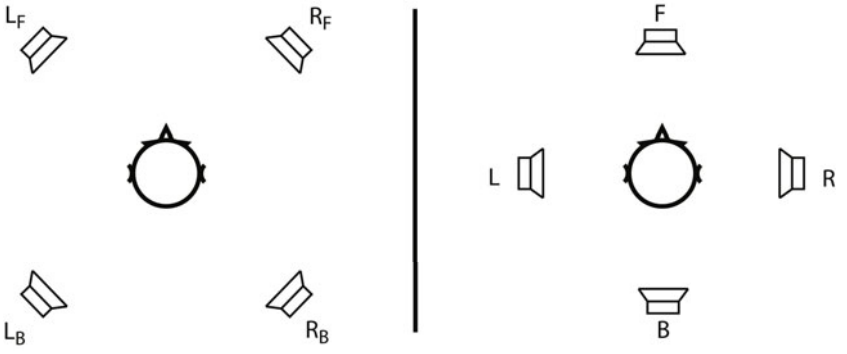

Fig. 2. Quadrophonic loudspeaker configurations: Scheiber array (left) and Dynaquad array (right)

reproduction systems. These include the classical two-channel stereophony which was studied at EMI already around 1930 [6], realised on two-channel magnetic tape from 1943 on in Germany [7], and distributed on stereo disc from 1958 on.

Quadrophonic loudspeaker configurations (Fig. 2) were first used in an experimental environment, for electronic and electroacoustic compositions such as *Symphonie pour un homme seul* (Pierre Schaeffer, 1951) or *Gesang der Jünglinge* (Karl-Heinz Stockhausen, 1956). The music industry's attempt to establish quadrophony as a successor to stereophony between 1971 and 1978 ultimately failed, due to the incompatibility of different technical solutions for 4-2-4 matrix systems which used conventional two-channel vinyl discs with additional encoding and decoding [8] offered by different manufacturers, as well as due to elementary psychoacoustic restrictions. No stable phantom sources can be established between the lateral pairs of loudspeakers [9], hence, the original promise of making the whole area of sound incidence around the listener accessible could not be fulfilled. In addition, while the sound source locations encoded in two-channel stereophony can be perceived largely correctly as long as the listener has an equal distance to the loudspeakers, a symmetrical listener position within all loudspeaker pairs of a quadrophonic loudspeaker array is only given in one central *sweet spot*.

In multichannel sound for cinema, where the spatial transmission of speech, music and environmental sounds has to be correct for a large audience area, these restrictions have been considered ever since the introduction of *Dolby Stereo* in 1976. The basic loudspeaker configuration for mix and reproduction has not changed since then (see Fig. 3). The center speaker (C) has always been used for dialog, thus providing a consistent central sound perception for the whole audience. The frontal stereo pair (L, R) is primarily used for music and sound effects. The surround speakers (S), distributed all over the walls around the audience area, are fed by a mono signal in the original *Dolby Stereo* format, while they are fed with two signals (*Left Surround, Right Surround*) or even three signals, including an additional *Back Surround*, in modern digital formats (Dolby Digital, DTS, SDDS).

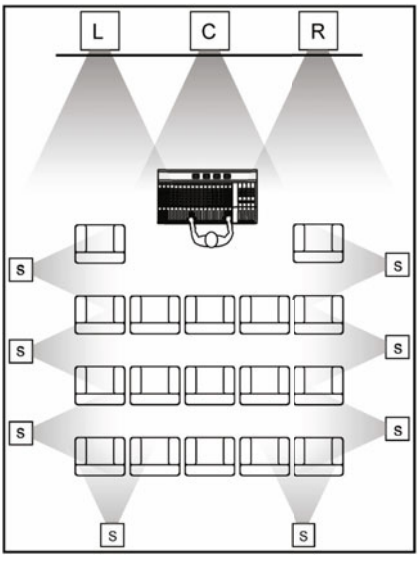

Fig. 3. Loudspeaker configuration for production and reproduction of 4.0, 5.1, 6.1 sound formats in cinema

3 Advanced Sound Spatialization Techniques

Stereophonic reproduction techniques are currently widespread in application areas like home entertainment and cinemas. However, these techniques exhibit a number of drawbacks, like e.g. the sweet spot, that have led to the development of advanced audio reproduction techniques. This section will give a brief overview on the major techniques, their physical background and their properties. Please refer to the cited literature for a more in-depth discussion of the particular methods.

3.1 Binaural Reproduction

Binaural reproduction aims at recreating the acoustic signals at the ears of a listener such that these are equal to a recorded or synthesized audio scene. Appropriately driven headphones are typically used for this purpose. The term *binaural reproduction* refers to various techniques following this basic concept. Audio reproduction using *Head-Related Transfer Functions* (HRTFs) is of special interest in the context of this contribution. The ability of the human auditory system to localize sound is based on exploiting the acoustical properties of the human body, especially the head and the outer ears [10]. These acoustical properties can be captured by HRTFs by measuring the transfer function between a sound source and the listener's ears. Potentially these measurements have to be undertaken for all possible listener positions and head poses. However, it is typically assumed that the listeners position is stationary. HRTF-based reproduction is then implemented by filtering a desired source signal with the appropriate HRTF

for the left and right ear. In order to cope for varying head orientations, head tracking has to be applied. Head-tracked binaural reproduction is also refereed to as dynamic binaural resynthesis [11].

Binaural reproduction has two major drawbacks: (1) it may not always be desired to wear headphones and (2) reproduction for large audiences or moving listeners is technically complex. The first drawback can be overcome, within limits, by using loudspeakers for binaural reproduction. In this case, appropriate crosstalk cancelation has to be employed in order that the signals at both ears of the listener can be controlled independently [12]. Such crosstalk cancelation typically exhibits a very pronounced sweet spot. Alternatives to binaural reproduction for potentially larger audiences will be introduced in the following. The first two are based on the physical reconstruction of a desired sound field within a given listening area.

3.2 Higher Order Ambisonics

Higher Order Ambisonics (HOA) and related techniques [13,14,15,16] base on the concept of single-layer potentials and theoretically provide a physically exact solution to the problem of sound field reproduction. The underlying theory assumes that a continuous single layer potential (secondary source distribution) surrounds the listening area. Appropriate weights (driving functions) applied to the secondary source distribution allow to reproduce almost any desired sound field within the listening area. Although arbitrary secondary source contours which enclose the receiver area are theoretically possible, explicit solutions are currently exclusively available for spherical and circular geometries.

The continuous distribution of secondary source is approximated in practice by a spatially discrete distribution of loudspeakers. This constitutes a spatial sampling process which may lead to spatial aliasing. The artifacts due to spatial sampling result in a pronounced artifact-free area in the center of the loudspeaker arrangement [17] that HOA and related techniques exhibit. The size of this area decreases with increasing frequency of the signal to be reproduced. For feasible loudspeaker setups the size of the artifact-free reproduction area is typically smaller than a human head at the upper end of the audible frequency range. Outside, spatial sampling artifacts arise that may be perceived as coloration of the desired sound field [18]. A number of HOA systems have been realized at various research institutes and other venues.

3.3 Wave Field Synthesis

Like HOA, *Wave Field Synthesis* (WFS) aims at physically recreating a desired sound field within a given listening area. However, the theoretical background of WFS differs in comparison to HOA [17]. WFS is based on the quantitative formulation of the *Huygens-Fresnel-Principle*, which states that a propagating wave front can be synthesized by a superposition of simple sources placed on the wave front.

WFS has initially been developed for linear secondary source distributions [19], where *Rayleigh integrals* describe the underlying physics. No explicit solution of

the reproduction problem is required in order to derive the secondary source driving function. The initial concept of WFS has been extended to arbitrary convex secondary source layouts [20] which may even only partly enclose the listening area. As for HOA, the continuous secondary source distribution is approximated by spatially discrete loudspeakers in practice. The resulting spatial sampling artifacts for typical geometries differ considerably from HOA [17]. WFS exhibits no pronounced sweet spot, the sampling artifacts are rather evenly distributed over the receiver area. The sampling artifacts may be perceived as coloration of the sound field [21].

The loudspeaker driving signals for WFS can be computed very efficiently by weighting and delaying the virtual source signals for the reproduction of virtual point sources and plane waves.

3.4 Numerical Methods

Besides HOA and WFS, several numerical methods of sound field reproduction [22,23,24] exist, the properties of which are typically somewhere between HOA and WFS. The advantage of these numerical methods are very flexible loudspeaker layouts. The drawback is the fact that they are numerically complex compared to the analytical solutions given by HOA and WFS, and that they do not provide such a high degree of flexibility.

3.5 Generalized Panning Techniques

Currently the most widely used method for creating a spatial sound impression is still based on the stereophonic approach described in Sect. 2. Panning techniques

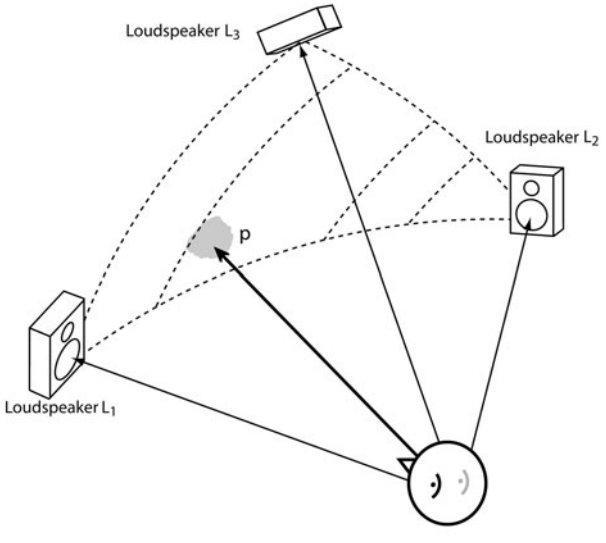

Fig. 4. Three-dimensional amplitude panning with amplitude factors calculated by projecting the target direction on a vector base spanned by loudspeaker triplets L_1, L_2, L_3

exploit the amplitude and delay differences between a low number of loudspeaker signals to create the illusion of a phantom source. For reproduction in a two-dimensional plane (e.g. 5.1 systems) normally a pair, for three-dimensional setups a triple or quadruple of loudspeakers is used. *Vector Base Amplitude Panning* (VBAP) [25] can be regarded as a generalization of amplitude panning. In the three-dimensional case, phantom sources can be placed on triangles spanned by the closest three loudspeakers (see Fig. 4).

Panning techniques allow flexible loudspeaker layouts but have a pronounced sweet spot. Outside this sweet spot, the spatial impression is heavily distorted and also some impairment in terms of sound color may occur.

4 Object-Based Rendering of Audio Scenes

Depending on how the loudspeaker driving signals are computed, two rendering techniques can be differentiated for most of the introduced audio reproduction approaches. On the one hand this is *data-based rendering* and on the other hand *model-based rendering*. In data-based rendering the entire audio scene is captured by a suitable number of microphones. Post-processing (encoding) is applied to the microphone signals if required for the particular technique under consideration. Typically circular or spherical microphone arrays are used for these approaches. The encoded signals are then stored or transmitted and the loudspeaker driving signals are computed from them. The benefit of this approach is that arbitrarily complex sound fields can be captured, the drawbacks are the technical complexity, the required storage and transmission capacity and the limited flexibility in post-processing.

Model-based rendering is based on the concept of using a parameterizable spatio-temporal model of a virtual source. Point sources and plane waves are the most frequent models used here, however, more complex models and superpositions of these simple sources are also possible. The signal of the virtual source together with its parameters is stored or transmitted. The receiver takes care that the desired audio scene is rendered appropriately for a given reproduction system. The benefits of this approach are the flexibility with respect to the reproduction system, the possibility for post-processing of the audio scene by modifying the parameters of the virtual source and the lower requirements for storage capacities for rather simple audio scenes. A drawback is that ambient sounds, like applause, are hard to model. Model-based rendering is termed as object-based audio since the sources in the scene are regarded as objects.

In practice, a combination of both approaches is used. The encoded microphone signals can be interpreted as a special source in model-based rendering and hence can be included in the scene description. Most of the discussed systems in Section 3 support both data-based and model-based rendering and also hybrid methods.

5 Formats

Along with the trend from channel-based to object-based audio reproduction there comes the necessity to store audio and metadata. Instead of storing a collection of audio tracks representing the loudspeaker signals, a so-called *audio scene* is created. Such an audio scene consists of sounding objects which are associated with input (source) signals, a certain position or trajectory in the virtual space and other source parameters. In addition to source objects there can also be objects describing the acoustical properties of a virtual room.

A wide variety of object-based audio reproduction systems is already in existence, both academic prototypes and commercially available implementations. Most of them use non-standard storage formats which are tailored to a single setup and in many cases contain implementation-specific data. Some systems are based on a specific digital audio workstation software – often using custom-made plugins – and use its native storage format including track envelope data for the dynamic control of source parameters. Although this may work very well on one single system, it is in most cases impossible to share audio scenes between different reproduction systems without major customization on the target system. It is crucial for a audio reproduction system to have content available, therefore it is desirable to establish a common storage format for audio scenes to be exchanged between different venues. Even if some fine-tuning is still necessary to adapt a scene to the acoustical conditions on location, this would very much facilitate the process.

There are two different paradigms for storing audio scenes, (1) to create a single file or stream which contains both audio data and scene data, or (2) to create one file for the description of the scene which contains links to audio data stored in separate audio files or streams. An advantage of the former method is its compactness and the possibility to transmit all data in one single stream. The latter method allows more flexibility as audio data can be edited separately or can be completely exchanged and several versions of an audio scene can be created with the same audio data. Another important aspect of a scene format is if it is stored in a binary file or in a text file. Binary files are typically smaller and their processing and transmission is more efficient. They are also the only feasible option if audio data and scene data are combined into one file. Text files have the advantage that they can be opened, inspected and edited easily with a simple text editor. Text-based formats can normally be extended more easily than binary formats. There are several markup languages which can be used to store data in a text file in a structured manner. One of the most widespread is the *eXtensible Markup Language* (XML). Many tools and software libraries to read, manipulate and store XML data are available.

Most of todays high resolution spatial audio reproduction systems have a means of controlling the audio scene parameters in realtime. This can be done by sending messages to the reproduction software, for example via network sockets. These messages can be collected, tagged with timestamps and written to a file. This way the realtime control format is also used as a storage format.

This paradigm is for example used in the *Spatial Sound Description Interchange Format* (SpatDIF) [26]. Because it is just an unstructured stream of messages, it is hard to make meaningful changes later.

In the following sections a few standardized formats are presented which could be suitable as an exchange format for spatial audio scenes. Thereafter, in Sect. 5.4, the *Audio Scene Description Format* (ASDF) is presented which is still in development and which tries to address the mentioned shortcomings of the other formats.

5.1 VRML/X3D

The *Virtual Reality Modeling Language* (VRML) is a format for three-dimensional computer graphics mainly developed for displaying and sharing of 3D models on the internet. Its scene description is based on a single scene graph, which is a hierarchical tree-like representation of all scene components. Geometrical objects are placed in local coordinate systems which can be translated/scaled/rotated and also grouped and placed in other coordinate systems and so on. Light sources, camera views and also audio objects have to be added to the same scene graph. To add an audio object to the scene graph, a `Sound` node has to be used. This node contains an `AudioClip` node which holds the information about the audio file or network stream to be presented. The format of the actual audio data is not specified by the standard. All elements of the scene graph can be animated with the so-called `ROUTE` element. This, however, is quite cumbersome for complex animations, therefore in most cases the built-in ECMAScript/JavaScript interpreter is used. To enable user interaction, mouse-events can be defined and can be bound to any visual element in the scene graph.

The use of a scene graph to represent a three-dimensional scene is very widespread in computer graphics applications. It is possible to combine very simple objects – mostly polygons – to more complex shapes and then combine those again and again to create high level objects. When transforming such a high level object, the transformation is automatically applied to all its components. In pure audio scenes, sounding objects normally consist of only one or a few parts and an entire scene often contains only a handful of sources. Using a scene graph in such a case would make the scene description overly complicated. The far worse disadvantage, however, is the distribution of the timing information. The timing of sound file playback is contained in the respective `Sound` node, the timing information of animations is spread over `ROUTE`s, interpolators and scripts. This makes it essentially impossible to edit the timing of a scene directly in the scene file with a text editor.

The VRML became an ISO standard in 1997 with its version 2.0, also known as VRML97. It has been superseded by *eXtensible 3D* (X3D) [27], which is an ISO standard since 2004. X3D consists of three different representations: the classic VRML syntax, a new XML syntax and a compressed binary format for efficient storage and transmission.

5.2 MPEG-4 Systems/AudioBIFS

The ISO standard MPEG-4 contains the *BInary Format for Scenes* (BIFS) which incorporates the VRML97 standard in its entirety and extends it with the ability to stream scene metadata together with audio data. The used audio codecs are also defined in the MPEG standard. The spatial audio capabilities – referred to as *(Advanced)AudioBIFS* [28] – were also extended by many new nodes and parameters.

Among the new features is the `AcousticMaterial` node, which defines acoustical properties like reflectivity (`reffunc`) and transmission (`transfunc`) of surfaces, the `AudioFX` node to specify filter effects in the *Structured Audio Orchestra Language* (SAOL) and the ability to specify virtual acoustics in both a physical and a perceptual approach. For the latter, the `PerceptualParameters` node with parameters like `sourcePresence` and `envelopment` can be used. Another new feature is the `DirectiveSound` node, used to specify source directivity.

AudioBIFS is a binary format which is designed to be streamed over a network. As a tool for easier creation and editing of scenes there is also a text-based representation, the *Extensible MPEG-4 Textual Format* (XMT). It comes in two variants: XMT-A has a syntax very similar to X3D (see Sect. 5.1), XMT-Ω is modeled after SMIL (see Sect. 5.3). However, the XMT is not a presentation language on its own, it has always to be converted to the binary format before it can be transmitted or played back.

AudioBIFS as part of MPEG-4 Systems became an ISO standard in 1999, but has evolved since. In its most recent update – AudioBIFS v3 [29] – several features were added, among them the `WideSound` node for source models with given shapes and the `SurroundingSound` node with the `AudioChannelConfig` attribute which allows to include Ambisonics signals and binaural signals into the scene.

AudioBIFS would definitely have all the features necessary to store spatial audio scenes. However, because of the huge size and complexity of the standard, it is very hard to implement an en- and decoder. No complete library implementation of MPEG-4 Systems is available.

5.3 SMIL

Contrary to the aforementioned formats, the XML-based *Synchronized Multimedia Integration Language* (SMIL, pronounced like "smile") is not able to represent three-dimensional content. Its purpose is the temporal control and synchronization of audio, video, images and text elements and their arrangement on a 2D screen. The SMIL is a recommendation of the *World Wide Web Consortium* since 1998, the current version (SMIL 3.0) was released in 2008 [30].

All SMIL functionality is organized in modules, for example *MediaDescription*, *PrefetchControl* and *SplineAnimation*. Different sets of modules are combined to language profiles tailored for different applications and platforms. With the *3GPP SMIL Language Profile* the SMIL is used for *Multimedia Messaging Service* (MMS) on mobile phones.

The central part of a SMIL document is a timeline where media objects can be placed either relative to other objects or by specifying absolute time values. The timing does not have to be static, interactive presentations can be created where the user dictates the course of events e.g. by mouse clicks. Animations along 2D-paths are possible with the `animateMotion` element. The temporal structure is mainly defined by `<seq>`-containers ("sequence"), whose content elements are played consecutively one at a time, and by `<par>`-containers ("parallel"), whose content elements start all at the same time. Of course, these containers can be arbitrarily nested giving possibilities ranging from simple slide shows to very complex interactive mega-presentations. Inside of the time containers, media files are linked to the SMIL file with `<img>`, `<audio>`, `<text>` and similar elements.

SMIL has very limited audio capabilities. Except for the temporal placement, the only controllable parameter of audio objects is the sound level, given as a percentage of the original volume. The SMIL format itself is especially not able to represent 3D audio scenes, but it can either be used as an extension to another XML-based format or it can be extended itself. To extend another XML-based format with SMIL timing features, the W3C recommendation *SMIL Animation* [31] can be utilized. This was done, for example, in the wide-spread *Scalable Vector Graphics* (SVG) format. However, *SMIL Animation* is quite limited because a "flat" timing model without the powerful time containers (like `<par>` and `<seq>`) is used. A more promising approach would be to extend the SMIL with 3D audio features. An example for such an extension is given in [32], where the SMIL was extended with the so-called *Advanced Audio Markup Language* (AAML).

5.4 ASDF

The *Audio Scene Description Format* (ASDF) [33] is an XML-based format which – contrary to the aforementioned formats – has a focus on pure audio scenes. It is still in an early development state, but basic functionality is already available in the *SoundScape Renderer* (SSR) [34].

The ASDF aims at being both an authoring and a storage format at the same time. In absence of a dedicated editing application, scene authors should still be able to create and edit audio scenes with their favorite text editor. The ASDF does not try to cover every single imaginable use case (like MPEG-4 does), but just follows the development of current audio reproduction techniques (see Sect. 3) and intends to provide a lowest-common-denominator description of scenes to be rendered by these techniques. To ensure the smooth exchange of scenes between different systems, it is independent of the rendering algorithm and contains no implementation-specific or platform-specific data. Requirements and implementation issues are discussed within the scientific community and with partners from the industry. The goal is to collaboratively develop an open format which is easy to implement even with limited resources. A reference implementation will be provided in form of a software library.

Although the ASDF is capable of representing three-dimensional audio scenes, there is also a simplified syntax available to describe two-dimensional scenes in a

horizontal plane. The ASDF does not have a hierarchical scene graph, it is rather using SMIL's timeline and time container concept (see Sect. 5.3). In its easiest form, the ASDF is used to represent static scenes, where source positions and other parameters do not change. If source movement is desired, trajectories can be assigned to sources or groups of sources.

In its current draft proposal, the ASDF is a stand-alone format, but it is planned to become an extension to SMIL. This way, a SMIL library can be used for media management and synchronisation and only the spatialization aspects have to be newly implemented. As a positive side effect of using SMIL, videos, images and texts can be easily synchronized with the audio scene and displayed on a screen.

When extending SMIL, it is important to separate the 3D audio description from SMIL's 2D display layout. Adding 3D audio features is not just a matter of adding a third dimension to the available 2D elements like `<layout>` and `<region>` (as done in [32]), because 2D has a different meaning in screen presentations and in spatial audio presentations. In the former case, the most natural choice for a 2D plane is of course the plane where the screen is part of. In the latter case, however, it makes much more sense to choose the horizontal plane as a 2D layout. If reproduction systems for spatial audio are limited to two dimensions, it will be in the most – if not all – cases the horizontal plane.

6 Applications

Most of the existing technical solutions for high resolution spatial audio reproduction are limited to one specific reproduction method. The required number and geometrical setup of loudspeakers differ considerably for different reproduction methods on the one hand, the digital signal processing for creation of the loudspeaker signals on the other.

In order to investigate the potential of object-based audio, the *SoundScape Renderer* (SSR) [34] has been implemented. The SSR supports a wide range of reproduction methods including binaural reproduction (Sect. 3.1), HOA (Sect. 3.2), WFS (Sect. 3.3) and VBAP (Sect. 3.5). It uses the ASDF as system independent scene description. The implementation allows for a direct comparison of different sound reproduction methods.

As one common interface is provided for different rendering backends, *system-independent mixing* can be performed, which means that a spatial audio scene can be created in one location (e.g. a studio with an 8-channel VBAP system) and performed in another venue (e.g. a concert hall with a several-hundred-channel WFS system). Figure 5 shows two screenshots of the graphical user interface of the SSR where the same scene is rendered with two different reproduction setups. With even less hardware requirements, the scene could also be created using binaural rendering and a single pair of headphones [35]. In any case, final adjustments may be necessary to adapt to the reproduction system and the room acoustics of the target venue.

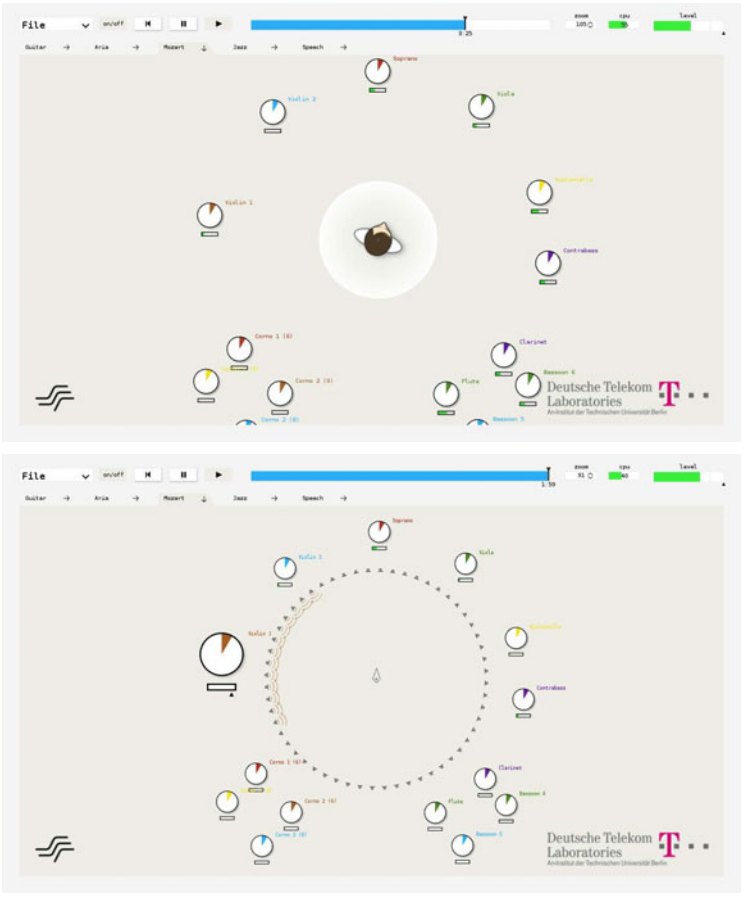

Fig. 5. The graphical user interface of the *SoundScape Renderer* reproducing the same scene with binaural rendering (top) and WFS (bottom)

7 Conclusions

Traditionally, techniques for capture, transmission and reproduction of spatial audio have been centered around a channel-based paradigm. Technically and practically this was feasible due to the low number of required channels and the restricted geometrical layout of the reproduction system. In practice, typical audio scenes contain more audio objects than reproduction channels which makes the channel-based approach also quite efficient in terms of coding and transmission.

Advanced loudspeaker based audio reproduction systems use a high number of channels for high-resolution spatial audio reproduction. It is almost impossible to predefine a geometrical setup for widespread deployment of such techniques. Here object-based audio scene representations are required that decouple the

scene to be reproduced from the geometrical setup of the loudspeakers. The rendering engine at the local terminal has to generate suitable loudspeaker signals from the source signals and the transmitted scene description.

Object-based audio also offers a number of benefits. Two major ones are: (1) efficient manipulation of a given audio scene, and (2) independence of the scene description from the actual reproduction system. In traditional channel-based audio, all audio objects that compose the scene are down-mixed to the reproduction channels. This makes it almost impossible to modify the scene when only having access to the down-mixed channels. The object-based approach makes such modifications much easier.

An variety of reproduction systems are currently being used or proposed for the future. Most likely none of these systems will be the sole winner in the future. Currently most audio content is either produced only for one reproduction system or has been produced separately for more than one. This situation leads to a number of proposals for the automatic up- and down-mixing of stereophonic signals. An object-based mixing approach will provide a number of improvements since it allows a very flexible rendering of audio scenes with respect to a given reproduction system.

A number of proposed formats for the object-based representation of audio(-visual) scenes exist. One of the most powerful formats, MPEG-4, is technically very demanding. This has lead to a number of other formats that are specialized for a specific application area. The ASDF focuses explicitly on the easy exchange of audio scenes.

In order to illustrate and investigate the benefits of object-based audio, the *SoundScape Renderer* has been developed. It fully separates the scene description from the audio reproduction approach. The SSR generates the loudspeakers signals for a variety of audio reproduction approaches in real-time from the scene description. Furthermore, real-time interaction with the audio scene is possible. Traditional channel-based approaches can not provide this degree of flexibility.

References

1. Leakey, D.: Further thoughts on stereophonic sound systems. Wireless World 66, 154–160 (1960)
2. Williams, M.: Unified theory of microphone systems for stereophonic sound recording. In: 82nd Convention of the Audio Engineering Society (March 1987)
3. Wendt, K.: Das Richtungshören bei Zweikanal-Stereophonie. Rundfunktechnische Mitteilungen 8(3), 171–179 (1964)
4. Theile, G.: Zur Theorie der optimalen Wiedergabe von stereophonen Signalen über Lautsprecher und Kopfhörer. Rundfunktechnische Mitteilungen 25, 155–169 (1981)
5. Gernemann-Paulsen, A., Neubarth, K., Schmidt, L., Seifert, U.: Zu den Stufen im "Assoziationsmodell". In: 24. Tonmeistertagung (2007)
6. Blumlein, A.: Improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems. British Patent Specification 394325 (1931)
7. Thiele, H.H.K. (ed.): 50 Jahre Stereo-Magnetbandtechnik. Die Entwicklung der Audio Technologie in Berlin und den USA von den Anfängen bis 1943. Audio Engineering Society (1993)

8. Woodward, J.: Quadraphony–A Review. Journal of the Audio Engineering Society 25(10/11), 843–854 (1977)
9. Theile, G., Plenge, G.: Localization of lateral phantom sources. Journal of the Audio Engineering Society 25, 196–200 (1977)
10. Blauert, J.: Spatial Hearing: The Psychophysics of Human Sound Localization. MIT Press, Cambridge (1996)
11. Lindau, A., Hohn, T., Weinzierl, S.: Binaural resynthesis for comparative studies of acoustical environments. In: 122nd Convention of the Audio Engineering Society (May 2007)
12. Møller, H.: Reproduction of artificial-head recordings through loudspeakers. Journal of the Audio Engineering Society 37, 30–33 (1989)
13. Daniel, J.: Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia. PhD thesis, Université Paris 6 (2000)
14. Poletti, M.: Three-dimensional surround sound systems based on spherical harmonics. Journal of the Audio Engineering Society 53(11), 1004–1025 (2005)
15. Ahrens, J., Spors, S.: An analytical approach to sound field reproduction using circular and spherical loudspeaker distributions. Acta Acoustica united with Acoustica 94(6), 988–999 (2008)
16. Fazi, F., Nelson, P., Christensen, J., Seo, J.: Surround system based on three dimensional sound field reconstruction. In: 125th Convention of the Audio Engineering Society (2008)
17. Spors, S., Ahrens, J.: A comparison of Wave Field Synthesis and Higher-Order Ambisonics with respect to physical properties and spatial sampling. In: 125th Convention of the Audio Engineering Society (October 2008)
18. Ahrens, J., Spors, S.: Alterations of the temporal spectrum in high-resolution sound field reproduction of different spatial bandwidths. In: 126th Convention of the Audio Engineering Society (May 2009)
19. Berkhout, A.: A holographic approach to acoustic control. Journal of the Audio Engineering Society 36, 977–995 (1988)
20. Spors, S., Rabenstein, R., Ahrens, J.: The theory of Wave Field Synthesis revisited. In: 124th Convention of the Audio Engineering Society (May 2008)
21. Wittek, H.: Perceptual differences between Wavefield Synthesis and Stereophony. PhD thesis, University of Surrey (2007)
22. Kirkeby, O., Nelson, P.: Reproduction of plane wave sound fields. Journal of the Acoustic Society of America 94(5), 2992–3000 (1993)
23. Ward, D., Abhayapala, T.: Reproduction of a plane-wave sound field using an array of loudspeakers. IEEE Transactions on Speech and Audio Processing 9(6), 697–707 (2001)
24. Hannemann, J., Leedy, C., Donohue, K., Spors, S., Raake, A.: A comparative study of perceptual quality between Wavefield Synthesis and multipole-matched rendering for spatial audio. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (April 2008)
25. Pulkki, V.: Virtual sound source positioning using Vector Base Amplitude Panning. Journal of the Audio Engineering Society 45(6), 456–466 (1997)
26. Peters, N.: Proposing SpatDIF – the spatial sound description interchange format. In: International Computer Music Conference (August 2008)
27. Web3D Consortium: eXtensible 3D, X3D (2004), `http://www.web3d.org/x3d/`
28. Väänänen, R., Huopaniemi, J.: Advanced AudioBIFS: Virtual acoustics modeling in MPEG-4 scene description. IEEE Transactions on Multimedia 6(5), 661–675 (2004)

29. Schmidt, J., Schröder, E.F.: New and advanced features for audio presentation in the MPEG-4 standard. In: 116th Convention of the Audio Engineering Society (May 2004)
30. World Wide Web Consortium: Synchronized Multimedia Integration Language, SMIL 3.0 (2008), http://www.w3.org/TR/SMIL3/
31. World Wide Web Consortium: SMIL Animation (2001), http://www.w3.org/TR/smil-animation/
32. Pihkala, K., Lokki, T.: Extending SMIL with 3D audio. In: International Conference on Auditory Display (July 2003)
33. Geier, M., Spors, S.: ASDF: Audio Scene Description Format. In: International Computer Music Conference (August 2008)
34. Geier, M., Ahrens, J., Spors, S.: The SoundScape Renderer: A unified spatial audio reproduction framework for arbitrary rendering methods. In: 124th Convention of the Audio Engineering Society (May 2008)
35. Geier, M., Ahrens, J., Spors, S.: Binaural monitoring of massive multichannel sound reproduction systems using model-based rendering. In: NAG/DAGA International Conference on Acoustics (March 2009)