The Reference Model Architecture for MPEG Spatial Audio Coding

J. Herre[1], H. Purnhagen[2], J. Breebaart[3], C. Faller[5], S. Disch[1], K. Kjörling[2], E. Schuijers[4], J. Hilpert[1], F. Myburg[4]

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

[2] Coding Technologies, 11352 Stockholm, Sweden

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

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

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

ABSTRACT

Recently, technologies for parametric coding of multi-channel audio signals have received wide attention under the name of "Spatial Audio Coding." In contrast to a fully discrete representation of multi-channel sound, these techniques allow for a backward compatible transmission at bitrates only slightly higher than rates commonly used for mono / stereo sound. Motivated by this prospect, the MPEG Audio standardization group started a new work item on Spatial Audio Coding. This paper reports on the reference model zero architecture, as emerged from the MPEG Call for Proposals (CfP) and the subsequent evaluation of the submissions. The architecture combines the strong features of the two submissions to the CfP that were found best in the evaluation process.

1. INTRODUCTION

With the ongoing quest for an improved consumer experience in the audio world, several dimensions of progress are addressed, such as higher sampling rates and word lengths ('high-resolution audio') and multi-channel sound (see SA-CD, DVD-Audio etc.). While the introduction of multi-channel sound into the consumer domain was initially driven largely by movie sound, nowadays there is also an increasing demand for audio-only program material. Many of today's consumer audio applications could benefit significantly from extending their traditional two-channel stereophonic capability towards surround sound rendering. It is, however, also clear that multi-channel capability comes at a price in terms of required data rate / storage capacity which may not be feasible for many

applications for which bandwidth is an important factor (e.g. Digital Broadcasting, Video on demand, Internet radio and streaming).

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 advantages that are significant in both terms of compression efficiency and user features. Firstly, it allows the transmission of multi-channel audio down to bitrates which so far are 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., spatial audio coding 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.

The paper describes the concepts of spatial audio coding and reports on the ongoing activities of the ISO/MPEG standardization group in this field. The main part of this publication is dedicated to describing the new MPEG reference model zero architecture, as it emerged from the MPEG Call for Proposals (CfP) [2] and the subsequent evaluation of the submissions. The most prominent features and capabilities of the system will be discussed in their context.

2. THE SPATIAL AUDIO CODING IDEA

This section explains the concepts behind the Spatial Audio Coding (SAC) approach.

The basic idea is to capture 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. This is illustrated in Figure 1. In the encoding process, the spatial 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 downmixes (‘artistic downmix‘) may be used. On the decoding side, the cue parameters are used to expand the downmix signal into a high quality multi-channel output.

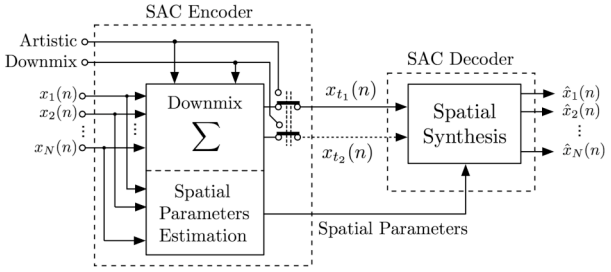

Figure 1: Principle of Spatial Audio Coding

As a result, this approach provides on one hand an extremely efficient representation of multi-channel audio signals due to the reduced number of audio channels to be transmitted (e.g. just one channel for a monophonic downmix signal). On the other hand, a receiver device without a spatial audio decoder will still be able to produce a meaningful output signal by simply presenting the downmix signal.

Conceptually, this is 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 the *parametric stereo* [4][5] to multi-channel application, and extension of the Binaural Cue Coding (BCC) scheme [6] towards using more than one transmitted downmix channel. From a different point of view, the Spatial Audio Coding approach can also be considered an extension of well-known matrixed surround schemes (Dolby Surround/Prologic, Logic 7, Circle Surround …) [7][8] by transmission of dedicated (spatial cue) side information to guide the multi-channel reconstruction process [1] .

Using the SAC technology, a large number of existing mono or stereo services can be enhanced to multi-channel in a backward compatible fashion, by using the existing audio transmission channel for the downmix signal, and sending the spatial parameter information 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 5 to 32 kb/s. Examples for such applications include music download services, streaming music services / Internet radios, Digital Audio Broadcasting, multi-channel teleconferencing and audio for games.

3. MPEG SPATIAL AUDIO CODING

Motivated by the potential of the SAC approach, ISO/MPEG started a new work item on SAC as a next step for MPEG-4 Audio standardization by issuing a CfP on Spatial Audio Coding in March 2004 [2]. A total of 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 cue bitrate and other parameters (additional functionality, computational complexity etc.).

As a result of these extensive evaluations, MPEG decided to define the basis of the subsequent standardization process, called Reference Model 0 (RM0), by combining the submissions of Fraunhofer IIS/Agere Systems and Coding Technologies/Philips [9]. These systems not only outperformed the other submissions but also showed complementary performance to each other in terms of per-item quality, bit-rate and complexity [10]. Consequently, the merged RM0 is designed to combine the best features of both individual systems and serve as the basis for the further technical development of Spatial Audio Coding within MPEG-4 Audio. At the time of writing this publication, the verification of the performance of the merged system is scheduled to be completed by the MPEG meeting in April 2005.

4. MPEG REFERENCE MODEL 0 ARCHITECTURE AND FEATURES

This section provides an overview of the most important features and characteristics of the merged reference model architecture.

Firstly, a general overview of the encoder and decoder structure is provided and the underlying filter bank structure is discussed. Secondly, the types of spatial parameters used are described. Among these parameters, the correlation / coherence of channel signals plays an important role for synthesizing sound images with a wide and enveloping characteristic. The synthesis of correlation/coherence aspects and other important aspects for optimizing coding performance in specific coding situations are discussed. These include the system's ability to handle an externally created (possibly hand-produced) downmix, and compatibility of the downmix signals with existing matrixed surround decoders (such as Dolby Prologic). Finally, some remarks are made about possible multi-channel channel configurations and the Spatial Audio bitstream structure.

4.1. General Overview

A high-level block diagram of the RM0 spatial encoder is outlined in Figure 2. A set of *N* input signals are fed to the encoder. These input channels may represent practically all common channel configurations from two-channel stereo (for simplicity just referred to as 'stereo' from now on within this paper) to complex configurations like a 10.3 setup. The input signals are processed by analysis filter banks to decompose the input signals into separate frequency bands. The frequency selectivity of these filter banks is tailored specifically towards mimicking the frequency resolution of the human auditory system. Furthermore, to enable processing of the frequency-domain signals without introducing audible aliasing distortion, the filter banks should be oversampled. A filter-bank design that fulfills these prerequisites is based on a complex-exponential modulated (Pseudo) Quadrature Mirror Filter (QMF) banks (see also Section 4.2), which enable flexible signal modifications at high computational efficiency [5]. An additional advantage is that this filter bank enables efficient (decoder) integration of spatial audio synthesis and Spectral Band Replication (SBR) [11].

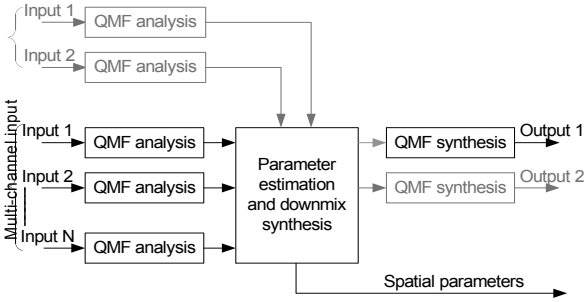

Figure 2: High-level overview of the RM0 Spatial Audio encoder

The sub-band signals resulting from the QMF analysis are subsequently analyzed in order to derive their perceptually relevant spatial properties. These properties are represented by a spatial parameter information stream that includes four basic types of spatial parameters (see also Section 4.3):

- Channel Level Differences (CLDs), representing level differences between pairs of input channels;

- InterChannel Correlations (ICCs), representing coherence or cross-correlation between pairs of input channels;

- Channel Prediction Coefficients (CPCs), representing coefficients to predict (combinations of) output channels from (combinations of) input channels;

- Prediction errors (or residual signals), representing the (waveform) differences between the parametric description of spatial properties described above and the actual waveforms.

In this context, the term *channel* may also denote signals formed from a combination of several audio channels.

A carefully designed downmix process strives to maximize the perceptual quality of the downmix as well as the multi-channel output. The downmix signals are transformed to the time domain by passing them through the QMF synthesis banks. Conventional mono, stereo or multi-channel coding techniques can then be used to encode the downmix signals.

In the application described above, the encoder automatically generates a downmix, which is optimized for mono or stereo playback or playback via a matrix-surround decoder (e.g. Dolby Prologic and compatible devices, see Section 4.8). If an external stereo downmix is presented to the encoder (represented by the additional artistic downmix input at the top in Figure 2), the encoder adjusts its spatial parameters to optimize the multi-channel reconstruction based on the external stereo downmix instead of the automatically generated downmix. Representative application scenarios for such external downmixes are so-called separate 'artistic stereo downmixes' provided by studio engineers / Tonmeisters, or modified downmixes resulting from post-processing algorithms as often deployed for radio transmission.

The corresponding spatial decoder is shown in Figure 3.The time-domain downmix signal(s) are processed by a QMF analysis bank. Subsequently, a spatial synthesis module generates a multi-channel output based on the transmitted spatial parameters. A set of QMF synthesis banks transforms the multi-channel output signals to the time-domain.

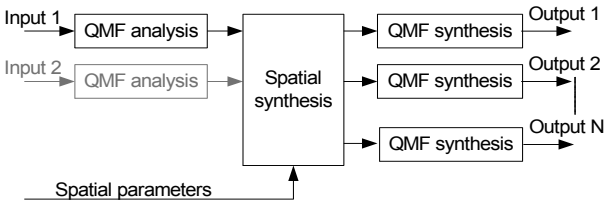

Figure 3: High-level overview of the RM0 Spatial Audio decoder

The 'spatial synthesis' stage consists of matrixing and de-correlation elements. A generalized spatial synthesis block is outlined in Figure 4. The QMF-domain input signals are first processed by a pre-mixing matrix M1. The resulting pre-mixed signals are either directly fed to a post-mixing matrix or fed to the post-mixing matrix via one of the de-correlation circuits D1 … Dm (see Section 4.4). Finally, the so called post-mixing matrix M2 generates the QMF-domain output signals.

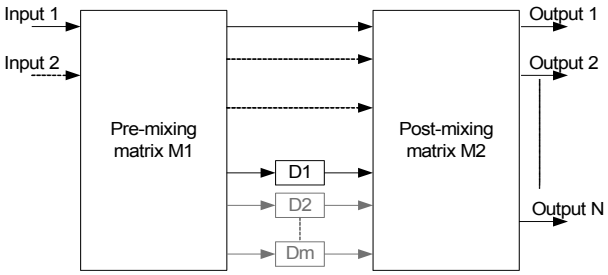

Figure 4: Generalized Spatial Audio synthesis block

The figures above show configurations with multi-channel input signals and a mono or stereo downmix. It should be noted, however, that the overall system is much more flexible in terms of channel configurations. The input channel configuration can range from stereo to 10.3 configurations or more channels. The downmix configuration can be mono, stereo, stereo with matrix surround compatibility, but may also consist of multi-channel formats. Finally, the decoder can synthesize multi-channel output for the same channel configuration as was presented to the encoder, but other channel configurations as well (see also Section 4.9).

4.2. Hybrid QMF Filter Banks

In the human auditory system, the processing of binaural cues is performed on a non-uniform frequency scale [12][13]. Hence, in order to estimate spatial parameters from a given input signal, it is desirable to transform its time-domain representations to a

representation that resembles this non-uniform scale. One way to achieve this is by applying non-uniform, e.g. warped, filter banks. There are, however, a number of disadvantages to these types of filter banks mostly related to the temporal allocation of frequency components.

Another possibility to arrive at a non-uniformly scaled representation is by means of grouping a number of bands of a *uniform* transform. A well-known example of such a suitable linear transform is the Discrete Fourier Transform (DFT).

Furthermore, it is desirable to restrain the complexity of the Spatial Audio Coding system. For applications including low bitrate audio coding, the spatial decoder is typically applied as a post-processing algorithm to a low bitrate mono or stereo decoder. Such a decoder typically also employs a spectral representation of the audio material which would be beneficial if it could be directly reused by the Spatial Audio Coding system. However, in practice, spectral representations for the purpose of audio coding are typically obtained by means of critically sampled filter banks (for example using MDCTs [14]) 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 [11] is an important exception in this respect. Similar to the Spatial Audio Coding system, the SBR algorithm is a post-processing algorithm on top of a conventional (band-limited) low bitrate audio decoder that allows to reconstruct 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. It has been shown previously [5] that this QMF bank can be extended to a hybrid structure to obtain an efficient non-uniform frequency resolution which matches the frequency resolution of the human auditory system.

As a first step, the complex-valued QMF sub-band domain signals $s_k[n]$ are obtained as:

$$s_k[n] = \sum_{l=0}^{L-1} x[n-l]\,p[l]\,e^{j\frac{\pi}{K}(k+\frac{1}{2})(l+\phi)},$$

where x[n] represents the input signal, p[n] the low-pass prototype filter impulse response of order L-1, ϕ represents a phase parameter, K represents the number of frequency bands (K=64) and k the sub-band index (k=0,…,K-1). Subsequently, the signals $s_k[n]$ are down-sampled by a factor K.

At a sampling rate of 44.1 kHz, the 64-bands analysis filter bank results in an effective bandwidth of approximately 344 Hz, which is considerably wider than the required spectral resolution at low frequencies. In order to further improve the frequency resolution, the lower QMF sub-bands are extended with an additional (secondary) filter bank based on oddly-modulated M^{th} band filter banks. The analysis filtering for QMF sub-band k is given by:

$$q_{k,m}[n] = \sum_{\lambda=0}^{\Lambda_k-1} \varsigma[n-\lambda]\,g_k[\lambda]\,e^{j\frac{2\pi}{M_k}(m+\frac{1}{2})(\lambda-\frac{\Lambda_k-1}{2})},$$

where $\varsigma_k[n]$ represents the down-sampled QMF output signal for sub-band k, Λ_k is the prototype length, g_k denotes the prototype filter, M_k the number of (hybrid) frequency bands for QMF band k, and m the hybrid filter band index (m=0,…,M_k). Given certain pass and stop band properties of the QMF analysis filter bank and the secondary filter bank, some of the secondary outputs can be combined by simple addition (see [5] for details). Those QMF bands that already exhibit a sufficient frequency resolution (at high frequencies) are delayed to compensate for the delay of the cascaded filter bank stages. The complete analysis filter bank structure for a certain configuration is outlined in Figure 5.

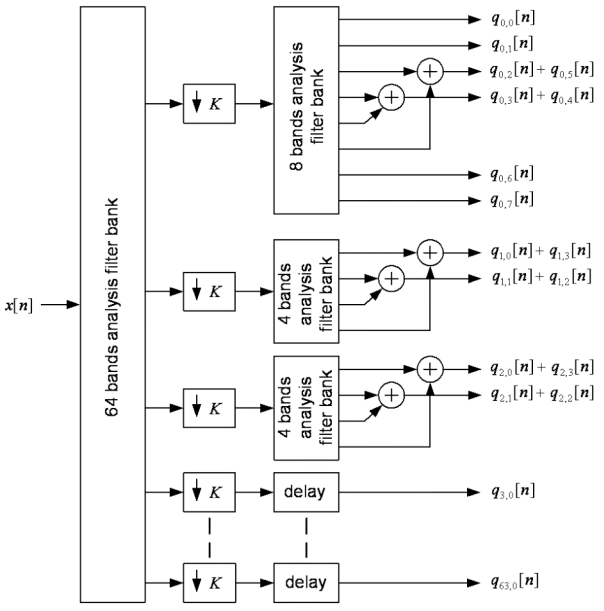

Figure 5: Hybrid analysis filter bank structure

4.3. Spatial Audio Parameters

The Spatial Audio Coding system employs two conceptual elements with which it is possible to describe any arbitrary mapping from M to N channels and back, with N<M. These elements are referred to as the One-To-Two (OTT) element and the Two-To-Three (TTT) element, both named in accordance to the corresponding decoder element. As an example, Figure 6 shows a block diagram of a 5.1 to stereo spatial audio encoder consisting of both a TTT element and a number of OTT elements. 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.

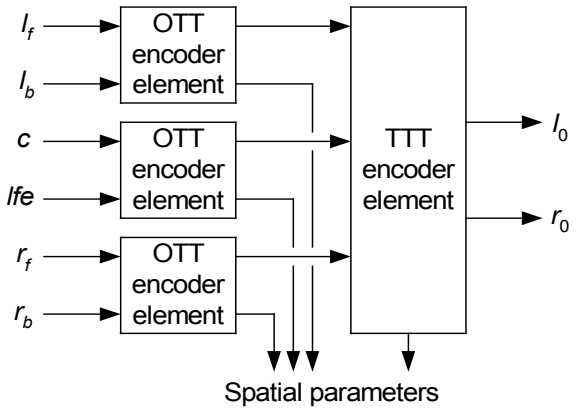

Figure 6: Block diagram of a 5.1 to stereo Spatial Audio encoder

4.3.1. OTT element

The purpose of the OTT encoder element is to create a mono downmix from a stereo input and extract relevant spatial parameters. The OTT element finds its history in the development of Parametric Stereo (PS, [5]) and Binaural Cue Coding (BCC, [6][15]). Similar to the signal models employed there, of the following parameters are extracted:

- Channel Level Differences (CLD) - these represent the time- and frequency-variant level differences between two input channels. The CLD parameters are quantized non-uniformly on a logarithmic scale. CLD parameters close to zero are quantized using a higher accuracy than CLD parameters that are not close to zero (i.e., one input signal is dominant over the other).
- Inter-channel coherence/cross-correlation (ICC) - these represent the time- and frequency-variant coherence or cross-correlation between two input channels. The ICC parameters are quantized on a non-uniform scale.

Furthermore, third output of OTT elements is the residual signal. This signal represents the errors of the underlying parametric model and enables full waveform reconstruction at the decoder side at a minimum bit cost. See Section 4.7 for more details.

4.3.2. TTT element

The TTT element is specifically tailored at resolving a third channel from another signal pair. Hence, the TTT element is appropriate for modeling the symmetrically downmixed center from a stereo downmix pair. The TTT element makes use of three types of parameters:

- Channel Prediction Coefficients (CPC) - the TTT element is based on the following linear signal model:

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

  This model assumes that the downmix signals l_0 and r_0 are a linear combination of the l, c and r signals. It can be shown that if the encoder downmix matrix H_{TTT} is known at the decoder side, only two independent parameters need to be transmitted to optimally recover the [l,c,r] signal triplet from the down-mix signal pair. However, as the signal triplet [l,c,r] will in general consist of only partially correlated signals, a prediction loss will occur.

- Inter-channel coherence/cross-correlation (ICC) Similarly to the OTT element, ICC parameters can be employed. In the case of the TTT element, these serve to model the prediction loss caused by the CPC parameters.

Similar to the OTT element, additional prediction errors (or residual signals) that would be required for full waveform reconstruction of a TTT element can be included in the parametric representation as well.

4.4. Decorrelation

As shown in Figure 4 the spatial synthesis stage of the parametric multi-channel decoder consists of matrixing and decorrelation units. Decorrelation units are required to synthesize output signals with a variable degree of correlation (dictated by the transmitted ICC parameters). To be more specific, each decorrelator should generate an output signal from an input signal according to the following requirements:

1. The coherence between input and output signal should be 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). Said differently, the coherence between input and output should be very small, even if analyzed in narrow frequency bands.

2. Both the spectral and temporal envelope of the output signal should be close to those of the incoming signals.

3. The outputs of multiple decorrelators should be mutually incoherent according to the same constraints as for their input/output relation.

A suitable implementation that meets these requirements is by using lattice all-pass filters, with additional spectral and temporal enhancement tools. A general decorrelator overview is shown in Figure 7.

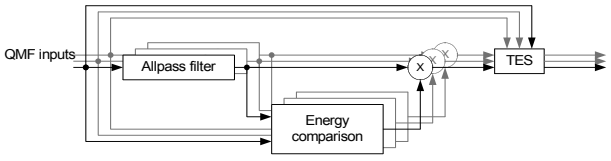

Figure 7: Overview of QMF-domain decorrelator unit

As can be observed from Figure 7, the QMF-domain input signals are first processed by lattice all-pass filters. Careful selection of the reflection coefficients ensures that constraints #1 and #3 are met. For semi-stationary signals, the all-pass behavior of the decorrelation filter already fulfils constraint #2. In the case of dynamically varying input signals, the spectral envelopes are matched by an (averaged) energy-comparison stage, while the temporal envelopes are matched using a Temporal Envelope Shaping (TES) algorithm or Temporal Processing (TP) algorithm (see Section 4.5).

4.5. Special Tools

In addition to the generic elements discussed so far, the Spatial Audio Coding reference Model 0 also includes certain additional tools which are designed to improve the performance of the system in the context of certain input signal types. At this point, three tools are discussed subsequently: Temporal Processing (TP), Temporal Envelope Shaping (TES) and Adaptive Parameter Smoothing.

4.5.1. Temporal Processing (TP)

In a spatial audio coding synthesis system, diffuse sound (see Section 4.4) is generated and mixed with the 'dry' signal in order to control the correlation of the synthesized output channels according to the transmitted ICC values. For transient signals, the diffuse sound generated in the decoder does not automatically match the fine temporal shape of the dry signals and does not fuse well perceptually with the dry signal. This results in poor transient reproduction (‚washed out attacks'), in analogy to the 'pre-echo problem' which is known from perceptual audio coding [16]. One possibility to address this problem is to employ temporal processing (TP) of the diffuse sound.

TP is applied in the time domain (see Figure 8). It basically consists of a temporal envelope estimation of dry and diffuse signal with a higher temporal resolution than that provided by the filter bank of the spatial audio coder. The diffuse signal is re-scaled in its temporal envelope to match the envelope of the dry signal. This results in a significant increase in sound quality for critical transient signals with a broad spatial image / low correlation between channel signals, such as applause.

The envelope shaping is done by matching the short time energy of the wet signal to that of the dry signal. This gives a time-varying gain function that is applied to the diffuse signal, and that will shape the time envelope of the diffuse signal to match that of the wet signal. Additionally, the envelope shaping process is controlled by side information transmitted by the spatial encoder.

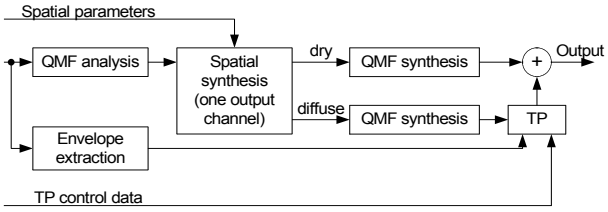

Figure 8: Principle of Temporal Processing (TP)

4.5.2. Temporal Envelope Shaping (TES)

The problem of precise control of the temporal shape of the diffuse sound can also be addressed by the so-called Temporal Envelope Shaping tool, which is designed to be a low complexity alternative to the Temporal Processing tool. While TP operates in the time domain by a time-domain scaling of the diffuse sound envelope, the TES approach achieves the same principal effect by controlling the diffuse sound envelope in a spectral domain representation. This is done in analogy to the Temporal Noise Shaping (TNS) approach [16][17], as it is known from MPEG-2/4 Advanced Audio Coding (AAC) [18]. Manipulation of the diffuse sound fine temporal envelope is achieved by convolution of its spectral coefficients across frequency with a suitable shaping filter derived from an LPC analysis of spectral coefficients of the dry signal. Due to the quite high time resolution of the spatial audio coding filter bank, TES processing requires only low-order filtering (1^{st} order complex prediction) and is thus low in its computational complexity. On the other hand, due to limitations e.g. related to temporal aliasing, it cannot provide the full extent of temporal control that the TP tool offers.

4.5.3. Adaptive Parameter Smoothing

In a low bitrate scenario, a coarse quantization of the spatial parameters (cues) is desirable in order to achieve a low parameter bitrate. In the decoded signal this may lead to audible artifacts, depending on the nature of the audio signals processed. Specifically long term stationary and tonal components, as well as slowly moving point sources require considerable precision of the transmitted parameter values and thus high parameter data rates in order to guarantee an unimpaired listening impression.

'Adaptive Parameter Smoothing' is a tool employed at the decoder side. It aims at adapting the continuity of the quantized and transmitted signal parameters to that of the original signal by temporally smoothing the steps in parameter values introduced by the quantization process. There might also be some audible rapid toggling between adjacent (coarse) quantizer steps, which is also successfully removed by use of this technique. The adaptive smoothing process can be controlled automatically within the decoder (for lowest bitrates) and explicitly controlled from the encoder.

4.6. Artistic Downmix

When looking at today's consumer media which deliver multi-channel audio (DVD-Video/Audio, SA-CD, …) it has become common practice to deliver both dedicated multi-channel and stereo mixes that are stored as separate data on the media. The stereo mix data may contain a 'manual' mix of the recorded sound sources into two-channel stereo created by a sound engineer using hand-optimized mixing parameters, and thus preserving a maximum amount of artistic freedom.

It is recognized that Spatial Audio Coding technology should be able to use such hand-optimized downmix signals (referred to here as 'artistic downmixes') in order to guarantee an optimum listening experience also for a user with a traditional stereo reproduction setup. While support for artistic downmix can certainly be seen as a topic of ongoing research, the MPEG Spatial Audio Coding RM0 already includes some provisions for this purpose.

The SAC decoder aims at reproducing the correct spatial properties that were present in the original multi-channel input. This spatial reconstruction depends predominantly on the transmitted spatial parameters. However, in some cases, the reconstruction quality also depends on (assumed) statistical properties of the transmitted downmix signal (such as the level differences and the cross-correlation). In the case of an artistic downmix, these assumptions on the downmix properties may no longer hold. For example, the panning of a certain sound source may be different in the automatically generated downmix and the corresponding artistic downmix. Furthermore, different effects may have been used, such as the addition of reverberation. Although the spatial audio coding principle is highly robust against downmix modifications, critical conditions may occur in which the reconstruction quality is improved if the actual transmitted downmix is known to the encoder. For this purpose, the encoder shown in Figure 2 has additional input means for artistic downmixes. Knowledge of the transmitted downmix enables the encoder to adjust its

spatial parameters accordingly. In this way, maximum multi-channel reconstruction quality is obtained.

4.7. Rate/Distortion Scalability

Spatial Audio Coding can be useful in a broad range of applications with very different requirements regarding the trade-off between bitrate of the parametric side information and multi-channel quality. 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 cover all conceivable application areas of Spatial Audio Coding, it is important for a such technology to provide sufficient *Rate/Distortion Scalability* which permits to flexibly select an operating point for the trade-off between side information rate and multi-channel audio quality while retaining its generic structure. This concept is illustrated in Figure 9.

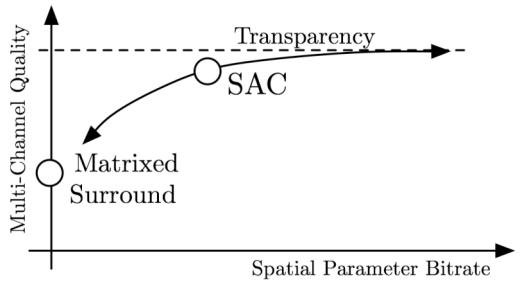

Figure 9: Rate/Distortion Scalability

The technology merging process that led to MPEG Spatial Audio Coding RM0 integrated a number of concepts some of which are designed to permit low spatial overhead operation, and others that aim at further increasing multi-channel audio quality towards transparency. Together, these provide the desired wide range of operating points. Subsequently, two important scalability mechanisms will be discussed, i.e. *parameter scalability* and *residual coding*.

4.7.1. Parameter Scalability

One important dimension of scalability comes from the flexibility of sending spatial parameters at different granularity and resolution. Specifically, different degrees of freedom are provided by the Spatial Audio Coding RM0 technology:

- Parameter frequency resolution
  Clearly, the frequency resolution of spatial audio processing provides an enormous amount of scalability. A high number of frequency bands ensures optimum separation between sound events occupying different frequency ranges, but leads to a higher side information rate. On the other hand, reducing the number of frequency bands saves on spatial overhead and may still provide good quality for most types of audio signals. Currently, a range between 40 and 5 parameter frequency bands is supported by RM0.

- Temporal resolution / parameter update rate
  As a second degree of freedom, the temporal resolution of the transmitted spatial audio parameters can be varied which provides another (approximate) factor of two and more in bitrate scalability. Reducing temporal resolution without unacceptable loss in audio quality is possible because of the use of a flexible adaptive segmentation approach which aligns the parameters' temporal range of impact to the signal characteristics.

- Parameter quantization granularity
  Thirdly, different resolutions for transmitted parameters can selected. Coarser resolution leads to a saving in spatial overhead at the expense of losing some detail in the spatial description. Use of low-resolution parameter descriptions is supported by dedicated tools, such as the *Adaptive Parameter Smoothing* described in Section 4.5.

- Parameter configuration
  For certain parameters, there is a choice as to how extensive the transmitted parametrization describes the original multi-channel signal. An example is the numbers of ICC values transmitted which 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 5kbit/s to 32kbit/s and above.

4.7.2. Residual Coding

While a precise parametric description of the spatial sound image is a sound basis for achieving a high multi-channel audio quality, it is also known that purely parametric coding schemes are not able to scale all the way up to a 'transparent' representation of sound, as this could only be achieved by using a fully discrete multi-channel transmission at a much higher bitrate. To bridge this gap between the audio quality of a parametric description and transparent audio quality, a hybrid coding approach, referred to as residual coding, is available within the Spatial Audio Coding system. In the spatial audio encoding process, a multi-channel signal is downmixed to a lower number of channels (mono or stereo) and spatial cues are extracted. After downmixing, the resulting number of 'dominant' channels are coded and transmitted, while the remaining 'residual' channels are discarded as their perceptually important aspects are covered by the spatial parameters. This operation is illustrated by the following equations. The OTT element generates a dominant (m) and a residual signal (s) from its two input signals l and r. The downmix matrix H_{OTT} minimizes the energy of the residual signal (s) given its modeling capabilities (CLD and ICC). A similar operation is performed by the TTT element, which generates two dominant signals (l_0, r_0) and a (minimized) residual signals.

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

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

The decoder derives the multi-channel output from the transmitted downmix signals, spatial cues, and de-correlated signals. To be more specific, the residual signals (s) are replaced by synthetic residual signals (i.e., the decorrelator outputs) and subsequently, the inverse matrixing is applied to generate OTT and TTT-element output signals (l, r, and c). Although a high audio quality can be attained with this parametric approach [10], this quality level is not always sufficient for very demanding applications. To improve the perceptual quality delivered by the parametric spatial audio coder, a hybrid coding approach is taken where, for a variable bandwidth, the residual (parametric error) signals (s) are transmitted as well. Preferably, the lower frequency range of one or more of the residual channels is coded and transmitted in the spatial bit stream in a scalable fashion. In the decoder, the residual signal replaces the synthetic residual signal that is generated by decorrelation circuits (see Section 4.4). For those frequency ranges where no residual signal is provided, the decorrelator output is used. This process is illustrated in Figure 10.

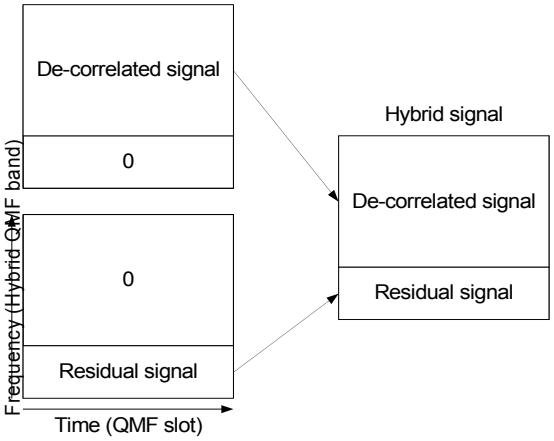

Figure 10: The complementary de-correlated and residual signals are combined into a hybrid signal

With the utilization of residual signals, the spatial audio bit stream is scalable in the sense that the residual-signal data can be stripped from the bit stream, thus lowering the bitrate, while the spatial audio decoder automatically reverts to the parametric operation (i.e., using decorrelator outputs for the full frequency range). The possibility to smoothly scale the residual bandwidth, and thus the total spatial bit rate, provides the flexibility to determine the operation point of the spatial audio coder on the rate-distortion curve. Experiments have shown that utilizing residual signals with a bandwidth of only 2 kHz, and coded at 8kbit/s per residual, provides a clear quality improvement over the parametric-only spatial audio coder operation.

4.8. Matrixed Surround Compatibility

As described in Section 4.1, the Spatial Audio encoder is 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). 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. Given that the transmitted stereo downmix is different for both cases, it is not trivial to meet this multi-channel quality constraint. Said differently, the spatial decoder should preferably be able to ‚invert' the changes made by the encoder to enable matrixed-surround (MTX) compatibility. An elegant solution for this problem is the use of a parameter-controlled post-processing unit that acts on the stereo downmix at the encoder side. A block diagram of a Spatial Audio encoder with this extension is shown in Figure 11.

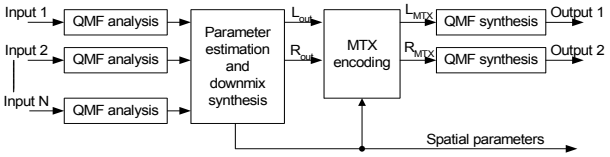

Figure 11: SAC encoder extended with post-processing to enable matrixed-surround compatible downmixes.

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. The processing consists of a matrixing operation applied to the stereo signal in which the actual matrixing equations depend on the spatial parameters. Furthermore, special care is taken to ensure that the inverse of the 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.

The corresponding decoder is shown in Figure 12. The stereo input signal is transformed to the QMF domain by two QMF analysis filter banks, resulting in the frequency-domain signals L_{MTX} and R_{MTX}. Given the constraints on the inverse of the matrixed-surround encoding matrix, the MTX inversion block applies the inverse matrix to recover the original (unmodified) stereo downmix (L_{out}, R_{out}). This stereo downmix is then processed by a regular spatial decoder.

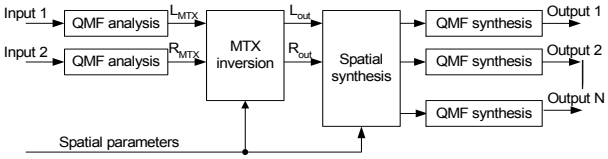

Figure 12: Multi-channel decoder structure for matrixed-surround compatible stereo downmix signals.

There are several advantages to this scheme. Firstly, the matrixed-surround compatibility comes without any additional spatial information. Secondly, the ability to invert the matrixed-surround compatibility processing means 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.

4.9. Channel Configurations

The first round of evaluations of Spatial Audio Coding systems within MPEG was performed with the 5.1 configuration with 3 front channels, 2 surround channels and an LFE (more precisely referred to as 3/2.1) that is common in the current marketplace and found e.g. in most home theater setups tests. In the listening tests, configurations with either a mono or a stereo downmix were assessed, as these are the main application scenarios envisaged. The Spatial Audio Coding approach is, however, flexible to support also other downmix channel configurations. For example, a 3/2.1 downmix could be extended into a 3/4.1 configuration. For further examples, see [19].

The MPEG Spatial Audio Coding RM0 is flexible with respect to the channel configuration of the multi-channel signal being conveyed. In addition to this 3/2.1 configuration, the Spatial Audio Coding system also supports other channel configurations with more or fewer front and side/surround channels and with no, one or several LFE channels. This includes common configurations like 2/2, 3/1, 3/3.1, and 3/4.1. If required, also channel configurations that include elevated ‘top' speakers can be supported. This flexibility w.r.t. multi-channel and downmix channel configurations is achieved by combining the OTT and TTT modules in appropriate structures which correspond the desired input/output formats.

4.10. Bitstream Structure

The Spatial Audio Coding system's bitstream structure is designed to fulfill three major goals:

- The capability of carrying Spatial Audio information for an arbitrary number of downmix and playback channels.
- The ability to scale the overall spatial bitrate according to the needs of the current application scenario.
- Achieving a significant amount of redundancy removal by use of entropy coding mechanisms.

The spatial audio bitstream contains two types of information blocks:

- The Spatial Audio Specific Configuration. This block contains information that is valid for the whole Spatial Audio bitstream and can be seen as a bitstream header. It reflects the channel configuration present at the encoder side and contains set up information for the different decoder processing modules, for instance presence of residual signals, operation modes for OTT and TTT boxes. General system parameters like e.g. time and frequency resolution are signaled in the Spatial Specific Configuration as well.
- The Spatial Audio Frame Data. The container for spatial cues belonging to one frame of the downmixed audio. It can be regarded as the bitstream payload. The data structure of one spatial frame corresponds to the structure reflected by corresponding numbers of OTT and TTT modules. For each available module the necessary set of parameters is transmitted. OTT modules carry CLD and ICC data. TTT modules are fed with CPC parameters and ICC data. Optionally, the coded residual signals can be included for OTT and TTT modules when aiming at very high quality spatial coding. Additionally, the spatial audio frame may contain time varying control information, e.g. for the temporal shaping and smoothing of cues.

CLD, ICC and CPC cues are entropy coded to further remove redundancy. Differential coding of these parameters can be used both in time and frequency direction. A 2-dimensional variable length code (Huffman code) is applied to these differentially coded quantized cue values. To achieve a deterministic upper border for the spatial information rate, PCM coding of the cues can be applied whenever the coding gain of the Huffman coder would be suboptimal.

By separating configuration (header) and payload (frame data), the bitstream structure follows the MPEG-4 paradigm of being transport agnostic. This allows to easily accommodate the varying requirements of different application scenarios for embedding the spatial audio information into ancillary data containers of the core coder bitstream or for transmitting it via the bitstream multiplex of a more complex system.

5. CONCLUSIONS

Spatial Audio Coding is a promising new technology for bitrate-efficient and backward compatible representation of multi-channel audio signals. The approach enables the carriage of such signals at data rates close to the rates used for the representation of two-channel (or even monophonic) audio. The standardization of Spatial Audio Coding technology is on its way within the ISO/MPEG group. The paper described the technical architecture and capabilities of the MPEG Spatial Audio Coding Reference Model 0 architecture, which results from a combination of the strongest features of the two systems found best during the MPEG CfP evaluation process. The architecture provides a wide range of scalability, which helps to cover almost any conceivable application scenario. Even though further improvements will undoubtedly be made in the course of the standards process, the performance of the contributing systems indicates that Spatial Audio Coding technology is able to offer a sound quality substantially beyond that of matrixed surround systems that are common in the market place today.

6. ACKNOWLEDGEMENTS

The authors would like to extend sincere thanks to Thorsten Kastner and Patrick Warmbold of Fraunhofer IIS for their invaluable help in preparing the manuscript.

7. 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] ISO/IEC JTC1/SC29/WG11 (MPEG), Document N6455, “Call for Proposals on Spatial Audio Coding”, Munich 2004.

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

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

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

[9] ISO/IEC JTC1/SC29/WG11 (MPEG), Document N6814, “Workplan for MPEG-4 Spatial Audio Coding”, Palma de Mallorca 2004.

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

[11] 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.

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

[13] 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).

[14] 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

[15] 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

[16] 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

[17] 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

[18] 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

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