Audio Engineering Society

Convention Paper

Presented at the 123rd Convention
2007 October 5–8 New York, NY, USA

A study of the MPEG Surround quality versus bit-rate curve

Jonas Rödén[1], Jeroen Breebaart[2], Johannes Hilpert[3], Heiko Purnhagen[1], Erik Schuijers[4], Jeroen Koppens[4], Karsten Linzmeier[3] and Andreas Hölzer[3]

[1] Coding Technologies, 113 30 Stockholm, Sweden

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

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

[4] Philips Applied Technologies, 5656 AE, Eindhoven, The Netherlands

ABSTRACT

MPEG Surround provides unsurpassed multi-channel audio compression efficiency by extending a mono or stereo audio coder with additional side information. This compression method has two important advantages. The first is its backward compatibility, which is important when MPEG Surround is employed to upgrade an existing service. Secondly, the amount of side information can be varied over a wide range to enable high-quality multi-channel audio compression at extremely low bit rates up to perceptual transparency at higher bit rates. The present paper provides a study of the performance of MPEG Surround, highlighting the various tradeoffs that are available when using MPEG Surround. Furthermore, a quality versus bit rate curve describing the MPEG Surround performance will be presented.

1. INTRODUCTION

Audio recording, storage, reproduction systems and processing methods have been a continuous topic for development during the last decades. Especially with the introduction of digital multi-channel content, the realism in terms of spatial reproduction has increased significantly. At the same time, the corresponding increase in the total amount of information, as well as compatibility issues between various existing formats and reproduction systems pose new challenges. This concerns both an efficient transmission of the content, as well as to ensure maximum quality during reproduction on a wide range of reproduction systems.

Based on these observations, the ISO/MPEG Audio standardization group had started a new work item on efficient and backward compatible coding of high-quality multi-channel sound using parametric coding

techniques in 2004. 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 bit rates, 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 [1].

This paper gives a brief introduction to the concepts and processing stages of MPEG Surround (Section 2). Special attention is given to the tools that allow MPEG Surround to operate at different bit rates while providing highest possible surround sound quality (Section 3). The large MPEG Surround span in terms of rate/distortion is described in Section 4, illustrated by listening test results.

2. MPEG SURROUND CONCEPT AND BASICS

2.1. Spatial Audio Concepts

Conventional subband or transform coders typically employ the concept of *monaural* perceptual masking to reduce the accuracy of an audio signal to result in a certain compression ratio. The repertoire of tools in such coders to exploit cross-channel redundancies and irrelevancies is rather limited [2,3,4]. However, recent developments in the area of audio compression provide means to exploit cross-channel relations in an extremely efficient manner. More specifically, Spatial Audio Coding (SAC) aims at capturing and reconstructing the spatial image of two or more audio channels by means of a compact set of perceptual parameters.

An audio encoder based on SAC captures the spatial image by means of perceptually-relevant spatial parameters and reduces the number of audio channels by a downmix process. This downmix is subsequently encoded by a conventional core coder, and the resulting bitstream is extended with the extracted parameters. The parameters may be stored in the ancillary data part of the core-coder bitstream to ensure backward compatibility with decoders that are not SAC enabled. The decoder basically performs the reverse process and reconstructs a high-quality spatial image based on the transmitted downmix and spatial parameters (see Figure 1).

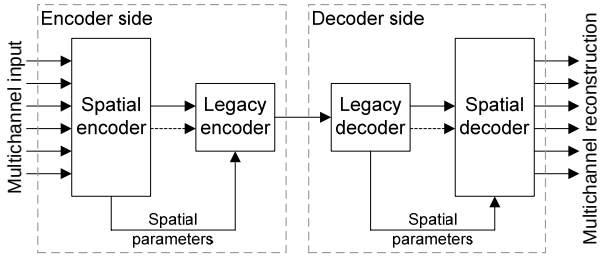

Figure 1 – Block diagram of generic SAC encoder decoder chain.

The concept of Spatial Audio Coding is found in techniques such as Binaural Cue Coding [5,6,7] and Parametric Stereo [8,9,10,11]. These techniques parameterize the perceptually most-prominent spatial localization and spatial quality cues such as inter-channel level differences, inter-channel time/or phase differences, and inter-channel coherence or correlations (see [12] for an extensive overview of the psychoacoustic background of spatial parameterization techniques). The major improvement of the Spatial Audio Coding implementation as employed in the recent MPEG Surround standard is that its SAC parameterization repertoire is not limited to localization cues only. The parametric approach has been extended with signal-driven prediction methods to exploit and efficiently model cross-channel relations, including means to overcome quality limitations of a purely parametric model [13]. As a result, MPEG Surround is not only capable of delivering unsurpassed compression efficiency at low bit rates, but is also capable of providing (near) transparent quality at bit rates that are very competitive if compared to conventional coders that employ discrete coding of all channels. This property makes MPEG Surround the codec of choice for a very wide range of applications and transmission bandwidths.

2.2. MPEG Surround

The MPEG Surround development and its various features and tools have been extensively described in a series of publications [14,15,16,17,18]. For completeness, the most important aspects will be summarized here.

2.2.1. Hybrid QMF bank

The MPEG Surround decoder operations are conducted in the so called hybrid QMF domain. This hybrid QMF domain is obtained by feeding the time domain signal through a cascade of a complex-modulated oversampled Pseudo QMF bank followed by an oddly-modulated Nyquist filter bank (see Figure 2 and [11,14]). The first QMF bank is identical to the filterbank used in High-Efficiency Advanced Audio Coding (HE-AAC). This filter bank was included in HE-AAC to provide additional compression efficiency by the process of Spectral Band Replication (SBR) [22]. Similar to the MPEG Surround approach, the SBR algorithm is a post-processing algorithm that works on top of a conventional (band-limited) low bit rate audio decoder and allows the reconstruction of a full-bandwidth audio signal by means of additional parameters. The extension of the QMF bank with a second oddly-modulated Nyquist filter bank has a number of advantages:

- The spectral resolution closely matches the perceptual Equivalent Rectangular Bandwidth (ERB) scale [13]. A high frequency resolution at low frequencies is obtained through the hybrid structure, essentially splitting the lower subbands into smaller subbands. At high frequencies, a relatively low frequency resolution is desired, which is obtained by grouping of QMF bands. Section 3.2.2 provides a more thorough discussion on the parameter resolution.

- Due to the oversampled nature of the hybrid QMF bank it is possible to apply time- and frequency-variant manipulations to the input signal without introducing audible aliasing distortion.

- As the QMF bank is already employed in the SBR tool it allows for a highly efficient decoder combination of HE-AAC and MPEG Surround in a similar way as employed for Parametric Stereo [11,12].

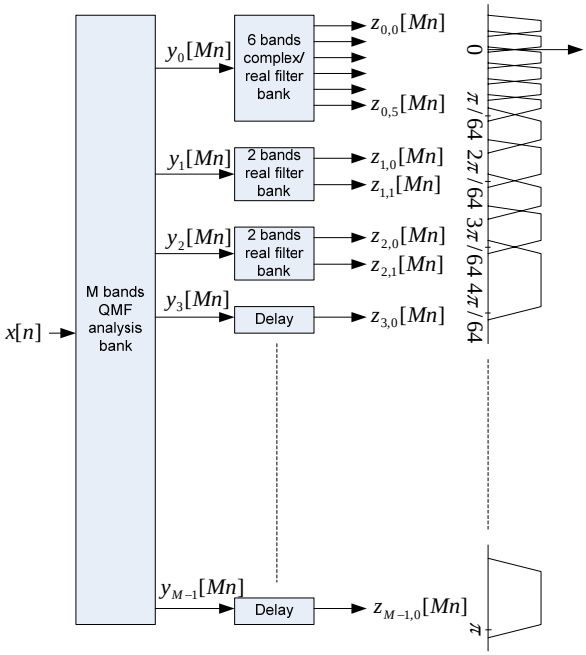

Figure 2 – Hybrid QMF analysis bank structure. The input signal x[n] gets filtered into M (M=64) uniform subbands. Three of the lower subbands are further split using additional small filter banks. The other subbands are delay compensated. As a result a non uniform frequency resolution is obtained.

2.2.2. Encoder structure

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 the 'One-To-Two' (OTT) and 'Two-To-Three' (TTT) elements, referring to their respective input and output channel configurations at the decoder side. In other words, an OTT element produces two output channels by means of a single input channel, based on spatial

parameters. Similarly, the TTT element creates three output channels by means of a stereo input signal and spatial parameters.

Each conceptual decoder element has a corresponding encoder element (Reverse-OTT and -TTT element, respectively) that generates a downmix and extracts the frequency-dependent spatial parameters required by the decoder elements.

By concatenating these basic building blocks into tree structures, many different channel configurations can be constructed. Figure 3 shows an example tree used in MPEG Surround to encode a 5.1 multi-channel signal into a stereo downmix with spatial parameters. This encoder configuration is therefore referred to as 5-2-5 indicating the encoder input, downmix and decoder output channel configurations respectively.

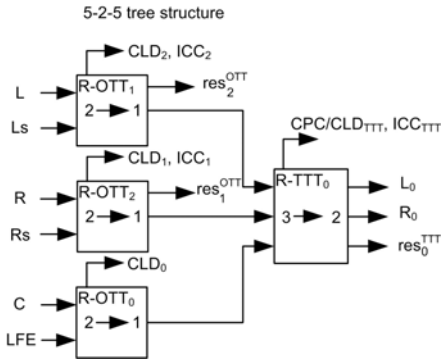

Figure 3 – Tree structure reducing a multi-channel input signal to a stereo downmix and spatial parameters.

The spatial parameters extracted in the R(everse)-OTT and R-TTT building blocks capture the perceptually dominant spatial attributes and include:

- Channel Level Difference (CLD) – Indicates the level difference between two input channels (or between two intermediate signals in the tree structure); a primary spatial localization cue.

- Inter-Channel Coherence (ICC) – Gives a measure of the resemblance (or correlation) between two input signals; an important property to achieve spaciousness.

- Channel Prediction Coefficient (CPC) – These parameters are used by the TTT element to estimate a third channel from the two input signals [13].

In addition to these spatial parameters, the R-OTT and R-TTT elements also generate a residual signal containing the modeling error associated with the parameterization of the input [12,13]. These residual signals may be transmitted as an addition to the spatial data to enable full waveform reconstruction at the decoder side.

2.2.3. Decoder signal processing

The OTT and TTT elements reconstruct output signals according to the transmitted ICC and CLD parameters, and provide a full waveform match if residual signals are transmitted. For those time/frequency tiles where no residual signal is available, reconstruction of the correct coherence requires an additional signal (as replacement of the omitted residual signal) that is statistically independent from the input, but with the same spectro-temporal structure and perceptual timbre. These signals are generated by decorrelation filters. These filters output "decorrelated" versions of their input signals while preserving the spectral and temporal envelopes [13,19].

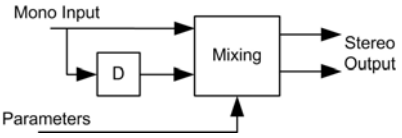

Figure 4 – Basic principle of OTT module; decorrelation (D) and consecutive mixing.

Compared to the conceptual tree-based views presented previously the actual decoder-side upmix to the multi-channel output does not take place in a tree-structured fashion. Instead, the decoder signal processing happens in a "flattened" way, i.e. the tree-like parameterization is converted into a three-step approach using parallel decorrelation filters. This achieves increased computational efficiency and minimizes degradations due to concatenated decorrelation operations that would result from a decoder-side tree structure [12,13]. As a result, the spatial synthesis process can be described by a pre-mixing matrix, decorrelation stage and a post-mixing matrix as illustrated in Figure 5.

The input signals are first processed by the pre-mixing matrix, to prepare the signals for the decorrelation filters. The post-mixing matrix performs the actual mixing with the decorrelated or residual signals, similar to the mixing in an individual OTT module.

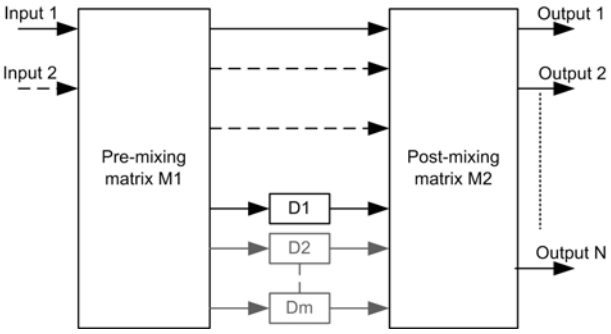

Figure 5 – Generalized decoder structure with one decorrelation stage.

The matrix entries of the pre- and post-mixing matrices are derived from the transmitted spatial parameters given the specific tree-structure used for the parameterization. The decorrelators correspond to the various decorrelators as part of the individual OTT and TTT modules in the tree structure, but now operating in parallel.

2.2.4. Features

The following system features make MPEG Surround very flexible for application in a broadcast environment.

Matrix-Compatible Downmix

In addition to an 'ITU-style' stereo downmix of the multi-channel input, an MPEG Surround Encoder is capable of producing a matrix-surround compatible downmix signal. This feature enables compatibility to receivers that are able to decode the stereo core signal, but do not support MPEG Surround and are supplied with a conventional matrix surround decoder instead.

Technically this matrix-compatible downmix is obtained by a post processing of the conventional stereo downmix [13]. The advantage of such a post-processing approach is that the MPEG Surround decoder is able to invert this process before applying its upmix process. Thus for MPEG Surround equipped receivers there are no inherent quality degradations when running in this mode. Additionally it is possible to retrieve the standard (or non-matrixed) downmix from an MPEG Surround decoder.

Enhanced Matrix Mode

For transmission scenarios where the additional MPEG Surround data cannot be conveyed, or when there is only a stereo signal available on the receiver side (without spatial parameters), the MPEG Surround decoder can be operated in the so-called 'enhanced matrix mode'. In this case, a parameter estimation block estimates spatial parameters from the inter-channel relations in the stereo downmix signal. The MPEG Surround upmix engine is subsequently fed with these parameter estimates. The 'enhanced matrix mode' outperforms conventional matrix-surround systems in terms of quality (see [13,15,20]).

Artistic Downmix Handling

During the production of multi-channel audio content, a sound engineer usually also provides a dedicated stereo ('artistic') mix of the recorded material. Often the balance between different instrument groups, the amount of effects, or the amount of natural ambience differs between the stereo and the 5.1 mix. For a broadcast scenario, the dedicated multi-channel mix is preferably reproduced on a multi-channel setup, while the separate artistic stereo mix is preferred for conventional stereo reproduction systems. Unfortunately, conventional coders require simulcast of both mixes, which is very undesirable in terms of transmission bandwidth.

With MPEG Surround, the automatic downmix from the MPEG Surround encoder can be replaced by the artistic stereo downmix. The MPEG Surround encoder is then able to add a small amount of data to the bitstream that allows an MPEG Surround decoder to reconstruct the dedicated 5.1 mix from the artistic stereo downmix instead of the automated downmix. The amount of additional information is highly scalable, depending on the desired bit rate and quality tradeoffs (see also [12,13,18]).

3. MPEG SURROUND RATE / DISTORTION TOOLS

3.1. Introduction

Due to the built-in flexibility, MPEG Surround covers a broad range of operation points both in terms of side information rate and multi-channel audio quality. This flexibility is important since different applications require individually optimized operation points. Furthermore the achieved multi-channel audio quality depends on both the audio quality delivered by the downmix coder and the amount of data used for the MPEG Surround side information. Hence one of the trade-offs for practical applications is the amount of data rate used by the downmix coder versus the amount used for the MPEG Surround side information.

Typically, the MPEG Surround side information rate is very small compared to the bit rate required by the downmix coder (in the order of 10%). However, if (near) transparency is required, the amount of information occupied by parameters and residual signals may be gradually increased and hence the amount of MPEG Surround side information increases accordingly. The concept of this side-information scalability is illustrated in Figure 6. The horizontal axis denotes the side information rate, the vertical axis the resulting quality. Without any side information, the quality of MPEG Surround is beyond conventional matrix-surround systems [15,18]. With increasing side information, the perceived quality increases as well. This increase in side information stems from an increase in parameter data rate, as well as an increased accuracy to represent residual signals. The various methods to control the side-information rate for parameters and residual signals are discussed in the following sections.

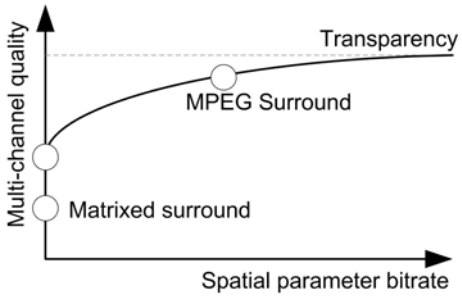

Figure 6 – MPEG Surround rate-distortion scalability

3.2. Parameter scalability

3.2.1. Parameter accuracy

The accuracy of the spatial representation can be controlled in various means.

One degree of freedom results from scaling the frequency resolution of the parameters. While a high frequency resolution 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.

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.

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 such low-resolution parameter description is accommodated by dedicated tools, such as the Adaptive Parameter Smoothing mechanism (see next Section).

3.2.2. Time and Frequency Resolution

Given the importance of parameter scalability and the influence of time and frequency resolution on bit rate,

the methods to control the spectro-temporal resolution are explained in more detail in this section.

Given the hybrid QMF domain in which MPEG Surround operates, it is possible to dynamically group and modify subband samples without introducing audible aliasing distortion. The frequency resolution of the parameter bands can be varied dynamically for each parameter individually. The maximum frequency resolution is defined by the *master frequency*, which defines the maximum number of parameters across frequency. The master frequency varies between 28 (maximum) and 4 (minimum). For a given master frequency, the number of parameters to span the full frequency range can be varied from one parameter for the full range up to the master frequency.

While the master frequency resolution typically stays the same for a given operation point, the dynamic frequency resolution can change the number of parameter bands for each parameter in a parameter set allowing adaptation to the audio content. The possible number of parameter bands is shown in Table 1.

Master frequency	Dynamic subsets			
28	28	14	6	1
20	20	10	4	1
14	14	7	3	1
10	10	5	2	1
7	7	4	2	1
5	5	3	1	1
4	4	2	1	1

Table 1 – Number of frequency bands for the different master frequency resolutions and the possible dynamic subset there of.

In Figure 7 the bandwidth of each parameter band as a function of its center frequency is illustrated for the case of 4 and 28 parameter bands. For comparison the Figure also illustrates the ERBs and the critical bandwidth CB. As can be seen the 28 parameter band grouping is closely matched to the ERB and CB but follows a staircase function in the low frequency range due to the hybrid QMF bands which have a certain fixed bandwidth.

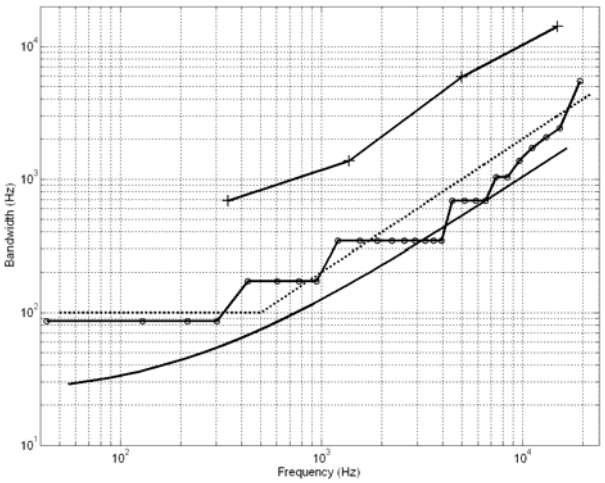

Figure 7 – Bandwidth for the 4 (+) and 28 (o) parameter band master frequency resolutions as a function of frequency. The dotted line represents the bandwidth CB which amounts 20% of the center frequency above 500 Hz and 100 Hz otherwise. The solid line represents the ERB defined as ERB(f) = 21.4 (0.00437 f +1).

Besides the spectral resolution of parameters, the temporal *position* at which a parameter set is valid can be controlled as well. Since one hybrid QMF sample corresponds to around 1,3 milliseconds at a samplingrate of 48 kHz, a parameter set can be signaled with the same temporal accuracy. One MPEG Surround frame may comprise up to 8 parameter sets, where the frame size would typically equal the downmix coder e.g. ~43 milliseconds for HE-AAC with MPEG Surround at 48 kHz. This allows for a flexible positioning of parameters where for example the number of parameter sets per frame is kept to a minimum for a really low bit rate application, but can be increased for an operation point where more bits can be spent on the side information data rate.

3.2.3. Low bit rate tools

MPEG Surround provides several tools to further lower the parameter bit rate demand with minimum impact on the subjective sound quality. The following paragraph gives a short overview.

Energy dependent Quantization
The 'Energy dependent Quantization' tool utilizes the psychoacoustic phenomenon that audio channels with

a relatively high energy or loudness should be described by a higher parameter resolution than audio channels with a relatively low energy. A coarser quantization of CLD parameters can be used if a parameterization stage represented by an R-OTT module is low in energy compared to the total energy.

Single ICC

For low side information bit rates, ICC values can be combined in the encoder into a single ICC parameter subset per parameter frequency band. The combined parameter is used in the decoder as a substitute for all individual ICC parameters. The encoder parameter combination process is carried out such that the sound image of the original multi-channel signal is preserved as closely as possible after reconstruction by the decoder.

Adaptive Parameter Smoothing

For low bit rate scenarios, it is desirable to employ a coarse quantization for the spatial parameters in order to reduce the required bit rate as much as possible. For certain kinds of signals, this may, however, result in audible artefacts. Especially in the case of stationary and tonal signals, modulation artefacts 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 artefact. The 'Adaptive Parameter Smoothing' tool, which is applied on the decoder side, is designed to address these artefacts by temporally smoothing the parameter values for signal portions with the described characteristics. The adaptive smoothing process can be controlled from the encoder by transmitting dedicated control flags.

3.2.4. Lossless parameter coding

For the sake of efficient storage or transmission of the quantized parameters, sophisticated entropy coding schemes are applied. Mostly, these consist of a combination of differential coding and one- or two-dimensional Huffman coding. Differential coding can take place between parameters neighboring in time or in frequency. In case of two-dimensional Huffman coding, one single Huffman code word represents pairs of differential values neighboring either in frequency or in time direction. For typical audio material this type of entropy coding gains approximately 40-55% of bit rate in case of fine quantization and 45-60% in case of coarse quantization when compared to uncompressed transmission.

3.3. Residual scalability

As described in Section 2.2.2, for low-bit rate applications MPEG Surround operates in a parametric mode for maximum compression efficiency. For most audio material this mode provides a faithful sound reproduction. However, since in this mode the decorrelated signals are a synthetic representation of part of the audio signal that has been discarded during encoder operation, a small loss in temporal and spectral detail may occur. The goal of residual coding is to overcome quality limitations imposed by a parametric model and provide exact waveform reconstruction for high-quality applications.

There are a number of options for residual coding to enable a large degree of scalability in terms of bit rate:

- The bandwidth of each residual signal can be varied from zero to the full signal bandwidth.
- The bit rate of each coded residual signal can be varied.
- Residual signals can be enabled or disabled independently for each element in the tree structure.

Given the hybrid QMF domain in which MPEG Surround operates, the residual signals should be provided in this domain as well. However, this hybrid QMF domain provides an oversampled signal representation and is hence not suited for coding purposes. Therefore the residual signals are coded in the MDCT domain. The MDCT domain representation is obtained by means of a direct transformation from the QMF domain representation. This is illustrated in Figure 8. A QMF domain residual analysis frame of length 32 QMF slots (effective length 16 QMF slots) and a bandwidth of 40 QMF bands is transformed to the MDCT domain resulting in a vector of $16 \cdot 40 = 640$ MDCT

coefficients. The residual MDCT coefficients are coded within the MPEG Surround bitstream by means of existing MPEG AAC technology. At the decoder side a direct transformation back to the QMF domain is employed.

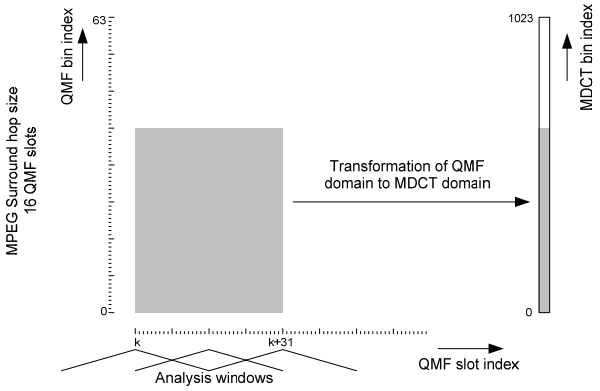

Figure 8 – Transformation of QMF domain representation residual signal to MDCT domain.

If the QMF domain analysis length differs from 16 QMF slots, the AAC sampling frequency index can be adjusted to allow scalability of the coding efficiency. This is illustrated in Figure 9. A QMF domain analysis window of length 64 QMF slots (effective length 32 QMF slots) by 24 QMF bands is transformed to the MDCT domain resulting in a vector of 24·32=768 MDCT coefficients. It is to be noted that these 768 MDCT coefficients represent only approximately 3/8 of the sampling frequency at which the MPEG Surround encoder is operating. For this particular example of an effective QMF domain analysis length of 32 slots, the residual signal will be coded using MPEG AAC at half the MPEG Surround sampling frequency.

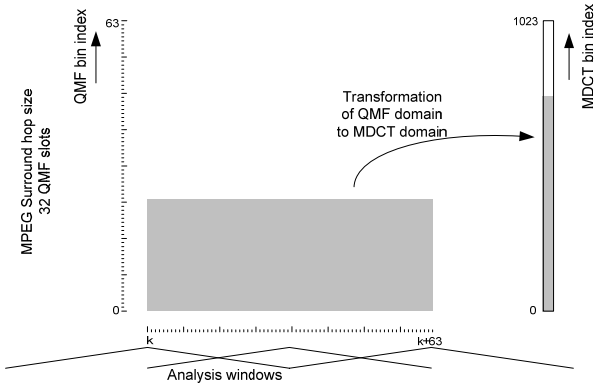

Figure 9 - Transformation of QMF domain representation residual signal to MDCT domain with change of sampling frequency.

3.4. Adaptations to downmix coders

3.4.1. Introduction

MPEG Surround operates as a pre- and post-processing extension on top of legacy coding schemes. It is therefore equipped with means to accommodate to virtually any downmix coder. In this section, the various tools to accommodate virtually every downmix coder are discussed.

3.4.2. Framing

Framing in MPEG Surround is highly flexible to ensure synchrony with a wide range of downmix coders. In theory all frame lengths between 1 and 128 QMF slots are supported, corresponding to 64 to 8192 time samples, with steps of 64 samples. In practice only frame sizes that are an integer multiple of downmix coder frame lengths will be useful. Table 2 contains a non-exhaustive list of core coders with their typical spatial frame lengths.

Downmix coder	Coder frame length (time samples)	Spatial frame lengths (QMF slots)
AAC	1024	16, 32, 64
HE-AAC	2048	32, 64
MPEG 1 – Layer II/III	1152	18, 36
AAC – LD	960	15, 30, 60

Table 2 – Typical spatial frame lengths for core coders.

Parametric core coder adaptation

Core coders may or may not provide a waveform reconstruction at their output. For example, core coders employing SBR employ parametric methods to reconstruct the high frequency range, and hence no waveform match is obtained between core coder input and output for that frequency range.

MPEG Surround provides means to optimize its parameterization depending on waveform-preserving properties of the core coder. For example, the CPC parameters calculated by the TTT element aim at waveform reconstruction and hence may result in suboptimal performance when using CPC parameters in the SBR range.

Therefore the TTT module supports a 'dual mode' operation where the frequency range is divided into a low band range and a high band range. The CPC parameters can still be transmitted for the low band range, while in the high band range an alternative parameterization scheme is used, based on statistical reconstruction rather than waveform reconstruction utilizing 2 CLD parameters instead of 2 CPC parameters per parameter band.

3.4.3. Buried data

Most legacy coders provide means to insert the spatial data as so-called ancillary data into the bitstream without affecting normal decoding. This way the MPEG Surround data can be conveyed using an existing infrastructure. However, uncompressed signals are often transmitted in a continuous stream without possibilities to add additional side information.

For the uncompressed PCM downmix format, MPEG Surround data can be inserted in the audio signal itself using a technique called *Buried data*. This technique employs psycho-acoustic masking to transfer MPEG Surround data in the least significant bits of the audio signal. Listening tests revealed that such addition of side information has no audible consequences [21].

4. PERFORMANCE

4.1. Introduction

In order to derive the rate/distortion curve of MPEG Surround in combination with different core coders (Layer II, AAC, HE-AAC), a listening test was conducted according to MUSHRA [23] methodology using high quality 5.1 loudspeaker setups. The tests were carried out at three separate sound labs with a total number of 13 subjects, all of which can be considered expert listeners. Besides the hidden reference and the mandatory 3.5 kHz bandwidth low pass filtered anchor, discrete multi-channel AAC at 320 kbps and Dolby Prologic II in combination with Layer II at 256 kbps for encoding of the stereo downmix were included as additional upper and lower quality anchors, respectively. A set of ten critical items covering a wide variety of content were employed in this test. Table 3 gives an overview of the test excerpts used for the listening test.

Item name	Description
Applaus	ambience (applause)
bach565	Single instrument, church organ Bach d-minor toccata
brassEX	orchestra (exodus)
Fleetwd	transient guitar
Harpsic	Single instrument, harpsichord
hornWag	orchestra (Lohengrien)
Mouthha	mouth organ
mtChoir	Choir
tenorRP	radio drama
Trumpet	Trumpet

Table 3 – Overview of test sequences.

4.2. Audio codecs under test

As described above, three core coders were evaluated in the listening tests. The first, MPEG-1 Layer II is currently used in several digital broadcast systems such as Digital Audio Broadcasting (DAB). By including this codec, the added value of MPEG Surround in a broadcast environment while maintaining backward compatibility with existing stereo receivers can be assessed. For DAB, common bit rates for stereo transmission are in the order of 200-250 kbps, and hence the total bit rates under test for MPEG Surround in combination with Layer II were 192 and 256 kbps.

The second coder that was included is MPEG-4 HE-AAC. MPEG-4 HE-AAC is currently considered as a state-of-the-art audio coder that has found its way in numerous application scenarios such as mobile devices and broadcasting environments. For this coder, total bit rates from 64 kbps to 160 kbps were tested, and a comparison is made between multi-channel HE-AAC discrete coding on the one hand and HE-AAC with MPEG Surround on the other hand. Using these configurations, the change in the rate-distortion curve can be examined that is introduced by MPEG Surround. Given the fact that HE-AAC employs SBR in the high frequency range, this test also provides insight on the effect of non-waveform preserving coders.

Finally, the combination of MPEG-4 AAC-LC and MPEG Surround at a total bit rate of 192 kbps was included as a reference for state-of-the-art waveform-preserving coders. This particular combination could be of interest for mobile music downloads. Table 4 gives an overview of all codecs and anchors used in the test.

4.3. Listening test results

The results of the listening test (averaged across subjects) for various excerpts and codecs are shown in Figure 10. The various excerpts are given along the abscissa; the various codecs are represented by different symbols. The errorbars denote the 95% confidence intervals of the mean. The test included 4 anchors:

- Hidden reference (squares). The scores are very close to 100 and very consistent across excerpts.
- Low-pass anchor (diamonds). The scores of this anchor are the lowest for all excerpts with values between 15 and 22.
- MPEG-1 Layer II at 256 kbps with Dolby Prologic II. This configuration has an average score of about 45 with per-item scores between 38 and 58.
- AAC discrete multi-channel at 320 kbps. This codec provides a consistent high quality at this bit rate with a score above 96.

The leftward and rightward triangles represent HE-AAC discrete multi-channel at a bit rate of 64 and 160 kbps, respectively. As can be observed from the results the quality increases with bit rate from an average of 56 at 64 kbps to 93 at 160 kbps.

Codec ID	Description	Bit rate (kbps)
Ref	Hidden reference	n.a.
LP anchor	3.5 kHz anchor	n.a.
L2+DPL2(256)	Layer-2 + DPL2 encoding and decoding	256
AAC(320)	MPEG-4 AAC-LC	320
HE-AAC(64)	MPEG-4 HE-AAC	64
HE-AAC(160)	MPEG-4 HE-AAC	160
HE-AAC+MPS(64)	MPEG-4 HE-AAC + MPEG Surround	64
HE-AAC+MPS(96)	MPEG-4 HE-AAC + MPEG Surround	96
HE-AAC+MPS(160)	MPEG-4 HE-AAC + MPEG Surround	160
L2+MPS(192)	MPEG-1 Layer 2 + MPEG Surround	192
L2+MPS(256)	MPEG-1 Layer 2 + MPEG Surround	256
AAC+MPS(192)	MPEG-4 AAC-LC + MPEG Surround	192

Table 4 – Codecs and anchors of the subjective listening test.

MPEG Surround applied to a HE-AAC core-coder at a total bit rate of 64 kbps (circles) has an average score of 72.6 and is therefore already in the middle of the 'good' region. The sound quality shows a monotonic increase with increasing bit rate. At 96 kbps (indicated by an x) the mean subjective quality of this combination crosses the border into the 'excellent' region with a mean score of 80.6. At a total bit rate of 160 kbps (asterisk) the mean score (89.5) resides in the middle of the excellent range. This scalability is in line with MPS verification test results [20].

Discrete HE-AAC at 160 kbps and the combination of HE-AAC with MPEG Surround at 160 kbps show overlapping confidence intervals. Therefore both codecs yield the same high quality in the middle of the "excellent" range. A similar excellent quality is provided if MPEG Surround is applied to an AAC core-coder at a total bit rate of 192 kbps (hexagon).

The performance of MPEG Surround using MPEG-1 Layer II as core coder at 192 kbps (cross) received a mean score of 73.7 located in the upper half of the 'good' range. A significant increase in perceptual quality at a slightly higher bit rate of 256 kbps (pentagon) is observed. The overall quality however is partially limited by the capabilities of the core coder (e.g. for tonal items like the harpsichord). The combination of MPEG-1 Layer II with MPEG Surround clearly outperforms its combination with Dolby Prologic II.

Figure 11 shows the average scores of the tested coders as a function of bit rate. This illustrates the rate-distortion curves of MPEG Surround in combination with HE-AAC or MPEG-1 Layer II, in comparison to the rate-distortion curve of HE-AAC discrete. The other codecs are shown as a reference.

5. DISCUSSION

The listening test results reveal that for many operating points and core coders, MPEG Surround ensures backward compatibility inherited from its spatial coding methodology, while at the same time provides a significant increase in coding efficiency. For example when using MPEG-1 Layer II as core coder, the perceived quality for multi-channel content increases from 'fair' to 'excellent' if MPEG Surround is employed instead of Dolby Prologic II. For HE-AAC, the backward compatibility also comes with an increase in efficiency, especially at low bit rates (below 160 kbps).

Even for complex signals such as applause, the perceived quality of MPEG Surround is very competitive compared to discrete coders over a wide bit rate range. For a bit rate of 64 kbps, HE-AAC with MPEG Surround clearly outperforms discrete multi-channel HE-AAC, while at higher bit rates, the confidence intervals all overlap indicating equal perceived quality.

6. CONCLUSIONS

The large repertoire for side-information scalability of the MPEG Surround standard for multi-channel audio coding has been described. Dedicated tools for extremely low side-information rates of only a few kbps up to tens of kbps for very high audio quality have been outlined. The resulting rate-distortion behavior has been shown by means of listening test results. From these results, it is evident that MPEG Surround provides a significant increase in compression efficiency for a wide range of audio codecs and operating points, while at the same time ensures backward compatibility with legacy codecs. The residual coding option of MPEG Surround enables excellent multi-channel audio quality with MUSHRA scores of 90 or higher, which makes MPEG Surround also applicable to environments that demand excellent performance.

7. ACKNOWLEDGEMENTS

The authors would like to thank Frans de Bont from Philips Applied Technologies and Claus-Christian Spenger from Fraunhofer Institute for Integrated Circuits for their support in the preparation of the subjective tests. Furthermore, the authors would like to express their gratitude to the listeners for their valuable time and effort.

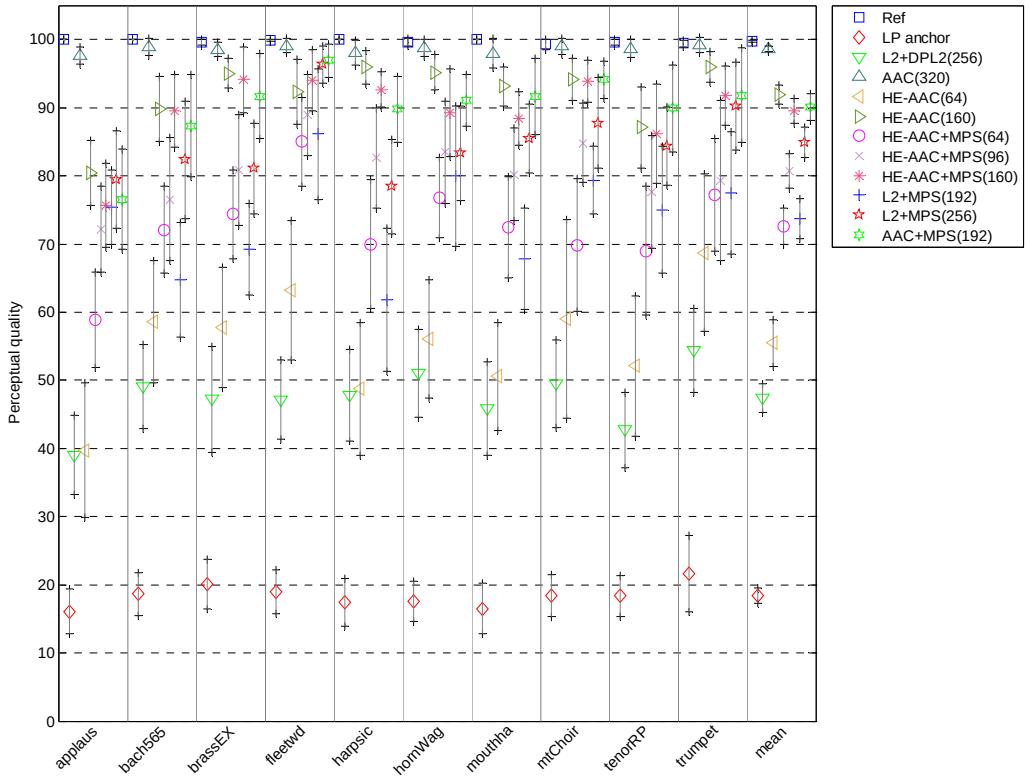

Figure 10 - Listening test results for various excerpts (abscissa) and codecs (symbols). Errorbars denote the 95% confidence interval of the mean.

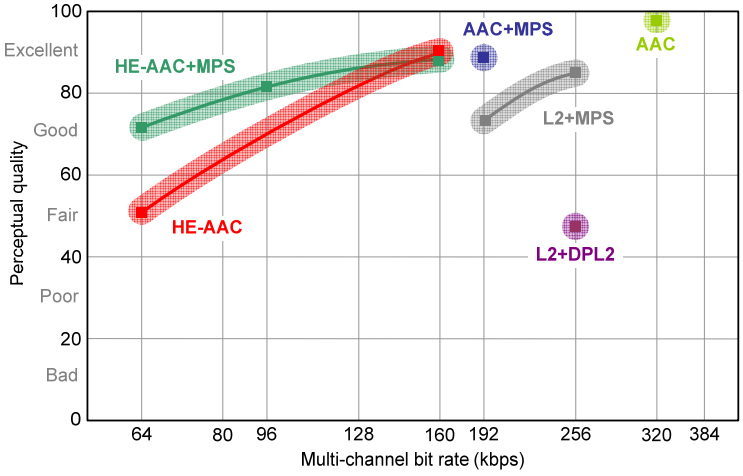

Figure 11 - Perceptual quality (averaged across excerpts) as a function of bit rate for various coder configurations.

REFERENCES

[1] ISO/IEC 23003-1:2007, "Information technology - MPEG audio technologies - Part 1: MPEG Surround", 2007

[2] J. D. Johnston and A. J. Ferreira, "Sum-Difference Stereo Transform Coding" in *Proc. ICASSP* (San-Francisco, CA, 1992).

[3] R. G. van der Waal and R. N. J. Veldhuis, "Subband Coding of Stereophonic Digital Audio Signals" in *Proc. ICASSP* (Toronto, Ont., Canada, 1991).

[4] J. Herre, K. Brandenburg, and D. Lederer, "Intensity Stereo Coding", presented at the 96th AES Convention, *J. Audio Eng. Soc*. Vol 42, p. 394 (May 1994).

[5] F. Baumgarte and C. Faller, "Binaural Cue Coding – Part I: Psychoacoustic Fundamentals and Design Principles" *IEEE Trans SAP*, vol 11, pp. 509-519 (2003).

[6] C. Faller and F. Baumgarte, "Binaural Cue Coding – Part II: Schemes and Applications", *IEEE Trans SAP*, vol 11, pp. 520-531 (2003).

[7] C. Faller and F. Baumgarte, "Efficient Parameterization of Spatial Audio Using Perceptual Parameterization" presented at WASPAA, Workshop on Applications of Signal Processing on Audio and Acoustics (2001).

[8] E. Schuijers, W. Oomen, B. den Brinker, and J. Breebaart, "Advances in Parametric Coding for High-Quality Audio", *Proc. 114th AES convention*, Amsterdam, The Netherlands (2003).

[9] J. Breebaart, S. van de Par, A. Kohlrausch and E. Schuijers, "Parametric Coding of Stereo Audio" *EURASIP J. Appl. Signal Process*., vol 9, pp. 1305-1322 (2004).

[10] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers, "High-Quality Parametric Spatial Audio Coding at Low Bit Rates", *Proc. 116th AES convention*, Berlin, Germany (2004).

[11] E. Schuijers, J. Breebaart, H. Purnhagen, and J. Engdegård, "Low Complexity Parametric Stereo Coding" *Proc. 11th AES convention*, Berlin, Germany (2004).

[12] J. Breebaart and C. Faller, "*Spatial Audio Processing: MPEG Surround and other Applications*" Wiley, London, 2007.

[13] J. Breebaart, G. Hotho, J. Koppens, E. Schuijers, W. Oomen, and S. van de Par, "Background, Concept and Architecture for the Recent MPEG Suround Standard on Multichannel Audio Compression" *J. Audio Eng. Soc*. vol 55, pp. 331-351 (2007).

[14] J. Herre, H. Purnhagen, J. Breebaart, C. Faller, S. Disch, K. Kjörling, E. Schuijers, J. Hilpert, and F. Myburg, "The Reference Model Architecture for MPEG Spatial Audio Coding" *Proc. 118th AES convention*, Barcelona, Spain (2005).

[15] J. Breebaart, J. Herre, C. Faller, J. Röden, F. Myburg, S. Disch, H. Purnhagen, G. Hotho, M. Neusinger, K. Kjörling, and W. Oomen, "MPEG Spatial Audio Coding / MPEG Surround: Overview and Current Status" *Proc. 119th AES convention*, New York, USA (2005).

[16] L. Villemoes, J. Herre, J. Breebaart, G. Hotho, S. Disch, H. Purnhagen, and K. Kjörling, "MPEG Surround: The Forthcoming ISO Standard for Spatial Audio Coding" *Proc. 28th AES conference*, Pitea, Sweden (2006).

[17] J. Breebaart, J. Herre, L. Villemoes, C. Jin, K. Kjörling, J. Plogsties, and J. Koppens, "Multi-channel goes Mobile: MPEG Surround Binaural Rendering" *Proc. 29th AES conference*, Seoul, South Korea (2006).

[18] J. Herre, K. Kjörling, J. Breebaart, C. Faller, S. Disch, H. Purnhagen, J. Koppens, J. Hilpert, J. Röden, W. Oomen, K. Linzmeier, and K. S. Chong, "MPEG Surround – The ISO/MPEG Standard for Efficient and Compatible Multi-Channel Audio Coding", *Proc. 122nd AES convention*, Vienna, Austria (2007).

[19] J. Engdegård, H. Purnhagen, J. Rödén, and L. Liljeryd: "Synthetic ambience in parametric

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

[20] Audio Subgroup, "Report on MPEG Surround Verification Test", ISO/IEC JTC 1/SC 29/WG 11 N8851, 2007 http://www.chiariglione.org/mpeg/working_documents/mpeg-d/sac/VT-report.zip

[21] F. de Bont, and W. Oomen, "Update to CE on MPEG Surround over PCM", ISO/IEC JTC 1/SC 29/WG 11 M13250, 2006.

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

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