MULTI-CHANNEL GOES MOBILE: MPEG SURROUND BINAURAL RENDERING

JEROEN BREEBAART[1], JÜRGEN HERRE[2], LARS VILLEMOES[3], CRAIG JIN[4], KRISTOFER KJÖRLING[3], JAN PLOGSTIES[2] AND JEROEN KOPPENS[5]

[1] *Philips Research Laboratories, 5656 AA, Eindhoven, The Netherlands*
jeroen.breebaart@philips.com
[2] *Fraunhofer Institute for Integrated Circuits IIS, 91058 Erlangen, Germany*
{hrr;pts}@iis.fraunhofer.de
[3] *Coding Technologies, 11352 Stockholm, Sweden*
{lv;kk}@codingtechnologies.com
[4] *Vast Audio, NSW 1430 Sydney, Australia*
craig@ee.usyd.edu.au
[5] *Philips Applied Technologies, 5616 LW Eindhoven, The Netherlands*
jeroen.koppens@philips.com

Surround sound is on the verge of broad adoption in consumers' homes, for digital broadcasting and even for Internet services. The currently developed MPEG Surround technology offers bitrate efficient and mono/stereo compatible transmission of high-quality multi-channel audio. This enables multi-channel services for applications where mono or stereo backwards compatibility is required as well as applications with severely bandwidth limited distribution channels. This paper outlines a significant addition to the MPEG Surround specification which enables computationally efficient decoding of MPEG Surround data into binaural stereo as is appropriate for appealing surround sound reproduction on mobile devices, such as cellular phones. The publication describes the basics of the underlying MPEG Surround architecture, the binaural decoding process, and subjective testing results.

INTRODUCTION

Approximately half a century after the broad availability of two-channel stereophony, multi-channel sound is finally on its way into consumer's homes as the next step towards higher spatial reproduction quality. While the majority of multi-channel audio consumption is still in the context of movie sound, consumer media for high-quality multi-channel audio (such as SACD and DVD-Audio) now respond to the demand for a compelling surround experience also for the audio-only market. Many existing distribution channels will be upgraded to multi-channel capability over the coming years if two key requirements can be met: a) As with the previous transition from mono to stereophonic transmission, the plethora of existing (now stereo) users must continue to receive high-quality service, and b) For digital distribution channels with substantially limited channel capacity (e.g. digital audio broadcasting), the introduction of multi-channel sound must not come at a significant price in terms of additional data rate required.

This paper reports on the forthcoming MPEG Surround specification that offers an efficient representation of high quality multi-channel audio at bitrates that are only slightly higher than common rates currently used for coding of mono / stereo sound. Due to the underlying principle the format is also completely backward compatible with legacy (mono or stereo) decoders and is thus ideally suited to introduce multi-channel sound into existing stereophonic or monophonic media and services. Specifically, the paper reports on a recent addition to the MPEG Surround specification which complements the original loudspeaker-oriented multi-channel decoding procedure by modes that enables a compelling surround sound reproduction on mobile devices. Using these *binaural rendering* modes, multi-channel audio can be rendered into a realistic virtual sound experience on a wide range of existing mobile devices, such as mp3 players and cellular phones.

Due to the nature of the task, MPEG Surround binaural rendering represents an intimate merge between binaural technologies, as they are known from virtual sound displays [1][2], and the parametric multi-channel audio coding that is at the heart of the MPEG Surround scheme. Thus, the paper is structured as follows: Sections 1 and 2 describe the basic concepts that classic binaural technology and the MPEG Surround specification are based on respectively. Then Section 3 introduces the new MPEG Surround binaural rendering modes as a synthesis between both worlds and characterizes it in terms of both sonic performance and implementation complexity. Finally, a number of

interesting applications for MPEG Surround binaural rendering are discussed.

1 BINAURAL RENDERING

Spatial hearing relies to a great extent on binaural cues like time-, level- and spectral differences between the left and right ear signals [3]. These cues are contained in the acoustic transfer function from a point in space to the ear canal of a listener, called a head-related transfer function (HRTF). HRTFs are measured under anechoic conditions on human or artificial heads with small microphones in the ear canal. They are strongly dependent on direction, but also on the head and ear shape [4]. If acoustic transfer functions are measured in an echoic room, i.e. in the presence of reflections and reverberation, they are referred to as binaural room transfer functions (BRTFs).

The well-known concept of *binaural rendering* makes use of the knowledge of transfer functions between sound sources and the listener's ear signals to create virtual sound sources which are placed around the listener. This is done by convolving a signal with a pair of HRTFs or BRTFs to produce ear signals as they would have resulted in a real acoustic environment. Such signals are typically reproduced via headphones. Alternatively, cross-talk cancelled loudspeakers can be used [5]. Typical applications of binaural rendering are auditory virtual displays, gaming and other immersive environments.

The input signals for binaural rendering are monophonic sounds to be spatialized. Most existing audio content is produced for loudspeaker reproduction. In order to render such material for headphone reproduction, each loudspeaker can be represented by a virtual source placed at a defined location. Each loudspeaker signal is filtered by a pair of HRTFs or BRTFs corresponding to this location. Finally, the filtered output signals for each ear are summed to form the headphone output channel.

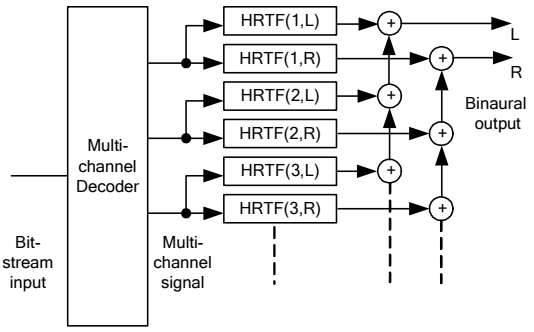

Figure 1: Decoding and binaural rendering of multi-channel signals.

In Figure 1 the straightforward process of decoding and binaural rendering of a discrete multi-channel signal is depicted. First, the audio bitstream is decoded to N channels. In the subsequent binaural rendering stage each loudspeaker signal is then rendered for reproduction via two ear signals, yielding a total number of 2xN filters. Depending on the number of channels and the length of the filters, this process can be demanding in terms of both computational complexity and memory usage.

Binaural rendering has several benefits for the user. Since the important cues for spatial hearing are conveyed, the user is able to localize sounds in direction and distance and to perceive envelopment. Sounds appear to originate somewhere outside the listener's head as opposed to the in-head localization that occurs with conventional stereo headphone reproduction. The quality of binaural rendering is mostly determined by the localization performance, front-back discrimination, externalization and perceived sound coloration.

Some studies show that there is a benefit in using individualized HRTF for binaural rendering, i.e. using the user's own HRTFs. However, careful selection of HRTFs allows good localization performance for many subjects [6]. In addition, tracking of the listener's head and updating the HRTF filtering accordingly can further reduce localization errors and create a highly realistic virtual auditory environment [7]. Another kind of interaction is the modification of the position of the virtual sources by the user. The aforementioned scenarios require that there is a defined interface such that different HRTFs/BRTFs can be applied to the binaural rendering process.

2 MPEG SURROUND TECHNOLOGY

This section provides a brief description of the MPEG Surround technology. First, the basics of MPEG Surround will be discussed. Afterwards, an introduction of the decoder structure will be given. The section will then conclude with an outline of a selection of important MPEG Surround features.

2.1 Spatial Audio Coding concept

MPEG Surround is based on a principle called Spatial Audio Coding (SAC). Spatial Audio Coding is a multi-channel compression technique that exploits the perceptual inter-channel irrelevance in multi-channel audio signals to achieve higher compression rates.

This can be captured in terms of spatial cues, i.e. parameters describing the spatial image of a multi-channel audio signal. Spatial cues typically include level/intensity differences, phase differences and measures of correlation/coherence between channels, and can be represented in an extremely compact way.

During encoding, spatial cues are extracted from the multi-channel audio signal and a downmix is generated. Typically, backwards-compatible downmix signals will be used like mono- or stereophonic signals. However, any number of channels that is smaller than that used for

the original audio can be used for the downmix. In the remainder of this section a stereophonic downmix will be assumed.
The downmix can then be compressed and transmitted without the need to update existing coders and infrastructures. The spatial cues (spatial side information) are transmitted in a low bitrate side channel, e.g. the ancillary data portion of the downmix bitstream.
For most audio productions both a stereo as well as a 5.1 multi-channel mix is produced from the original multi-track recording by an audio engineer. Naturally, the *automated* downmix produced by the spatial audio coder can differ significantly from the *artistic stereo downmix* as intended by the audio engineer. For optimal backward compatibility, this artistic downmix can be transmitted instead of the automated downmix. The difference between this artistic stereo downmix and the automated stereo downmix signal, required by the decoder for optimal multi-channel reconstruction, can be coded as part of the spatial side information stream either in a parametric fashion for low bitrate applications or as a wave-form coded difference signal.
On the decoder side, a multi-channel up-mix is created from the transmitted downmix signal and spatial side information. In this respect, the Spatial Audio Coding concept can be used as a pre- and post-processing step to upgrade existing systems. Figure 2 illustrates this for a 5.1 original with a stereo downmix.

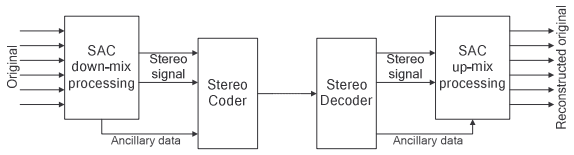

Figure 2: Typical SAC encoder/decoder chain.

Conceptually, this approach can be seen as an enhancement of several known techniques, such as an advanced method for joint stereo coding of multi-channel signals [8], a generalization of Parametric Stereo [9] [10] [11] to multi-channel application, and an extension of the Binaural Cue Coding (BCC) scheme [12] [13] towards using more than one transmitted downmix channel [14].
From a different viewpoint, the Spatial Audio Coding approach may also be considered an extension of well-known matrix surround schemes (Dolby Surround/Prologic, Logic 7, Circle Surround etc.) [15] [16] by transmission of dedicated side information to guide the multi-channel reconstruction process and thus achieve improved subjective audio quality [17].

2.2 MPEG Surround decoder

2.2.1 Decoder structure

In an MPEG Surround decoder, the decoded (i.e. PCM) downmix signal is up-mixed by a spatial synthesis process using the transmitted spatial cues. Due to the frequency dependence of the spatial side information the downmix is analyzed by a hybrid filterbank before spatial synthesis, and the multi-channel reconstruction is re-synthesized by a hybrid synthesis filterbank. This is shown in Figure 3.

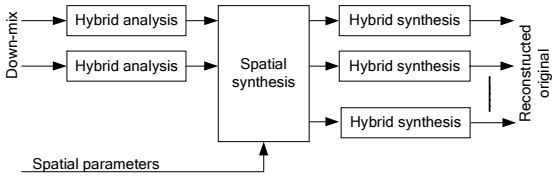

Figure 3: High level MPEG Surround decoder structure.

Spatial synthesis is applied to this time-frequency representation by matrix transformations where the matrices are calculated for defined ranges in time and frequency (*tiles*) parameterized by the spatial side information.
A more detailed overview focusing on the spatial synthesis is shown in Figure 4. In order to be able to reconstruct the inter-channel coherence in the up-mix, decorrelated signals are required. Therefore, the up-mix process is split up into three steps. A first matrix performs pre-processing and initial mixing of the downmix signals. Subsequently, some of these signals are decorrelated by independent decorrelators. Finally a second matrix mixes the decorrelated signals with the linearly pre-processed signals into a reconstruction of the original multi-channel signal.

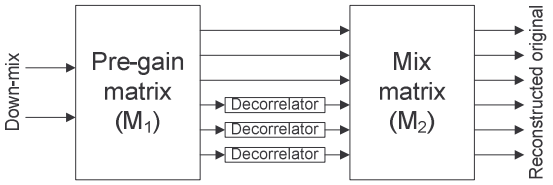

Figure 4: Spatial synthesis process.

2.2.2 Decoder concept

The underlying concept of the MPEG Surround up-mixing process is based on a tree structure consisting of conceptual up-mixing elements. Figure 5 shows the tree for decoding a stereophonic downmix to a 5.1 reconstruction. There are two basic elements,

- One-To-Two (OTT) element, up-mixes one channel into two,

- Two-To-Three (TTT) element, up-mixes two channels into three.

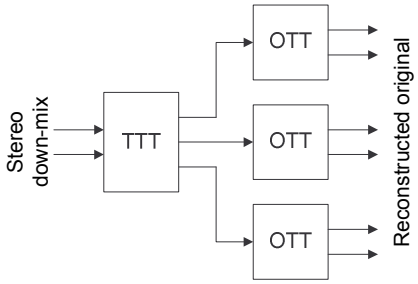

Figure 5: Conceptual synthesis structure.

The up-mixing in these elements is done by a simple matrix operation. In order to recreate coherence, a decorrelator is available in each block. Figure 6 shows the operation of the OTT module. The incoming downmix signal is decorrelated and up-mixed by matrix W_{umx} for each time-frequency tile.

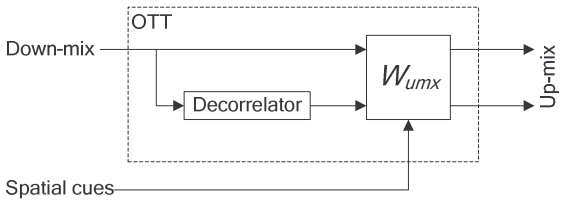

Figure 6: Operation of an OTT element.

The up-mix matrix is calculated based on the spatial side information. The relevant spatial cues for the OTT element are

- Channel Level Difference (CLD) – the level difference between the two input channels,
- Inter-channel Coherence/cross-correlation (ICC) – represents the coherence or cross-correlation between the two input channels.

The up-mixing in the TTT element is slightly different. It estimates a third channel from two input channels and two Channel Prediction Coefficients (CPC). Additionally, a parameter is transmitted that can be used to compensate for a possible prediction loss, e.g. by means of a decorrelated signal.
The above-described conceptual decoding process of separate processing blocks is lumped together into the structure shown in Figure 4.

2.3 MPEG Surround features

This section highlights some important features of MPEG Surround. For a thorough description of features the reader is referred to [18][19][20].

2.3.1 Rate/Distortion scalability

In order to make MPEG Surround useable in as many applications a possible, it is important to cover a broad range, both in terms of side information rates and multi-channel audio quality. There are two different focus areas in this trade-off:

- Downmix,
- Spatial side information.

Although not completely orthogonal, the perceptual quality of the downmix largely determines the *sound* quality of the multi-channel reconstruction whereas the spatial side information mainly determines the quality of the spatialization of the up-mix.
In order to provide the highest flexibility on the part of the spatial side information and to cover all conceivable application areas, the MPEG Surround technology was equipped with a number of provisions for rate/distortion scalability. This approach permits one to flexibly select the operating point for the trade-off between side information rate and multi-channel audio quality without any change in its generic structure. This concept is illustrated in Figure 7 and relies on several dimensions of scalability that are briefly described in the list below.

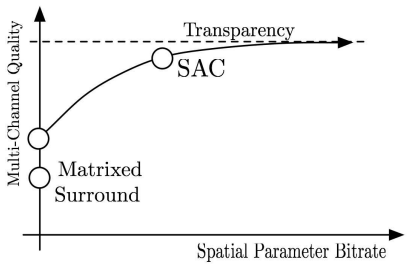

Figure 7: Rate/Distortion scalability.

- Parameter frequency resolution
  A first degree of freedom results from scaling the frequency resolution of spatial audio processing. Currently the MPEG Surround syntax covers between 28 and a single parameter frequency band.
- Parameter time resolution
  Another degree of freedom is available in the temporal resolution of the spatial parameters, i.e., the parameter update rate. The MPEG Surround syntax covers a wide range of update rates and also allows for adapting the temporal grid dynamically to the signal structure.
- Parameter quantization resolution
  As a third possibility, different granularities for transmitted parameters can be used. Using low-resolution parameter descriptions is accommodated by dedicated tools, such as

the *Adaptive Parameter Smoothing* mechanism [20].

- Parameter choice

  Furthermore, there is a choice as to how extensive the transmitted parameterization describes the original multi-channel signal. As an example, the number of ICC values transmitted to characterize the wideness of the spatial image may be as low as a single value per time-frequency tile, applicable to all OTT instances.

- Residual coding

  Finally, it is recognized that the quality level of the multi-channel reconstruction is limited by the limits of the parametric model used. Therefore, the MPEG Surround system supports "residual coding" which is a wave-form coding extension that codes the error-signal originating from the limits of the parametric model.

Together, these scaling dimensions enable operation at a wide range of rate/distortion trade-offs from side information rates below 3 kbit/s to 32 kbit/s and above.

Although MPEG Surround offers the most efficient multi-channel coding to date while at the same time allowing for a backwards compatible downmix signal, there are applications where, due to the construction of the transmission infrastructure, no transmission of additional (however small) side information is possible. In order to account for this, the MPEG Surround decoder can be operated in a *non-guided* mode. This means that the multi-channel signal is recreated based solely on the available downmix signal without the MPEG Surround spatial data, and no spatial side information is transmitted in this mode. However, due to its adaptive nature, this mode still provides better quality than matrix surround based systems.

2.3.2 Matrix surround capability

Besides a conventional stereo downmix, the MPEG Surround encoder is also capable of generating a matrix surround compatible stereo downmix signal. This feature ensures backward-compatible 5.1 audio playback on matrix surround decoders not supporting MPEG Surround. In this context, it is important to ensure that the perceptual quality of the multi-channel reconstruction is not affected by enabling the matrix surround feature.

The matrix surround capability is achieved by using a parameter-controlled post-processing unit that acts on the stereo downmix at the encoder side. A block diagram of an MPEG Surround encoder with this extension is shown in Figure 8.

The matrix surround enabling post-processing unit, implemented as a matrix transformation, operates in the time-frequency domain on the output of the spatial analysis block and is controlled by the spatial parameters. The transformation matrix is guaranteed to have an inverse which can be uniquely determined from the spatial parameters in the bitstream.

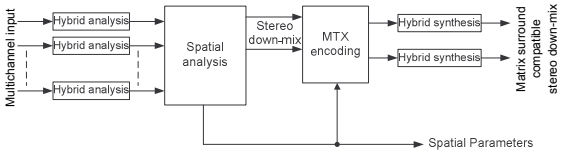

Figure 8: MPEG Surround encoder with post-processing for matrix surround (MTX) compatible downmix.

In the MPEG Surround decoder the process is reversed, i.e., a complementary pre-processing step is applied to the downmix signal before entering the spatial synthesis process.

The matrix surround compatibility comes without any significant additional spatial information (1 bit indicates whether it is enabled). The ability to invert the matrix surround compatibility processing guarantees that there is no negative effect on the multi-channel reconstruction quality. Furthermore, this feature enables optimal performance of the before-mentioned non-guided mode within the MPEG Surround framework.

2.3.3 Binaural rendering

One of the most recent extensions of MPEG Surround is the capability to render a 3D/binaural stereo output. Using this mode, consumers can experience a 3D virtual multi-channel loudspeaker setup when listening over headphones. This extension is of especially significant interest for mobile devices (such as DVB-H receivers) and will be outlined in more detail below.

3 MPEG SURROUND BINAURAL RENDERING

3.1 Application scenarios

Two distinct use-cases are supported. In the first use case, referred to as '*3D*', the transmitted (stereo) downmix is converted to a 3D headphone signal at the *encoder* side, accompanied by spatial parameters. In this use case, legacy stereo devices will automatically render a 3D headphone output. If the same (3D) bitstream is decoded by an MPEG Surround decoder, the transmitted 3D downmix can be converted to (standard) multi-channel output optimized for loudspeaker playback.

In the second use case, a conventional MPEG Surround downmix / spatial parameter bitstream is decoded using a so-called '*binaural decoding*' mode. Hence the 3D/binaural synthesis is applied at the *decoder* side.

Within MPEG Surround, both use cases are covered using a new technique for binaural audio synthesis. As described previously in Section 1, the synthesis process

of conventional 3D synthesis systems comprises convolution of each virtual sound source with a pair of HRTFs (e.g., 2N convolutions, with N being the number of sound sources). In the context of MPEG Surround, this method has several disadvantages:

- Individual (virtual) loudspeaker signals are required for HRTF convolution. Within MPEG surround this means that multi-channel decoding is required as an intermediate step.
- It is virtually impossible to 'undo' or 'invert' the encoder-side HRTF processing at the decoder (which is needed in the first use case for loudspeaker playback).
- Convolution is most efficiently applied in the FFT domain while MPEG Surround operates in the QMF domain.

To circumvent these potential problems, MPEG Surround 3D synthesis is based on new technology that operates in the QMF domain without intermediate multi-channel decoding. The incorporation of this technology in the two different use cases is outlined in the sections below.

3.2 HRTF parameters

MPEG Surround facilitates the use of HRTF *parameters*. Instead of describing HRTFs by means of a transfer function, the perceptually relevant properties of HRTF pairs are captured by means of a small set of statistical properties. The parameterization is especially suitable for anechoic HRTFs and works in a similar way as the spatial parameterization of multi-channel content that is used in MPEG Surround. Parameters are extracted as a function of frequency (i.e., using the concept of non-uniformly distributed parameter bands) and describe the spectral envelopes of an HRTF pair, the average phase difference and optionally the coherence between an HRTF pair. This process is repeated for the HRTFs of every sound source position of interest.

Using this compact representation, the perceptually relevant localization cues are represented accurately, while the perceptual irrelevance of fine-structure detail in HRTF magnitude and phase spectra is effectively exploited [21][22]. More importantly, the HRTF parameters facilitate low-complexity, parameter-based binaural rendering in the context of MPEG Surround.

3.3 Parameter-based binaural rendering

As described in the previous sections, the spatial parameters describe the perceptually relevant properties of multi-channel content. In the context of binaural rendering, these parameters describe relations between virtual sound sources (e.g., virtual loudspeakers). The HRTF parameters, on the other hand, describe the relation between a certain sound source position and the resulting spatial properties of the signals that are presented over headphones. Consequently, spatial parameters and HRTF parameters can be combined to estimate so-called 'binaural' parameters (see Figure 9). These binaural parameters represent binaural properties (e.g., binaural cues) that are the result of simultaneous playback of all virtual sound sources. Said differently, the binaural parameters represent the changes that a mono or stereo downmix signal must undergo to result in a binaural signal that represents all virtual sound sources simultaneously, but *without* the need for an intermediate 5.1 signal presentation. This shift of HRTF processing from the traditional signal domain to the parameter domain has the great advantage of a reduced complexity, as will be outlined below.

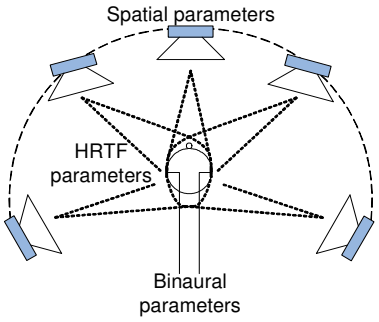

Figure 9: Binaural parameters result from spatial parameters and HRTF parameters

In Figure 10, a conceptual spatial decoder (by means of a single OTT element) is shown which generates two output signals from a mono input signal using a Channel Level Difference (CLD) and Inter-Channel Correlation (ICC) parameter. For each parameter band, the input signal has a power given by σ^2, and the two OTT output signals have powers given by σ_1^2 and σ_2^2, respectively. Since we are only interested in relative changes with respect to the input signal, we assume

$$\sigma^2 = 1,$$

and given the energy preservation property of OTT elements:

$$\sigma_1^2 + \sigma_2^2 = 1.$$

The transmitted CLD parameter is given by

$$\mathrm{CLD} = 10\log_{10}\left(\frac{\sigma_1^2}{\sigma_2^2}\right),$$

which gives the solution for by σ_1^2 and σ_2^2:

$$\sigma_1^2 = \frac{10^{\mathrm{CLD}/10}}{1+10^{\mathrm{CLD}/10}},$$

$$\sigma_2^2 = 1-\sigma_1^2 .$$

The two output signals of the OTT element are subsequently subject to HRTF processing. Each of the two output signals is processed by sets of HRTF parameters which change the (sub-band) level and phase of the input signals, determined by the (mean) amplitude parameters p and the average phase difference parameters ϕ of each HRTF pair. The resulting modified powers σ^2_{xy} after application of the HRTF parameters are given by

$$\sigma_{xy}^2 = \sigma_x^2 p_{xy}^2 ,$$

with p_{xy} being the HRTF amplitude parameter for a sound source position corresponding to OTT output channel x, and y the index for each ear of the HRTF pair.

In a last step, the signals are summed across virtual sound sources for each ear signal. This addition results in the estimated relative sub-band powers of the left and right-ear signals σ_L^2 and σ_R^2:

$$\sigma_L^2 = \sigma_{1L}^2 + \sigma_{2L}^2 + 2\mathrm{ICC}\cos(\phi_L)\sigma_{1L}\sigma_{2L} ,$$
$$\sigma_R^2 = \sigma_{1R}^2 + \sigma_{2R}^2 + 2\mathrm{ICC}\cos(\phi_R)\sigma_{1R}\sigma_{2R} .$$

In a similar fashion, the average phase difference and the coherence between the two binaural output signals can be estimated in the parameter domain. If all relevant binaural cues are known, the binaural rendering system only needs to re-instate these parameters given the mono input signal. This synthesis process, which is essentially a 'parametric stereo' decoder, is described in detail elsewhere [9][10][11]. In essence, it comprises a 2x2 sub-band matrix operation:

$$\begin{bmatrix} L_{bin} \\ R_{bin} \end{bmatrix}_b = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix}_b \begin{bmatrix} M \\ D \end{bmatrix}_b ,$$

with M being the mono input signal, D the output of a decorrelator, h_{xy} the upmix matrix elements, b the parameter band index, and L_{bin}, R_{bin} the binaural output signal.

The approach of parameter-domain estimation of binaural cues can be extended to arbitrary configurations of OTT and TTT boxes while retaining the resulting 2x2 matrix operation in the signal domain. In other words, the computational complexity of binaural rendering is largely independent of the number of simultaneous sound sources.

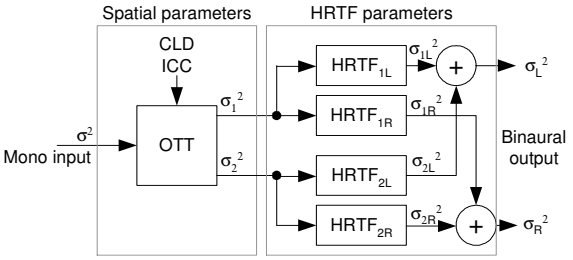

Figure 10: Concept of spatial and HRTF parameter combination.

3.4 Extension to high quality and echoic HRTFs

A starting point for the accurate (i.e. non-parametric) modeling of HRTFs/BRTFs of arbitrary length is the observation that any FIR filter can be implemented with high accuracy in the subband domain of the QMF filter bank used in MPEG Surround. The resulting subband filtering consists of simply applying one FIR filter per subband.

An N-tap filter in the time domain is converted into a collection of 64 complex K-tap subband filters, where

$$K = \left\lceil \frac{N}{64} \right\rceil + 2 .$$

In fact, the filter conversion algorithm itself consists of a complex modulated analysis filter bank very similar to the MPEG Surround analysis bank, albeit with a different prototype filter.

It is important to note that a straightforward polyphase implementation [23] of filtering in a subband filterbank would result in cross filtering between different subbands. The absence of cross filter terms is the key enabling factor for the combination of HRTF/BRTF data with the MPEG Surround parameters, as it allows this combination to be performed independently in each of the MPEG Surround parameter frequency bands. The details of the combination algorithm are beyond the scope of this paper, but the concept is very close to that described for the parametric case in the previous section. The final result is a 2x2 synthesis matrix which is populated by time varying subband filters.

3.5 Binaural Decoding

The binaural decoding scheme is outlined in Figure 11. The MPEG surround bitstream is decomposed into a downmix bitstream and spatial parameters. The downmix decoder produces conventional mono or stereo signals which are subsequently converted to the hybrid QMF domain by means of the MPEG Surround hybrid QMF analysis filter bank. A binaural synthesis stage generates the hybrid QMF-domain binaural output by means of a 2-in, 2-out matrix operation. Hence no

intermediate multi-channel up-mix is required. The matrix elements result from a combination of the transmitted spatial parameters and HRTF data. The hybrid QMF synthesis filter bank generates the time-domain binaural output signal.

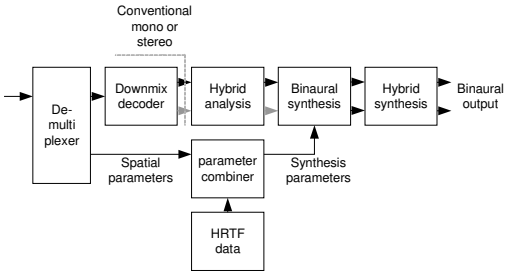

Figure 11: Binaural decoder schematic.

In the case of a mono downmix, the 2x2 binaural synthesis matrix has as inputs the mono downmix signal, and the same signal processed by a decorrelator. In case of a stereo downmix, the left and right downmix channels form the input of the 2x2 synthesis matrix.

The parameter combiner that generates binaural synthesis parameters can operate in two modes. The first mode is a high-quality mode, in which HRTFs of arbitrary length can be modelled very accurately, as described in Section 3.4. The resulting 2x2 synthesis matrix for this mode can thus have multiple taps in the time (slot) direction. The second mode is a low-complexity mode using the parameter-based rrendering as discussed in Section 3.3. In this mode, the 2x2 synthesis matrix has therefore only a single tap in the time direction. Furthermore, since interaural fine-structure phase synthesis is not employed for frequencies beyond approximately 2.5 kHz, the synthesis matrix is real-valued for approximately 90% of the signal bandwidth. This is especially suitable for low-complexity operation and/or representing short (e.g., anechoic) HRTFs. An additional advantage of the low-complexity mode is the fact that the 2x2 synthesis matrix can be inverted, which is an interesting property for the '3D' use case, as outlined subsequently.

3.6 3D-Stereo

In this use case, the binaural processing is applied in the encoder, resulting in a binaural stereo downmix that can be played over headphones on legacy stereo devices. A binaural synthesis module is applied as a post-processing step after spatial encoding in the hybrid QMF domain, in a similar fashion as the matrixed-surround compatibility mode (see Section 2.3.2). The 3D encoder scheme is outlined in Figure 12. The 3D post-processing step comprises the same invertible 2x2 synthesis matrix as used in the low-complexity binaural decoder, which is controlled by a combination of HRTF data and extracted spatial parameters. The HRTF data can be transmitted as part of the MPEG Surround bitstream using a very efficient parameterized representation.

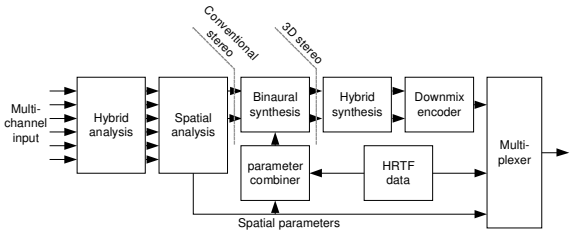

Figure 12: 3D encoder schematic

The corresponding decoder for multi-channel loudspeaker playback is shown in Figure 13. A 3D/binaural inversion stage operates as a pre-processing step before spatial decoding in the hybrid QMF domain, ensuring uncompromised quality for multi-channel reconstruction.

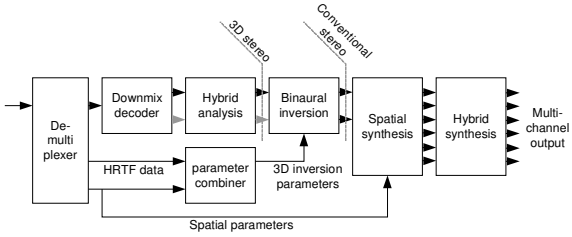

Figure 13: 3D decoder for loudspeaker playback.

3.7 3D Stereo using individual HRTFs

In the 3D-stereo use case, HRTF processing is applied in the encoder. It is therefore difficult to facilitate 3D rendering on legacy decoders with HRTFs that are matched to the characteristics of each listener (i.e., using individual HRTFs). MPEG Surround, however, does facilitate 3D rendering with individual HRTFs, even if a 3D-stereo downmix was transmitted using generic (i.e., non-individualized) HRTFs. This is achieved by replacing the spatial decoder for loudspeaker playback (see Figure 13) by a spatial decoder for binaural synthesis, controlled by the individual's personal HRTF data (see Figure 14). The binaural inversion stage re-creates conventional stereo from the transmitted 3D downmix, the transmitted spatial parameters and the non-individualized HRTF data. Subsequently, the binaural re-synthesis stage creates a binaural stereo version based on individual HRTFs, supplied at the decoder side.

In the low-complexity mode, binaural inversion and binaural synthesis both comprise a 2x2 matrix. Hence the cascade of binaural inversion and binaural re-synthesis is again a 2x2 matrix (resulting from a matrix product) and can thus be implemented very efficiently.

As a result, the decoder complexity using the combined binaural inversion and re-synthesis is similar to the complexity of the low-complexity binaural decoder alone.

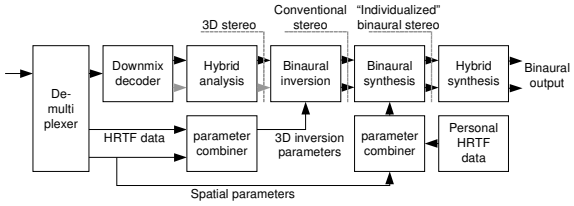

Figure 14: Binaural (re)synthesis using individual HRTFs based on a 3D-stereo downmix.

3.8 3D sound for stereo loudspeaker systems

MPEG Surround provides standardized interfaces for HRTF data (either in the parametric domain or as an impulse response), and thus ensures maximum flexibility for content providers, broadcasters and consumers to optimize playback according to their needs and demands. Besides the use of personal HRTFs for binaural rendering, this flexibility facilitates additional functionality. One example of such functionality is the possibility of creating 3D sound using a conventional stereo playback system. A well-known approach for creating 3D sound over stereo loudspeakers is based on the concept of *crosstalk cancellation* [24]. Technology based on this principle aims at extending the possible range of sound sources outside the stereo loudspeaker base by cancellation of inherent crosstalk (see Figure 15). In practice, this means that for every sound source, two filters (H_{xy}) are applied to generate two signals that are fed to the two loudspeakers. In most cases, these filters differ between sound sources to result in a different perceived position of each sound source.

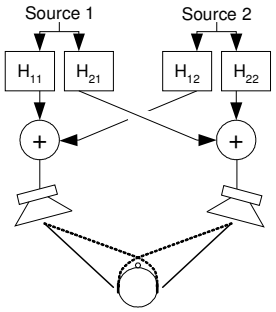

Figure 15: Crosstalk (dashed lines) and crosstalk cancellation filters (H_{xy}) for 2 sound sources.

The processing scheme for multiple simultaneous sound sources is, in fact, identical to the (conventional) HRTF processing scheme depicted in Figure 1, with the only modification that the HRTFs are replaced by crosstalk-cancellation filters. As a result, the crosstalk cancellation principle can be exploited in MPEG Surround decoders by re-using binaural technology. For each audio channel (for example in a 5.1 setup), a set of crosstalk cancellation filters can be provided to the binaural decoder. Playback of the decoded output over a stereo loudspeaker pair will result in the desired 3D sound experience. Compared to conventional crosstalk cancellation systems, the application of these filters in the context of MPEG Surround has the following important advantages:

- All processing is performed in a 2x2 processing matrix without the need of an intermediate 5.1 signal representation.
- Only 2 synthesis filterbanks are required.
- Freedom of crosstalk-cancellation filter designs; the filters can be optimized for each application or playback device individually.

3.9 Performance

This section presents results from recent listening tests conducted within the context of the MPEG standardization process. Two types of tests were conducted: (1) individualized sound localization tests [25] were conducted to examine the spectral resolution required in the QMF domain for an adequate parameter-based representation of HRTF information; and (2) several MUSHRA [26] tests were conducted to perceptually evaluate the performance of the binaural decoding and 3D stereo encoding technologies. A brief description of the tests and their results follow.

3.9.1 Sound Localization Test

Stimuli. A sound localization experiment was conducted in virtual auditory space to examine the spatial fidelity of a Gaussian broadband noise source (150ms duration with raised-cosine time envelope). Two normally-hearing subjects performed the localization task for two sound conditions: (1) a control sound condition in which the binaural stimuli were prepared using normal HRTF filter convolution; (2) a test sound condition in which the binaural stimuli were prepared using a 28 QMF band parameter-based binaural rendering as described in Section 3.3. The subject's individualized HRTF filters were measured for 393 positions evenly spaced around an imaginary sphere one meter in radius about the subject's head.

Experimental paradigm. Localization performance was assessed using a nose pointing task. For this task, an electromagnetic tracking system (Polhemus Fastrak) is used to measure the subject's perceived sound source location relative to the centre of the subject's head. The sensor is mounted on top of a rigid headband worn by the subject. Prior to each stimulus presentation, the subject aligns his/her head to a calibrated start position with the aid of an LED display. After pressing a

handheld pushbutton, the stimulus is played over in-ear tube phones (Etymotic Research ER-2). The subject responds by turning and pointing his/her nose to the perceived position of the sound source and once again presses the handheld pushbutton. The controlling computer records the orientation of the electromagnetic sensor and the subject returns to the calibrated start position for the next stimulus presentation. Localization performance is assessed for three repeats of 76 test locations spaced around the subject.

Localization results. The positions on a sphere can be described using a lateral and polar angle coordinate system. The lateral angle is the horizontal angle away from the midline where negative lateral angles (down to -90°) define the left hemisphere and positive lateral angles (up to +90°) define the right hemisphere. The lateral angle describes positions for which binaural cues, such as interaural time and level differences are very similar. The polar angle is the angle on the circle around the interaural axis, for a given lateral angle, with 0° representing the horizontal plane in front, 90° representing directly above, 180° representing behind and 270° representing directly below. Localization on the polar angle depends on the spectral cues generated by the directionally dependent filtering of the outer ear. Figure 17 and 18 show localization data for both subjects for the lateral and polar angle, respectively, and the data indicate that there were no substantial differences in localization performance across conditions.

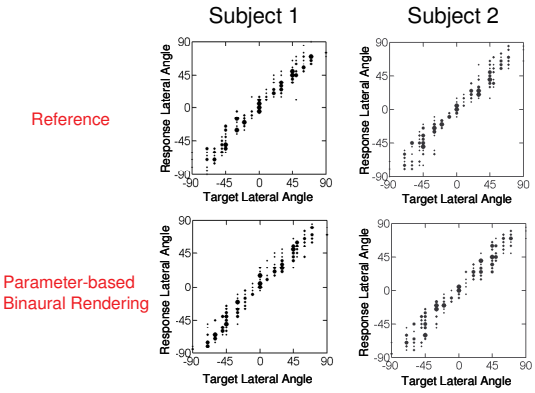

Figure 16. The lateral angle component of the localization performance data is shown for both subjects using a scatter plot.

3.9.2 MUSHRA Tests

Test Setup: Subjective listening tests were conducted to evaluate the fidelity of the binaural rendering process within MPEG Surround. The experiments were conducted in virtual auditory space using the MUSHRA testing methodology.

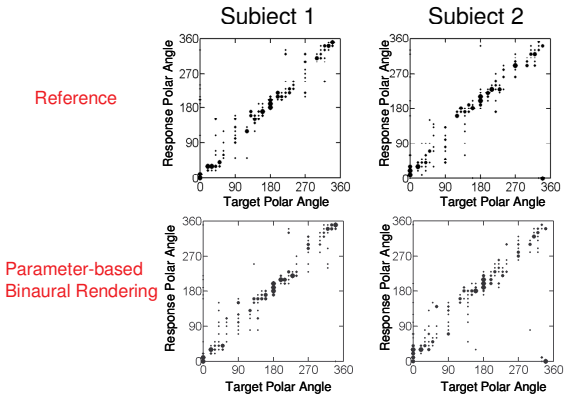

Figure 17. The polar angle component of the localization performance data is shown for both subjects using a scatter plot.

There were two sets of tests, one for the binaural rendering technique that occurs within the decoder, referred to as the binaural decoder, and one for the binaural rendering technique that occurs within the encoder, referred to as the 3D stereo encoder. For both binaural rendering techniques, tests were carried out using both a TC1 and TC3 configuration. For TC1, a stereo AAC core coder is used operating at 160 kbps stereo. For TC3, a monaural HE-AAC core coder is employed such that the total bitrate of core coding and spatial side information amounts to 48 kbps. For the HE-AAC core coder, KEMAR HRTF filters were used and for the AAC core coder, 1000 tap BRTF filters were used. A number of reference signals were employed during the testing and these are listed in Table 1. Table 2 shows the size of the MUSHRA tests.

Table 1 –Signals under test

Label	Description
Ref	Original 5.1 item downmixed to binaural with common HRTF set
Ref-3.5k	Anchor, 3.5 kHz low-pass filtered reference
RMB	MPEG Surround (RM) decoder 5.1 output downmixed to binaural with common HRTF set
ADG	RM 5.1 decoding of a 3D/binaural down mix including Artistic Downmix Gains
RMS	MPEG Surround (RM) decoder 5.1 output downmixed to stereo

Table 2 – Size of the MUSHRA tests.

test		Number of stimuli	Number of subjects	Number of rejected subjects
Bin.stereo decoder	TC3	9108	92	8
	TC1	8316	84	11
3D stereo encoder	TC3	3036	46	7
	TC1	4235	77	17

Test results: The MUSHRA test results for the binaural decoder are shown in Figures 18 and 19. Similarly, the MUSHRA test results for the 3D stereo encoder are shown in Figures 21 and 22.

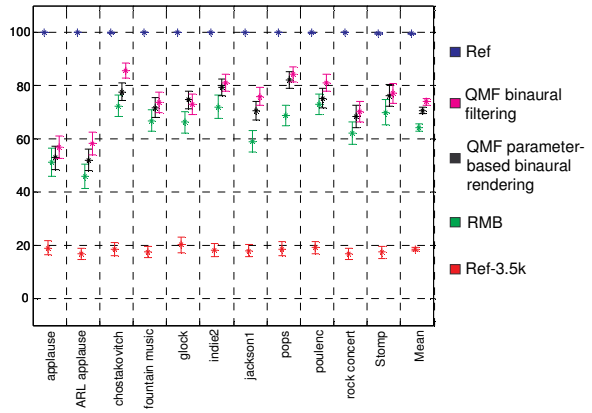

Figure 18. MUSHRA test results for the binaural decoder in TC3 configuration. Mean opinion score is shown on the vertical axis and the audio test items are shown on the horizontal axis.

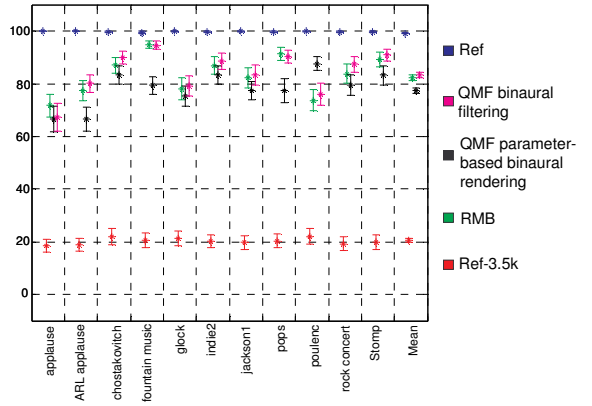

Figure 19. MUSHRA test results for the binaural decoder in TC1 config. Other details as in Fig. 18.

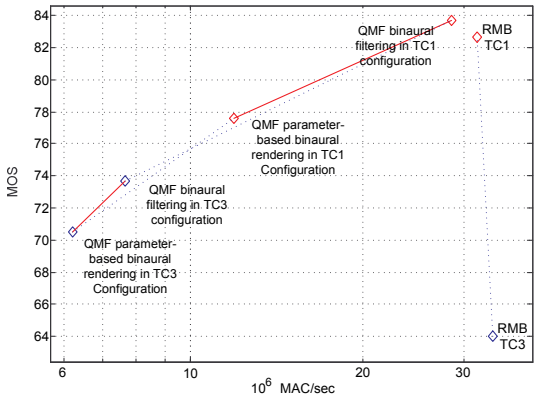

Figure 20. Quality versus complexity for MPEG Surround binaural rendering techniques.

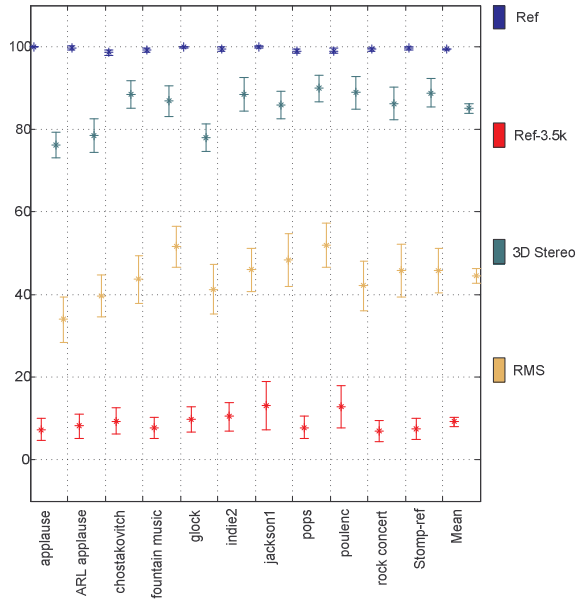

Figure 21. MUSHRA tests results for the 3D stereo encoder in TC1 config. Other details as in Fig. 18.

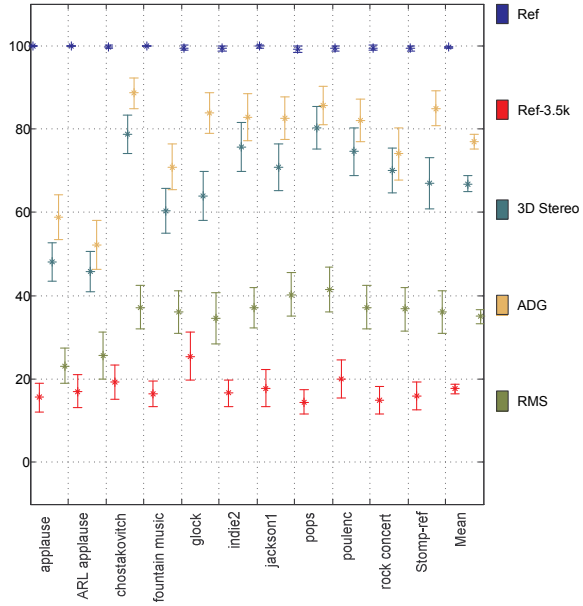

Figure 22. MUSHRA tests results for the 3D stereo encoder in TC3 config. Other details as in Fig 18

For benchmarking, the complexity of the MPEG Surround binaural rendering techniques has been analyzed in terms of multiply-accumulate operations. Figure 20 shows a 2-D representation of the quality versus complexity numbers. Note that the RMB reference condition (MPEG Surround decoding

followed by efficient external HRTF filtering in the FFT domain) has been included.

4 APPLICATIONS

The binaural rendering capability of MPEG Surround brings surround sound to portable devices to an extent not previously possible. MPEG Surround offers, by means of the binaural decoding functionality, a surround sound format that is suitable for stationary home devices, car radios etc., as well as portable devices. The following section gives a few examples of interesting applications that can be envisioned for the binaural rendering modes of MPEG Surround.

Digital Radio Broadcasting. Surround sound in radio broadcasting is particularly interesting for car-radio applications since the listener's position is fixed with respect to the loudspeakers. Hence, MPEG Surround with its inherent stereo backwards compatibility is ideal for this application. A legacy "kitchen radio" device (where typically the listener is moving around doing other things while listening to the radio) will play the stereo signal, while the car-radio can decode the MPEG Surround data and render the multi-channel signal. In a similar manner a legacy portable radio receiver will play the backwards compatible stereo part of the MPEG Surround stream while a binaural MPEG Surround equipped portable radio-receiver, will operate in the binaural decoding mode and provide a surround sound listening experience over headphones for the listener.

Digital Video Broadcasting. The MPEG Surround binaural rendering capability is particularly attractive for TV/Movie consumption on portable devices. Since surround sound has an important place in TV/Movie consumption it is interesting to maintain this for portable TV/Movie consumption such as with DVB-H. For this application MPEG Surround is ideal, not only because it enables surround sound at a very low bitrate, but also because it enables the surround sound experience over headphones on portable devices.

Music Download Services. There are several popular music store services available as of today, either for download of music over the Internet, e.g. "iTunes Music Store", or for download of music directly to the mobile-phone, e.g. KDDI's EZ "Chaku-Uta Full™" service. These make a very interesting application for MPEG Surround. Since the MPEG Surround data adds a minimal overhead to the existing downmix data, storing surround files on devices limited by disk space poses no problems.
One could envision a portable music player that, when connected to the home stereo equipment, decodes the MPEG Surround files stored on the player into surround sound played over the home speaker set-up. However, when the player is "mobile" i.e. carried around by the user, the MPEG Surround data is decoded into binaural stereo enabling the surround sound experience on the mobile device. Finally, the legacy player can store the same files (with very limited penalty on storage space), and play the stereo backwards compatible part.

5 CONCLUSIONS

Recent progress in the area of parametric coding of multi-channel audio has led to the MPEG Surround specification which provides an efficient and backward compatible representation of high quality audio at bitrates comparable to those currently used for representing stereo (or even mono) audio signals. While the technology was initially conceived for use with conventional loudspeaker reproduction, the idea of accommodating multi-channel playback on small mobile devices led to a number of interesting additions to the specification. These extensions combine traditional approaches for binaural rendering with the MPEG Surround framework in an innovative way to achieve high-quality binaural rendering of surround sound even with very limited computational resources, as they are typically available on mobile devices like cell phones, mp3 players or PDAs. Several options for binaural rendering are available, including approaches that allow binaural rendering even on legacy devices. The blending of binaural technology with parametric modeling techniques enables a wide range of attractive applications for both mobile and stationary home use and makes MPEG Surround an attractive format for the unified bitrate-efficient delivery of multi-channel sound.

REFERENCES

[1] D. Begault. 3D Sound for Virtual Reality and Multimedia. Academic Press, Cambridge, 1994

[2] Gilkey, R. H. and Anderson, T. R., "Binaural and spatial hearing in real and virtual environments." Lawrence Erlbaum Associates, 1997

[3] J. Blauert, "Spatial hearing: The psychophysics of human sound localization", MIT Press, Revised edition, 1997.

[4] H. Møller, M. F. Sørensen, D. Hammershøi, C. B. Jensen, "Head-related transfer functions of human subjects", J. Audio Eng. Soc., Vol. 43, No. 5, pp. 300-321, 1995.

[5] W. G. Gardner, 3-D Audio Using Loudspeakers. Kluwer Academic Publishers, 1998.

[6] P. J. Minnaar, S. K. Olesen, F. Christensen, H. Møller, “Localization with binaural recordings from artificial and human heads”, J. Audio Eng. Soc., May 2001

[7] P. J. Minnaar, S. K. Olesen, F. Christensen, H. Møller: "The importance of head movements for binaural room synthesis", Proceedings of ICAD 2001, Espoo, 2001

[8] J. Herre: “From Joint Stereo to Spatial Audio Coding - Recent Progress and Standardization“, Sixth International Conference on Digital Audio Effects (DAFX04), Naples, Italy, October 2004

[9] H. Purnhagen: “Low Complexity Parametric Stereo Coding in MPEG-4”, 7th International Conference on Audio Effects (DAFX-04), Naples, Italy, October 2004

[10] E. Schuijers, J. Breebaart, H. Purnhagen, J. Engdegård: “Low complexity parametric stereo coding”, Proc. 116th AES convention, Berlin, Germany, 2004, Preprint 6073

[11] J. Breebaart, S. van de Par, A. Kohlrausch, E. Schuijers: “Parametric coding of stereo audio”, EURASIP J. Applied Signal Proc. 9:1305-1322 (2005)

[12] C. Faller, F. Baumgarte: “Efficient Representation of Spatial Audio Using Perceptual Parametrization”, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York 2001

[13] C. Faller and F. Baumgarte, “Binaural Cue Coding - Part II: Schemes and applications,” IEEE Trans. on Speech and Audio Proc., vol. 11, no. 6, Nov. 2003

[14] C. Faller: “Coding of Spatial Audio Compatible with Different Playback Formats“, 117th AES Convention, San Francisco 2004, Preprint 6187

[15] Dolby Publication: ““Dolby Surround Pro Logic II Decoder – Principles of Operation”, http://www.dolby.com/assets/pdf/tech_library/209_Dolby_Surround_Pro_Logic_II_Decoder_Principles_of_Operation.pdf

[16] D. Griesinger: “Multichannel Matrix Decoders For Two-Eared Listeners”, 101st AES Convention, Los Angeles 1996, Preprint 4402

[17] J. Herre, C. Faller, S. Disch, C. Ertel, J. Hilpert, A. Hoelzer, K. Linzmeier, C. Spenger, P. Kroon: “Spatial Audio Coding: Next-Generation Efficient and Compatible Coding of Multi-Channel Audio”, 117th AES Convention, San Francisco 2004, Preprint 6186

[18] J. Herre, H. Purnhagen, J. Breebaart, C. Faller, S. Disch, K. Kjörling, E. Schuijers, J. Hilpert, F. Myburg: “The Reference Model Architecture for MPEG Spatial Audio Coding”, Proc. 118th AES convention, Barcelona, Spain, May 2005, Preprint 6477

[19] J. Breebaart, J. Herre, C. Faller, J. Rödén, F. Myburg, S. Disch, H. Purnhagen, G. Hotho, M. Neusinger, K. Kjörling, W. Oomen: “MPEG spatial audio coding / MPEG Surround: overview and current status”, Proc. 119th AES convention, New York, USA, October 2005, Preprint 6447

[20] L. Villemoes, J. Herre, J. Breebaart, G. Hotho, S. Disch, H. Purnhagen, K. Kjörling: “MPEG Surround: The forthcoming ISO standard for spatial audio coding”, AES 28th International Conference, Piteå, Sweden, 2006

[21] A. Kulkarni, S. K. Isabelle, H. S. Colburn: “Sensitivity of human subjects to head-related transfer-function phase spectra”, J. Acoust. Soc. Am. 105: 2821-2840, 1999.

[22] J. Breebaart, A. Kohlrausch: “The perceptual (ir)relevance of HRTF magnitude and phase spectra”, Proc. 110th AES convention, Amsterdam, The Netherlands, 2001.

[23] C. Lanciani, R.W. Schafer: “Subband-domain filtering of MPEG audio signals”. ICASSP’ 99. Proceedings, 1999.

[24] M. R. Schroeder: “Models of hearing”, Proc. IEEE 63 (9): 1332-1350 (1975).

[25] S. Carlile, P. Leong, and S. Hyams (1997). "The nature and distribution of errors in sound localisation by human listeners," Hearing Research, 114: 179-196.

[26] ITU-R Recommendation BS.1534-1, “Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA)”, International Telecommunications Union, Geneva, Switzerland, 2001