Audio Engineering Society

Convention Paper

Presented at the 122nd Convention
2007 May 5–8 Vienna, Austria

The papers at this Convention have been selected on the basis of a submitted abstract and extended precis that have been peer reviewed by at least two qualified anonymous reviewers. This convention paper has been reproduced from the author's advance manuscript, without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. Additional papers may be obtained by sending request and remittance to Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA; also see www.aes.org. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

MPEG Surround – The ISO/MPEG Standard for Efficient and Compatible Multi-Channel Audio Coding

Jürgen Herre[1], Kristofer Kjörling[2], Jeroen Breebaart[3], Christof Faller[4], Sascha Disch[1], Heiko Purnhagen[2], Jeroen Koppens[5], Johannes Hilpert[1], Jonas Rödén[2], Werner Oomen[5], Karsten Linzmeier[1], and Kok Seng Chong[6]

[1] Fraunhofer Institute for Integrated Circuits, Am Wolfsmantel 33, 91058 Erlangen, Germany

[2] Coding Technologies, 113 30 Stockholm, Sweden

[3] Philips Research, 5656 AE Eindhoven, The Netherlands

[4] Agere Systems, Allentown, PA 18109, USA

[5] Philips Applied Technologies, 5616 LW, Eindhoven, The Netherlands

[6] Panasonic Singapore Laboratories Pte. Ltd., Singapore

ABSTRACT

In 2004, the ISO/MPEG Audio standardization group started a new work item on efficient and backward compatible coding of high-quality multi-channel sound using parametric coding techniques. Finalized in the fall of 2006, the resulting MPEG Surround specification allows the transmission of surround sound at bitrates that have been commonly used for coding of mono or stereo sound. This paper summarizes the results of the standardization process by describing the underlying ideas and providing an overview of the MPEG Surround technology. The performance of the scheme is characterized by the results of the recent verification tests. These tests include several operation modes as they would be used in typical application scenarios to introduce multi-channel audio into existing audio services.

1. INTRODUCTION

In 2004, the ISO/MPEG Audio standardization group started a new work item on efficient and backward compatible coding of high-quality multi-channel sound using parametric coding techniques. Specifically, the technology to be developed should be based on the Spatial Audio Coding (SAC) approach that extends traditional approaches for coding of two or more channels in a way that provides several significant advantages, both in terms of compression efficiency and features. Firstly, it allows the transmission of multi-channel audio at bitrates, which so far only allowed for the transmission of monophonic audio. Secondly, by its underlying structure, the multi-channel audio signal is transmitted in a backward compatible way. As such, the technology can be used to upgrade existing distribution infrastructures for stereo or mono audio content (radio channels, Internet streaming, music downloads etc.) towards the delivery of multi-channel audio while retaining full compatibility with existing receivers. After an intense development process, the resulting MPEG Surround specification was finalized in the second half of 2006 [30].

This paper summarizes the results of the standardization process by describing the underlying ideas and providing an overview of the MPEG Surround technology. Special attention is given to the results of the recent MPEG Surround verification tests that assess the technology's performance in several operation modes as they would be used in typical application scenarios introducing multi-channel audio into existing audio services.

Sections 2 and 3 illustrate the basic approach and the MPEG standardization process that was executed to develop the MPEG Surround specification. The core of the MPEG Surround architecture and its further extensions are described in Sections 4 and 5. Finally, system performance and applications are discussed.

2. SPATIAL AUDIO CODING BASICS

In a nutshell, the general underlying concept of Spatial Audio Coding can be outlined as follows: Rather than performing a discrete coding of the individual audio input channels, a system based on Spatial Audio Coding captures the spatial image of a multi-channel audio signal into a compact set of parameters that can be used to synthesize a high quality multi-channel representation from a transmitted downmix signal. Figure 1 illustrates this concept. During the encoding process, the spatial parameters (cues) are extracted from the multi-channel input signal. These parameters typically include level/intensity differences and measures of correlation/coherence between the audio channels and can be represented in an extremely compact way. At the same time, a monophonic or two-channel downmix signal of the sound material is created and transmitted to the decoder together with the spatial cue information. The downmix can be conveyed to the receiver using known audio coders for monophonic or stereophonic signals. On the decoding side, the transmitted downmix signal is expanded into a high quality multi-channel output based on the spatial parameters.

Due to the reduced number of audio channels to be transmitted (e.g. just one channel for a monophonic downmix signal), the Spatial Audio Coding approach provides an extremely efficient representation of multi-channel audio signals. Furthermore, it is backward compatible on the level of the downmix signal: A receiver device without a spatial audio decoder will simply present the downmix signal.

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 [4], a generalization of Parametric Stereo [5] [6] to multi-channel application, and an extension of the Binaural Cue Coding (BCC) scheme [7] [8] towards using more than one transmitted downmix channel [9]. From a different viewing angle, 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.) [10] [11] by transmission of dedicated (spatial cue) side information to guide the multi-channel reconstruction process and thus achieve improved subjective audio quality [1].

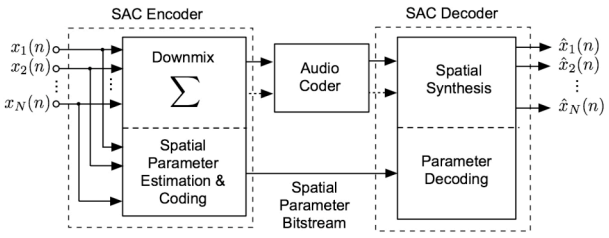

Figure 1: Principle of Spatial Audio Coding

Due to the combination of bitrate-efficiency and backward compatibility, SAC technology can be used to enhance a large number of existing mono or stereo services from stereophonic (or monophonic) to multi-channel transmission in a compatible fashion. To this aim, the existing audio transmission channel carries the downmix signal, and the spatial parameter information is conveyed in a side chain (e.g. the ancillary data portion of an audio bitstream). In this way, multi-channel capability can be achieved for existing audio

distribution services for a minimal increase in bitrate, e.g. between 3 to 32kbit/s.

3. MPEG SURROUND DEVELOPMENT PROCESS

In March of 2004, the ISO/MPEG standardization group started a new work item on SAC by issuing a "Call for Proposals" (CfP) on Spatial Audio Coding [12]. Four submissions were received in response to this CfP and evaluated with respect to a number of performance aspects including the subjective quality of the decoded multi-channel audio signal, the subjective quality of the downmix signals generated, the spatial parameter bitrate and other parameters (additional functionality, computational complexity etc.).

As a result of these extensive evaluations, MPEG decided that the basis for the subsequent standardization process, called Reference Model 0 (RM0), would be a system combining the submissions of Fraunhofer IIS/Agere Systems and Coding Technologies/Philips. These systems outperformed the other submissions and, at the same time, showed complementary performance in terms of other parameters (e.g. per-item quality, bitrate) [13]. The merged RM0 technology (now called MPEG Surround) combines the best features of both individual submissions and was found to fully meet (and even surpass) the performance expectation [2] [14]. The successful development of RM0 set the stage for the subsequent improvement process of this technology that was carried out collaboratively within the MPEG Audio group. After a period of active technological development, the MPEG Surround specification was frozen in the second half of 2006 and its performance confirmed by the final verification test report in January of 2007.

4. BASIC CONCEPTS OF MPEG SURROUND

While a detailed description of the MPEG Surround technology is beyond the scope of this paper, this section provides a brief overview of the most salient underlying concepts. An extended description of the technology can be found in [2] [3] [27] [28]. Refinements of this basic framework will be described in the subsequent section.

4.1. Filterbank and Top-Level Structure

In the human auditory system, the processing of binaural cues is performed on a non-uniform frequency scale [15] [16]. Hence, in order to estimate spatial parameters from a given input signal, it is important to transform its time-domain representation to a representation that resembles this non-uniform scale by using an appropriate filterbank.

For applications including low bitrate audio coding, the MPEG Surround decoder is typically applied as a post-processor to a low bitrate (mono or stereo) decoder. In order to minimize computational complexity, it would be beneficial if the MPEG Surround system could directly make use of the spectral representation of the audio material provided by the audio decoder. In practice, however, spectral representations for the purpose of audio coding are typically obtained by means of critically sampled filterbanks (for example using a Modified Discrete Cosine Transform (MDCT) [17]) and are not suitable for signal manipulation as this would interfere with the aliasing cancellation properties associated with critically sampled filterbanks. The Spectral Band Replication (SBR) algorithm [18] is an important exception in this respect. Similar to the Spatial Audio Coding / MPEG Surround approach, the SBR algorithm is a post-processing algorithm that works on top of a conventional (band-limited) low bitrate audio decoder and allows the reconstruction of a full-bandwidth audio signal. It employs a complex-modulated Quadrature Mirror Filter (QMF) bank to obtain a uniformly-distributed, oversampled frequency representation of the audio signal. The MPEG Surround technology takes advantage of this QMF filterbank which is used as part of a hybrid structure to obtain an efficient non-uniform frequency resolution [6] [19]. Furthermore, by grouping filterbank outputs for spatial parameter analysis and synthesis, the frequency resolution for spatial parameters can be varied extensively while applying a single filterbank configuration. More specifically, the number of parameters to cover the full frequency range (number of "parameter bands") can be varied from only a few (for low bitrate applications) up to 28 (for high-quality processing) to closely mimic the frequency resolution of the human auditory system. A detailed description of the hybrid filterbank in the context of MPEG Surround can be found in [2].

The top-level structure of the MPEG Surround decoder is illustrated in Figure 2, showing a three-step process that converts the supplied downmix into the multi-channel output signal. Firstly, the input signal is decomposed into frequency bands by means of a hybrid QMF analysis filterbank. Next, the multi-channel output signal is generated by means of the spatial synthesis process, which is controlled by the spatial parameters conveyed to the decoder. This synthesis is carried out on the subband signals obtained from the hybrid filterbank in order to apply the time- and frequency dependent spatial parameters to the corresponding time/frequency region (or "tile") of the signal. Finally, the output subband signals and converted back to time domain by means of a set of hybrid QMF synthesis filterbanks.

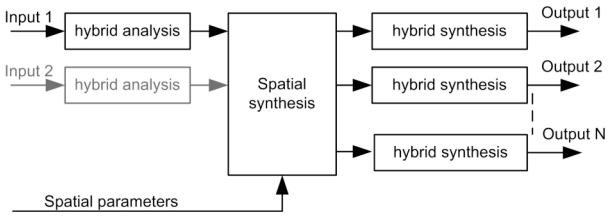

Figure 2: High-level overview of the MPEG Surround synthesis

4.2. Structure of Spatial Synthesis

MPEG Surround provides great flexibility in terms of the input, downmix and decoder channel configurations. This flexibility is obtained by using relatively simple conceptual elements that can be grouped to build more complex coder structures. The two most important elements are referred to as 'One-To-Two' (OTT) and 'Two-To-Three' (TTT) elements. The numbers refer to the input and output channel configurations of each element at the decoder side. In other words, an OTT element describes two output channels by means of a single input channel, accompanied by spatial parameters. Similarly, the TTT element characterizes three output channels by means of a stereo input signal and parameters. Each conceptual decoder element has a corresponding encoder element that extracts spatial parameters and generates a downmix from its input signals. Using the two building blocks many encoder and decoder configurations can be built that encode a multi-channel signal into a downmix and parameters, or conversely decode a downmix into a multi-channel signal based on spatial parameters. An example decoder is outlined in Figure 9, illustrating the basic idea of connecting the building blocks into an MPEG Surround decoder. The various encoding and decoding elements are described in more detail subsequently.

4.2.1. OTT Elements

At the encoder side, the OTT encoder element (also referred to as Reverse-One-To-Two module, R-OTT) extracts two types of spatial parameters, and creates a downmix (and a residual) signal. The OTT encoder element is virtually identical to a Parametric Stereo coder ([6] [19]) and is based on similar principles as Binaural Cue Coding (BCC, [7] [8]). The following spatial parameters are extracted for each parameter band:

- Channel Level Difference (CLD) – this is the level difference between the two input channels. Non-uniform quantization on a logarithmic scale is applied to the CLD parameters, where the quantization has a high accuracy close to zero dB and a coarser resolution when there is a large difference in level between the input channels (as is in line with psychoacoustics).
- Inter-channel coherence/cross-correlation (ICC) – represents the coherence or cross-correlation between the two input channels. A non-uniform quantization is applied to the ICC parameters.

The residual signal represents the error associated with representing the two signals by their downmix and associated parameters and, in principle, enables full multi-channel waveform reconstruction at the decoder side (see section on residual coding).

At the decoder side, the OTT element recreates two channels from the single downmix channel and the spatial parameters. This is visualized in Figure 3 below.

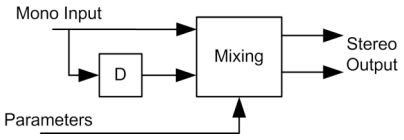

Figure 3: Basic principle of the OTT module

The OTT module takes the single input signal and creates a decorrelated version of the same, by means of

a decorrelator D. These two signals are mixed together based on the CLD parameter and the ICC parameter. The CLD parameter controls the energy distribution of the input signal between the two output signals, and the ICC parameter controls the amount of decorrelated signal mixed into the two output signals. If a residual signal is present, the decorrelated signal is replaced by this residual.

4.2.2. TTT Elements

The TTT encoder element (R-TTT) generates a stereo downmix signal from three input channels, accompanied by spatial parameters. More specifically, the stereo downmix l_0, r_0 is a linear combination of the three input signals l, c and r, in which the center input signal c is represented as phantom-center in the stereo downmix. To provide 3-channel reconstruction at the decoder side, two Channel Prediction Coefficients (CPCs) are transmitted. An additional ICC-coded parameter provides compensation for the prediction-loss at the decoder side based on statistical properties rather than on waveform reconstruction principles. Similar to the OTT element, the TTT encoder element also provides a residual signal that may be transmitted to enable full waveform reconstruction at the decoder side. A detailed description of the TTT element and its various operation modes is provided in [20].

Conversely, the TTT decoder module re-creates three channels based on the two downmix channels and the corresponding parameters. The ICC codec parameter quantifying the prediction loss can be used to compensate for the prediction loss either by means of gain adjustment of the output signal, or by means of adding a decorrelated signal, or by means of a residual signal, corresponding to the prediction loss.

4.2.3. Tree-structured Parameterization

Since OTT and TTT elements can be combined in a tree-structured manner to build more complex coder structures, many different channel configurations can be supported in a flexible way. The following two example configurations describe the encoding of 5.1 surround sound into a downmix and its decoding back to 5.1 multi-channel for a mono and a stereo downmix, respectively.

Figure 4 shows how to combine several R-OTT modules into a multi-channel encoder that encodes a multi-channel into a mono downmix and corresponding parameters. For every R-OTT module CLD and ICC parameters are derived as well as a possible residual signal (the residual may be omitted if lowest bitrate overhead possible is desired). The signals L, Ls, C, LFE, R and Rs denote the left front, left surround, center, Low Frequency Effects, right front and right surround channels, respectively.

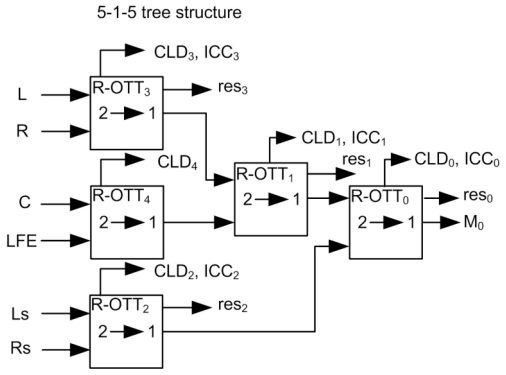

Figure 4: OTT tree forming a 5.1-to-mono encoder

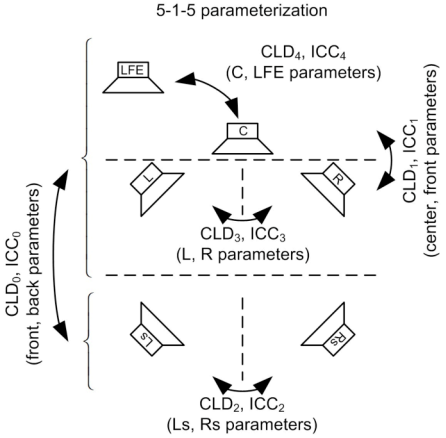

Figure 5: Multi-channel parameterization for a mono-to-5.1 MPEG Surround signal

In Figure 5 the parameterization corresponding to the tree-structure of Figure 4 is visualized. The tree-structured parameterization combines the channels into larger groups of channels, and for every such combination of groups it derives parameters describing how to separate the groups given the parameters and the combined group. Hence, the CLD and ICC parameters

with the subscript 2 describe how to separate the Ls and Rs channels from each other given the combination of the two and the corresponding parameters. Similarly, the CLD and ICC parameters with the subscript 0 describe how to separate a combination of the surround channel (Ls and Rs) from a combination of all the front channels (C, LFS, L and R) given the combination of the two groups and the corresponding parameters.

Among the many conceivable configurations of MPEG Surround, the encoding of 5.1 surround sound into two-channel stereo is particularly attractive in view of its backward compatibility with existing stereo consumer devices. Figure 6 shows a block diagram of an encoder for such a typical system consisting of one R-TTT and three R-OTT encoder elements, and Figure 7 visualizes the parameterization for the underlying tree structure.

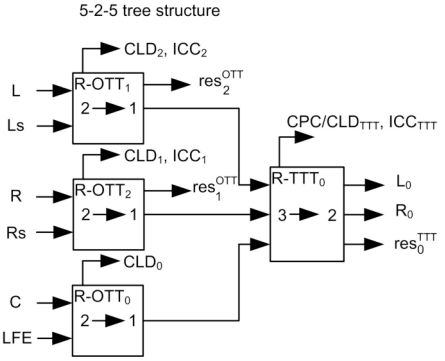

Figure 6: OTT/TTT tree forming a 5.1-to-stereo encoder

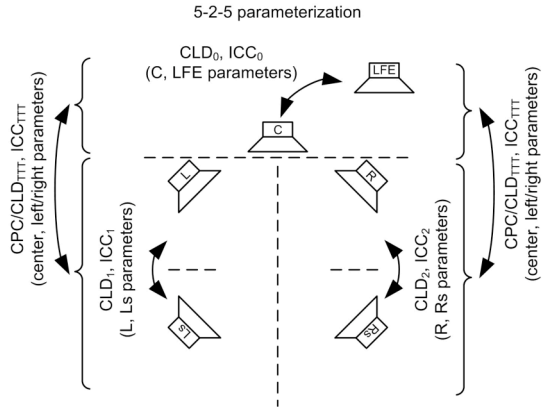

Figure 7: Multi-channel parameterization for a stereo-to-5.1 MPEG Surround signal.

Similarly, the OTT and TTT elements can be used to build up a multitude of different encoder/decoder structures, supporting arbitrary downmixing / upmixing configurations such as e.g. 7.1-5.1-7.1 (i.e. 7.1 input channels downmixed to 5.1 downmix channels and subsequently upmixed again to 7.1 output channels).

On the decoder side, the parameterization of the multi-channel signal can be used to visualize a conceptual upmix of the downmix signal based on the spatial parameters (the following subsection will clarify how the actual signal flow in the decoder differs from the conceptual up-mix outlined in the following). Figure 8 visualizes a conceptual mono-to-5.1 decoder using OTT modules. The subscripts of the OTT modules and the parameters match those of the corresponding encoder picture (Figure 4), and parameterization visualization (Figure 5).

Similarly, Figure 9 displays a conceptual decoder counterpart to the encoder visualized in Figure 6.

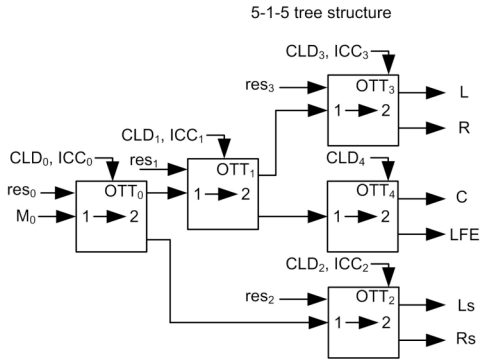

Figure 8: Conceptual mono-to-5.1 decoder

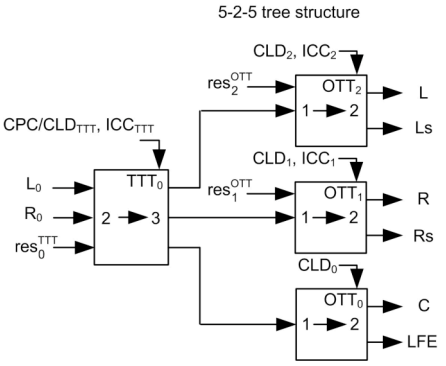

Figure 9: Conceptual stereo-to-5.1 decoder

4.2.4. Flat-structured MPEG Surround Decoder Signal Processing

Compared to the conceptual views presented previously the actual upmix to the multi-channel output does not take place in a tree-structured fashion in an MPEG Surround decoder. Instead, the decoder signal processing happens in a "flattened" way, i.e. the tree-like parameterization visualized by a tree-structured upmix process comprising several subsequent stages is transformed into a single-stage operation. This achieves increased computational efficiency and minimizes possible degradations due to multiple decorrelation operations, i.e. sending the output from one OTT module comprising a decorrelated signal component into another OTT module deriving a new decorrelated signal component.

As a result, the spatial synthesis process, as shown in the previous tree-structured decoder visualizations, can be described by two matrices and decorrelators, as illustrated in Figure 10. These matrices maps a lower number of input channels to a higher number of output channels, i.e. if the input signal is a mono downmix, the input signal multiplied by the pre-mix matrix is a scalar, and if the downmix input is a stereo signal the input is a vector of two elements. Similarly, the output is a vector containing one element for each output channel. The processing takes place in the subband domain of a filterbank, and hence the matrix operations are carried out for every sample in every subband.

The matrix elements are derived from the transmitted spatial parameters given the specific tree-structure used for the parameterization. The decorrelators correspond to the decorrelators as part of the OTT and TTT modules.

The input signals are first processed by a pre-mix matrix, in order to ensure that the input to the different decorrelators are of the same level as if the up-mix would be carried out in a tree-structured fashion. The post-mix matrix mixes the decorrelated signals with input signals similarly to the OTT modules, albeit for all OTT modules in a single step.

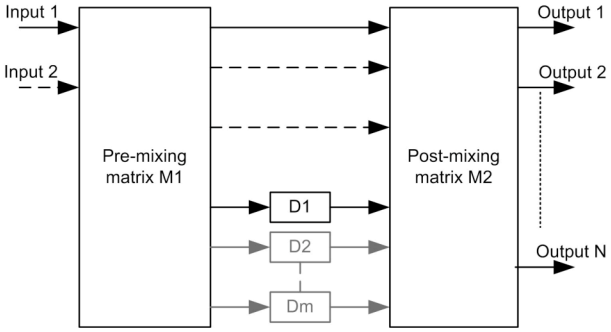

Figure 10: Generalized structure of the spatial synthesis process, comprising two mixing matrices; M1, M2, and a set of decorrelators, D_1, D_2, … D_m

4.3. Decorrelation

The spatial synthesis stage of the MPEG Surround decoder consists of matrixing and decorrelation units. The decorrelation units are required to synthesize output signals with a variable degree of correlation between each other (as dictated by the transmitted ICC parameters) by a weighted summation of original signal and decorrelator output [21]. Each decorrelator unit generates an output signal from an input signal according to the following properties:

- The coherence between input and output signal is sufficiently close to zero. In this context, coherence is specified as the maximum of the normalized cross-correlation function operating on band-pass signals (with bandwidths sufficiently close to those estimated from the human hearing system).
- Both the spectral and temporal envelopes of the output signal are close to those of the incoming signal.
- The outputs of all decorrelators are mutually incoherent according to the same constraints as for their input/output relation.

The decorrelator units are implemented by means of lattice all-pass filters operating in the QMF domain, in combination with spectral and temporal enhancement tools. More information on QMF-domain decorrelators can be found in [21] [2] and a brief description of the enhancement by means of temporal envelope shaping tools is given subsequently.

4.4. Rate/Distortion Scalability

In order to make MPEG Surround useable in as many applications a possible, it covers a broad range of operation points in terms of both side information rate and multi-channel audio quality. Naturally, there is a trade-off between a very sparse parametric description of the signal's spatial properties and the desire for the highest possible sound quality. This is where different applications exhibit different requirements and, thus have their individual optimal "operating points". For example, in the context of multi-channel audio broadcasting with a compressed audio data rate of ca. 192kbit/s, emphasis may be given on achieving very high subjective multi-channel quality and spending up to 32kbit/s of spatial cue side information is feasible. Conversely, an Internet streaming application with a total available rate of 48kbit/s including spatial side information (using e.g. MPEG-4 HE-AAC) will call for a very low side information rate in order to achieve best possible overall quality.

In order to provide highest flexibility and cover all conceivable application areas, the MPEG Surround technology was equipped with a number of provisions for rate/distortion scalability. This approach permits 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 11 and relies on several dimensions of scalability that are discussed briefly in the following.

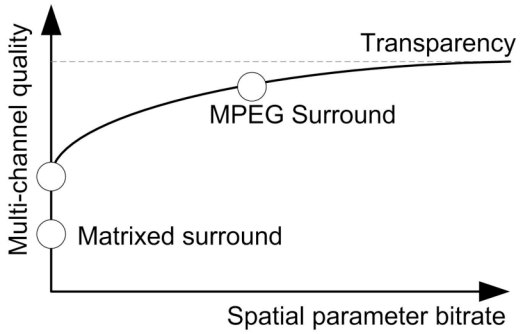

Figure 11: Rate/Distortion Scalability

Several important dimensions of scalability originate from the capability of sending spatial parameters at different granularity and resolution:

- Parameter frequency resolution
  One degree of freedom results from scaling the frequency resolution of spatial audio processing. While a high number of frequency bands ensures optimum separation between sound events occupying adjacent frequency ranges, it also leads to a higher side information rate. Conversely, reducing the number of frequency bands saves on spatial overhead and may still provide good quality for most types of audio signals. 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 large range of update rates and also allows to adapt the temporal grid dynamically to the signal structure.
- Parameter quantization resolution
  As a third possibility, different resolutions for transmitted parameters can be used. Choosing a coarser parameter representation naturally saves in spatial overhead at the expense of losing some detail in the spatial description. Using low-resolution parameter descriptions is accommodated by dedicated tools, such as the *Adaptive Parameter Smoothing* mechanism (as outlined in Section 5.6.4).
- Parameter choice
  Finally, 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 parameter frequency band.

Furthermore, in order to not be limited in audio quality by the parametric model used to describe the multi-channel signal, a residual coding element is available, that enables the MPEG Surround system to offer quality at the level of discrete multi-channel coding algorithms. The residual coding tool is outlined in Section 5.5.

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

In addition, MPEG Surround also supports a matrix surround mode called *Enhanced Matrix Mode* which will be described in more detail in Section 5.2.

4.5. Low Power Processing

The MPEG Surround decoder can be implemented in a High Quality (HQ) version and a Low Power (LP) version. The LP version is realized by simplifying the most computationally intensive modules of the HQ version, i.e. the QMF filterbanks and the decorrelators. Both versions operate on the same data streams, but the LP version consumes considerably less computational power.

The HQ version employs complex-valued QMF analysis and synthesis filterbanks to perform time/frequency transforms. The LP version halves the complexity by using real-valued QMF filterbanks. Since the real-valued QMF filterbanks are critically sampled, real-valued hybrid data at parameter band borders are susceptible to aliasing when they are independently modified by spatial parameters of large difference. The aliasing problem is especially pronounced if it occurs in the low frequency portion of signal spectrum, or if the signal at a parameter band border has a tonal characteristic.

To alleviate the aliasing effect, the low frequency portion of the QMF data is converted to complex values before spatial synthesis, and converted back to real values before QMF synthesis. This is achieved by connecting a 'real to complex' converter to the low frequency output from the QMF analysis filterbank, and a 'complex to real' converter to the low frequency input to the QMF synthesis filterbank. These new 'partially complex' filterbanks significantly suppress aliasing in the low frequency spectrum. This is displayed in Figure 12 where the real-valued QMF analysis is followed by a real-to-complex converter for the lower subbands, and a delay for the higher subbands. The real-to-complex converter creates a complex-valued signal from the real-valued signal by means of filtering in this way introducing a delay. Hence, the higher subbands not subdued to the real to complex processing need to be delayed as well in order to maintain synchronization between the low frequency part and the high frequency part of the signal. A similar procedure is done prior to the real-valued QMF synthesis.

For the high frequency portion of QMF data, the tonality of signals at parameter band borders is estimated. If the signal is found to be potentially tonal at a parameter band border, the spatial parameters of the adjacent bands are replaced by the average of both. This processing for the real-valued QMF subbands is more or less identical to that used in the Low Power version of SBR in High Efficiency AAC [22]

While the HQ version employs high quality Lattice IIR decorrelators (optionally including fractional delays) to achieve perceptual separation among the upmix signals, the LP version employs a mixture of the real-valued version of the above decorrelators and the low complexity decorrelators used in Parametric Stereo [21]. For certain decoder configurations where the downmix signal is mono, a simpler decorrelator structure is used, where the pre- and post- matrixing operations are merged to facilitate the re-use of some decorrelated signals in the spatial synthesis process.

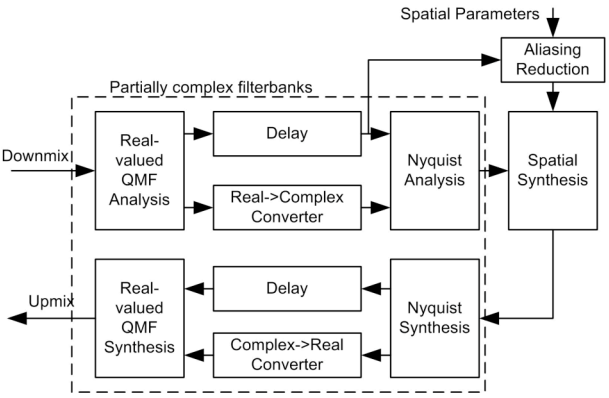

Figure 12: Low Power MPEG Surround

5. ADDITIONAL SYSTEM FEATURES

5.1. Matrix Surround Compatibility

Besides a mono or conventional stereo downmix, the MPEG Surround encoder is also capable of generating a matrix-surround (MTX) compatible stereo downmix signal. This feature ensures backward-compatible 5.1 audio playback on decoders that can only decode the stereo core bitstream (i.e., without the ability to interpret the spatial side information) but are equipped with a matrix-surround decoder. Moreover, this feature also enables the so-called 'Enhanced Matrix' MPEG

Surround mode (i.e., a mode without transmission of spatial parameters as side information), which is discussed further in the next subsection. Special care was taken to ensure that the perceptual quality of the parameter-based multi-channel reconstruction is not affected by this 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 13.

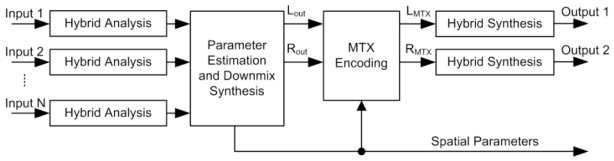

Figure 13: MPEG Surround encoder with post-processing for matrix-surround (MTX) compatible downmix

The MTX-enabling post-processing unit operates in the QMF-domain on the output of the downmix synthesis block (i.e., working on the signals L_{out} and R_{out}) and is controlled by the encoded spatial parameters. Special care is taken to ensure that the inverse of the post-processing matrix exists and can be uniquely determined from the spatial parameters. Finally, the matrix-surround compatible downmix (L_{MTX}, R_{MTX}) is converted to the time domain using QMF synthesis filterbanks. In the MPEG Surround decoder, the process is reversed, i.e. a complementary pre-processing step is applied to the downmix signal before entering into the upmix process.

There are several advantages to the scheme described above. Firstly, the matrix-surround compatibility comes without any additional spatial information (the only information that has to be transmitted to the decoder is whether the MTX-processing is enabled or disabled). Secondly, the ability to invert the matrix-surround compatibility processing guarantees that there is no negative effect on the multi-channel reconstruction quality. Thirdly, the decoder is also capable of generating a 'regular' stereo downmix from a provided matrix-surround-compatible downmix. Last but not least, this feature enables an operation mode where MPEG Surround can operate without the use of spatial side information (see below).

5.2. Enhanced Matrix Mode

In some application scenarios, the transmission of spatial side information is undesirable, or even impossible. For example, a specific core coder may not provide the possibility of transmitting an additional parameter stream. Also in analog systems, transmission of additional digital data can be cumbersome. Thus, in order to broaden the application range of MPEG Surround even further, the specification also provides an operation mode that does not rely on any explicit transmission of spatial parameters. This mode is referred to as 'enhanced matrix mode' and uses similar principles as matrix-surround systems to convey multi-channel audio.

The MPEG Surround encoder is used to generate a matrix-surround compatible stereo signal (as described previously in the section on matrix-surround compatibility). Alternatively, the stereo signal may be generated using a conventional matrix-surround encoder. The MPEG Surround decoder is then operated without externally provided side information. Instead, the parameters required for spatial synthesis are derived from an analysis stage working on the received downmix. In particular, these parameters comprise Channel Level Difference (CLD) and Inter-channel Cross Correlation (ICC) cues estimated between the left and right matrix-surround compatible downmix signals. Subsequently, these downmix parameters are mapped to spatial parameters according to the MPEG Surround format, which can then be used to synthesize multi-channel output by an MPEG Surround spatial synthesis stage. Figure 14 illustrates this concept. The MPEG Surround decoder analyzes the downmix and maps the downmix parameters to the parameters needed for the spatial synthesis.

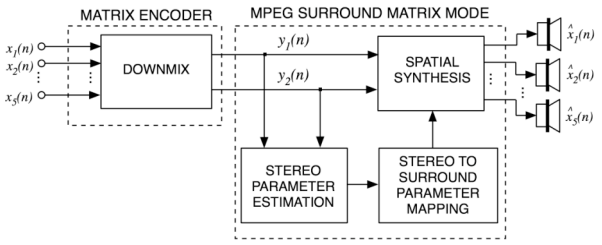

Figure 14: MPEG Surround decoding without side information

The enhanced matrix mode thus extends the MPEG Surround operating range in terms of 'bitrate versus quality'. This is illustrated in Figure 10. Without transmission of spatial parameters, the enhanced matrix mode outperforms conventional matrix-surround systems in terms of quality (see also Section 6).

5.3. Artistic Downmix Handling

Contemporary consumer media of multi-channel audio (DVD-Video/Audio, SA-CD etc.) in practice deliver both dedicated multi-channel and stereo audio mixes that are separately stored on the media. Both mixes are created by a sound engineer who expresses his artistic creativity by 'manually' mixing the recorded sound sources using different mixing parameters and audio effects. This implies that a stereo downmix, such as the one produced by the MPEG Surround coder (henceforth referred to as spatial downmix), may be quite different from the sound engineer's stereo downmix (henceforth referred to as artistic downmix).

In the case of a multi-channel audio broadcast using the MPEG Surround coder, there is a choice as to which downmix to transmit to the receiver. Transmitting the spatial downmix implies that all listeners not in the possession of a multi-channel decoder would listen to a stereo signal that does not necessarily reflect the artistic choices of a sound engineer. In contrast to matrix-surround systems, however, MPEG Surround allows the artistic downmix to be transmitted and thus guarantees optimum sound quality for stereo listeners. In order to minimize potential impairments of the reproduced multi-channel sound resulting from using an artistic downmix signal, several provisions have been introduced into MPEG Surround which are described subsequently.

A first layer of decoder parameters transforms the artistic downmix such that some of the statistical properties of the transformed artistic downmix match those of the MPEG Surround downmix. Additionally, a second enhancement layer transforms the artistic downmix such that a waveform match with the spatial downmix is achieved.

A match of the statistical properties is obtained by computing frequency dependent gain parameters for each downmix channel at the encoder side. These parameters match the energy of the artistic downmix channels to the energy of the corresponding spatial downmix channels. These so-called Artistic Downmix Gains (ADGs) employ the same time-frequency grid as the spatial parameters.

In order to obtain a (complete or bandlimited) waveform reconstruction of the MPEG Surround downmix in the decoder, the encoder computes enhancement signals. These signals are very similar to the residual signals in the spatial synthesis in the sense that they complement the parametric reconstruction by the ADGs, obtaining a waveform match. Therefore, these enhancement signals, or artistic downmix residuals, are coded as residual signals (see Section 5.5 for a more in-depth description of residual coding). The support for artistic downmix residuals is not part of the "Baseline MPEG Surround Profile", as outlined in Section 5.7.3.

5.4. MPEG Surround 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. Especially for mobile devices (such as mobile DVB-H receivers), this extension is of significant interest.

Two distinct use-cases are supported. The first use case is referred to as 'binaural decoding'. In this case, a conventional MPEG Surround downmix / spatial parameter bitstream is decoded using a so-called 'binaural decoding' mode. This mode generates a stereo signal that evokes a (virtual) multi-channel audio experience when played over legacy stereo headphones.

In the second use case, referred to as '3D', the binaural rendering process is applied at the *encoder* side. As a result, legacy stereo devices will automatically render a virtual multi-channel setup over headphones. If the same (3D) bitstream is decoded by an MPEG Surround decoder attached to a multi-channel loudspeaker system, the transmitted 3D downmix can be converted to (standard) multi-channel signals optimized for loudspeakers.

Within MPEG Surround, both use cases are covered using a new technique for binaural synthesis. Conventional binaural synthesis algorithms typically employ Head-Related Transfer Functions (HRTFs). These transfer functions describe the acoustic pathway from a sound source position to both ear drums. The

synthesis process 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 intermediate step;
- It is virtually impossible to 'undo' or 'invert' the encoder-side HRTF processing at the decoder (which is needed in the second 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 binaural 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. A more detailed description can be found in [27].

5.4.1. Binaural Decoding in MPEG Surround

The binaural decoding scheme is outlined in Figure 15. 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 QMF analysis filterbank. A binaural synthesis stage generates the (hybrid QMF-domain) binaural output by means of a 2-channels-in, 2-channels-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 filterbank generates the time-domain binaural output signal.

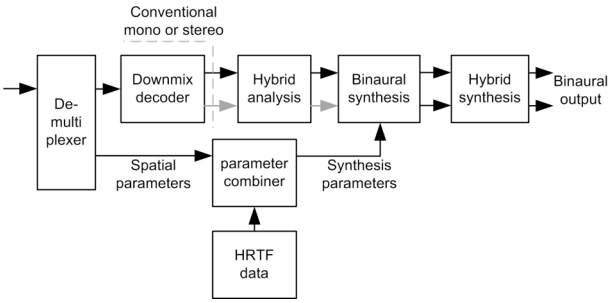

Figure 15: Binaural decoder schematic

In 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 modeled very accurately. The resulting 2x2 synthesis matrix for this mode can have multiple taps in the time (slot) direction. The second mode is a low-complexity mode. In this mode, the 2x2 synthesis matrix has a single tap in the time direction, and is real-valued for approximately 90% of the signal bandwidth. It is especially suitable for low-complexity operation and/or short (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 second use case, as outlined subsequently.

5.4.2. MPEG Surround and 3D-Stereo

In this use case, the 3D processing is applied in the encoder, resulting in a 3D stereo downmix that can be played over headphones on legacy stereo devices. A binaural synthesis module is applied as a post-process after spatial encoding in the hybrid QMF domain, in a similar fashion as the matrix-surround compatibility mode (see Section 3.5). The 3D encoder scheme is outlined in Figure 16. The 3D post-process 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.

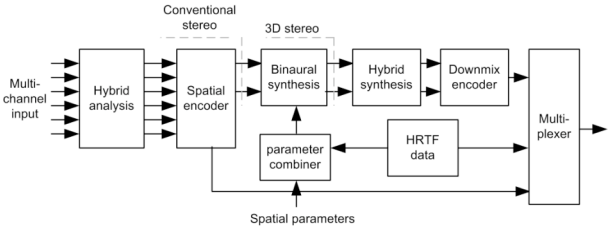

Figure 16: 3D encoder schematic

The corresponding decoder for loudspeaker playback is shown in Figure 17. A 3D/binaural inversion stage operates as pre-process before spatial decoding in the hybrid QMF domain, ensuring maximum quality for multi-channel reconstruction.

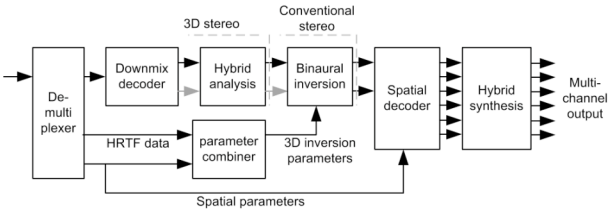

Figure 17: 3D decoder for loudspeaker playback.

5.5. Residual Coding

While a precise parametric model of the spatial sound image is a sound basis for achieving a high multi-channel audio quality at low bitrates, it is also known that parametric coding schemes alone are usually not able to scale up all the way in quality to a 'transparent' representation of sound, as this can only be achieved by using a fully discrete multi-channel coding technique, requiring a much higher bitrate.

In order to bridge this gap between the audio quality of a parametric description and transparent audio quality, the MPEG Surround coder supports a hybrid coding technique. The non-parametric part of this hybrid coding scheme is referred to as residual coding.

As described above, a multi-channel signal is downmixed and spatial cues are extracted. During the process of downmixing, residual signals are calculated which represent the error signal. These signals can be discarded, as their perceptual relevance is rather low. In the hybrid approach, these residual signals, or a band-limited version thereof, are encoded and transmitted to (partially) replace the decorrelated signals in the decoder, as illustrated in Figure 18. This provides a waveform match between the original and decoded multi-channel audio signal for the transmitted bandwidth.

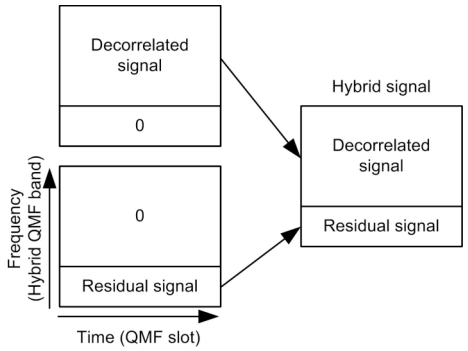

Figure 18: The complementary decorrelated and residual signals are combined into a hybrid signal

The (band-limited) residual signals are encoded by an encoder conforming to MPEG-2 AAC low-complexity profile [23]. The resulting AAC frames are embedded in the spatial bitstream as individual channel stream elements which is illustrated in Figure 19. Transients in the residual signals are handled by utilizing the AAC block switching mechanism and Temporal Noise Shaping (TNS) [24]. For arbitrary downmix residuals (Section 5.3), channel pair elements as defined for MPEG-2 AAC low-complexity profile [23] are used as well.

A trade-off between bitrate increase and multi-channel audio quality can be made by selecting appropriate residual bandwidths and corresponding bitrates. After encoding, the MPEG Surround bitstream is scalable in the sense that the residual signal related data can be stripped from the bitstream, thus lowering the bitrate, such that an MPEG Surround decoder reverts back to the fully parametric operation (i.e., using decorrelator outputs for the entire frequency range).

Listening test results have shown that a significant quality gain is obtainable by utilizing residual signals.

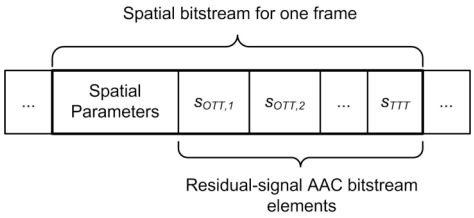

Figure 19: Embedding of residual-signal bitstream elements for each OTT and TTT element in the spatial audio bitstream

5.6. Other Tools

5.6.1. Quantization and Lossless Coding

MPEG Surround provides quantization mechanisms tailored individually to each type of parameter in combination with a set of sophisticated entropy coding techniques, thus striving to minimize the amount of spatial side information while at the same time offering best possible multi-channel audio quality.

Depending on the parameter type, specific quantization schemes are used. Whereas e.g. the Inter-Channel Correlation (ICC) parameters are quantized with as few as eight quantization steps, the Channel-Prediction Coefficients (CPC) used in the TTT box call for a much higher precision of up to 51 quantization steps. Both uniform quantization (e.g. for CPCs), and non-uniform quantization are used (e.g. for Channel Level Differences, CLD) for some parameter types. Some of the quantization mechanisms can be switched in their resolution to accommodate further reduction of the side information in case of very low bitrate application scenarios. For instance it is possible to switch to a coarse CPC quantization mode, where the coder only makes use of every second quantization step (see also subsection on rate/distortion scalability).

For further saving of side information, entropy coding is applied to the majority of the quantized parameters, generally as a combination of differential coding and Huffman coding (only the Guided Envelope Shaping tool makes use of a combination of run length coding and Huffman coding). Differential coding is conducted relative to neighbor parameter values in either frequency or in time direction. Differential coding over time can be conducted relative to the value in the predecessor frame or between values within the same frame. A special case of differential coding is the so-called 'pilot based coding', where the differences of a whole set of parameters are calculated relative to one separately transmitted constant value called the 'pilot value'.

The differentially encoded coefficients are subsequently entropy coded using one- or two-dimensional Huffman code books. In case of a two-dimensional Huffman table, the pair of values which is represented by one Huffman code word belong to parameters neighboring in either frequency or time direction. In order to keep the code books as small as possible, symmetries within the statistical distribution of the differentially encoded parameters are exploited by applying the same Huffman code words to groups of equally likely parameter tuples. Each of these entropy coding schemes can be combined with any of the differential encoding schemes mentioned above, and there are separate Huffman code books trained for every parameter type and coding scheme.

Finally, for the rare case of outlier signal statistics where none of these entropy coding schemes results in a sufficiently low bit consumption, a 'grouped PCM coding' scheme is available as a fall-back strategy which consumes a fix average number of bits per transmitted coefficient (which is only determined by the number of quantization steps). Thus an upper limit of the side information bitrate per MPEG Surround frame can be guaranteed.

5.6.2. Subband Domain Temporal Processing (STP)

In order to synthesize decorrelation between output channels a certain amount of diffuse sound is generated by the spatial decoder's decorrelator units and mixed with the 'direct' (non-decorrelated) sound. In general, the diffuse signal temporal envelope does not match the 'direct' signal envelope resulting in a temporal smearing of transients.

For low bitrate applications, the Subband Domain Temporal Processing (STP) tool can be activated to render temporal shaping since only a binary shaping decision needs to be coded for each upmix channel. STP mitigates the afore-mentioned 'smeared transient' effect by shaping the envelope of the diffuse signal portion of

each upmix channel to approximately match the temporal shape of the transmitted downmix signal.

A schematic of the tool is shown in Figure 20. If STP is activated, the 'direct' and the 'diffuse' signal contribution to the final upmix signal is synthesized separately. To compute the shaping factors for the diffuse signal portion, the temporal energy envelope of the downmixed direct portion of the upmix signal and the temporal energy envelope of the diffuse portion of each upmix channel are estimated. The shaping factors are computed as the ratio between the two energy envelopes. The shaping factors are subjected to some post-processing to achieve a compromise between restoring a 'crisp' transient effect and avoiding potential distortions, before being applied to the high frequency part of the diffuse signals. Finally, each shaped diffuse signal is combined with its corresponding direct signal to reconstruct the particular upmix channel.

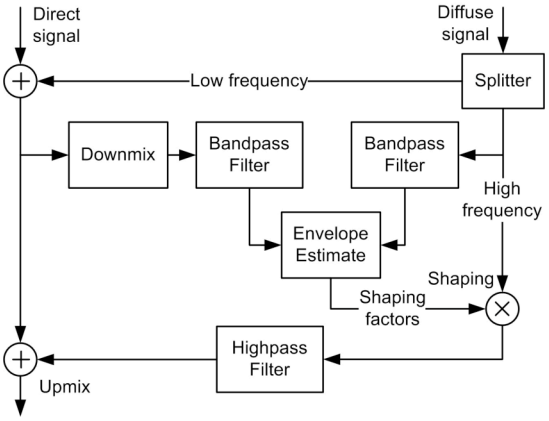

Figure 20: Subband Temporal Processing (STP)

5.6.3. Guided Envelope Shaping (GES)

While the STP tool is suitable to enhance the subjective quality of, for example, applause-like signals, it is still subject to some limitations:

- The spatial re-distribution of single, pronounced transient events in the soundstage is limited by the temporal resolution of the spatial upmix which may span several attacks at different spatial locations.
- The temporal shaping of diffuse sound may lead to characteristic distortions when applied in a rigorous way.

The Guided Envelope Shaping (GES) tool provides enhanced temporal and spatial quality for such signals while avoiding distortion problems. An overview of the GES tool placement in the processing chain is provided in Figure 21.

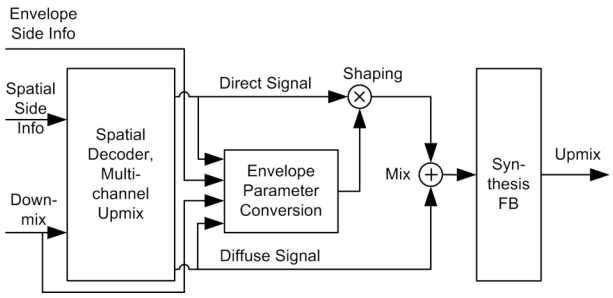

Figure 21: Guided Envelope Shaping (GES)

Additional side information is transmitted by the encoder to describe the broadband fine grain temporal envelope structure of the individual channels, and thus allow for sufficient temporal/spatial shaping of the upmix channel signals at the decoder side. In the decoder, the basic upmix is performed separately for the 'direct' and 'diffuse' signal parts. The associated GES processing only alters the 'direct' part of the upmix signal in a channel, thus promoting the perception of transient direction (precedence effect) and avoiding additional distortion.
Since the diffuse signal also contributes to the overall energy balance of the upmixed signal, GES accounts for this by calculating a modified broadband scaling factor from the transmitted information that is applied solely to the direct signal part. The factor is chosen such that the overall energy in a given time interval is approximately the same as if the original factor had been applied to both the direct and the diffuse part of the signal. Finally, direct and diffuse signal parts are mixed and passed on to the synthesis filterbank stage.

Using GES, best subjective audio quality for applause-like signals is obtained if a coarse spectral resolution of the spatial cues is chosen. In this case, use of the GES tool does not necessarily increase the average spatial side information bitrate, since spectral resolution is advantageously traded for temporal resolution.

5.6.4. Adaptive Parameter Smoothing

For low bitrate scenarios, it is desirable to employ a coarse quantization for the spatial parameters in order to reduce the required bitrate as much as possible. This may result in artifacts for certain kinds of signals. Especially in the case of stationary and tonal signals, modulation artifacts may be introduced by frequent toggling of the parameters between adjacent quantizer steps. For slowly moving point sources, the coarse quantization results in a step-by-step panning rather than a continuous movement of the source and is thus usually perceived as an artifact. The 'Adaptive Parameter Smoothing' tool, which is applied on the decoder side, is designed to address these artifacts by temporally smoothing the dequantized parameters for signal portions with the described characteristics. The adaptive smoothing process is controlled from the encoder by transmitting additional side information.

5.6.5. Channel Configurations

In Section 4.2.3, the tree-based spatial parameterization for the standard 5.1 channel configuration was described. In addition to this, MPEG Surround supports almost arbitrary channel configurations by extending the tree structures shown in Figures 8 and 9 by additional OTT elements. For 7.1 material (with either 4 surround channels or, alternatively, 5 front channels), dedicated 7-2-7 tree configurations are defined using a stereo downmix. Furthermore, there are 7-5-7 trees defined that convey a 5.1 downmix of the original 7.1 channel configuration.

In order to enable flexible support of arbitrary channel configurations, a so-called "arbitrary tree" extension mechanism is available in MPEG Surround that can extend any of the previously defined tree configurations by additional sub-trees built from only OTT modules. For reasons of complexity, decorrelation is not included in these additional OTT modules.

When considering playback of MPEG Surround coded material on a device that only provides stereo output (e.g. a portable music player), a stereo downmix, if available, can be played back directly without any MPEG Surround decoding. In case of a 5-1-5 configuration using a mono downmix, it is however more desirable to play back a stereo dowmix of the coded 5.1 signal than to play just the mono downmix. In order to avoid the computational complexity of a full MPEG Surround decoding to 5.1 channels, followed by an downmix from this 5.1 signal to stereo, a dedicated "stereo output" decoding mode is defined for MPEG Surround coded material using a 5-1-5 configuration. In this mode, spatial parameters of a "virtual" stereo downmix are calculated directly in the parameter domain, based on the received spatial side information. This results in just a single pair of CLD and ICC parameters for one OTT module that generates the desired stereo output from the mono downmix, similar to plain Parametric Stereo coding.

5.7. System Aspects

5.7.1. Carriage of Side Information

In order to achieve backward compatibility with legacy devices that are not aware of MPEG Surround, the spatial side information needs to be embedded in the transmitted downmix signal in a backwards compatible manner, such that the downmix itself can still be decoded by the devices. Depending on the technology used to code and transmit the downmix, different embedding mechanisms for MPEG Surround are defined.

In case of the downmix being coded with MPEG-1/2 Layer I/II/III perceptual audio coders, the spatial side information is embedded in the ancillary data part of the downmix bitstream. Detection of the embedded spatial side information is achieved by means of a dedicated syncword, and the consistency of the embedded data is verified using a CRC mechanism.

In case of the downmix being coded with MPEG-2/4 AAC, the extension_payload data container of the fill_elements available in the AAC bitstream syntax is used to identify and convey the spatial side information. If required, the spatial side information associated with one MPEG Surround frame can be split and carried in several fill_elements, and also distributed over several subsequent AAC frames. If HE-AAC (MPEG-4 High Efficiency AAC) [22] is used as the downmix coder, both SBR and MPEG Surround side information can be embedded at the same time.

When MPEG Surround is used in an MPEG-4 Systems environment, the spatial side information can either be embedded in the downmix bitstream (as described

above) or, alternatively, conveyed as a separate elementary stream (ES) that depends on the ES carrying the coded downmix itself.

Furthermore, it is possible to embed the spatial side information as buried data [29] in a PCM waveform representation of the downmix. For this, a randomized version of the spatial side information is embedded in the least significant bits of the PCM downmix. This bitstream embedding is rendered inaudible by employing subtractively dithered noise shaping controlled by the masked threshold of the downmix signal. This PCM buried data technique allows to store MPEG Surround side information e.g. on a regular audio CD, and to transmit it over digital audio interfaces like S/P-DIF and AES/EBU in a backward compatible manner.

While the MPEG Surround standard defines the carriage of the spatial side information for the MPEG family of perceptual audio codecs (and for PCM waveforms), MPEG Surround can be used in combination with any other downmix coding and transmission system. To this end, it is merely necessary to define a way how the spatial side information is embedded in the downmix or conveyed in parallel to the downmix signal.

5.7.2. Efficient Combination with HE-AAC

When MPEG Surround is used in combination with a downmix coded with HE-AAC, it is of interest to note that both systems make use of a 64 band QMF bank of the same type. Therefore, it is possible to connect the MPEG Surround decoder directly to the output of a HE-AAC downmix decoder in the QMF signal domain. This direct connection avoids the QMF synthesis at the output of the HE-AAC decoder and the QMF analysis at the input of the MPEG Surround decoder that would be necessary if both decoders were connected using the normal time domain representation of the downmix signal. In this way, the overall computational complexity of an integrated HE-AAC and MPEG Surround decoder is significantly reduced both for High-Quality and for Low-Power processing.

5.7.3. MPEG Surround Profiles and Levels

MPEG defines decoder profiles as technology sets to be used for certain application fields and ensuring interoperability on a bitstream basis. Currently there is a single “Baseline MPEG Surround Profile” available.

The following tools are contained in this profile:

- artistic downmix functionality
- matrix compatibility
- enhanced matrix mode decoding
- temporal shaping
- residual coding (not including artistic downmix residuals)
- binaural decoding
- 3D audio decoding
- low power decoding

This profile is further subdivided into five hierarchical levels that come with increasing number of output channels, range of sampling rates and bandwidth of the residual signal when advancing to a higher level. The following table lists the supported tree configurations, maximum number of audio output channels, availability of residual coding and complexity for each level.

Level	Tree config	Max. output channels	Fs max	Residual coding
1	515, 525	2.0	48 kHz	n/a
2	515, 525	5.1	48 kHz	n/a
3	515, 525	5.1	48 kHz	Yes
4	515, 525, 757, 727	7.1	48 kHz	Yes
5	515, 525, 757, 727, plus arbitrary tree extension	32 incl. 4 LFE	96 kHz	Yes

Table 1: Levels of Baseline Profile

Level 1 allows for stereo output of multi-channel content with and without the use of binaural rendering techniques. It is ideally suited for mobile devices that are equipped with stereo loudspeakers or headphones and exhibits very low computational complexity. This

level includes the ability to upmix from a mono based operation to stereo loudspeaker reproduction, the 5-1-2 mode with 5-1-5 streams. Levels 2 and 3 additionally offer discrete 5.1 output for loudspeakers. Level 4 extends to 7.1 output, while Level 5 can handle up to 32 channels and operate at up to 96 kHz sampling rate.

Independently from this level hierarchy, an MPEG Surround decoder can be implemented either as a high quality (HQ) decoder or as a low power (LP) decoder in order to optimally support both stationary equipment as well as battery powered mobile devices.

The worst-case decoding complexity in terms of Processor Complexity Units (PCU) and RAM Complexity Units (RCU) is listed in the following table for the different levels. Processor Complexity Units are specified in MOPS, and RAM Complexity Units are expressed in kWords (1000 words). Both High Quality (HQ) and Low Power (LP) operation is shown.

Level	PCU (HQ)	RCU (HQ)	PCU (LP)	RCU (LP)
1	12	5	6	4
2	25	15	12	11
3	25	15	12	11
4	34	21	17	15
5 (fs = 48 kHz)	70	38	44	32

Table 2: Processor and RAM complexity depending on decoder level and operation

For the combination of MPEG Surround decoding with a HE-AAC core decoder, the complexity savings mentioned in the previous section apply. This leads to low total complexity numbers for a combined 525 system (in PCU/RCU assuming Level 2):

High Quality: 28/23 (saves 6/2)
Low Power: 15/17 (saves 4/2)

Here, the Low Power operation for a combined HE-AAC + MPEG Surround decoder comes with a significant reduction in processing complexity of almost 50% when compared to High Quality operation. The performance differences between both modes are discussed in the next section.

6. PERFORMANCE

In the period between August 2006 and January 2007, MPEG conducted a formal verification test for the MPEG Surround specification [25]. Such test is not only a validation of the standardized technology but also provides relevant test data on MPEG surround to the interested industry. For this purpose, two use-cases relevant to industry were considered, i.e. a DVB oriented use-case, and a music-store / portable player use-case. An additional test was included to evaluate the difference between the High Quality (HQ) and Low Power (LP) decoding modes of MPEG Surround. A total of 148 subjects at 8 different test sites participated in a total of 5 different tests. All tests were conducted conforming to the MUSHRA [26] test methodology and included a reference and a 3.5kHz low pass anchor.

6.1. DVB Oriented Use-case

The test conditions for the DVB oriented use-case strive to compare MPEG Surround in DVB-like applications with other potential MPEG Audio solutions for that space. With reference to labels to the results for this multi-channel test in Figure 22, two scenarios are covered:

- The establishment of a new service, for which a stereo backwards compatible part at highest quality per bit is desired. For this scenario, MPEG-4 High Efficiency AAC in combination with MPEG Surround (HE-AAC_MPS) is compared with discrete multi-channel MPEG-4 HE AAC (HE-AAC_MC) at bitrates of 64 and 160kbit/s total.

- The extension of an existing stereo service with surround sound in a backwards compatible fashion. For this scenario, MPEG-1 Layer 2 (as is currently employed in DVB) is extended with MPEG Surround (Layer2_MPS) at a total rate of 256kbit/s. This mode is compared to Dolby Prologic II encoding and decoding using MPEG-1 Layer 2 at 256kbit/s for coding the downmix (Layer2_DPL2).

In addition to this multi-channel test a corresponding downmix test has been performed to verify the MPEG Surround downmix quality.

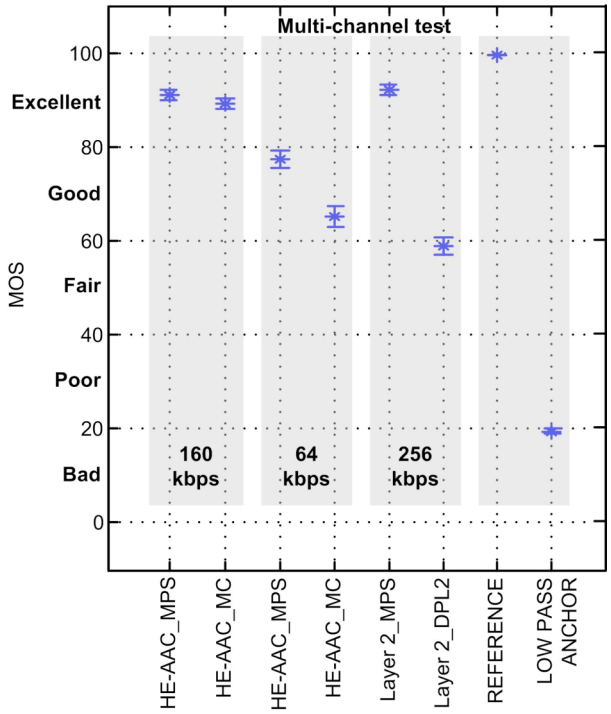

Figure 22: Test results for the DVB oriented use-case

From the results in Figure 22 it can be concluded that MPEG Surround provides improved coding efficiency at low bitrates and same quality, while including backwards compatible downmix at higher bitrates. This makes it a perfect candidate for upgrading existing DVB broadcasting services to surround sound. In addition MPEG Surround is vastly superior to the matrixed system at the same total bitrate.

6.2. Music-store / Portable Player Use-case

The test conditions for this use-case strive to test the performance of MPEG Surround in "music-store-like" applications. With reference to the labels to the results for this multi-channel test in Figure 23, a scenario is envisioned where an online music store selling stereo material using AAC at 160kbit/s, extends the service to provide multi-channel content in a backwards compatible way. This is achieved by additionally providing low bitrate MPEG Surround data (AAC_MPS). Hence, a consumer can play the backwards compatible part on his/her legacy stereo player, while enjoying a multi-channel experience when playing the content at home using an MPEG Surround enabled multi-channel setup.

This point of operation is compared to MPEG Surround decoding of the AAC_MPS bitstreams in enhanced matrix mode (AAC_MPS_EMM, not employing the spatial side information) and Dolby Prologic encoding and decoding using AAC at 160kbit/s for coding the downmix (AAC_DPLII).

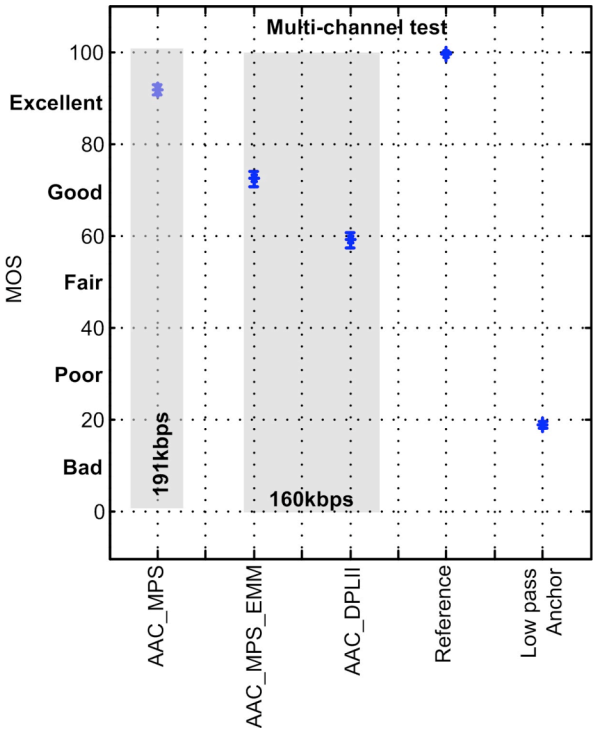

Figure 23: Test results for the music-store / portable player use-case

From the results in Figure 23 it is concluded that MPEG Surround at 192kbit/s provides an excellent multi-channel quality thus being very attractive for upgrading existing electronic music store services. In the enhanced Matrix decoding mode, where the side information is not exploited, the results are still well into the "good" quality range. Nevertheless, MPEG Surround enhanced matrix mode shows to be superior to legacy matrix technology (Dolby Prologic II).

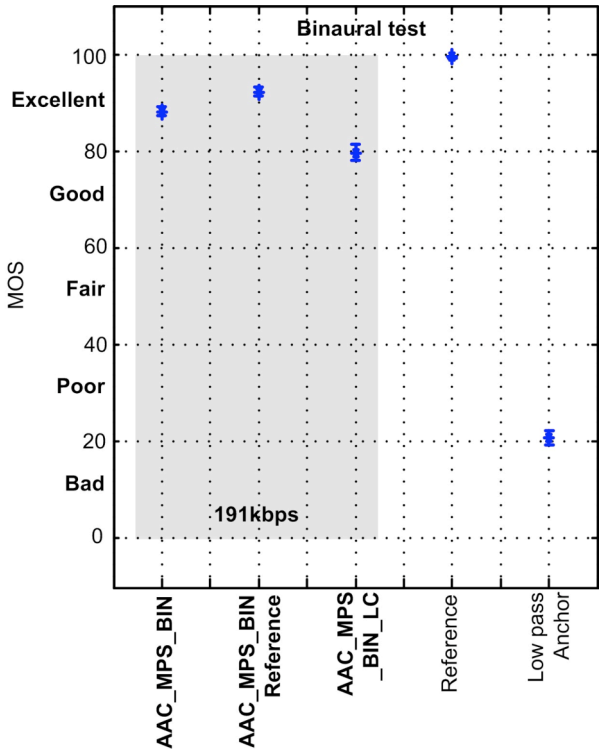

Figure 24: Binaural test results for the music-store / portable player use-case

An additional headphone test has been conducted in order to assess the quality of the binaural decoding capability of MPEG Surround, particularly for this music-store use case. With reference to the labels to the results for this binaural test in Figure 24, three configurations were tested. High quality binaural decoding (AAC_MPS_BIN) and low complexity binaural decoding combined with low power MPEG Surround decoding (AAC_MPS_BIN_LC) at a total of 191kbit/s for stereo head-phone listening at home. Reference signal have been created by applying HRTF filtering to the MPEG Surround multi-channel output (AAC_MPS_BIN_Reference) and the multi-channel original (Reference). The Low pass anchor is the 3.5kHz bandwidth limitation of the 'Reference'. Note that residual signals are not employed in binaural decoding mode.

The results in Figure 24 show that MPEG Surround binaural rendering offers excellent quality.

6.3. Additional Tests

An additional 'technology driven' test shows that the sound quality of the LP version is statistically comparable to the sound quality of the HQ version. Moreover, as shown in Table 2, the LP version requires only about half of the power consumption and three quarters of the RAM requirement of the HQ version.

A number of further test results from earlier evaluations of other operating points within the wide range of possible MPEG Surround operating conditions can be found e.g. in [28].

7. APPLICATIONS

MPEG Surround enables a wide range of applications, Among the many conceivable applications are music download services, streaming music services / Internet radios, Digital Audio Broadcasting, multi-channel teleconferencing and audio for games.

MPEG Surround is perfectly suited for digital audio broadcasting. Given its inherent backwards compatibility and low overhead of the side information, MPEG Surround can be introduced in existing systems without rendering legacy receivers obsolete. Legacy and stereo-only receivers will simply play the stereo downmix, while an MPEG Surround enabled receiver will play the true multi-channel signal. Test transmissions using over the air transmission of MPEG Surround and Eureka Digital Audio Broadcasting (DAB) have been carried out at several occasions, as well as test transmissions using MPEG Surround with HDC (the US terrestrial digital radio system).

Given the ability to do surround sound with low overhead and the ability to do binaural rendering, MPEG Surround is ideal for the introduction of multi-channel sound on portable devices. As one very attractive example, an Internet music store can upgrade its content to surround, and - using an MPEG Surround enabled portable player - the user can get a surround sound experience over headphones. When connecting the portable player to a surround sound AV receiver by means of a cradle or docking station, the content can be played over speakers in true surround sound. As is always the case with MPEG Surround, legacy devices

will not take notice of the MPEG Surround data and play the backwards compatible downmix.

The existing DVB-T system uses MPEG-1 Layer-2, and can only be expanded to multi-channel by means of simulcast, at high bitrate penalty, or by matrix-surround systems, resulting in low multi-channel audio quality. MPEG Surround effectively solves these problems. An MPEG Surround decoder can be introduced in the set-top box, decoding the Layer 2 downmix and the MPEG Surround data into a multi-channel signal. This enables a DVB-T system to be upgraded to surround sound by simply offering new MPEG Surround enabled set-top boxes to customers who wishes to have surround sound TV. Legacy receivers will decode the Layer 2 downmix as usual, not affected by the MPEG Surround data. Test-transmissions have successfully been carried out using MPEG Surround over DVB-T.

8. CONCLUSIONS

After several years of intense development, the Spatial Audio Coding approach has proven to be extremely successful for bitrate-efficient and backward compatible representation of multi-channel audio signals. Based on these principles, the MPEG Surround technology has been under standardization within the ISO/MPEG group for about two years. This paper describes the technical architecture and capabilities of the MPEG Surround technology and its most recent extensions.

Most importantly, MPEG Surround enables the transmission of multi-channel signals at data rates close to the rates used for the representation of two-channel (or even monophonic) audio. It allows for a wide range of scalability with respect to the side information rate, which helps to cover almost any conceivable application scenario. Listening tests confirm the feasibility of this concept: Good multi-channel audio quality can be achieved down to very low side information rates (e.g. 3kbit/s). Conversely, using higher rates allows approaching the audio quality of a fully discrete multi-channel transmission. Along with the basic coding functionality, MPEG Surround provides a plethora of useful features that further increase its attractiveness (e.g. support for artistic downmix, full matrix-surround compatibility, binaural decoding) and may promote a quick adoption in the marketplace.

Finally, MPEG Surround enables the use of multi-channel audio on portable devices due to the low overhead of the spatial data and the binaural rendering capabilities.

9. ACKNOWLEDGEMENTS

A great number of people made this project a reality. The following persons have contributed significantly to the development of MPEG Surround:

- J. Engdegård, Coding Technologies
- A. Gröschel, Coding Technologies
- L. Sehlström, Coding Technologies
- L. Villemoes, Coding Technologies

- B. Grill, Fraunhofer IIS
- A. Hölzer, Fraunhofer IIS
- M. Multrus, Fraunhofer IIS
- H. Mundt, Fraunhofer IIS
- M. Neusinger, Fraunhofer IIS
- J. Plogsties, Fraunhofer IIS
- C. Spenger, Fraunhofer IIS
- R. Sperschneider, Fraunhofer IIS
- L. Terentiev, Fraunhofer IIS

- T. Norimatsu, Panasonic

- E. Schuijers, Philips
- M. Ostrovskyy, Philips
- G. Hotho, Philips
- F. de Bont, Philips
- F. Myburg, Philips
- M. van Loon, Philips

The authors would furthermore like to express their gratitude to the MPEG Audio Committee and its chairman, Schuyler Quackenbush, for the support provided as well as to the many other contributors to the standards effort.

REFERENCES

[1] J. Herre, C. Faller, S. Disch, C. Ertel, J. Hilpert, A. Hölzer, 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

[2] 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

[3] 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

[4] 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

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

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

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

[8] 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

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

[10] Dolby Publication, Roger Dressler: "Dolby Surround Prologic Decoder – Principles of Operation", http://www.dolby.com/assets/pdf/tech_library/209_Dolby_Surround_Pro_Logic_II_Decoder_Principles_of_Operation.pdf

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

[12] ISO/IEC JTC1/SC29/WG11 (MPEG), Document N6455, "Call for Proposals on Spatial Audio Coding", Munich 2004

[13] ISO/IEC JTC1/SC29/WG11 (MPEG), Document N6813, "Report on Spatial Audio Coding RM0 Selection Tests", Palma de Mallorca 2004

[14] ISO/IEC JTC1/SC29/WG11 (MPEG), Document N7138, "Report on MPEG Spatial Audio Coding RM0 Listening Tests", Busan, Korea, 2005. Available at http://www.chiariglione.org/mpeg/working_documents/mpeg-d/sac/RM0-listening-tests.zip

[15] B. R. Glasberg and B. C. J. Moore. Derivation of auditory filter shapes from notched-noise data. Hearing Research, 47: 103-138 (1990)

[16] J. Breebaart, S. van de Par, A. Kohlrausch. Binaural processing model based on contralateral inhibition. I. Model setup. J. Acoust. Soc. Am. 110:1074-1088 (2001)

[17] J. Princen, A. Johnson, A. Bradley: "Subband/ Transform Coding Using Filter Bank Designs Based on Time Domain Aliasing Cancellation", IEEE ICASSP 1987, pp. 2161 - 2164

[18] M. Dietz, L. Liljeryd, K. Kjőrling, O. Kunz: "Spectral band replication, a novel approach in audio coding", Proc. 112th AES convention, Munich, Germany, May 2002, Preprint 5553

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

[20] G. Hotho, L. Villemoes, J. Breebaart: "A stereo backward compatible multi-channel audio codec", IEEE Trans. on audio, speech & language processing, submitted (2007)

[21] J. Engdegård, H. Purnhagen, J. Rödén, L. Liljeryd: "Synthetic ambience in parametric stereo coding",

Proc. 116th AES convention, Berlin, Germany, 2004, Preprint 6074

[22] M. Wolters, K. Kjörling, D. Homm, H. Purnhagen: "A closer look into MPEG-4 High Efficiency AAC", 115th AES Convention, New York 2003, Preprint 5871

[23] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, Oikawa, "ISO/IEC MPEG-2 Advanced Audio Coding", Journal of the AES, Vol. 45, No. 10, October 1997, pp. 789-814

[24] J. Herre, J. D. Johnston: "Enhancing the Performance of Perceptual Audio Coders by Using Temporal Noise Shaping (TNS)", 101st AES Convention, Los Angeles 1996, Preprint 4384

[25] ISO/IEC JTC1/SC29/WG11 (MPEG), Document N8851, "Report on MPEG Surround Verification Test", Marrakech, Morocco, 2007. Available at http://www.chiariglione.org/mpeg/working_documents/mpeg-d/sac/VT-report.zip

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

[27] J. Breebaart, J. Herre, L. Villemoes, Craig Jin, K. Kjörling, J. Plogsties, J. Koppens: "Multi-Channel Goes Mobile: MPEG Surround Binaural Rendering", 29th AES International Conference, Seoul, Korea, 2006

[28] L. Villemoes, J. Herre, J. Breebaart, G. Hotho, S. Disch, H. Purnhagen, K. Kjörling: "MPEG Surround: The Forthcoming ISO Standard For Spatial Audio Coding", 28th AES International Conference, Piteå, Sweden, 2006

[29] A. W. J. Oomen, M. E. Groenewegen, R. G. van der Waal, and R. N. J. Veldhuis: "A Variable-Bit-Rate Buried-Data Channel for Compact Disc", Journal of the AES, Vol. 43, No. 1/2, January/February, pp. 23-28, 1995

[30] ISO/IEC 23003-1:2006, "Information technology - MPEG audio technologies - Part 1: MPEG Surround", 2003