Audio Engineering Society

Convention Paper

Presented at the 119th Convention
2005 October 7–10 New York, New York USA

MPEG Spatial Audio Coding / MPEG Surround: Overview and Current Status

J. Breebaart[1], J. Herre[2], C. Faller[3], J. Rödén[4], F. Myburg[5], S. Disch[2], H. Purnhagen[4], G. Hotho[1], M. Neusinger[2], K. Kjörling[4], W. Oomen[5]

[1] Philips Research Laboratories, 5656 AA, Eindhoven, The Netherlands

[2] Fraunhofer Institute for Integrated Circuits IIS, 91058 Erlangen, Germany

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

[4] Coding Technologies, 11352 Stockholm, Sweden

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

ABSTRACT

Recently, the MPEG Audio standardization group started a new work item on Spatial Audio Coding. This new approach allows for a fully backward compatible representation of multi-channel audio at bit rates that are only slightly higher than common rates currently used for coding of mono / stereo sound. This paper briefly describes the underlying idea and reports on the current status of the MPEG standardization activities. It provides an overview of the resulting "MPEG Surround" technology, and discusses its capabilities. The current level of performance will be illustrated by listening test results.

1. INTRODUCTION

Recently, a new approach in perceptual coding of multi-channel audio has emerged [1]. This approach, commonly referred to as Spatial Audio Coding (SAC), 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 user features. Firstly, it allows the transmission of multi-channel audio at bitrates, which so far have been used for the transmission of monophonic audio. Secondly, by its underlying structure, the multi-channel audio signal is transmitted in a backward compatible way, i.e., 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.

This paper briefly describes the concepts behind the idea of spatial audio coding and reports on the status of

the ongoing activities of the ISO/MPEG standardization group in this field which recently were renamed into *MPEG Surround*. Specifically, it describes the new MPEG Surround reference model architecture [2], and its manifold capabilities along with some extension work that is currently under development in MPEG. Finally, in order to illustrate the performance of the technology, the results of several recent listening tests are discussed.

2. THE SPATIAL AUDIO CODING APPROACH

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. Also, externally created downmix signals ('artistic downmix') may be used instead. On the decoding side, the transmitted downmix signal is expanded into a high quality multi-channel output based on the spatial parameters.

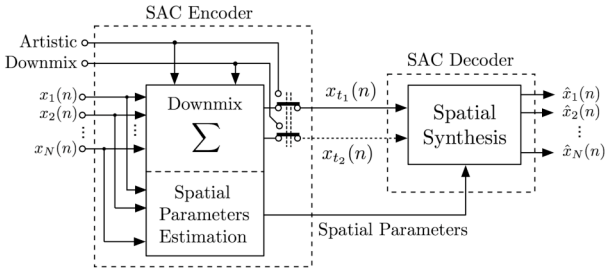

Figure 1 Principle of Spatial Audio Coding

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 [3], a generalization of *Parametric Stereo* [4] [5] to multi-channel application, and an extension of the *Binaural Cue Coding* (BCC) scheme [6] [7] towards using more than one transmitted downmix channel [8]. From a different viewing angle, the Spatial Audio Coding approach may also be considered an extension of well-known matrixed surround schemes (Dolby Surround/Prologic, Logic 7, Circle Surround etc.) [9] [10] by transmission of dedicated (spatial cue) side information to guide the multi-channel reconstruction process and thus achieve improved subjective audio quality [1].

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. around 3 to 32 kb/s. Among the manifold conceivable applications are music download services, streaming music services / Internet radios, Digital Audio Broadcasting, multi-channel teleconferencing and audio for games.

Stimulated by the potential of the SAC approach, the ISO/MPEG standardization group recently started a new work item on SAC by issuing a "Call for Proposals" (CfP) on Spatial Audio Coding in March 2004 [11]. 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 of the subsequent standardization

process, called Reference Model 0 (RM0), will 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) [12]. 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 [13]. Details from this verification test will be presented in the section discussing MPEG Surround performance. The successful development of RM0 sets the stage for the subsequent improvement process of this technology that is currently being carried out collaboratively within the MPEG Audio group.

3. OVERVIEW OF THE MPEG SURROUND REFERENCE MODEL

While a detailed description of the MPEG Surround RM0 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].

3.1. General Structure of SAC Synthesis

The general structure of the MPEG Surround decoder is illustrated in Figure 2, showing a three-step process that converts the downmix signal supplied as input into the multichannel output signal. Firstly, the input signal is decomposed into frequency bands by means of a hybrid QMF analysis filter bank (see below). Next, the multichannel 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 filter bank 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 are combined and converted back to time domain by means of a set of hybrid QMF synthesis filter banks.

The spatial synthesis process is shown in more detail in Figure 3. The input signals are processed by an upmix matrix, where the matrix elements (i.e., gain factors) depend on the transmitted spatial parameters in frequency and time. In addition, decorrelator modules are employed to enable reconstruction of spaciousness in the output signal. Therefore, the upmix matrix is decomposed into a pre-matrix, M1, and a post-matrix, M2.

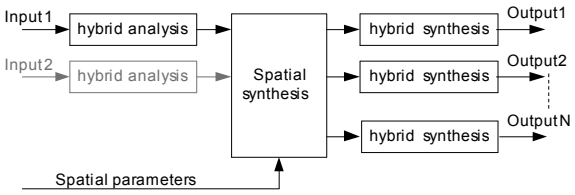

Figure 2 High-level overview of the MPEG Surround synthesis

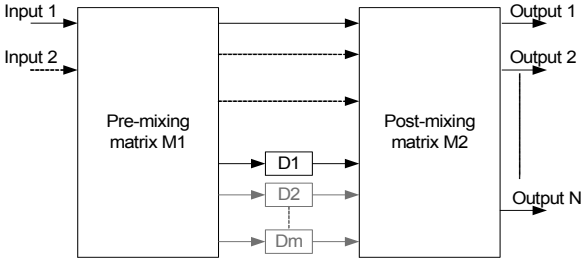

Figure 3 Generalized structure of the spatial synthesis process, comprising two mixing matrices; M1, M2, and a set of decorrelators, D1, D2, … Dm

In the following sections, the hybrid QMF filter banks, the signal flow in the upmix matrix (which can be characterized by a tree structure composed of smaller processing blocks called OTT and TTT), and the decorrelator modules are described in more detail. This is followed by the description of additional tools for temporal envelope shaping and for adaptive parameter smoothing that further enhance the performance of the spatial audio coding system.

3.2. Hybrid QMF Filter Banks

In the human auditory system, the processing of binaural cues is performed on a non-uniform frequency scale [14] [15]. 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 filter bank.

For applications including low bitrate audio coding, the SAC 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 filter banks (for example using a Modified Discrete Cosine Transform (MDCT) [16]) and are not suitable for signal manipulation as this would interfere with the aliasing cancellation properties associated with critically sampled filter banks. The Spectral Band Replication (SBR) algorithm [17] is an important exception in this respect. Similar to the Spatial Audio Coding 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 [5] [18]. Furthermore, by grouping filter bank outputs for spatial parameter analysis and synthesis, the frequency resolution for spatial parameters can be varied extensively while applying a single filter bank configuration. More specifically, the number of parameters to cover the full frequency range can be varied from only a few (for low bitrate applications) up to 40 (for high-quality processing) to closely mimic the frequency resolution of the human auditory system. A detailed description of the hybrid filter bank in the context of MPEG Surround can be found in [2].

3.3. OTT and TTT Elements

Generally speaking, the MPEG Surround approach can be used to map from M to N channels and back again, where N < M. This is possible due to the flexible module-based approach that makes use of two conceptual elements, i.e. the One-To-Two (OTT) element and the Two-To-Three (TTT) element where the names imply the number of input and output channels of the corresponding decoder element. For better understanding, the corresponding encoder elements and combinations thereof are discussed first.

3.3.1. OTT Encoding

On the encoder side, the OTT encoder element extracts two spatial parameters, and creates a downmix (together with a residual) signal. Thus a mono downmix signal and spatial parameters are output from a stereo input signal while the residual signal is discarded. The OTT element has a history from Parametric Stereo (PS, [5] [18]) and Binaural Cue Coding (BCC, [6] [7]). The following spatial parameters are extracted in a time- and frequency-varying grid.

- 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.

- 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 of the parameterization and enables full waveform reconstruction at the decoder side (see section on residual coding).

3.3.2. TTT Encoding

In analogy to the OTT encoder element, the TTT encoder element mixes down three audio signals into two output channels, i.e. a stereo downmix (plus a residual signal).

$$\begin{bmatrix} l_0 \\ r_0 \end{bmatrix} = H_{TTT} \begin{bmatrix} l \\ c \\ r \end{bmatrix}$$

In addition, it extracts two parameters called Channel Prediction Coefficients (CPC). Conversely, on the decoder side, the TTT element estimates a third channel from two channels and the CPC parameters, which makes it a perfect candidate to extract the center channel from a stereo downmix

This model assumes that the stereo downmix l_0 and r_0 is a linear combination of the three-channel input signal l, c and r. By transmitting two independent CPC parameters, the [l, c, r] signal can be optimally recovered from the stereo downmix signal [l_0, r_0]. Since the original [l ,c, r] signals often only contain partially correlated signals there will be a prediction loss.

The ICC parameter can also be used in the TTT element and will then indicate the amount of prediction loss for the given CPC parameters as additional information. A residual signal can also be used in the TTT element to enable perfect waveform reconstruction at the decoder.

3.3.3. Hierarchical Encoding

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 4 shows a block diagram of an encoder for such a typical system consisting of three OTT and a TTT encoder element. The signals l_f, l_b, c, lfe, r_f and r_b denote the left front, left back, center, LFE, right front and right back channels, respectively.

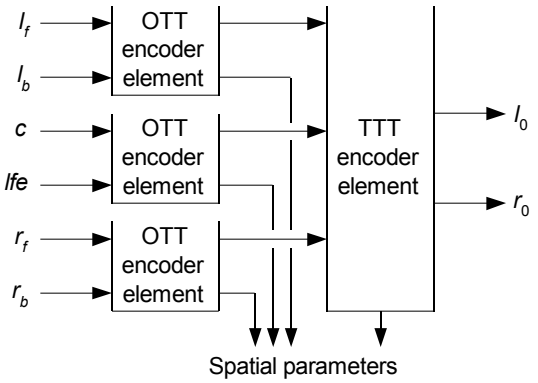

Figure 4 Block diagram of a 5.1-to-stereo MPEG Surround encoder

Several OTT elements can be cascaded and hence easily support surround systems with more channels. Figure 5 exemplifies how OTT elements can be connected in a tree structure, forming a 5.1-to-mono encoder.

Another example is illustrated in Figure 6 where a 7.1 surround signal is encoded into a 5.1 surround signal and spatial information is obtained from two OTT elements. The signals l_b, l_s, l_f, c, lfe, r_f, r_s, r_b denote the left back, left side, left front, center, LFE, right front, right side and right back respectively.

From these examples, it becomes clear how arbitrary downmixing / upmixing configurations can be addressed using OTT and TTT elements.

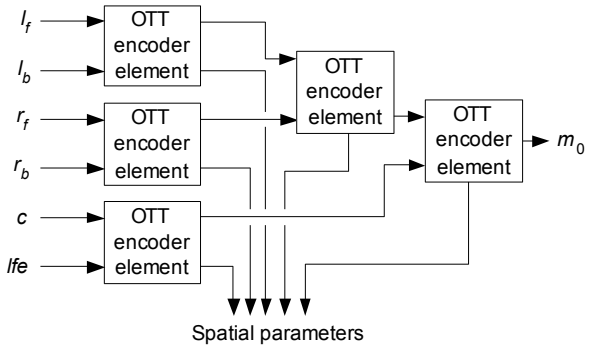

Figure 5 Block diagram of a 5.1-to-mono MPEG Surround encoder

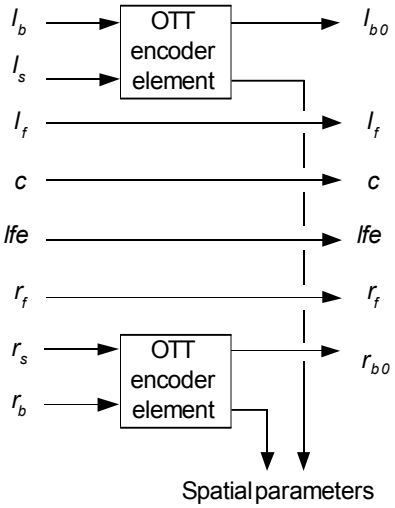

Figure 6 Block diagram of a 7.1-to-5.1 MPEG Surround encoder

3.3.4. Hierarchical Decoding

From a signal flow point of view, the inverse of the encoder is used to create the gain values in the two mixing matrices M1 and M2. In Figure 7 a conceptual block diagram of a stereo-to-5.1 decoder is shown. Each OTT and TTT decoder element contains a decorrelator and hence the order of the OTT/TTT elements in the tree describes how the mixing matrices are structured. The actual gain values for each element in the mixing matrices are calculated by combining the decoded spatial parameters from one or several of the OTT/TTT elements.

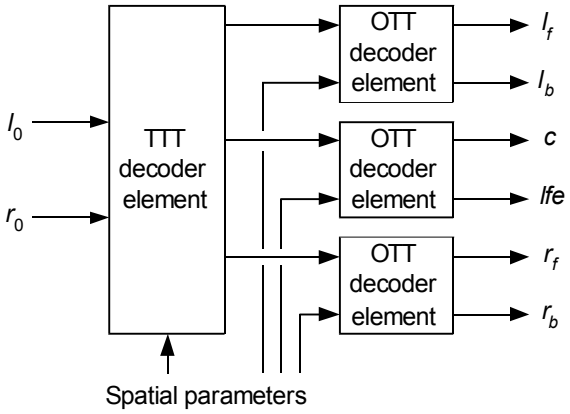

Figure 7 Block diagram of a stereo-to-5.1 MPEG Surround decoder

3.4. 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. 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 signals.

- 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 [19] [2]; a brief description of the enhancement by means of temporal envelope shaping tools is given subsequently.

3.5. Temporal Shaping Tools

In order to synthesize correlation between output channels a certain amount of diffuse sound is generated by the spatial decoder's decorrelator units and mixed with the 'dry' (non-decorrelated) sound. In general, the diffuse signal envelope does not match the 'dry' signal envelope resulting in a weak transient reproduction degraded by pre-echo type of artifacts. The TP and the TES tools are designed to address this problem by shaping the temporal envelope of the diffuse sound.

3.5.1. Time Domain Temporal Processing (TP)

The TP processing operates in the time domain by shaping the diffuse signal to match the temporal envelope of the dry signal. This is accomplished by using the dry signal for deriving a target envelope to be imposed on the diffuse signal. The shaping of the diffuse signal is done at the higher frequency bands only. Therefore a frequency selective splitting of the signal is done in the QMF domain by using a modified upmix ('splitter') providing separate outputs for dry and diffuse signal. Subsequently, these two sets of hybrid subband domain signals are passed through the hybrid synthesis, resulting in two sets of time-domain signals. The first holds the dry signals for the full frequency range combined with the low frequency range of the diffuse signals that does not require temporal shaping. The second signal set holds the high pass filtered diffuse signals, which are subjected to temporal shaping. This is done by estimating the target temporal envelope from suitable dry signals and imposing this envelope on each of the diffuse signals by means of scaling with a smoothed gain function. Finally, the dry and diffuse signal portions of each channel are mixed to form the output. Figure 8 provides a schematic block diagram of the processing steps for TP.

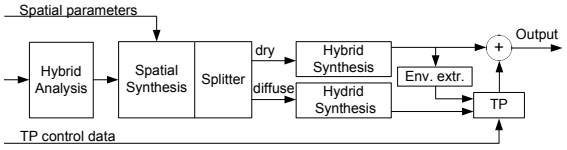

Figure 8 Temporal Processing Tool

3.5.2. Temporal Envelope Shaping (TES)

An alternative way to address the diffuse signal envelope shaping problem is exploited by the temporal

envelope shaping tool (TES): As opposed to TP, the TES approach achieves the same effect by manipulating the diffuse signal envelope in the subband domain representation, analogous to the Temporal Noise Shaping (TNS) [20] [21] known from MPEG-2/4 Advanced Audio Coding (AAC) [22]. By convolving the spectral coefficients of the diffuse signal with a shaping filter derived from an LPC analysis of the spectral coefficients of the dry signal, the envelope of the former is matched to the envelope of the latter. Due to the rather high time resolution of the spatial audio coding QMF filter bank, TES filtering requires only low-order processing and is thus a low computational complexity alternative to the TP tool, yet not offering the full extent of temporal control due to QMF subband processing artifacts.

3.6. 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 can be controlled automatically by the decoder or explicitly from the encoder by transmitting additional side information.

4. GENERAL SYSTEM FEATURES

This section provides a short description of the most salient features of the MPEG Surround RM0 technology.

4.1. Mono vs. Stereo Based Operation

In bandwidth-constrained applications, such as broadcasting, an efficient transmission of program material is of high importance. Given that the spatial side information only amounts to a small fraction of the overall required transmission capacity, the transmission of the stereo downmix signal occupies the major part of the transmission capacity. In this context, MPEG Surround technology offers an interesting option for boosting bandwidth efficiency further: Multi-channel audio output can be obtained even with the transmission of a monophonic downmix signal (which requires considerably less bitrate than a stereo signal). While the perceived multi-channel audio quality for a Spatial Audio Coding system based on a monophonic audio transmission does not reach the level of performance offered by a stereo-based system, the overall quality is still competitive with matrixed surround systems (see section on MPEG Surround performance for recent test results). Note that this is an option, which – by definition – cannot be offered by a matrixed surround approach.

4.2. Matrixed Surround Compatibility

Besides a mono or conventional stereo downmix, the MPEG Surround encoder is also capable of generating a matrixed-surround (MTX) compatible stereo downmix signal. This feature ensures backwards-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 matrixed-surround decoder. Moreover, this feature also enables a so-called 'non-guided' MPEG Surround mode (i.e., a mode without transmission of spatial parameters), which is discussed further in the section on ongoing extension work. Special care was taken to ensure that the perceptual quality of the parameter-based multi-channel reconstruction does not depend on whether the matrixed-surround feature is enabled or disabled. The matrixed-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 9.

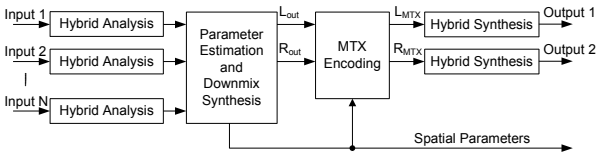

Figure 9 MPEG Surround encoder with post-processing for matrixed-surround (MTX) compatible downmix

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 matrixed-surround compatible downmix (L_{MTX}, R_{MTX}) is converted to the time domain using QMF synthesis filter banks. 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 matrixed-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 matrixed-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 matrixed-surround-compatible downmix. Last but not least, this feature enables a non-guided mode within the MPEG Surround framework (see below).

4.3. Artistic Downmix Capability

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 stereo and multi-channel mixes are created by a sound engineer, who expresses his 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 Audio stereo downmix), can be quite different from the sound-engineer's stereo downmix (henceforth referred to as artistic stereo downmix).

In the case of a multi-channel audio broadcast, using the stereo-based MPEG Surround coder, there is a choice as to which stereo downmix to transmit to the receiver. Transmitting the spatial stereo 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. If, however, the artistic stereo downmix is chosen for transmission, an impairment of the quality of the reproduced multi-channel sound may result. In order to address this challenge, we are looking for a method in which both the artistic downmix can be transmitted and no (or few) concessions are made with respect to the perceived multi-channel sound quality.

One approach to this problem is to derive additional parameters that describe how the artistic stereo downmix would need to be modified such that it closely resembles the Spatial Audio stereo downmix. The MPEG Surround RM0 technology uses two parameters for this purpose. The first parameter describes the energy ratio of the left channels of the MPEG Surround and the artistic stereo downmix. The second parameter describes this energy ratio for the right channels. These modification parameters are typically analyzed as a function of time and frequency and are included in the parameter bitstream.

At the decoder, the modification parameters are obtained from the bitstream and the artistic stereo downmix is transformed using these parameters, resulting in a modified artistic stereo downmix. The modification parameters are applied such that the energy of the left and the right channel of the modified artistic stereo downmix equal the energy of the left and right channel of the MPEG Surround stereo downmix. Finally, the MPEG Surround upmix process is applied to this modified artistic stereo downmix resulting in the reproduced multi-channel sound.

4.4. 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. 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 RM0 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 10 and relies on several dimensions of scalability that are discussed briefly in the following.

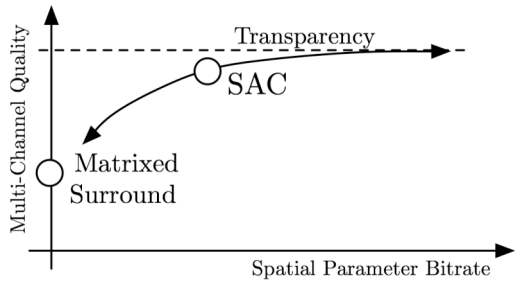

Figure 10 Rate/Distortion Scalability

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 RM0 syntax covers between 40 and a single parameter frequency band.

- Parameter quantization resolution
  As a third possibility, different resolutions for transmitted parameters can be used. Chosing 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.

- Parameter choice
  Finally, there is a choice as to how extensive the transmitted parametrization 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.

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.

4.5. 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 bit rates, 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 could 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, referred to as residual coding. In this approach, residual signals are encoded and transmitted to the decoder, and replace the decorrelated signals, providing a waveform match between the original and decoded multi-channel audio signal.

As described above, a multi-channel signal is downmixed to a lower number of channels (mono or stereo) and spatial cues are extracted in the spatial audio encoding process. During the process of downmixing, the resulting downmix channels are kept, while the 'residual' channels are discarded, as their perceptually important aspects are described by the extracted spatial cues. This operation is illustrated by the following encoding equations:

$$\begin{bmatrix} m \\ s_{OTT} \end{bmatrix} = H_{\mathrm{OTT}} \begin{bmatrix} l \\ r \end{bmatrix}$$

$$\begin{bmatrix} l_0 \\ r_0 \\ s_{TTT} \end{bmatrix} = H_{\mathrm{TTT}} \begin{bmatrix} l \\ c \\ r \end{bmatrix}$$

The encoding process for an OTT element generates a dominant (m) and a residual signal (s_{OTT}) from its two input signals, l and r. The elements of the downmix matrix H_{OTT} are chosen such that the energy of the residual signal (s_{OTT}) is minimized, given its modeling capabilities (based on the CLD and ICC parameters). A

similar operation is performed by the TTT element, for which the encoding process derives two dominant signals (*l0*, *r0*) and a residual signal (s_{TTT}) with minimal energy from the three input signals *l*, *c*, and *r*.

A corresponding residual signal can be derived for each OTT and TTT element in the MPEG Surround encoder. Furthermore, the residual-signal bandwidth can be chosen independently for each OTT and TTT element. The overall audio quality is controlled by selecting the appropriate trade-off between residual-signal bandwidth and bit rate, the amount of bits allocated to the core, and the remaining spatial side information.

In order to be independent from the mono or stereo core coder, while achieving the highest possible audio quality for the residual signals, the (band-limited) residual signals are represented as MPEG-2 AAC low-complexity profile individual channel stream elements [22]. The residual-signal AAC bitstreams are embedded in the spatial bitstream, as illustrated in Figure 11. Transients in the residual signals are handled by utilizing block switching and Temporal Noise Shaping (TNS) [20]. The MPEG Surround bit stream is scalable in the sense that the residual-signal AAC bitstreams can be stripped from the bit stream, thus lowering the bitrate, while the MPEG Surround decoder reverts back to the fully parametric operation (i.e., using decorrelator outputs for the entire frequency range).

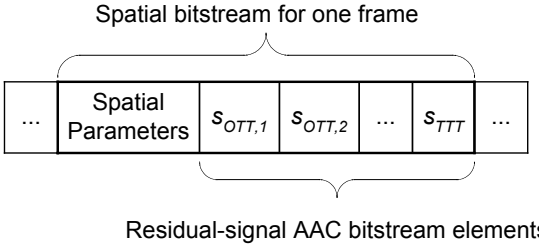

Figure 11 Embedding of residual-signal bitstream elements for each OTT and TTT element in the spatial audio bitstream

In the MPEG Surround decoder, the residual-signal AAC bitstreams are decoded into MDCT coefficients, which are transformed to the hybrid QMF domain where further processing of residual signals takes place. This decoded residual signal is used to replace the synthetic residual signals (i.e., the decorrelator outputs), within the bandwidth where transmitted residuals are available. This is illustrated in Figure 12.

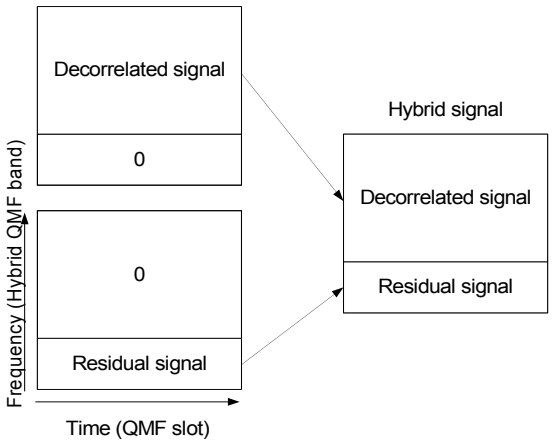

Figure 12 The complementary decorrelated and residual signals are combined into a hybrid signal

Inverse matrixing is applied to generate OTT and TTT element output signals from the (decoded) dominant and hybrid signals. Listening test results have shown the quality gain obtained by utilizing residual signals, as described in the section on MPEG Surround performance.

5. ONGOING WORK

As described in the previous sections, the MPEG Surround system is designed to provide a large range of rate/distortion scalability, starting at a few kbit/s parameter bitrate for low bitrate applications, up to near-transparency. In some cases, however, an even lower parameter bit rate may be required, or transmission of an additional parameter layer may not be feasible at all. 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, there is ongoing work within MPEG Audio to extend the current RM0 system with a mode that does not rely on any explicit transmission of spatial parameters. In the following, this mode will be referred to as *non-guided MPEG Surround Decoding* in contrast to the regular mode of operation, in which the decoding process is carried out (guided) by the transmitted spatial side information.

In non-guided operation mode, only a stereo downmix signal is transmitted from the encoder to the decoder, without a need for a spatial cue transmission. The MPEG Surround encoder is used to generate a matrixed-

surround compatible stereo signal (as described previously in the section on matrixed-surround compatibility). Alternatively, the stereo signal may be generated using a conventional matrixed-surround encoder. The MPEG Surround decoder is then operated without external side information input. Instead, the parameters needed for spatial synthesis are derived from an analysis stage working on the received downmix. In particular, these parameters are determined as a function of Channel Level Difference (CLD) and Inter-channel Cross Correlation (ICC) cues estimated between the left and right matrixed-surround compatible stereo input signal. Figure 13 illustrates this concept. The MPEG Surround encoder (or, alternatively, a conventional matrixed-surround encoder) generates a stereo downmix. The MPEG Surround decoder estimates the properties mentioned above for this downmix and maps these to the parameters needed for the spatial synthesis. Said differently, all required parameters for SAC synthesis (CLDs, ICCs, prediction coefficients) are generated as a function of the properties of the stereo downmix.

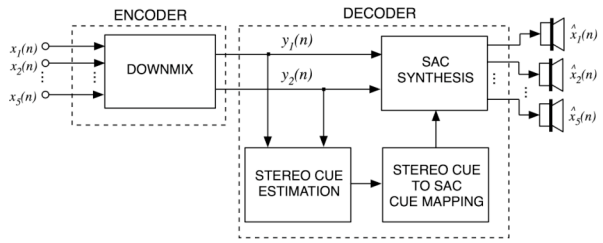

Figure 13 Spatial Audio Coding system without side information

A listening test carried out for a recently submitted MPEG core experiment indicates that non-guided MPEG Surround (without side information), as described previously, performs significantly better than conventional matrixed-surround systems. This is illustrated in Figure 14 and gives an indication of the high quality of the underlying spatial rendering engine. The two circles on the y-axes conceptually correspond to conventional matrixed-surround systems and the non-guided MPEG Surround system. Starting from this operating mode, it is attractive to gradually add side information for increasing the quality and in this way scale up towards regular (parameter-guided) mode.

Earlier work on "Hybrid BCC" [23] indicates that the perceived overall spatial quality mostly depends on the accuracy of the waveforms and spatial cues at the lower frequencies. By being precise up to about 1-2kHz while being less precise above these frequencies one gets a benefit much larger than what this small fraction of lower frequency range would imply. Therefore, in order to gradually improve (scale up) the quality of non-guided MPEG Surround system, it is one possible strategy to transmit spatial parameters for bands only below a certain cutoff frequency and in this way fade smoothly between non-guided mode without side information and the regular (fully parameter-guided) mode in which side information for the whole frequency range is transmitted. More work on this promising approach is carried out in the context of the MPEG development work.

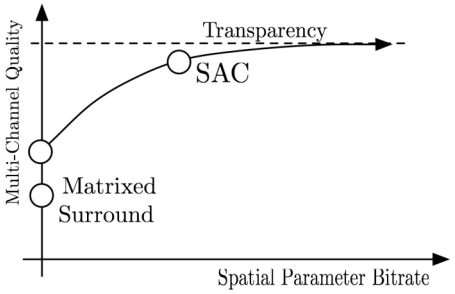

Figure 14 Extended SAC scalability down to non-guided operation without side information

6. PERFORMANCE

This section presents a number of recent listening tests done within the context of the MPEG standardization of the Spatial Audio Coding technology. The results illustrate the current level of performance of the MPEG Surround RM0 technology. The tests strive to evaluate the technology at several points on the rate / distortion curve. Firstly, the results of a general test are shown using three different MPEG Surround configurations that address different application scenarios. Subsequently, data from four more tests is provided exploring the rate / distortion scalability capability to scale to higher quality and lower rates, and also exploring the ability to handle artistic downmix signals.

For the different tests, a total of 11 items were used as listed in Table 1. The items are the same as used in the Call for Proposals (CfP) on Spatial Audio Coding [11], and range from pathological signals (designed to be critical items for the technology at hand) to movie sound and multi-channel productions. All input and output items were sampled at 44.1kHz. The playback

was done using multi-channel speaker setups conforming to ITU-R BS.1116.

Table 1 Items under test

No.	Name	Category	LFE chan.
1	BBC applause	pathological & ambience	
2	ARL applause	pathological & ambience	
3	Chostakovitch	music (back: direct)	
4	fountain music	pathological & ambience	
5	Glock	pathological & ambience	
6	indie2	movie sound	
7	jackson1	music (back: ambience)	
8	Pops	music (back: direct)	
9	Poulenc	music (back: direct)	
10	Rock concert	music (back: ambience)	
11	Stomp	movie sound	yes

All tests were conducted using the MUSHRA test. For this test methodology, a quality scale is used where the intervals are labeled ”bad“, ”poor“, ”fair“, ”good“ and ”excellent“. The subjective response is recorded on a scale ranging from 0 to 100, with no decimals digits.

6.1. Verification Test Results

The following test [13] was carried out as a verification of the Reference Model zero technology, as defined by MPEG in response to the Call for Proposals on Spatial Audio Coding [11]. For this verification, four tests were performed (see Table 2).

The aim of the first verification test (test t1, see Table 3) was to show the performance of the MPEG Surround system, when operating on a stereo signal coded by AAC at 160kbit/s. The bitrate for the spatial parameter data was 12kbit/s.

The second verification test (test t2, see Table 4) intended to show the performance of the MPEG Surround system when operating on a mono signal coded by AAC at 80kbit/s. The spatial parameter bitrate for this test was again 12kbit/s.

Table 2 Verification tests

Label	Config.	Core bitrate [kbit/s]	Spatial bitrate [kbit/s]	Comment
t1	5-2-5	160	12	
t2	5-1-5	80	12	
t3	5-1-5	43	5	Total bitrate limited to 48kbit/s
t1_LrHq	5-2-5	160	6, 12, 32	

Table 3 Codecs under test (test t1)

Label	Core bitrate [kbit/s]	Spatial bitrate [kbit/s]	Comment
RM0_160	160	12	
DPL2	160	Not Appli-cable	The Dolby Prologic 2 signals were en/decoded with a professional Dolby en/decoder
Href			Hidden reference
BW_35			3.5kHz anchor

Table 4 Codecs under test (test t2)

Label	Core bitrate [kbit/s]	Spatial bitrate [kbit/s]	Comment
RM0_80	80	12	
DPL2	160	Not Applicable	See Table 3
Href			Hidden reference
BW_35			3.5kHz anchor

The third verification test (test t3, see Table 5) intended to show the performance of the MPEG Surround system when operating on a mono signal coded by AAC at a total bitrate of 48kbit/s. The spatial parameter bitrate was for this test 5kbit/s, and since the total bitrate of the

system was limited to be 48kbit/s, the bitrate used by the underlying core coder was 43kbit/s.

Table 5 Codecs under test (test t3)

Label	Core bitrate [kbit/s]	Spatial bitrate [kbit/s]	Comment
RM0_48	43	5	
Href			Hidden reference
BW_35			3.5kHz anchor

A fourth verification test (test t1LrHq) intended to show the performance of the MPEG Surround system for a higher quality configuration and low side information configuration. Therefore, three configurations of the MPEG Surround system were included, (see Table 6) operating at different bitrates, 6kbit/s, 12kbit/s and 32kbit/s. The bitrate for the underlying core coder was 160kbit/s.

Table 6 Codecs under test (test t1LrHq)

Label	Core bitrate [kbit/s]	Spatial bitrate [kbit/s]	Comment
RM0_6	160	6	
RM0_12	160	12	
RM0_32	160	32	
DPL2	160	Not Applicable	See Table 3
Href			Hidden reference
BW_35			3.5kHz anchor

The results of all four tests are combined in Figure 15. The figure shows the mean results and 95% confidence intervals over all items and subjects (after post-screening). For the MPEG Surround system, the spatial bitrates in kbit/s are listed. For benchmarking purposes also the results of Dolby Prologic II are included when applicable.

The test results show that MPEG Surround RM0 provides an audio quality vastly better than that obtained with Dolby Prologic 2. Even when the system operates on a mono signal it is clearly better than the Dolby Prologic output operating on a stereo signal. Furthermore, the test indicates that the quality of the MPEG Surround system can be increased by increasing the spatial parameter bitrate. This is explored further in an additional test below.

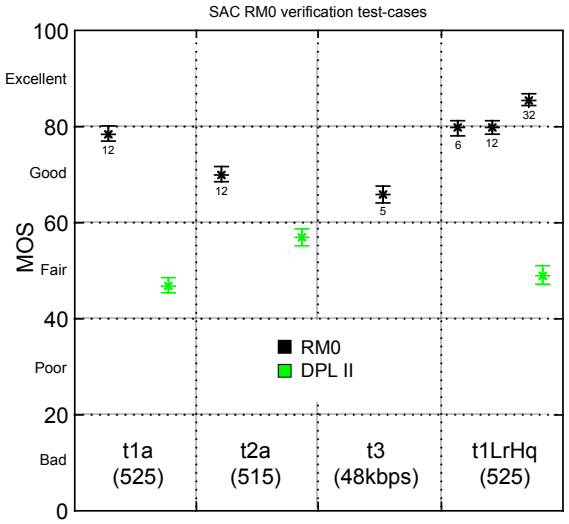

Figure 15 RM0 verification test results for the 4 test cases that have been tested.

6.2. Scalability to High Audio Quality

Given that the Spatial Audio Coding concept is based on parametric coding techniques, it is important to ensure that the highest achievable audio quality is not limited by the assumptions of the underlying parametric model. As described in the corresponding section, residual coding is a technique utilized by MPEG Surround to bridge the gap between the audio quality of a parametric description and transparent audio quality. To this end, subjective tests were carried out in order to assess the performance of RM0 operating in high quality mode.

For testing of RM0 in high-quality mode, three configurations have been selected, covering spatial audio bitrates ranging from 32kbit/s, an operating point that roughly corresponds to the high quality operating point (RM0_32 in the t1LrHq test) as tested during the RM0 verification tests, up to 192kbit/s. The stereo downmix has been coded at 128kbit/s, ensuring a high audio quality for the backwardly compatible stereo. For benchmarking purposes, MPEG-2 AAC LC 5.1 multi-channel coding has been added at 192 and 320kbit/s total. Finally, a hidden reference and low bandwidth anchor have been included. The same 11 items as in the RM0 verification test were used (listed in Table 1). A

total of 13 subjects participated in the test. The results from this test are provided in Figure 16

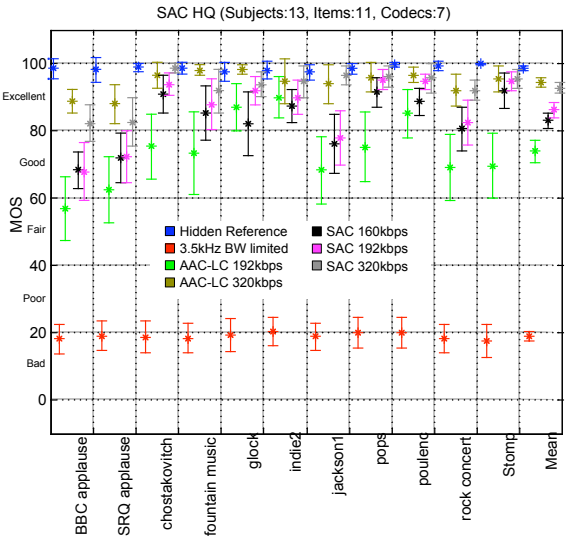

Figure 16 Results of high quality test

The test results show that MPEG Surround RM0 provides an improvement in audio quality for increasing spatial side information bitrate. Furthermore, the results show that the system at 160 kbit/s total is statistically significantly better than AAC 5.1 multi-channel at 192 kbit/s.

6.3. Scalability to Lower Side Information Bitrate

In order to further explore the possibility for scaling down the RM0 system to even lower side information rates, a listening test was performed. This scaling process was addressed by selecting an alternative time/frequency tiling in the encoder. This can be done fully compatible within specified bitstream syntax.

Four different configurations of MPEG Surround RM0 were included in the test, together with the standard hidden reference and 3.5kHz band-limited anchor condition, as mandated by the MUSHRA specification. The different configurations employ a spatial parameter bitrate of 6.6, 4.1, 2.8 and 1.8kbit/s respectively. The first configuration is comparable to the low rate condition (RM0_6) tested in the t1LrHq RM0 verification test.

Similarly to the tests described in the previous section, the 11 items from the MPEG spatial set of test signals were used. The subjective test was carried out by 8 expert listeners. The results are shown in Figure 17.

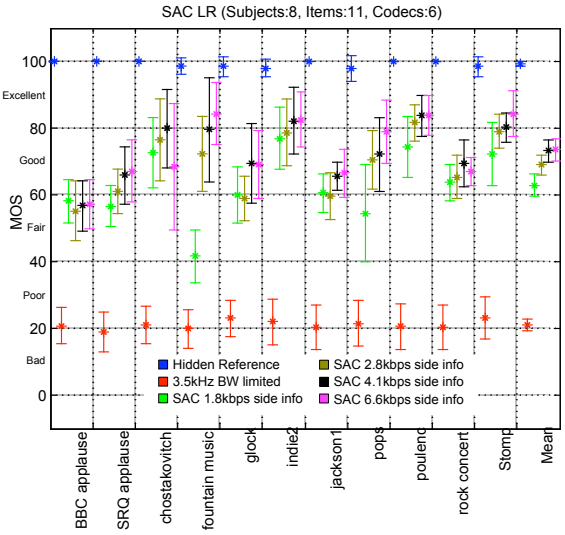

Figure 17 Results of low bitrate test

From the results it is seen that the average sound quality increases gradually and monotonically, as the side information rate is increased from below 2kbit/s up to 6.6kbit/s. It is, however, interesting to observe that this happens in a surprisingly graceful manner and thus leads to additional attractive operating points for MPEG Surround applications.

6.4. Non-Guided MPEG Surround Decoding

A MUSHRA test was performed to investigate the quality of a non-guided MPEG Surround configuration, based on a matrixed-surround compatible downmix. This mode is described in the corresponding section on non-guided decoding technology. For benchmarking purposes, Dolby Prologic II encoding/decoding was added. Both the MPEG Surround stereo downmix and the stereo downmix signals of the Dolby Prologic II algorithm were encoded using AAC-LC encoding at 160 kbit/s. Again, the 11 items from the RM0 verification test were used for the test (see Table 1). In total 12 subjects participated in the test. The results are provided in Figure 18.

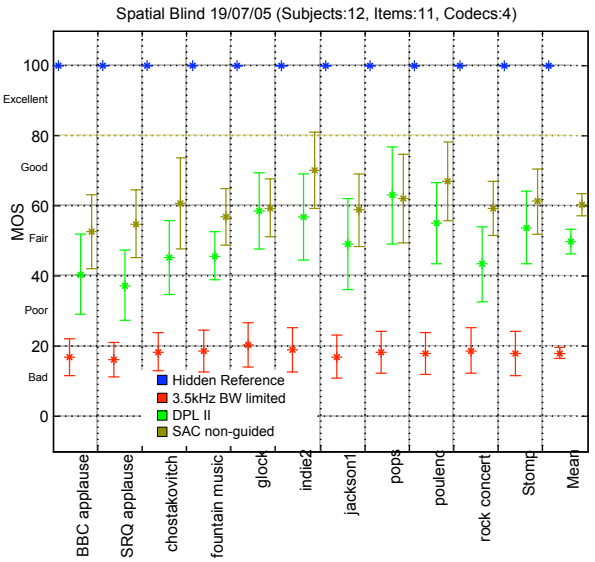

Figure 18 Test results for non-guided mode

The subjective test results show that a non-guided mode of MPEG Surround RM0 offers a performance that is statistically significantly better than state-of-the-art matrixing technology.

6.5. Support For Artistic Downmix

A listening test was performed in order to evaluate the MPEG Surround performance in the context of an artistic stereo downmix as described in Section 4.3. In this listening test, three different coder settings were used:

1. the stereo-based MPEG Surround coder,
2. the stereo-based MPEG Surround coder where the spatial stereo downmix was replaced by the modified artistic downmix and
3. the stereo-based MPEG Surround coder where the stereo downmix was replaced by the artistic downmix.

For the third configuration, the following modification was made: It was ensured that the overall signal energy is preserved with respect to the original signal. This was done in order to enhance the coder's quality to such an extent that comparison with the other coders is facilitated. The bitrate of these modification parameters amounts to approximately 1 kbit/s.

For this test, 10 test items were used, of which both a multi-channel mix and an artistic downmix were available. Comparing the artistic downmix to the MPEG Surround stereo downmix of the corresponding multi-channel mix, the following main differences are observed in the test items: additional reverberation, a voice-over, different panning of sources, flanger, phasing and multi-band compression.

The MUSHRA test methodology was used, except for the fact that no anchor was present. In the test, the listeners had to rate the perceived quality of the test items against the original excerpt on a 100-point scale. The listening panel consisted of 7 subjects, each of them experienced in the field of multi-channel audio.

The results of the listening test are shown in Figure 19. On the horizontal axis, the different ways of processing are listed. These are, from left to right, the hidden reference, the stereo-based MPEG Surround coder, the same coder where the spatial stereo downmix is replaced by the modified artistic downmix, and the same coder where the spatial downmix is replaced by the (unmodified) artistic downmix. From Figure 19 it is noted that the stereo-based MPEG Surround coder performs significantly better than both other coders. Although the mean coder performance for the modified artistic downmix is not statistically significantly better than for the unmodified artistic downmix, a clear trend in favor of the modified artistic downmix is visible when observing the results of the individual items.

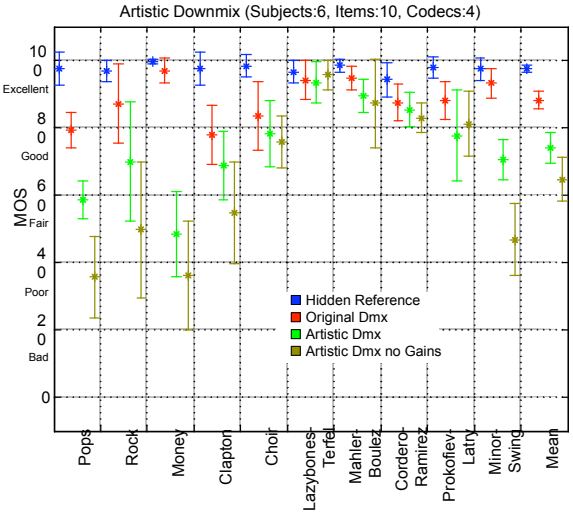

Figure 19 Results of the listening test on artistic downmix

7. CONCLUSIONS

Within the field of low-bitrate audio coding, Spatial Audio Coding is a promising approach for both bitrate-efficient and backward compatible representation of multi-channel audio signals. It enables the transmission of such signals at data rates close to the rates used for the representation of two-channel (or even monophonic) audio. Currently, Spatial Audio Coding is under standardization within the ISO/MPEG group. The paper described the technical architecture and capabilities of the MPEG Surround Reference Model 0 technology and some ongoing extension work. As one of the most prominent features, the technology 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.

8. REFERENCES

[1] 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

[2] J. Herre, H. Purnhagen, J. Breebaart, C. Faller, S. Disch, K. Kjőrling, 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. Herre: "From Joint Stereo to Spatial Audio Coding - Recent Progress and Standardization", Sixth International Conference on Digital Audio Effects (DAFX04), Naples, Italy, October 2004

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

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

[6] 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

[7] 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

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

[9] Dolby Publication, Roger Dressler: "Dolby Surround Prologic Decoder – Principles of Operation", http://www.dolby.com/tech/whtppr.html

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

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

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

[13] 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/explorations/sac/RM0-listening-tests.zip

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

[15] 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)

[16] 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

[17] 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

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

[19] H. Purnhagen, J. Engdegård, J. Rődén, L. Liljeryd: “Synthetic ambience in parametric stereo coding”, Proc. 116th AES convention, Berlin, Germany, 2004, Preprint 6074

[20] 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

[21] J. Herre, J. D. Johnston: "Exploiting Both Time and Frequency Structure in a System that Uses an Analysis/Synthesis Filterbank with High Frequency Resolution" (invited paper), 103rd AES Convention, New York 1997, Preprint 4519

[22] 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

[23] F. Baumgarte, C. Faller, P. Kroon: “Audio Coder Enhancement using Scalable Binaural Cue Coding with Equalized Mixing”, Proc. 116th AES convention, Berlin, Germany, 2004, Preprint 6060