Background, Concept, and Architecture for the Recent MPEG Surround Standard on Multichannel Audio Compression*

JEROEN BREEBAART,[1] *AES Member,* **GERARD HOTHO,**[1] **JEROEN KOPPENS,**[2] **ERIK SCHUIJERS,**[2]
(jeroen.breebaart@philips.com) (gerard.hotho@hotmail.com) (jeroen.koppens@philips.com) (erik.schuijers@philips.com)

WERNER OOMEN,[2] *AES Member,* **AND STEVEN VAN DE PAR,**[1] *AES Member*
(werner.oomen@philips.com) (steven.van.de.par@philips.com)

[1]*Philips Research Laboratories, 5656 AE Eindhoven, The Netherlands*
[2]*Philips Applied Technologies, 5616 LW Eindhoven, The Netherlands*

An overview of the recently finalized ISO/MPEG standard for multichannel audio compression MPEG Surround is provided. This audio compression scheme enables backward-compatible multichannel audio coding and transmission at unsurpassed coding efficiency. This is achieved by generating a mono, stereo, or matrixed-surround compatible down mix, which can be transmitted using any existing mono or stereo service, extended with a small amount of parametric side information that describes the perceptually relevant spatial properties of the original multichannel content. The concepts behind spatial parameterization are outlined, and the architecture of the MPEG Surround system is explained. Results of subjective evaluations are included to demonstrate its efficiency.

0 INTRODUCTION

Approximately half a century after the introduction of two-channel stereophony, multichannel sound is now on its way into consumers' homes as the next step toward a more realistic audio reproduction. Initially multichannel audio was present predominantly in the movie domain on consumer media (DVD, for example). The widespread availability of movie material with multichannel sound tracks led to a fast penetration of multichannel playback devices in consumers' homes. Recently, probably in part due to the increased popularity of multichannel movie material, the demand for a compelling surround experience has extended to the audio-only market as well (such as SACD and DVD-audio).

In contrast, the traditional broadcast services (such as radio and television) are still operating in stereo mode due to bandwidth and compatibility constraints. In audio transmission systems the required bandwidth (or amount of information) of a six-channel broadcast would require approximately three times as much bandwidth as a conventional stereo broadcast. In many cases this increased amount of information is undesirable or unavailable. Even if the increased bandwidth would be available, the upgrade process of a stereo service to multichannel audio should ensure that existing stereo receivers will still operate as before. With the existing technology this means an even larger increase in bandwidth for simulcast of stereo and multichannel audio.

Subband or transform coders such as those standardized in MPEG typically employ the concept of monaural perceptual masking to introduce quantization noise in time and frequency tiles, where this noise is (just) inaudible. The quantization noise is introduced either in each channel independently [1], or on a mid/side projection [2], [3] in case two channels contain a significant amount of mutual information. In other words, except for a method referred to as "intensity stereo" [4], any spatial perceptual irrelevancies are hardly exploited. However, recent developments in the field of audio compression have resulted in significant increases in efficiency for stereo and multichannel audio. More specifically, techniques such as parametric stereo [5]–[7] and binaural cue coding [8]–[13]

*Manuscript received 2006 November 17; revised 2007 April 2.

model the perceptually relevant properties using a parametric approach. For multichannel audio this approach is often referred to as spatial audio coding [14]–[17]. An important difference of this parametric approach in comparison to traditional signal quantization methods is its focus on modeling perceptually relevant spatial information in a parametric domain instead of removal of irrelevant information from the signal (subband or transform) domain. This leads to two important advantages over traditional compression methods. First the compression efficiency of parametric methods is significantly higher than that of traditional methods. Second it enables full backward compatibility with existing mono or stereo services.

In this paper the psychophysical basis and concepts of spatial audio coding are explained. Subsequently the incorporation of spatial audio coding technology in the recently finalized ISO/MPEG standard, MPEG Surround, is outlined. Given the complexity and the large variety of features of this standard, the focus is on describing the basic processing stages and their relations, rather than giving a detailed system description and overview of all possible configurations. (Such is provided in [18].) Finally its performance in terms of compression efficiency for multichannel audio is demonstrated.

The MPEG Surround standard emerged from activities of the MPEG Audio standardization group. In March 2004 MPEG issued a call for proposals (CfP) requesting technology in the field of spatial audio coding [19]. In response to this CfP various companies responded with a total of four submissions. The subjective evaluation of these submissions was concluded in October 2004. The test results revealed that there were two out of the four submissions that showed complementary performance. One of these systems was submitted by Coding Technologies/Philips and embodied a multichannel extension to earlier developments on parametric stereo as employed in HE-AAC v2. The other system was developed by Fraunhofer IIS/Agere and was based on binaural cue coding as employed in MP3 Surround. The proponents of both systems decided to cooperate and define a single system, to combine the best of both propositions. Beginning 2005, this resulted in reference model 0 (RM0), the starting point for the collaborative phase within the MPEG Audio group. Numerous core experiments have been conducted by various companies in order to improve and extend the MPEG Surround system, including a low-complexity mode and a dedicated binaural decoding mode to simulate a virtual multichannel loudspeaker setup over stereo headphones. The standard specification of MPEG Surround [20] has been finalized in July 2006.

1 PSYCHOACOUSTIC BACKGROUND

Spatial perception of audio is mediated by a limited set of cues that are created in a natural way due to the properties of sound propagation. For example, a sound source that is placed toward the left side of a listener will result in different acoustical pathways toward the left and right ears. As a result the sound arriving at the left ear will be leading in time compared to the sound arriving at the right ear, creating an interaural time difference (ITD) [21]. Due to the acoustic shadow effect of the head, the signal at the right ear will also tend to be lower in intensity than at the left ear, especially at high frequencies, creating an interaural level difference (ILD) [21]. In line with these acoustical laws, it has been observed that ITDs and ILDs are binaural cues that influence the perceived direction of a sound source in the horizontal plane (see, for example, [22]).

A sound source that is placed in an echoic environment will create numerous reflections that, together with the direct sound, arrive at both ears with many different time delays and amplitudes. As a result the signals at the left and right ears will be (partially) incoherent, that is, the maximum of the normalized interaural correlation function is smaller than 1. This reduction in interaural correlation is perceived as a widening of the sound source [23].

Besides the ITDs and ILDs, additional localization cues result from the direction-dependent acoustical filtering of the outer ear. Specifically in the perceptually relevant region from 6 to 10 kHz, sharp peaks and valleys are found, which result from the acoustical filtering of the head and pinna [24]–[26]. These spectral features allow listeners to differentiate between sounds arriving from the back and front directions, and to perceive the elevation of a sound source.

When listening to a multichannel loudspeaker setup, all these spatial cues play a role in creating the perceived spatial sound image. Under most practical circumstances signals that are played through one loudspeaker can be localized accurately at the position of the loudspeaker using these binaural cues. When identical signals are played simultaneously on the left and right loudspeakers, a phantom source is created in between the two loudspeakers, assuming that the listener is sitting at an equal distance from both loudspeakers [27]. The reason that a single image is perceived in the middle instead of two separate images at the two loudspeakers is that the left and right loudspeaker sounds are mixed at the entrance of the ear canal in a very similar way in both ears. As a result no effective interaural time or level differences are perceived; only the pinna cues contribute to the perceived elevation.

When identical sounds are played on the left and right front loudspeakers, with the left signal having a higher intensity, there will be differences in the signals entering the ear canals. Both the left and the right ears will receive the signals from the left and right loudspeakers. At low frequencies, the left loudspeaker signal will be dominating in both ears due to its higher level (because of the absence of a head-shadow effect) and predominantly determine the arrival time of the signal. Since the left loudspeaker is closer to the left ear and the right loudspeaker closer to the right ear, the composite signal at the left ear will be leading in time, whereas the right ear receives a delayed left loudspeaker signal and therefore the composite signal will tend to be lagging in the right ear. As a result binaural ITD cues are present at low frequencies that will create a localization of the sound toward the left loudspeaker while at high frequencies head shadow effects will create ILD cues resulting from cross-channel level differences (CLDs) since the left signal will arrive attenuated at the right ear and vice versa [28].

Often signals will be played over two loudspeakers that result from the same source, but will have gone through different acoustical pathways before being recorded with two microphones. For example, this occurs when recording a single sound source in an echoic room with two microphones placed at different positions. When playing these microphone signals one to one through left and right frontal loudspeakers, the mixed signals at the ear canal will tend to have an interaural correlation that is reduced significantly compared to the situation where identical signals would be played on both loudspeakers. As discussed earlier, a reduction in interaural correlation will result in an increase of the perceived source width.

In general the interchannel level differences (ICLDs) and interchannel time differences (ICTDs), together with the interchannel correlation (ICC), will be transformed into binaural ITDs, ILDs, and interaural correlation cues at the entrance of the two ears. The exact transformation will depend on the loudspeaker placement, the room acoustic properties, and the relevant anthropometric aspects of the listener. Nevertheless it is clear that the across-channel differences define the binaural cues and are therefore also defining the spatial image.

In practical situations binaural cues will not be constant across time nor frequency. The spectral resolution for perceiving binaural cues seems to be mainly determined by the resolution imposed by the peripheral auditory system [29], [30]. A good approximation of this resolution is given by the ERB scale derived from various monaural masking experiments [31], [32].

The human hearing system can track sound source positions that change over time given certain restrictions. For example, the perception of temporal changes in binaural cues has been shown to be rather sluggish. For ITDs, already at a rate of fluctuation of 10 to 20 Hz, listeners cannot follow the movement at all and hear a spatially widened image [33] reflecting that the long-term interaural correlation of the fluctuating stimulus is less than 1. For ILDs, the binaural system seems to be less sluggish, although it still tends to become less sensitive to dynamic ILDs above rates of fluctuation of 50 Hz [34] for low frequencies. The perception of changes in interaural correlation has also been reported to be very sluggish [35].

If a binaural signal has an interaural correlation that is less than 1 (or more precisely, a coherence less than 1 if temporal alignment is taken into account), it implies that there is a difference between the two signals. The relative intensity of the difference signal compared to the common signal determines the reduction in interaural correlation [36] and contributes in this way to the perceived widening of the sound source. Although the presence of the difference signal is highly detectable, it has been shown that listeners are not very sensitive to the character of the difference signal [37].

The binaural ITDs, ILDs, and interaural correlation cues provide simple statistical relations between the acoustic signals that arrive in the left and right ears, which together form in fact the basic cues for the spatial perception of sound. Therefore it should be possible to reinstate the original spatial illusion that is present in a two-channel recording by imposing the proper binaural cues on a mono down mix of a two-channel recording taking into account the spectral and temporal resolution of the binaural hearing system. Breebaart et al. [5] showed that this is indeed possible, maintaining a high audio quality for stereo recordings. In their work a two-channel input signal was down-mixed to a mono signal, and in addition the spectrotemporal patterns of binaural cues were analyzed. The spatial parameters derived from this analysis were encoded at a very low bit rate, creating a significant reduction in overall bit rate because only a single instead of two audio signals needed to be encoded in the bit stream. With this information it was possible at the decoder side to recreate a high-quality spatial stereophonic audio signal.

The current work extends this concept toward multichannel conditions, where spatial parameters are derived from the multichannel audio signal such that across channels differences in level and correlation are extracted accurately, and can be imposed on a down mix at the decoder side. By creating a multichannel up mix in this way, the multichannel reconstruction will result in binaural cues at the two ears very similar to those that would result from the original multichannel signal.

2 SPATIAL AUDIO CODING

2.1 Concept

The concept of spatial audio coding as employed in the MPEG Surround standard [20] is outlined in Fig. 1. A multichannel input signal is converted to a down mix by an MPEG Surround encoder. Typically the down mix is a mono or a stereo signal, but more down-mix channels are also supported (for example, a 5.1 down mix from a 7.1 input channel configuration). The perceptually relevant spatial properties of the original input signals that are lost by the down-mix process are captured in a spatial parameter bit stream. The down mix can subsequently be encoded with an existing compression technology. In the last encoder step the spatial parameters are combined with the down-mix bit stream by a multiplexer to form the output bit stream. Preferably the parameters are stored in an ancillary data portion of the down-mix bit stream to ensure backward compatibility.

Fig. 1(b) outlines the MPEG Surround decoding process. In a first stage the transmitted bit stream is split into a down-mix bit stream and a spatial parameter stream. The down-mix bit stream is decoded using a legacy decoder. Finally the multichannel output is constructed by an MPEG Surround decoder based on the transmitted spatial parameters.

The use of an MPEG Surround encoder as a preprocessor for a conventional (legacy) codec (and a corresponding postprocessor in the decoder) has important advantages over existing multichannel compression methods.

- The parametric representation of spatial properties results in a significant compression gain over conventional multichannel audio codecs, as will be shown in Section 5.
- The use of a legacy codec with an additional spatial parameter stream allows for backward compatibility with existing compression schemes and broadcast services.
- The spatial parameterization enables novel techniques to process or modify certain aspects of a down mix. Examples are matrixed-surround compatible down mixes, support for so-called artistic down mixes or the generation of a three-dimensional/binaural signal to evoke a multichannel experience over legacy headphones.
- The channel configuration at the spatial encoder can be different from that of the spatial decoder without the need of full multichannel decoding as intermediate step. For example, a decoder may directly render an accurate four-channel representation from a 5.1 signal configuration without having to decode all 5.1 channels first.

2.2 Elementary Building Blocks

The MPEG Surround spatial coder structure is composed of a limited set of elementary building blocks. Each elementary building block is characterized by a set of input signals, a set of output signals, and a parameter interface. A generic elementary building block is shown in Fig. 2. An elementary building block can have up to three input and output signals (as shown left and right, respectively), as well as an input or output for (sets of) spatial parameters.

Different realizations of elementary building blocks serve different purposes in the spatial coding process. For example, a first type of building block may decrease the number of audio channels by means of spatial parameterization. Hence if such a block is applied at the encoder side, the block will have fewer output channels than input channels, and has a parameter output. The corresponding block at the decoder side, however, has a parameter input and more output channels than input channels. The encoder and decoder representations of such an encoding/decoding block are shown in Fig. 3(a) and (b). Two different realizations of the encoding/decoding blocks exist. The first realization is a block that describes two signals as one down-mix signal and parameters. The corresponding encoding block is referred to as two-to-one (TTO), whereas the decoding block is termed one-to-two (OTT). In essence, these blocks are similar to a parametric stereo encoder/decoder [5]–[7], [38]–[40]. The second realization is a so-called three-to-two (TTT) encoding block, which generates two output signals and parameters from three input signals. The corresponding two-to-three decoding block generates three signals from a stereo input accompanied by parameters.

A second type of building block is referred to as signal converter. For example, a stereo input signal may be con-

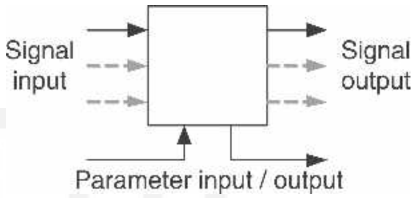

Fig. 2. Generic elementary building block for MPEG Surround coding process.

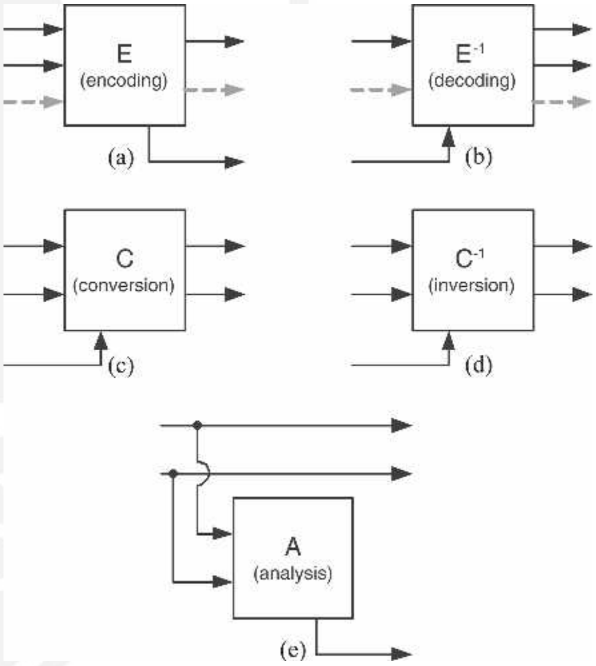

Fig. 3. Elementary building blocks for MPEG Surround coding process.

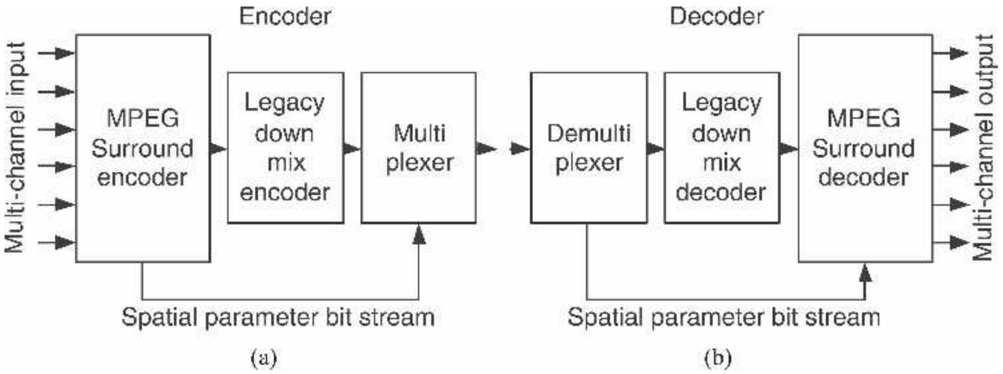

Fig. 1. Multichannel encoder and decoder according to spatial audio coding concept.

verted into a stereo output signal that has different spatial properties, and the processing of which is controlled by parameters. This is shown in Fig. 3(c). The corresponding decoder-side operation [as shown in Fig. 3(d)] inverts the processing that is applied at the encoder to retrieve the original (unmodified) stereo input signal. Examples of signal converters are the conversion from conventional stereo to matrixed-surround compatible stereo or to three-dimensional/binaural stereo for playback over headphones.

The third type of building block is an analysis block. This type generates parameters from a signal stream without modifying the actual signals or signal configuration. This block, which can be applied at both the spatial encoder and the decoder sides, is shown in Fig. 3(e).

3 MPEG SURROUND ENCODER

3.1 Structure

The structure of the MPEG Surround encoder is shown in Fig. 4. A multichannel input signal is first processed by a channel-dependent pregain. These gains enable adjustment of the level of certain channels (for example, LFE and surround) within the transmitted down mix. Subsequently the input signals are decomposed into time or frequency tiles using an analysis filter bank. A spatial encoder generates a down-mix signal and (encoded) spatial parameters for each time or frequency tile. These parameters are quantized and encoded into a parameter bit stream by a parameter encoder Q. The down mix is converted to the time domain using a synthesis filter bank. Finally a postgain is applied to control the overall signal level of the down mix.

3.2 Pre- and Postgains

In the process of down-mixing a multichannel signal to a stereo signal, it is often desirable to have nonequal weights for the different input channels. For example, the surround channels are often attenuated by 3 dB prior to the actual down-mix process. MPEG Surround supports user-controllable pregains between 0 and −6 dB, in steps of 1.5 dB. For the LFE, these weights are adjustable between 0 and −20 dB in steps of 5 dB.

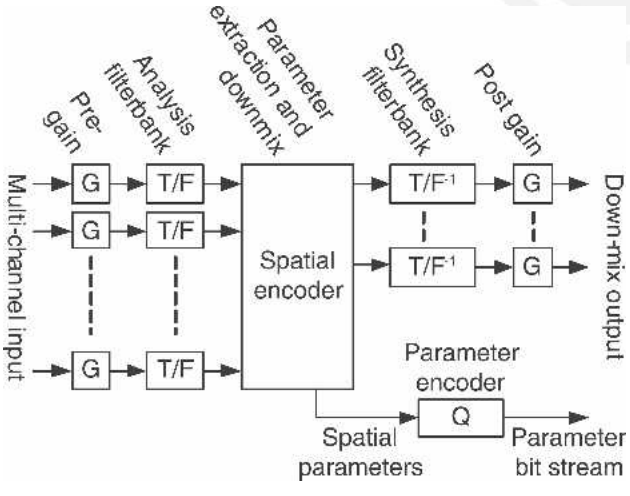

Fig. 4. Structure of MPEG Surround encoder.

The level of the generated down mix can also be controlled using (postencoder) gains to prevent clipping in the digital signal domain. The down mix can be attenuated between 0 and −12 dB in steps of 1.5 dB.

The applied pre- and postgain factors are signaled in the MPEG Surround bit stream to enable their inverse scaling at the decoder side.

3.3 Time–Frequency Decomposition

As outlined in Section 1, the human auditory system determines spatial properties based on a certain time and frequency decomposition. Therefore spatial audio parameterization cannot be employed directly on time-domain signals, but requires a filter bank to mimic the temporal and spectral resolution of the human listener. Moreover, given the need for time-variant processing (especially at the spatial decoder side), the filter bank used is preferably oversampled to reduce aliasing artifacts that would otherwise result from a critically sampled structure.

3.3.1 Analysis Filter Bank

The applied filter bank is a hybrid complex-modulated quadrature mirror filter bank (QMF), which is an extension of the filter bank used in spectral band replication (SBR) techniques [41]–[43]. The hybrid QMF analysis filter bank consists of a cascade of two filter banks. The structure is shown in of Fig. 5(a).

The first filter bank (QMF analysis) is compatible with the filter bank as used in SBR. The subband signals, which are generated by this first filter bank, are obtained by convolving the input signal $x[n]$ with a set of analysis filter impulse responses G_{m_0} $[n]$ given by

$$G_{m_0}[n] = g_0[n] \exp\left[j \frac{\pi}{4M_0} (2m_0 + 1)(2n - 1) \right] \quad (1)$$

with $g_0[n]$, for $n = 0, \ldots, N_0 - 1$, being the prototype window of the filter, $M_0 = 64$ the number of output channels, m_0 the subband index ($m_0 = 0, \ldots, M_0 - 1$), and $N_0 = 640$ the filter length. The filtered outputs are subsequently down sampled by a factor M_0 to result in a set of down-sampled QMF outputs (or subband signals) $X_{m_0}[k]$,

$$X_{m_0}[k] = (x[n] * G_{m_0}[n])[M_0 k]. \quad (2)$$

The down-sampled subband signals $X_{m_0}[k]$ of the first three QMF subbands are subsequently fed through a second complex-modulated filter bank (subfilter bank) of order N_1 to further enhance the spectral resolution in the low-frequency region. The remaining 61 subband signals are delayed to compensate for the delay that is introduced by the subfilter bank. The output of the hybrid (that is, combined) filter bank is denoted by $X_m[k]$, with m the index of the hybrid QMF bank. To allow easy identification of the two filter banks and their outputs, the index m_0 of the first filter bank will be called subband index, and the index m_1 of the subfilter bank is called subsubband index. The subfilter bank has a filter order of $N_1 = 12$, and an impulse response $G_{m_1}[k]$ given by

$$G_{m_1}[k] = g_1[k] \exp\left[j \frac{2\pi}{M_1} \left(m_1 + \frac{1}{2} \right) \left(k - \frac{N_1}{2} \right) \right] \tag{3}$$

with $g_1[k]$ the prototype window, k the sample index, and M_1 the number of subsubbands. Table 1 gives the number of subsubbands $M_1(m_0)$ as a function of the QMF band m_0. Although eight or four subsubbands are used in the second filter bank, some of the pass bands of the subsubfilters coincide with the stop band of the first QMF filter bank. Consequently such subsubband outputs are combined (summed) with complex conjugated counterparts, resulting in six or two subsubfilter output channels, respectively.

As a result of this hybrid QMF filter-bank structure, 71 down-sampled filter outputs $X_m[k]$ are available for further processing, with m the subband index of the complete filter bank, $m = 0, \ldots, 70$.

3.3.2 Segmentation

The subband signals are split into (time) segments. The analysis window length (or the corresponding parameter update rate) matches the lower bound of the measured time constants of the binaural auditory system (that is, between 23 and 100 ms). Dynamic window switching is used in the case of transients to account for the precedence effect, which dictates that only the first 2 ms of a transient in a reverberant environment determine its perceived location. The parameter set(s) resulting from each segment and their temporal positions are organized in frames. A frame has a fixed length (for example, 16, 32, or 64 QMF samples) and is typically aligned with the frame length of the down-mix encoder. Each frame may comprise multiple sets of parameters, each with its own temporal position and analysis window length, depending on the segmentation process at the encoder side.

3.3.3 Parameter Bands

The 71 subsubband signals are grouped into so-called parameter bands, which share common spatial parameters. Each parameter band comprises one or a set of adjacent subsubbands to form the corresponding time or frequency tiles for which spatial parameters are estimated. For the highest frequency resolution supported by MPEG Surround, the number of parameter bands amounts to 28, resulting in a frequency resolution that is closely related to the ERB scale. Bit-rate or quality tradeoffs are supported by coarser frequency resolutions, resulting in different combinations of subsubband signals into respective parameter bands. The following alternative numbers of parameter bands are supported: 4, 5, 7, 10, 14, and 20.

3.3.4 Synthesis Filter Bank

The spatial encoding process is followed by a set of hybrid QMF synthesis filter banks (one for each output channel), also consisting of two stages [see Fig. 5(b)]. The first stage comprises the summation of the subsubbands $\hat{X}_{i,m_1}$, which stem from the same subband m_0,

$$\hat{X}_{i,m_0}[k] = \sum_{m_1=0}^{M_1(m_0)-1} \hat{X}_{i,m_1}[k]. \tag{4}$$

Finally, up sampling, convolution with synthesis filters [which are similar to the QMF analysis filters as specified by Eq. (1)], and summation of the resulting subband signals results in the final outputs $\hat{x}_i[n]$.

3.4 Spatial Encoder

3.4.1 Tree Structures

The elementary building blocks (as described in Section 2.2) are combined to form a spatial coding tree. Depending on the number of (desired) input and output channels, and additional features that are employed, different tree structures may be constructed. The most common tree structures for 5.1-channel input will be outlined here. First two tree structures for a mono down mix will be described, followed by the preferred tree structure for a stereo down mix.

The first tree structure supports a mono down mix and is outlined in Fig. 6(a). The six input channels, left front, right front, left surround, right surround, center, and low-frequency enhancement, labeled L_f, R_f, L_s, R_s, C, and LFE, respectively, are combined pairwise using encoding blocks (TTO type) until a mono down mix is obtained. Each TTO block produces a set of parameters P. As a first step the two front channels (L_f, R_f) are combined into a TTO encoding blocks E_3, resulting in parameters P_3. Simi-

Table 1. Specification of M_1 and resulting number of output channels for first three QMF subbands.

QMF Subband (m_0)	$M_1(m_0)$	Output Channels
0	8	6
1	4	2
2	4	2

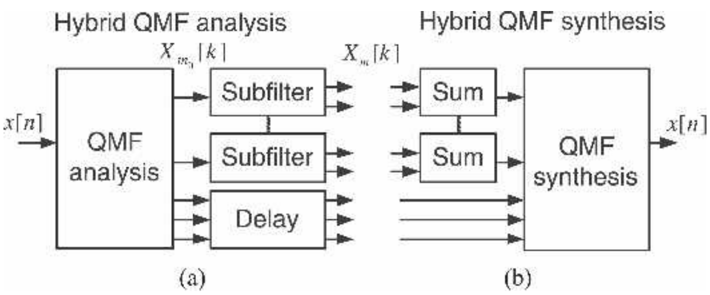

Fig. 5. Structure of hybrid QMF analysis and synthesis filter banks.

larly, the pairs C, LFE and L_s, R_s are combined by TTO encoding blocks E_4 and E_2, respectively. Subsequently the combination of L_f, R_f on the one hand, and C, LFE on the other hand are combined using TTO encoding block E_1 to form a "front" channel F. Finally this front channel is merged with the common surround channel in encoding block E_0 to result in a mono output S.

One of the advantages of this structure is its support for configurations with only one surround channel. In that case L_s and R_s are identical, and hence the corresponding TTO block can be omitted (that is, the tree can be pruned).

The second tree structure for 5.1 input combined with a mono down mix is shown in Fig. 6(b). In this configuration the L_f and L_s channels are first combined into a common left channel (L) using a TTO encoding block E_3. The same process is repeated for the R_f and R_s channels (E_4). The resulting common left and common right channels are then combined in E_1, and finally merged (E_0) with the combination of the center and LFE channels (E_2). The advantage of this scheme is that a front-only channel configuration (that is, only comprising L, R, and C) is simply obtained by pruning the tree.

For a stereo down mix the preferred tree configuration is given in Fig. 7. As for the second mono-based tree, this tree also starts by the generation of common left and right channels, and a combined center/LFE channel. These three signals are combined into a stereo output signal S_L, S_R using a TTT encoding block (E_3).

3.4.2 TTO Encoding Block

The TTO encoding block transforms two input channels X_1, X_2 into one mono output channel X_s plus spatial parameters. Its concept is identical to a parametric stereo encoder ([5]–[7], [38]–[40]). For each parameter band two spatial parameters are extracted. The first comprises the channel level difference (CLD) between the two input channels for each parameter band b,

$$\mathrm{CLD}_b = 10 \log_{10} \frac{\sigma^2_{X_1,b}}{\sigma^2_{X_2,b}} \tag{5}$$

with $\sigma^2_{x_i,b}$ the energy of signal X_i in parameter band b,

$$\sigma^2_{X_i,b} = \sum_k \sum_{m=m_b}^{m_{b+1}-1} X_{i,m}[k] X^*_{i,m}[k] \tag{6}$$

where m_b represents the hybrid start band of parameter band b (subsubband sample index) and k is the time slot of the windowed segment. The second parameter is the interchannel correlation (ICC),

$$\mathrm{ICC}_b = \mathrm{Re}\left\{ \frac{\sum_k \sum_{m=m_b}^{m_{b+1}-1} X_{1,m}[k] X^*_{2,m}[k]}{\sigma_{X_1,b}\sigma_{X_2,b}} \right\}. \tag{7}$$

The mono down mix X_s comprises a linear combination of the two input signals. The associated down-mix weights for each input channel are determined based on the following decomposition of the two input signals:

$$X_{1,m}[k] = c_{1,b} X_{\mathrm{S},m}[k] + X_{\mathrm{D},m}[k] \tag{8}$$

$$X_{2,m}[k] = c_{2,b} X_{\mathrm{S},m}[k] - X_{\mathrm{D},m}[k]. \tag{9}$$

Hence the two input signals are described by a common component $X_{\mathrm{S},m}$, which may have a different contribution to $X_{1,m}$ and $X_{2,m}$ (represented by the coefficients $c_{i,b}$), and an out-of-phase component X_D, which is, except for the sign, identical in both channels. Furthermore, energy preservation is imposed by demanding the signal X_S to have an energy that is equal to the sum of the energies of both input signals. The signal X_S, the desired mono down-mix signal, is given by

$$X_{\mathrm{S},m}[k] = \frac{X_{1,m}[k] + X_{2,m}[k]}{c_{1,b} + c_{2,b}}. \tag{10}$$

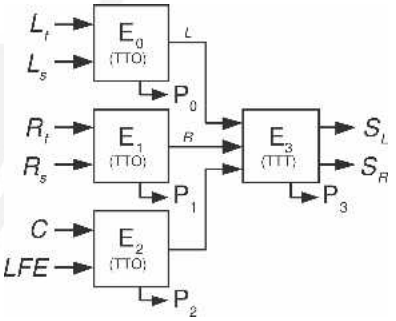

Fig. 7. Preferred tree configuration for stereo down mix.

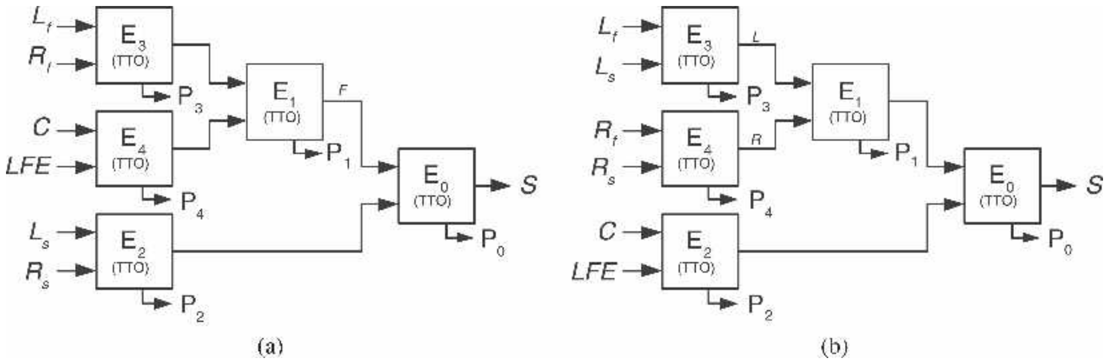

Fig. 6. Tree configurations for mono down mix.

The energy preservation constraint results in

$$(c_{1,b}+c_{2,b})^2 = \frac{\sigma^2_{X_1,b}+\sigma^2_{X_2,b}+2\sigma_{X_1,b}\sigma_{X_2,b}\mathrm{ICC}_b}{\sigma^2_{X_1,b}+\sigma^2_{X_2,b}}. \tag{11}$$

The signal $X_{\mathrm{D},m}$ is the residual signal. This signal is either discarded at the encoder side (in the case of a fully parametric description of the input signals, where synthetic residual signals are used at the decoder side) or can be transmitted to enable full waveform reconstruction at the decoder side. A hybrid approach is also facilitated: a specified low-frequency part of the residual signals can be selected for transmission, while for the remaining signal bandwidth the residual signals are substituted by synthetic signals at the decoder. This option makes the system very flexible in terms of quality and bit-rate tradeoffs.

3.4.3 TTT Encoding Block Using Prediction Mode

The TTT encoding block has three inputs (X_L, X_R, X_C), two down-mix outputs ($X_{\mathrm{S_L}}$, $X_{\mathrm{S_R}}$), and an auxiliary signal ($X_{\mathrm{S_C}}$). The two outputs and the auxiliary signal form a linear combination of the input signals according to

$$\begin{bmatrix} X_{\mathrm{S_L},m}[k] \\ X_{\mathrm{S_R},m}[k] \\ X_{\mathrm{S_C},m}[k] \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & -1 \end{bmatrix} \begin{bmatrix} X_{\mathrm{L},m}[k] \\ X_{\mathrm{R},m}[k] \\ X_{\mathrm{C},m}[k]\,\frac{1}{2}\sqrt{2} \end{bmatrix}. \tag{12}$$

The center signal X_C is attenuated by 3 dB to ensure preservation of the center-channel power in the down mix. The auxiliary output signal $X_{\mathrm{S_C}}$, which has orthogonal down-mix weights, would in principle allow full reconstruction of the three input signals by applying the inverse of the down-mix matrix as up-mix matrix. However, this third signal $X_{\mathrm{S_C}}$ is discarded at the encoder side and replaced by two prediction coefficients that enable an estimation $\hat{X}_{\mathrm{S_C}}$ from the two down-mix channels $X_{\mathrm{S_L}}$, $X_{\mathrm{S_R}}$,

$$\hat{X}_{\mathrm{S_C},m}[k] = \gamma_{1,b}X_{\mathrm{S_L},m}[k] + \gamma_{2,b}X_{\mathrm{S_R},m}[k] \tag{13}$$

with $\gamma_{1,b}$, $\gamma_{2,b}$ two channel prediction coefficients (CPCs) for each parameter band b. The prediction error X_D,

$$X_{\mathrm{D},m}[k] = X_{\mathrm{S_C},m}[k] - \hat{X}_{\mathrm{S_C},m}[k] \tag{14}$$

may be either transmitted or discarded, depending on the desired quality or bit-rate tradeoff. If the residual signal X_D is discarded, the corresponding energy loss is described by an ICC parameter,

$$\mathrm{ICC}_b^2 = 1 - \frac{\sigma^2_{X_\mathrm{D},b}}{\sigma^2_{X_\mathrm{L},b}+\sigma^2_{X_\mathrm{R},b}+\frac{1}{2}\sigma^2_{X_\mathrm{C},b}}. \tag{15}$$

3.4.4 TTT Encoding Block Using Energy Mode

The predictive mode for the TTT encoding block requires a reliable estimate of the signal $X_{\mathrm{S_C}}$ at the decoder side. If waveform accuracy cannot be guaranteed (for example, in the high-frequency range of an audio coder employing SBR), a different TTT encoding mode is supplied, which does not rely on specific waveforms but only describes the relative energy distribution of the three input signals using two CLD parameters,

$$\mathrm{CLD}_{1,b} = 10\log_{10}\frac{\sigma^2_{X_\mathrm{L},b}+\sigma^2_{X_\mathrm{R},b}}{\frac{1}{2}\sigma^2_{X_\mathrm{C},b}}, \tag{16}$$

$$\mathrm{CLD}_{2,b} = 10\log_{10}\frac{\sigma^2_{X_\mathrm{L},b}}{\sigma^2_{X_\mathrm{R},b}}. \tag{17}$$

The prediction and energy mode can be used independently in different bands. In that case, parameter bands of a specified (lower) frequency range apply prediction parameters, whereas the remaining (upper) parameter bands apply the energy mode.

3.4.5 MTX Conversion Block

Matrixed surround (MTX) refers to a method to create a pseudo surround experience based on a stereo down mix with specific down-mix properties. In conventional matrixed-surround systems, the down mix ($X_{\mathrm{S_{L,MTX}}}$, $X_{\mathrm{S_{R,MTX}}}$) is created such that signals of the surround channels are down-mixed in antiphase. The down-mix process in matrix form is given by

$$\begin{bmatrix} X_{\mathrm{S_{L,MTX}},m}[k] \\ X_{\mathrm{S_{R,MTX}},m}[k] \end{bmatrix} = \begin{bmatrix} 1 & 0 & \frac{1}{2}\sqrt{2} & j\sqrt{\frac{2}{3}} & j\sqrt{\frac{1}{3}} \\ 0 & 1 & \frac{1}{2}\sqrt{2} & -j\sqrt{\frac{1}{3}} & -j\sqrt{\frac{2}{3}} \end{bmatrix} \times \begin{bmatrix} X_{\mathrm{L_f},m}[k] \\ X_{\mathrm{R_f},m}[k] \\ X_{\mathrm{C},m}[k] \\ X_{\mathrm{L_s},m}[k] \\ X_{\mathrm{R_s},m}[k] \end{bmatrix}. \tag{18}$$

The antiphase relationship of the surround channels in the down mix enables a matrixed-surround decoder to control its front or surround panning. The drawback of this static down-mix matrix is that it is impossible to retrieve the original input channels, nor is it possible to reconstruct a conventional stereo down mix from the matrixed-surround-compatible down mix. In MPEG Surround, however, a matrixed-surround mode is supplied for compatibility with legacy matrixed-surround devices, and hence this option must not have any negative impact on any MPEG Surround operation. Therefore the approach of MPEG Surround to create a matrixed-surround-compatible down mix is different from the static down-mix approach as given by Eq. (18). A conversion from a conventional down mix to a matrixed-surround-compatible down mix is facilitated by an MTX conversion block applied as a postprocessing stage of the encoding tree.

The MTX conversion block has two inputs and two outputs. The two output signals are linear combinations of the two input signals. The resulting 2 × 2 processing matrix is dynamically varying and depends on the spatial parameters resulting from the spatial encoding process. If the surround channels contain relatively little energy, the two output signals of the MTX processing stage are (almost) identical to the two input signals. If, on the other hand, there is a significant surround activity, the 2 × 2 matrix creates negative crosstalk to signal surround activity to a matrixed-surround decoder. The advantage of employing this process on a stereo down mix rather than on the multichannel input, is that the 2 × 2 processing matrix is invertible. In other words, the MPEG Surround decoder can "undo" the processing by employing the inverse of the encoder matrix. As a result, the matrixed-surround compatibility has no negative effect on the 5.1-channel reconstruction of an MPEG Surround decoder.

The matrixed-surround conversion block is outlined in Fig. 8. Both down-mix signals, X_{S_L} and X_{S_R}, are split into two parts using parameters q_L and q_R. These parameters represent the relative amount of surround energy in each parameter band of X_{S_L} and X_{S_R}, respectively, and are derived from the encoded spatial parameters. For nonzero q part of the input signal is processed by a 90-degree phase shifter (indicated by the j block). The phase-shifted signal is subsequently mixed out of phase to both output channels $X_{S_{L,MTX}}$, $X_{S_{L,MTX}}$, including a (fixed) weight $G = 1/\sqrt{3}$ for the cross term.

The scheme depicted in Fig. 8 can be described in matrix notation using a conversion matrix $\boldsymbol{V}_b$,

$$\begin{bmatrix} X_{S_{L,MTX},m}[k] \\ X_{S_{R,MTX},m}[k] \end{bmatrix} = \boldsymbol{V}_b \begin{bmatrix} X_{S_L,m}[k] \\ X_{S_R,m}[k] \end{bmatrix} = \begin{bmatrix} v_{11,b} & v_{12,b} \\ v_{21,b} & v_{22,b} \end{bmatrix} \begin{bmatrix} X_{S_L,m}[k] \\ X_{S_R,m}[k] \end{bmatrix} \tag{19}$$

with

$$v_{11,b} = \frac{1 - q_{L,b} + \mathrm{j}q_{L,b}}{\sqrt{1 - 2q_{L,b} + 2q_{L,b}^2}} \tag{20}$$

$$v_{12,b} = \frac{\mathrm{j}q_{R,b}}{\sqrt{3(1 - 2q_{R,b} + 2q_{R,b}^2)}} \tag{21}$$

$$v_{21,b} = \frac{-\mathrm{j}q_{L,b}}{\sqrt{3(1 - 2q_{L,b} + 2q_{L,b}^2)}} \tag{22}$$

$$v_{22,b} = \frac{1 - q_{R,b} - \mathrm{j}q_{R,b}}{\sqrt{1 - 2q_{R,b} + 2q_{R,b}^2}}. \tag{23}$$

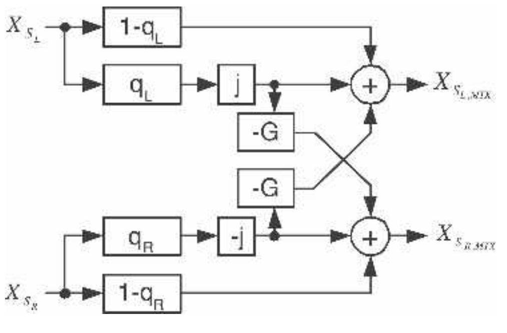

Fig. 8. Matrixed-surround conversion block.

3.4.6 External Down-Mix Analysis Block

In some cases the use of an externally provided down mix may be preferred over an automated down mix. For example, a studio engineer might produce separate stereo and multichannel mixes from the same (multitrack) recording. MPEG Surround provides the possibility to transmit such an externally provided down mix instead of the automated down mix. In order to minimize potential differences in the resulting multichannel reconstruction, the external down-mix analysis block parameterizes the differences between automated and externally provided down mixes. The external down-mix analysis block is used as a postprocessor of the full spatial encoder tree. For each internal, automated down-mix channel X_{S_i} and the corresponding externally provided down-mix channel X_{E_i}, the energy ratio within each parameter band is extracted according to

$$\mathrm{CLD}_{i,b} = 10 \log_{10}\left(\frac{\sigma^2_{X_{S_i}}}{\sigma^2_{X_{E_i}}}\right). \tag{24}$$

This down-mix gain parameter describes the level adjustment in each parameter band that should be applied to the externally provided down mix to result in a down mix that is equal to the automated down mix from a (statistical) energy point of view. On top of this CLD parameter, residual signals X_{R_i} can be transmitted for a user-selectable bandwidth to obtain waveform reconstruction of the automated down mix from the (transmitted) external down mix,

$$X_{R_i,m}[k] = X_{S_i,m}[k] - \eta \frac{\sigma_{X_{S_i}}}{\sigma_{X_{E_i}}} X_{E_i,m}[k]. \tag{25}$$

The parameter η controls the method of coding of the residual signal; $\eta = 0$ results in absolute coding of the automated down mix X_{S_i}, whereas for $\eta = 1$ the difference between the automated down-mix X_{S_i} and the gain-adjusted externally provided down mix X_{E_i} is used as residual signal. The latter method is especially beneficial if there exists a high correlation between the externally provided down mix and the automated down mix.

3.5 Parameter Quantization and Coding

3.5.1 Parameter Quantization

The CLD, ICC, and CPC parameters are quantized according to perceptual criteria. The quantization process aims at introducing quantization errors that are practically inaudible. For the CLD, this constraint requires a nonlinear quantizer given the fact that the sensitivity to changes in CLD depends on the reference CLD. For CLD and ICC parameters the same quantizer is used as applied in parametric stereo coders [5]. The CPC coefficients are quantized linearly with a step size of 0.1 and a range of −2.0 and +3.0, as they do not have a clear perceptual meaning.

3.5.2 Further Bit-Rate Reduction Techniques

The quantizer described in Section 5.1 aims at practically inaudible differences in spatial properties. An additional quantization strategy is also supplied based on a reduced number of quantizer steps to reduce the entropy per transmitted spatial parameter. This “coarse” quantization comprises only every even quantizer index of the quantizer described in Section 3.5.1

If such coarse quantization steps are applied, there is a risk that the relatively large discrete steps in changes in spatial properties give rise to audible artifacts. For example, if a certain sound object in the multichannel content is slowly moving from one loud speaker location to another, the smooth movement in the original content may be reproduced at the decoder side as a sequence of discrete positions, each perceived position corresponding to a quantizer value. To resolve such artifacts, the encoder may signal a “smoothing flag” in the bit stream, which signals the decoder to apply a low-pass filter on the discrete parameter values to result in a smooth transition between different quantizer values.

A related technique for further bit-rate reduction is referred to as energy-dependent quantization. This method allows for combinations of fine and coarse parameter quantization, depending on the amount of signal energy within the tree structure. If the amount of signal energy in a certain part of the parameter tree is significantly lower than the overall signal energy, large quantization errors in that specific part are in most cases inaudible, since they will be masked by signal components from other channels. In such cases a very coarse parameter quantization can be applied for relatively weak channel pairs, whereas a fine quantization may be applied for strong (loud) channel pairs.

Besides changes in quantizer granularity, MPEG Surround also features the possibility to transmit only a selected number of parameters. More specifically, only a single ICC parameter may be transmitted instead of a separate ICC value for each TTO block. If this single ICC mode is enabled, the same transmitted ICC value is used in each OTT decoding block.

Finally the resulting quantizer indexes are encoded differentially over time and frequency. Entropy coding is employed on the differential quantizer indexes to exploit further redundancies.

3.6 Coding of Residual Signals

As described in Sections 3.4.2 and 3.4.3, TTO and TTT encoding blocks can generate residual signals. These residual signals can be encoded in a bit-efficient manner and transmitted along with the corresponding down mix and spatial parameters.

Residual data do not necessarily need to be transmitted since MPEG Surround decoders are capable of reconstructing decorrelated signals with similar properties as the residual signals without requiring any additional information (see Section 4.2.1). However, if full waveform reconstruction at the decoder side is desired, residual signals can be transmitted. The bandwidth of the residual signals can be set at the encoder side, so that a tradeoff can be made between bit-rate consumption and reconstruction quality. The residual signals are transformed to an MDCT representation and subsequently encoded into an AAC bitstream element.

4 MPEG SURROUND DECODER

4.1 Structure

The MPEG Surround decoder structure is outlined in Fig. 9. The down mix is first processed by a pregain, which is the inverse of the postgain of the MPEG Surround encoder. Subsequently the input signals are processed by an analysis filter bank that is identical to the filter bank described in Section 3.3. A spatial decoder regenerates multichannel audio by reinstating the spatial properties described by the decoded parameters. Finally, applying a set of synthesis filter banks and postgains (the inverse of the encoder pregains) results in the time-domain multichannel output signals.

4.2 Spatial Decoder

4.2.1 Operation Principle

The spatial decoder generates multichannel output signals from the downmixed input signal by reinstating the spatial cues captured by the spatial parameters. The spatial synthesis of OTT decoding blocks employs so-called decorrelators and matrix operations in a similar fashion as parametric stereo decoders [5]. In an OTT decoding block two output signals with the correct spatial cues are generated by mixing a mono input signal with the output of a decorrelator that is fed with that mono input signal.

Given the tree structures that were explained in Section 3.4.1, a first attempt at building a multichannel decoder could be to simply concatenate OTT decoding blocks according to the tree structure at hand. An example of such a concatenation of OTT decoding blocks for three-channel output is shown in Fig. 10. A mono input signal X_{S} is processed by a first decorrelator D_1 and an up-mix matrix $\boldsymbol{W}(P_1)$ to obtain two output signals $\hat{X}_{11}$, $\hat{X}_{12}$ with spatial parameters P_1,

$$\begin{bmatrix} \hat{X}_{11} \\ \hat{X}_{12} \end{bmatrix} = \boldsymbol{W}(P_1) \begin{bmatrix} X_{\mathrm{S}} \\ D_1(X_{\mathrm{S}}) \end{bmatrix} \tag{26}$$

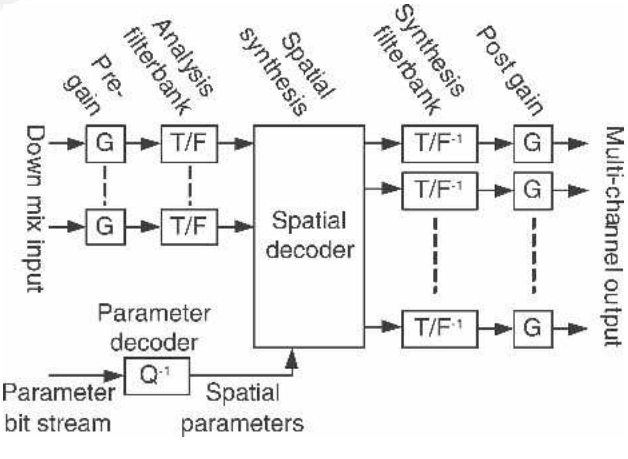

Fig. 9. Structure of MPEG Surround decoder.

with

$$\boldsymbol{W}(P_i) = \begin{bmatrix} w_{11}(P_i) & w_{12}(P_i) \\ w_{21}(P_i) & w_{22}(P_i) \end{bmatrix}. \tag{27}$$

Signal $\hat{X}_{12}$ is subsequently processed by a second decorrelator D_2 and mixed with $\hat{X}_{12}$ itself based on a second spatial parameter set P_2 to generate two output signals $\hat{X}_{21}$, $\hat{X}_{22}$,

$$\begin{bmatrix} \hat{X}_{21} \\ \hat{X}_{22} \end{bmatrix} = \boldsymbol{W}(P_2) \begin{bmatrix} \hat{X}_{12} \\ D_2(\hat{X}_{12}) \end{bmatrix}. \tag{28}$$

The up-mix matrices $\boldsymbol{W}$ ensure that their output pairs have the correct level difference as well as the correct correlation.

This scheme has the important drawback of decorrelators connected in series: the output of decorrelator D_1 is (partly) fed into decorrelator D_2. Given the most important requirement of decorrelators to generate output that is statistically independent from its output, its processing will result in a delay and temporal or spectral smearing of the input signals. In other words, the spectral and temporal envelopes of an input signal may be altered considerably, especially if the decorrelator contains reverberation-like structures. If two decorrelators are connected in series, the degradation of signal envelopes will be substantial. Moreover, since spatial parameters are temporally varying, temporal smearing and delays will cause an asynchrony between the signals and their parameters. This asynchrony will become larger if decorrelators are connected in series. Thus concatenation of decorrelators should preferably be avoided.

Fortunately the problem of concatenated decorrelators can be solved without consequences for spatial synthesis. Decorrelator D_2 should generate a signal that is statistically independent from $\hat{X}_{12}$, which is a combination of X_S and the output of decorrelator D_1. In other words, the output of D_2 should be independent of both X_S and the output of decorrelator D_1. This can be achieved by feeding decorrelator D_2 with mono input signal X_S instead of $\hat{X}_{12}$, if the decorrelators D_1 and D_2 are mutually independent. This enhancement is outlined in Fig. 11.

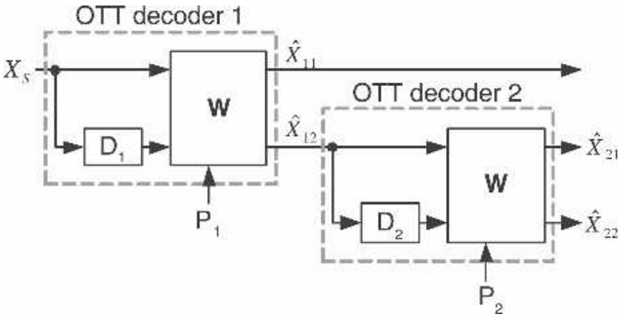

Fig. 10. Concatenation of two OTT decoding blocks to achieve three-channel output.

The input of decorrelator D_2 is now obtained directly from X_S with a gain $\lambda_2(P_1)$, which compensates for the change in energy that would otherwise be caused by matrix $\boldsymbol{W}(P_1)$,

$$\lambda_i^2(P_1) = w_{i1}^2(P_1) + w_{i2}^2(P_1). \tag{29}$$

Furthermore it can be observed that signal $\hat{X}_{12}$, which is a linear combination of X_S and the output of decorrelator D_1, is processed by matrix $\boldsymbol{W}(P_2)$ without any intermediate decorrelation process. Given the linear properties of the two matrix operations, the contribution of $\hat{X}_S$ within $\hat{X}_{21}$ and $\hat{X}_{22}$ can be obtained by a single (combined) matrix operation by multiplication of the respective elements from $\boldsymbol{W}(P_1)$ and $\boldsymbol{W}(P_2)$. The statistical equivalence of both schemes can be shown by computing the covariance matrices of the output signals in both cases, which are identical. In summary, cascaded decorrelators can be shifted through preceding OTT decoding blocks without changing statistical properties such as signal levels and mutual correlations, under the assumption that the different decorrelators are mutually independent.

The process of transforming spatial parameterization trees from cascaded decorrelator structures to decorrelators in parallel, extended with combined matrix multiplications, leads to the generalized spatial decoder structure shown in Fig. 12. Any encoder tree configuration can be mapped to this generalized decoder structure. The input signals are first processed by a preprocess matrix $\boldsymbol{M}_{\text{pre}}$, which applies decorrelator input gains as outlined in Fig. 11, TTT-type decoding (in case of a stereo down mix), as well as any decoder-side inversion processes that should be applied on the down mix (see Section 2.2). The outputs

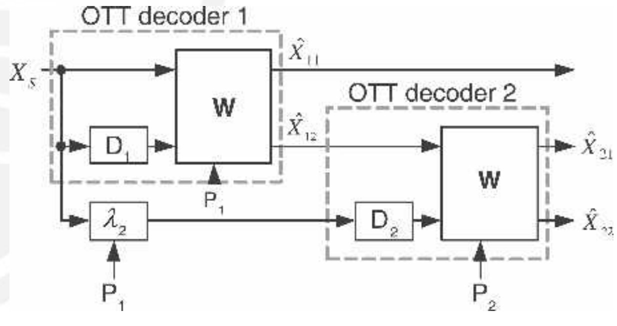

Fig. 11. Enhanced concatenation of two OTT decoding blocks to achieve three-channel output with decorrelators in parallel.

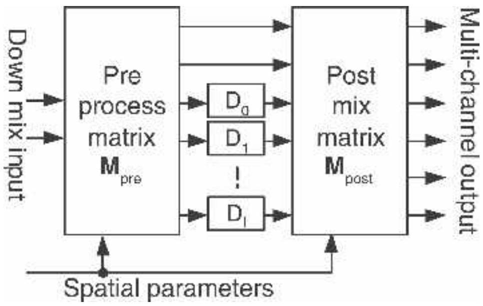
Down mix input
Pre process matrix $\mathbf{M}_{\text{pre}}$
D_0
D_1
D_l
Post mix matrix $\mathbf{M}_{\text{post}}$
Multi-channel output
Spatial parameters

Fig. 12. Generic spatial decoder.

of the prematrix are fed to a decorrelation stage with one or more mutually independent decorrelators. Finally a postmix matrix $\boldsymbol{M}_{\text{post}}$ generates the multichannel output signals. In this scheme both the preprocess matrix as well as the postmix matrix are dependent on the transmitted spatial parameters.

4.2.2 Decorrelators

In all tree configurations some outputs of the mix matrix $\boldsymbol{M}_{\text{pre}}$ are fed into decorrelators. These decorrelators create an output that is uncorrelated with their input. Moreover, in the case multiple decorrelators are used, they are conditioned such that their outputs will also be mutually uncorrelated (see Section 4.2.1). Fig. 13 shows a diagram of the decorrelator processing that is performed on the hybrid domain signals.

The decorrelators comprise a delay (which varies in different frequency bands), a lattice all-pass filter, and an energy adjustment stage. The configuration for the delay and all-pass filter are controlled by the encoder using decorrelator configuration data. The all-pass coefficients of the different decorrelators were selected such that their outputs are mutually independent (even if the same signal is used as input).

In order to avoid audible reverberation in the case of transients, an energy adjustment stage scales the output of the decorrelator to match the energy level of the input signal in all frequency (processing) bands.

If residual signals are transmitted for certain OTT or TTT decoding blocks, the outputs of the corresponding decorrelators are replaced by the decoded residual signals. This replacement is only applied for the frequency range of the transmitted residual signal. For the remaining bandwidth, the decorrelator output is maintained.

4.2.3 OTT Decoding Block

The up-mix matrix $\boldsymbol{W}$ for an OTT decoding block is determined by the following constraints:

1) The correlation of the two output signals must obey the transmitted ICC parameter.

2) The power ratio of the two output signals must obey the transmitted CLD parameter.

3) The sum of the energies of the output signals must be equal to the energy of the input signal.

Given these three constraints, the 2 × 2 matrix $\boldsymbol{W}$ has one degree of freedom. One interpretation of this degree of freedom is a common rotation angle of the two output signals in a two-dimensional space spanned by the two input signals. The mix matrix $\boldsymbol{W}$ can be expressed using a common rotation angle β, a differential rotation angle α, and two vector lengths λ_1 and λ_2,

$$\begin{bmatrix} \hat{X}_{1,m}[k] \\ \hat{X}_{2,m}[k] \end{bmatrix} = \begin{bmatrix} \lambda_1 \cos(\alpha+\beta) & \lambda_1 \sin(\alpha+\beta) \\ \lambda_2 \cos(-\alpha+\beta) & \lambda_2 \sin(-\alpha+\beta) \end{bmatrix} \times \begin{bmatrix} X_{\text{S},m}[k] \\ D(X_{\text{S},m}[k]) \end{bmatrix}. \quad (30)$$

A unique relation exists between the ICC parameter and the differential rotation angle α, which is given by

$$\alpha = \frac{1}{2} \arccos(\text{ICC}). \quad (31)$$

Thus the ICC value is independent of the overall rotation angle β. In other words, there exist an infinite number of solutions to linearly combine two independent signals to create two output signals with a specified ICC and CLD value and the additional constraint on the summed energy of the output signals. This degree of freedom is represented by the angle β. The angle β is chosen to minimize the total amount of decorrelation signal in the (summed) output signals (that is, minimize $w_{12} + w_{22}$). This leads to the following solution for β:

$$\beta = \tan\left[\frac{\lambda_2 - \lambda_1}{\lambda_2 + \lambda_1} \arctan(\alpha)\right]. \quad (32)$$

The variables λ_1 and λ_2, representing the relative amplitudes of the two output signals with respect to the input, are given by

$$\lambda_1 = \sqrt{\frac{10^{\text{CLD}/10}}{1 + 10^{\text{CLD}/10}}} \quad (33)$$

$$\lambda_2 = \sqrt{\frac{1}{1 + 10^{\text{CLD}/10}}}. \quad (34)$$

The solution for β implies that $w_{12,i} = -w_{22,i}$. In other words, the synthesis matrix can also be written for each parameter band b as

$$\boldsymbol{W}_b = \begin{bmatrix} \lambda_{1,b} \cos(\alpha_b + \beta_b) & +1 \\ \lambda_{2,b} \cos(-\alpha_b + \beta_b) & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \lambda_{1,b} \sin(\alpha_b + \beta_b) \end{bmatrix}. \quad (35)$$

Stated differently, the decorrelation signal level is identical in both output signals but the contribution to both output channels is in antiphase. Hence this decoder synthesis matrix employs the same decomposition that was used at the encoder side (see Section 3.4.2), with the exception that the common out-of-phase component is now synthetically generated by decorrelation and scaling [with $\lambda_1 \sin(\alpha + \beta)$].

4.2.4 OTT Decoding Block Using Residual Coding

If for a certain parameter band residual signals X_{S_D} are transmitted, the decorrelator output is replaced by the

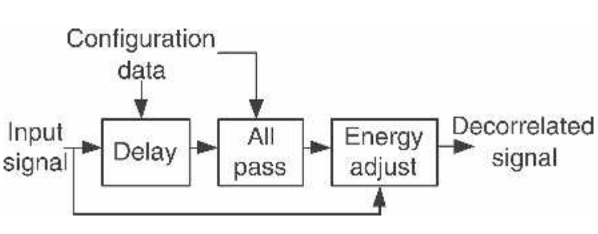

Fig. 13. Diagram of decorrelator processing on hybrid QMF domain signals.

transmitted residual signal and the corresponding matrix elements are set to +1 and −1, respectively, according to the corresponding signal decomposition at the encoder (see Section 3.4.2),

$$\boldsymbol{W}_b = \begin{bmatrix} \lambda_{1,b}\cos(\alpha_b + \beta_b) & +1 \\ \lambda_{2,b}\cos(-\alpha_b + \beta_b) & -1 \end{bmatrix}. \tag{36}$$

4.2.5 TTT Decoding Block Using Prediction Mode

Three output signals $\hat{X}_\mathrm{L}$, $\hat{X}_\mathrm{R}$, $\hat{X}_\mathrm{C}$ are synthesized according to the inverse encoder-side down-mix matrix using an estimated signal $\hat{S}_\mathrm{C}$,

$$\begin{bmatrix} \hat{X}_{\mathrm{L},m}[k] \\ \hat{X}_{\mathrm{R},m}[k] \\ \hat{X}_{\mathrm{C},m}[k] \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 2 & -1 & 1 \\ -1 & 2 & 1 \\ \sqrt{2} & \sqrt{2} & -\sqrt{2} \end{bmatrix}\begin{bmatrix} X_{\mathrm{S_L},m}[k] \\ X_{\mathrm{S_R},m}[k] \\ \hat{X}_{\mathrm{S_C},m}[k] \end{bmatrix} \tag{37}$$

with

$$\hat{X}_{\mathrm{S_c},m}[k] = \gamma_{1,b}X_{\mathrm{S_L},m}[k] + \gamma_{2,b}X_{\mathrm{S_R},m}[k] + X_{\mathrm{D},m}[k] \tag{38}$$

m being the filter band index, b the processing band index, and X_D the residual signal. The resulting up-mix matrix $\boldsymbol{W}$ is then given by

$$\begin{bmatrix} \hat{X}_{\mathrm{L},m}[k] \\ \hat{X}_{\mathrm{R},m}[k] \\ \hat{X}_{\mathrm{C},m}[k] \end{bmatrix} = \frac{1}{3}\begin{bmatrix} \gamma_{1,b}+2 & \gamma_{2,b}-1 & 1 \\ \gamma_{1,b}-1 & \gamma_{2,b}+2 & 1 \\ \sqrt{2}(1-\gamma_{1,b}) & \sqrt{2}(1-\gamma_{2,b}) & -\sqrt{2} \end{bmatrix} \times \begin{bmatrix} X_{\mathrm{S_L},m}[k] \\ X_{\mathrm{S_R},m}[k] \\ X_{\mathrm{D},m}[k] \end{bmatrix}. \tag{39}$$

If no residual signal was transmitted, the resulting energy loss can be compensated for in two ways, depending on the complexity of the decoder. The first, low-complexity solution is to apply a gain to the three output signals according to the prediction loss. In that case the up-mix matrix is given by

$$\boldsymbol{W}_b = \frac{1}{3\mathrm{ICC}_b}\begin{bmatrix} \gamma_{1,b}+2 & \gamma_{2,b}-1 & 0 \\ \gamma_{1,b}-1 & \gamma_{2,b}+2 & 0 \\ \sqrt{2}(1-\gamma_{1,b}) & \sqrt{2}(1-\gamma_{2,b}) & 0 \end{bmatrix}. \tag{40}$$

This method does ensure correct overall power, but the relative powers of the three output signals, as well as their mutual correlations, may be different from those of the original input signals.

Alternatively, the prediction loss can be compensated for by means of a decorrelator signal. In that case the (synthetic) residual signal X_D of Eq. (39) is generated by decorrelators fed by the two down-mix signals (only for those frequency bands for which no transmitted residual signal is available). This more complex method reconstructs the full covariance structure of the three output signals.

4.2.6 TTT Decoding Block Based on Energy Reconstruction

TTT decoding based on energy reconstruction (henceforth called energy mode) supports two methods. These methods are characterized by the way the up-mix matrix is derived, using the same (transmitted) parameters. The bitstream header signals which method should be used.

In the energy mode without center subtraction, the left and right output signals are calculated from the left and right down-mix signals, respectively. In other words, the left output signal is generated independently from the right input channel and vice versa. The center signal is a linear combination of both down-mix signals. This method should be used if at least in a certain frequency range the legacy stereo coder does not have waveform-preserving properties (for example, when using SBR). The up-mix process is given by

$$\begin{bmatrix} \hat{X}_{\mathrm{L},m}[k] \\ \hat{X}_{\mathrm{R},m}[k] \\ \hat{X}_{\mathrm{C},m}[k] \end{bmatrix} = \begin{bmatrix} w_{11,b} & 0 \\ 0 & w_{22,b} \\ w_{31,b} & w_{32,b} \end{bmatrix} \cdot \begin{bmatrix} X_{\mathrm{S_L},m}[k] \\ X_{\mathrm{S_R},m}[k] \end{bmatrix}. \tag{41}$$

The derivation of the solution for the matrix elements is provided in [44]. The solution is given by

$$w_{11,b} = \sqrt{\frac{\kappa_{1,b}\cdot\kappa_{2,b}}{\kappa_{1,b}\cdot\kappa_{2,b} + \kappa_{2,b} + 1}} \tag{42}$$

$$w_{22,b} = \sqrt{\frac{\kappa_{1,b}}{\kappa_{1,b} + \kappa_{2,b} + 1}} \tag{43}$$

$$w_{31,b} = \frac{1}{2}\sqrt{2\cdot\frac{\kappa_{2,b}+1}{\kappa_{1,b}\cdot\kappa_{2,b} + \kappa_{2,b} + 1}} \tag{44}$$

$$w_{32,b} = \frac{1}{2}\sqrt{2\cdot\frac{\kappa_{2,b}+1}{\kappa_{1,b} + \kappa_{2,b} + 1}} \tag{45}$$

with

$$\kappa_{i,b} = 10^{\mathrm{CLD}_{i,b}/10} \tag{46}$$

The energy mode with center subtraction, on the other hand, tries to improve the reconstruction of the left and right signals by utilizing cross terms. This method is especially beneficial if the core coder is at least partly preserving the waveforms of its input. The up-mix matrix is given by

$$\boldsymbol{W}_b = \begin{bmatrix} w_{11,b} & w_{12,b} \\ w_{21,b} & w_{22,b} \\ w_{31,b} & w_{32,b} \end{bmatrix}. \tag{47}$$

The elements of the up-mix matrix $\boldsymbol{W}$ are calculated using the linear least-squares optimization technique. This technique tries to minimize the squared Euclidian norm of the difference between the original signals X_L, X_R, and X_C, and their decoder-side reconstructions $\hat{X}_L$, $\hat{X}_R$, and $\hat{X}_C$, under the constraint of correct energy ratios. The solution for the up-mix matrix is also derived in [44].

4.2.7 MTX Inversion Block

If the transmitted downmix is encoded using a matrixed-surround conversion block (see Section 3.4.5), the stereo input signal is processed by a matrixed-surround inversion matrix $\boldsymbol{W}$**,** which is the inverse of the encoder-side conversion matrix $\boldsymbol{V}$**,**

$$\boldsymbol{W}_b = \boldsymbol{V}_b^{-1}. \tag{48}$$

4.2.8 External Down-Mix Inversion Block

If an external down mix was provided, the external down-mix inverter aims at reconstructing the (discarded) automated down mix from the transmitted external down mix. The reconstructed down-mix signal $\hat{X}_{S_i}$ for channel i is given by

$$\hat{X}_{S_i,m}[k] = \left[\eta\sqrt{\kappa_{i,b}} \quad 1\right]\begin{bmatrix} X_{E_i,m}[k] \\ X_{R_i,m}[k] \end{bmatrix} \tag{49}$$

with $\kappa_{i,b}$ dependent on the external down-mix gain parameter $\mathrm{CLD}_{i,b}$ according to Eq. (46) for parameter band b and down-mix channel i, X_{Ei} the transmitted external down mix, X_{Ri} the external down-mix residual for channel i (if available), and η is computed using the decision regarding absolute or relative coding of the residual signals (if available).

4.2.9 Matrix Elements for a Mono Down Mix

The construction of pre- and postmix matrices for the mono-based tree as outlined in Fig. 6(a) is shown in Fig. 14. The gain compensation factors for decorrelator inputs resulting from cascaded OTT blocks are applied in the premix matrix $\boldsymbol{M}_{\mathrm{pre}}$. The LFE signal is not subject to decorrelation, and hence its output signal is solely constructed using gain factors resulting from all respective OTT blocks. If an external down mix was provided, the external down-mix inversion block is combined with $\boldsymbol{M}_{\mathrm{pre}}$ as well (not shown in Fig. 14).

The mixing matrices $\boldsymbol{W}$ for each OTT decoding block are combined in a single postmix matrix $\boldsymbol{M}_{\mathrm{post}}$. This process can be performed for any OTT tree structure, including trees with more than six input or output channels.

4.2.10 Matrix Elements for a Stereo Down Mix

The construction of the pre- and postmix matrices for a stereo-based tree is shown in Fig. 15. The premix matrix comprises the combined effect of matrixed-surround inversion (MTX) or external-down mix inversion (EXT) and the TTT decoding process. The left and right outputs of the TTT output signals are subsequently fed to parallel decorrelators. The postmix matrix is then composed of three parallel OTT blocks. The OTT decoding block for the center and the LFE channel does not have a decorrelator input since no correlation synthesis between center and LFE is applied (that is, the respective ICC values are set to +1).

4.2.11 Parameter Positions and Interpolation

For each transmitted parameter set the mixing matrices are determined as described previously. However, these matrices correspond in most cases to a single time instance, which depends on the segmentation and windowing procedure of the encoder. The sample index k at which a parameter set is valid is denoted by k_p, which is referred to as the parameter position. The parameter positions are

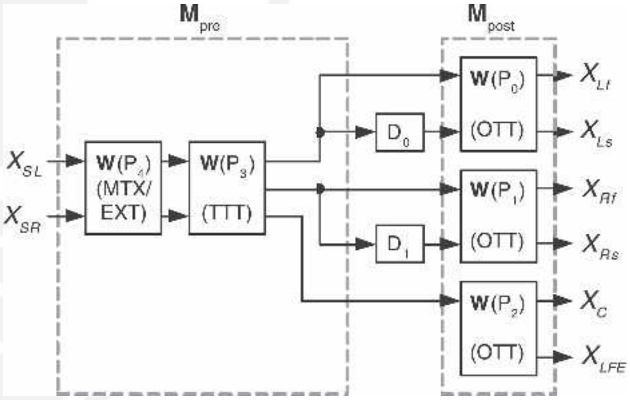

Fig. 15. Pre- and postmatrix construction for stereo-based tree configuration.

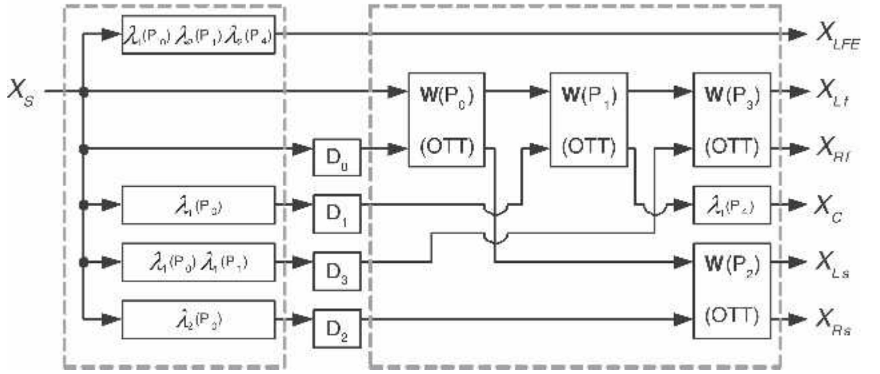

Fig. 14. Pre- and postmatrix construction for mono-based tree configuration.

transmitted along with the corresponding parameters themselves. For that particular QMF sample index ($k = k_p$), the mixing matrices are determined. For QMF sample indices k in between parameter positions, the mixing matrices are interpolated linearly (that is, its real and imaginary parts are interpolated individually). The interpolation of mixing matrices has the advantage that the decoder can process each "slot" of hybrid QMF samples (that is, one sample from each subband) one by one, without the need of storing a whole frame of subband samples in memory. This results in a significant memory reduction compared to frame-based synthesis methods.

4.3 Enhanced Matrix Mode

MPEG Surround features an analysis element that is capable of estimating spatial parameters based on a conventional or matrixed-surround compatible down mix. This element enables MPEG Surround to work in a mode that is similar to matrixed-surround systems, that is, by means of a matrixed-surround compatible down mix without transmission of additional parameters, or alternatively, to generate multichannel representations from legacy stereo material. For such a mode, the MPEG Surround decoder analyzes the transmitted (stereo) down mix and generates spatial parameters that are fed to the spatial decoder to up-mix the stereo input to multichannel output. Alternatively this analysis stage can be employed already in the encoder to enable multichannel audio transmission in MPEG Surround format based on conventional stereo source material.

A spatial decoder using this enhanced matrix mode is shown in Fig.16 The spatial parameters required to compute the matrix elements of the pre- and postmix matrix are generated by an analysis module A. The analysis module measures two parameters of the down mix received for each parameter band. These parameters are the down-mix level difference $CLD_{S,b}$ and the down-mix cross correlation $ICC_{S,b}$. To avoid analysis delays, these parameters are estimated using first-order filtering involving data from the past.

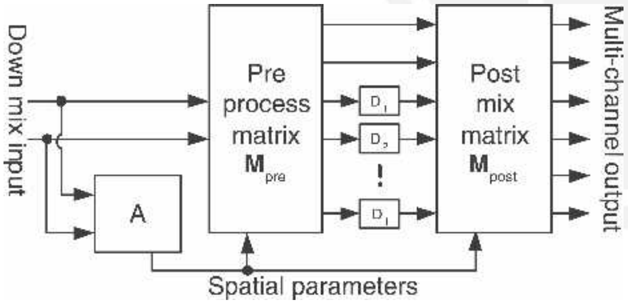
Down mix input
Pre process matrix $\mathbf{M}_{pre}$
D_1
D_2
!
D_l
Post mix matrix $\mathbf{M}_{post}$
A
Multi-channel output
Spatial parameters

Fig. 16. MPEG Surround spatial decoder using enhanced matrix mode.

5 SUBJECTIVE EVALUATION

During the MPEG Surround development the progress and corresponding performance have been documented in detail in several publications [15], [16]–[18] and in an MPEG verification report [45]. The results published in those papers primarily focused on bit-rate scalability, different channel configurations, support for external down mixes, and binaural decoding (not covered in this paper). A comparison between matrixed-surround systems such as Dolby Prologic II and MPEG Surround with and without additional side information was provided in [14].

The purpose of the listening test described in this paper is to demonstrate that existing stereo services can be upgraded to high-quality multichannel audio in a fully backward compatible fashion at transmission bit rates that are currently used for stereo.

5.1 Stimuli and Method

A list of the codecs that were used in the test is given in Table 2. The total employed bit rate (160 kbps) was set to a value that is commonly used for high-quality stereo transmission.

Configuration 1 represents stereo AAC at 128 kbps in combination with 32 kbps of MPEG Surround (MPS) parametric data. Configuration 2 is based on a different core coder (MP3 in combination with MPEG Surround) using a slightly lower parametric bit rate (and consequently a slightly higher bit rate for the core coder; informal listening indicated that this resulted in a higher overall quality). Configuration 3 is termed MP3 Surround [46], which is a proprietary extension to the MPEG-1 layer 3 (MP3) codec. This extension also employs parametric side information to retrieve multichannel audio from a stereo down mix, but is not compatible with MPEG Surround. Configuration 4 employs the Dolby Prologic II matrixed-surround system (DPLII) for encoding and decoding in combination with stereo AAC at a bit rate of 160 kbps. Configuration 5 is AAC in multichannel mode, which represents state-of-the-art discrete channel coding.

For configurations 1, 4, and 5 state-of-the-art AAC encoders were used. For configurations 2 and 3 an encoder and decoder available from www.mp3surround.com have been used (version April 2006). Dolby Prologic II encod-

Table 2. Codecs under test.

Configuration	Codec	Core Bit Rate (kbps)	Spatial Bit Rate (kbps)	Total Bit Rate (kbps)
1	AAC stereo + MPS	128	32	160
2	MP3 + MPS	149	11	160
3	MP3 Surround	144	16	160
4	AAC stereo + DPLII	160	n/a	160
5	AAC multichannel	160	n/a	160

ing and decoding was performed using the Dolby-certified Minnetonka Surcode for Dolby Prologic II package (version 2.0.3) using its default settings.

Eight listeners participated in this experiment. All listeners had significant experience in evaluating audio codecs and were specifically instructed to evaluate the overall quality, consisting of the spatial audio quality as well as any other noticeable artifacts. In a double-blind MUSHRA test [47] the listeners had to rate the perceived quality of several processed excerpts against the original (unprocessed) excerpts on a 100-point scale with five anchors, labeled “bad”, “poor”, “fair”, “good,” and “excellent.” A hidden reference and a low-pass-filtered anchor (cutoff frequency at 3.5 kHz) were also included in the test. The subjects could listen to each excerpt as often as they liked and could switch in real time between all versions of each excerpt. The experiment was controlled from a PC with an RME Digi 96/24 sound card using ADAT digital out. Digital-to-analog conversion was provided by an RME ADI-8 DS eight-channel digital-to-analog converter. Discrete preamplifiers (Array Obsydian A-1) and power amplifiers (Array Quartz M-1) were used to feed a 5.1 loudspeaker setup employing B&W Nautilus 800 speakers in a dedicated listening room according to ITU recommendations [48].

A total of 11 critical excerpts were used, as listed in Table 3. The excerpts are the same as were used in the MPEG call for proposals (CfP) on spatial audio coding [19] and range from pathological signals (designed to be critical for the technology at hand) to movie sound and multichannel productions. All input and output excerpts were sampled at 44.1 kHz.

5.2 Results

The subjective results of each codec and excerpt are shown in Fig. 17. The horizontal axis denotes the excerpt under test, the vertical axis the mean MUSHRA score averaged across listeners, and different symbols indicate different codecs. The error bars denote the 95% confidence intervals of the means.

For all excerpts, the hidden reference (square symbols) has scores virtually equal to 100, with a very small con-

Table 3. Test excerpts.

Excerpt	Name	Category
1	BBC applause	Pathological/ambience
2	ARL applause	Pathological/ambience
3	Chostakovitch	Music
4	Fountain music	Pathological/ambience
5	Glock	Pathological
6	Indie2	Movie sound
7	Jackson1	Music
8	Pops	Music
9	Poulenc	Music
10	Rock concert	Music
11	Stomp	Music (with LFE)

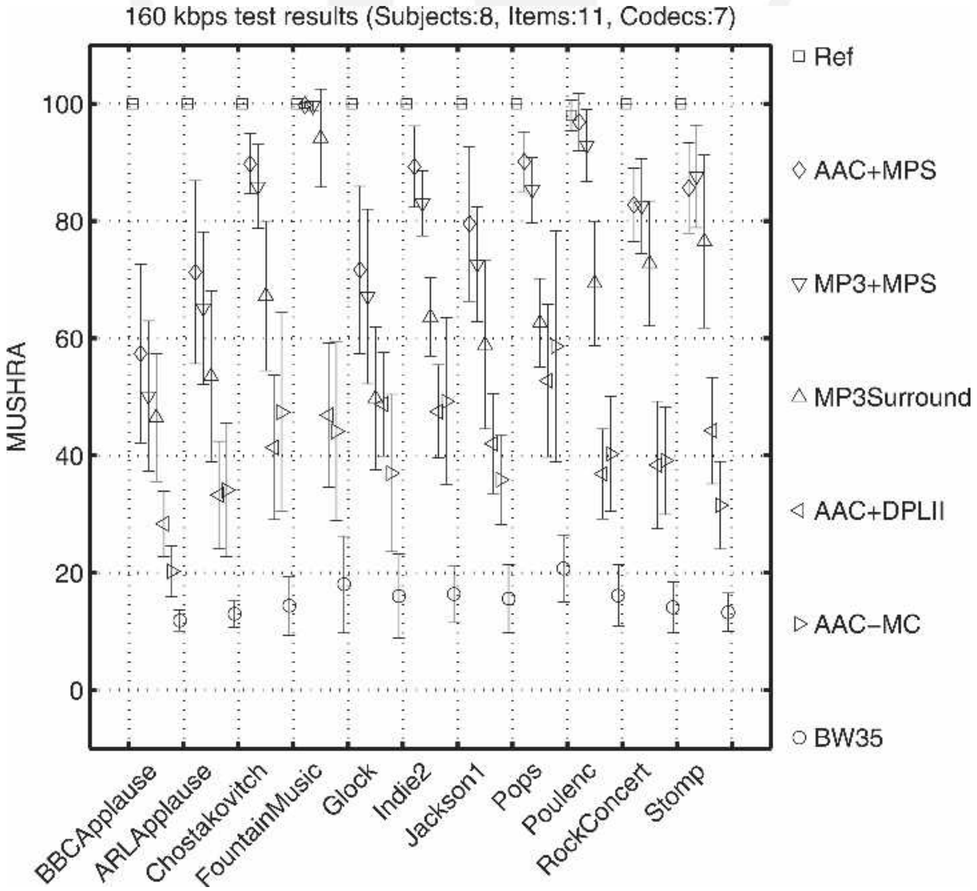

Fig. 17. Mean subjective results averaged across listeners for each codec and excerpt. Error bars denote 95% confidence intervals.

fidence interval. The low-pass anchor (circles), on the other hand, consistently has the lowest scores, around 10 to 20. The scores for AAC multichannel (rightward triangles) are between 20 and 60 for the individual excerpts, and its average rates approximately 40. Stereo AAC in combination with Dolby Prologic II (leftward triangles) scores only slightly higher on average. For 10 out of the 11 excerpts, the combination of stereo AAC and MPEG Surround has the highest scores (diamonds).

The overall scores (averaged across subjects and excerpts) are given in Fig. 18. AAC with MPEG Surround scores approximately 5 points higher than MP3 with MPEG Surround. MP3 Surround scores approximately 15 points lower than MPEG Surround when combined with MP3.

5.3 Discussion

The results indicate the added value of parametric side information with a stereo transmission channel (configurations 1, 2, and 3 versus configurations 4 and 5). The increase in quality for MPEG Surround compared to discrete multichannel coding or matrixed-surround methods amounts to more than 40 MUSHRA points (using AAC as the core coder), which is a considerable improvement. All three parameter-enhanced codecs demonstrated such a clear benefit, enabling high-quality audio transmission at bit rates that are currently used for high-quality stereo transmission. The two core coders tested seem to have only a limited effect since the difference between AAC with MPEG Surround and MP3 with MPEG Surround is reasonably small. On the other hand, given the large difference in quality between configurations 4 and 5 which are based on the same core coder using virtually the same bit rate, the two different parametric enhancements (MPEG Surround and MP3 Surround, respectively) seem to differ significantly in terms of quality and compression efficiency. MPEG Surround delivers significantly higher quality while using only 69% of the parameter bit rate of MP3 Surround.

6 CONCLUSIONS

In this paper a parametric extension to mono or stereo audio codecs has been described, which shows high-quality multi-channel capabilities at bit rates that are equal to those currently employed for stereo transmission. The subjective listening test revealed superior perceptual quality of MPEG Surround over conventional multichannel AAC, matrixed-surround, and MP3 Surround coders at an overall bit rate of 160 kbps.

Full backward compatibility is guaranteed with legacy receivers by storing parametric side information in the ancillary data part of existing compression schemes. The spatial side information is scalable between 0 and (typically) 32 kbps, although higher rates are supported for applications demanding (near) transparency.

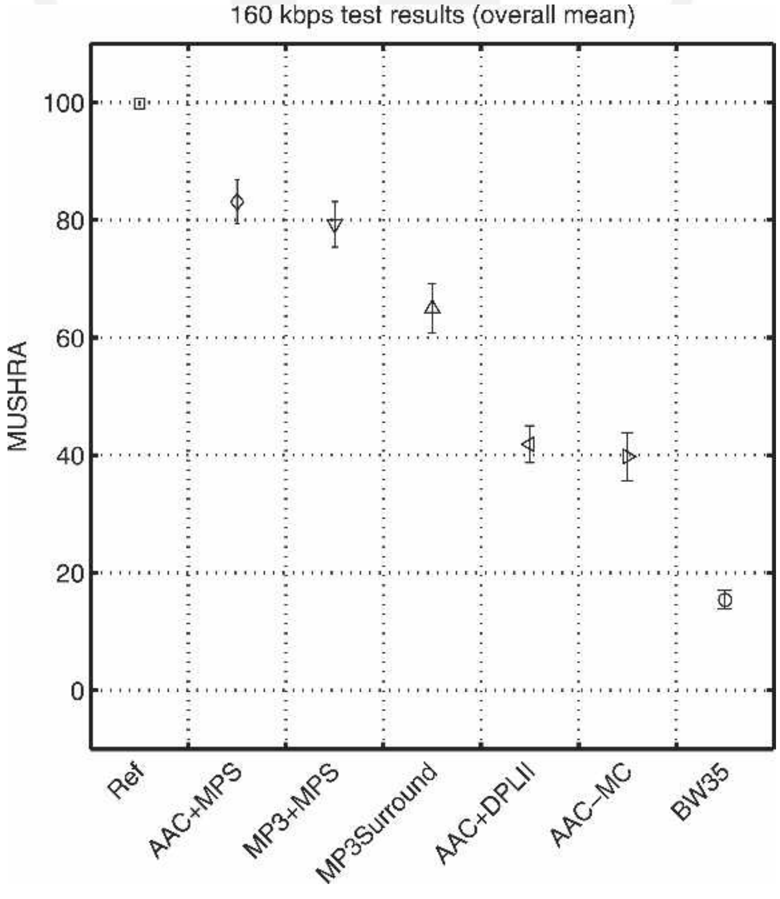

Fig. 18. Overall mean subjective results for each codec.

The system described is very flexible in terms of input and output channels; all common loudspeaker configurations are supported. The flexibility also extends to the down-mix domain. MPEG Surround features automated down mixes that can be mono, stereo, or matrixed-surround compatible stereo. Even multichannel 5.1 can be used as a down mix for configurations with a higher number of audio channels (such as 7.1 or 10.2). Furthermore, support is provided for externally provided down mixes as well. Last but not least, MPEG Surround features an enhanced matrix mode that enables up conversion of legacy stereo material to high-quality multichannel content.

7 ACKNOWLEDGMENT

The authors would like to thank their colleagues from Coding Technologies, Fraunhofer IIS, and Agere for their fruitful cooperation in developing the MPEG Surround standard. They would like to thank their colleagues Armin Kohlrausch and Bert den Brinker for their valuable input on earlier versions of this manuscript. They would also like to thank the anonymous reviewers and the editor of the AES *Journal* for their helpful comments on improving the manuscript.

8 REFERENCES

[1] A. J. M. Houtsma, C. Trahiotis, R. N. J. Veldhuis, and R. van der Waal, "Bit Rate Reduction and Binaural Masking Release in Digital Coding of Stereo Sound," *Acustica/acta acustica,* vol. 92, pp. 908–909 (1996).

[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 Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 42, p. 394 (1994 May), preprint 3799.

[5] 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).

[6] E. Schuijers, J. Breebaart, H. Purnhagen, and J. Engdegard, "Low Complexity Parametric Stereo Coding," presented at the 116th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 52, p. 800 (2004 July/Aug.), convention paper 6073.

[7] E. Schuijers, W. Oomen, B. den Brinker, and J. Breebaart, "Advances in Parametric Coding for High-Quality Audio," presented at the 114th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 51, pp. 440, 441 (2003 May), convention paper 5852.

[8] F. Baumgarte and C. Faller, "Why Binaural Cue Coding Is Better than Intensity Stereo Coding," presented at the 112th Convention of the Audio Engineering Society, *T. Audio Eng. Soc.* (*Abstracts*), vol. 50, p. 515 (2002 June), convention paper 5575.

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

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

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

[12] C. Faller and F. Baumgarte, "Binaural Cue Coding: A Novel and Efficient Representation of Spatial Audio," in *Proc. ICASSP* (2002).

[13] C. Faller and F. Baumgarte, "Binaural Cue Coding Applied to Stereo and Multichannel Audio Compression," presented at the 112th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 50, p. 515 (2002 June), convention paper 5574.

[14] 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," presented at the 119th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 53, p. 1228 (2005 Dec.), convention paper 6599.

[15] J. Breebaart, J. Herre, L. Villemoes, C. Jin, K. Kjörling, and J. Plogsties, "Multichannel Goes Mobile: MPEG Surround Binaural Rendering," in *Proc. 29th AES Int. Conf.* (Seoul, Korea, 2006).

[16] 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," presented at the 118th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 53, pp. 693, 694 (2005 July/Aug.), convention paper 6447.

[17] 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," in *Proc. 28th AES Int. Conf.* (Pitea, Sweden, 2006).

[18] J. Herre, K. Kjörling, J. Breebaart, C. Faller, K. S. Chon, S. Disch, H. Purnhagen, J. Koppens, J. Hilpert, J. Röden, W. Oomen, K. Linzmeier, and L. Villemoes, "MPEG Surround—The ISO/MPEG Standard for Efficient and Compatible Multichannel Audio Coding," presented at the 122nd Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 55, to be published (2007).

[19] Audio Subgroup, "Call for Proposals on Spatial Audio Coding," ISO/IEC JTC1/SC29/WG11 N6455 International Standards Organization, Geneva, Switzerland (2004).

[20] "MPEG Audio Technologies—Part 1: MPEG Sur-

round," ISO/IEC FDIS 23003-1:2006(E), International Standards Organization, Geneva, Switzerland (2004).

[21] W. E. Feddersen, T. T. Sandel, D. C. Teas, and L. A. Jeffress, "Localization of High Frequency Tones," *J. Acoust. Soc. Am.,* vol. 29, pp. 988–991 (1957).

[22] W. A. Yost, "Lateral Position of Sinusoids Presented with Interaural Intensive and Temporal Differences," *J. Acoust. Soc. Am.,* vol. 70, pp. 397–409 (1981).

[23] D. W. Grantham, *Spatial Hearing and Related Phenomena. Handbook of Perception and Cognition,* (Academic Press, London, 1995), hearing ed. B. C. J. Moore, Ed.

[24] F. L. Wightman and D. J. Kistler, "Headphone Simulation of Free-Field Listening. I. Stimulus Synthesis," *J. Acoust. Soc. Am.,* vol. 85, pp. 858–867 (1989).

[25] F. L. Wightman and D. J. Kistler, "Headphone Simulation of Free-Field Listening. II: Psychophysical Validation," *J. Acoust. Soc. Am.,* vol. 85, pp. 868–878 (1989).

[26] F. L. Wightman and D. J. Kistler, "Resolution of Front–Back Ambiguity in Spatial Hearing by Listener and Source Movement," *J. Acoust. Soc. Am.,* vol. 105, pp. 2841–2853 (1999).

[27] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization* (MIT Press, Cambridge, MA, 1997).

[28] S. P. Lipshitz, "Stereo Microphone Techniques: Are the Purists Wrong?," *J. Audio Eng. Soc.* (*Features*), vol. 34, pp. 716–744 (1986 Sept.)

[29] J. Breebaart, S. van de Par, and A. Kohlrausch, "An Explanation for the Apparently Wider Critical Bandwidth in Binaural Experiments," in *Physiological and Psychophysical Bases of Auditory Function* (Shaker Publ., Maastricht, The Netherlands, 2001), pp. 153–160.

[30] A. Kohlrausch, "Auditory Filter Shape Derived from Binaural Masking Experiments," *J. Acoust. Soc. Am.,* vol. 84, pp. 573–583 (1988).

[31] J. Breebaart, S. van de Par, and A. Kohlrausch, "Binaural Processing Model Based on Contralateral Inhibition. I. Model Setup," *J. Acoust. Soc. Am.,* vol. 110, pp. 1074–1088 (2001).

[32] B. R. Glasberg and B. C. J. Moore, "Derivation of Auditory Filter Shapes from Notched-Noise Data, *Hearing Res.,* vol. 47, pp. 103–138 (1990).

[33] D. W. Grantham and F. L. Wightman, "Detectability of Varying Interaural Temporal Differences," *J. Acoust. Soc. Am.,* vol. 63, pp. 511–523 (1978).

[34] D. W. Grantham, "Discrimination of Dynamic Interaural Intensity Differences," *J. Acoust. Soc. Am.,* vol. 76, pp. 71–76 (1984).

[35] D. W. Grantham and F. L. Wightman, "Detectability of a Pulsed Tone in the Presence of a Masker with Time-Varying Interaural Correlation," *J. Acoust. Soc. Am.,* vol. 65, pp. 1509–1517 (1979).

[36] M. van der Heijden and C. Trahiotis, "A New Way to Account for Binaural Detection as a Function of Interaural Noise Correlation," in *Proc. 20th Midwinter Meeting of the Association for Research in Otolaryngology,* D. J. Lim and G. R. Popelka, Eds. (St. Petersburg Beach, FL, 1997), (in press).

[37] S. van de Par, A. Kohlrausch, J. Breebaart, and M. McKinney, "Discrimination of Different Temporal Envelope Structures of Diotic and Dichotic Target Signals within Diotic Wide-Band Noise," in *Auditory Signal Processing: Physiology, Psychoacoustics, and Models,* D. Pressnitzer, A. de Cheveigné, S. McAdams, and L. Collect, ed., *Proc. 13th Int. Symp. on Hearing* (Springer, New York, 2004).

[38] J. Engdegard, H. Purnhagen, J. Röden, and L. Liljeryd, "Synthetic Ambience in Parametric Stereo Coding," presented at the 116th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 52, pp. 800, 801 (2004 July/Aug.), convention paper 6074.

[39] H. Purnhagen, "Low Complexity Parametric Stereo Coding in MPEG-4," in *Proc. Digital Audio Effects Workshop (DAFX)* (Naples, Italy, 2004 Oct.); available at http://dafx04.na.infn.it/.

[40] H. Purnhagen, J. Engdegard, W. Oomen, and E. Schuijers, "Combining Low Complexity Parametric Stereo with High Efficiency AAC," ISO/IEC JTC1/SC29/WG11 MPEG2003/M10385, International Standards Organization, Geneva, Switzerland (2003 Dec.).

[41] M. Dietz, L. Liljeryd, K. Kjörling, and O. Kunz, "Spectral Band Replication—A Novel Approach in Audio Coding," presented at the 112th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 50, pp. 509, 510 (2002 June), convention paper 5553.

[42] O. Kunz, "Enhancing MPEG-4 AAC by Spectral Band Replication," in *Tech. Sessions Proc. of Workshop and Exhibition on MPEG-4 (WEMP4)* (San Jose Fairmont, US, 2002), pp. 41–44.

[43] M. Wolters, K. Kjörling, D. Homm, and H. Purnhagen, "A Closer Look into MPEG-4 High Efficiency AAC," presented at the 115th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 51, pp. 1221 (2003 Dec.), convention paper 5871.

[44] G. Hotho, L. Villemoes, and J. Breebaart, "A Stereo Backward Compatible Multichannel Audio Codec, "*IEEE Trans. Audio, Speech, Language Process* (submitted 2007).

[45] "Report on the Verification Tests of MPEG-D MPEG Surround," ISO/IEC JTC1/SC29/WG11 N8851, International Standards Organization, Geneva, Switzerland (2007 Jan.).

[46] J. Herre, C. Faller, C. Ertel, J. Hilpert, A. Hoelzer, and C. Spenger, MP3 Surround: Efficient and Compatible Coding of Multichannel Audio," presented at the 116th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 52, p. 793 (2004 July/Aug.), convention paper 6049.

[47] ITU-R BS.1534, "Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems (MUSHRA)," International Standards Organization, Geneva, Switzerland (2001).

[48] ITU-R BS.1116-1, "Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems," International Standards Organization, Geneva, Switzerland (1997).

THE AUTHORS

J. Breebaart

G. Hotho

J. Koppens

E. Schuijers

W. Oomen

S. van de Par

Jeroen Breebaart was born in The Netherlands in 1970. He studied biomedical engineering at the Eindhoven University of Technology, The Netherlands, and received a Ph.D. degree in 2001 from the Institute for Perception Research (IPO) in the field of mathematical models of human spatial hearing.

Currently he is a researcher in the Digital Signal Processing Group at Philips Research Laboratories, Eindhoven. His main fields of interest and expertise are spatial hearing, parametric stereo and multichannel audio coding, automatic audio content analysis, and audio signal processing tools.

Dr. Breebaart published several papers on binaural detection, binaural modeling, and spatial audio coding. He also contributed to the development of parametric stereo coding algorithms as currently standardized in MPEG-4 and 3GPP and the recently finalized MPEG Surround standard.

●

Gerard Hotho was born in ’s-Hertogenbosch, The Netherlands, in 1969. He graduated in information technology and electrical engineering from Eindhoven University of Technology in 1993 and 1995, respectively.

As a researcher he is very much inspired by the ideas of J. Goethe and R. Steiner. Professionaly he has worked on digital signal processing topics for 10 years, initially in the field of sonar, later in the field of audio coding, where he tries to combine his passion for music with the inner beauty he occasionally experiences from mathematics.

●

Jeroen Koppens was born in The Netherlands in 1980. He received an M.Sc. degree in electrical engineering from the Eindhoven University of Technology, The Netherlands, in 2005. He did his graduation project in the Signal Processing Group of Philips Applied Technologies, where he worked on a state-of-the-art psychoacoustic model. After graduation he joined Philips Applied Technologies and contributed to the development of the MPEG Surround standard.

●

Erik Schuijers was born in The Netherlands in 1976. He received an M.Sc. degree in electrical engineering from the Eindhoven University of Technology, The Netherlands, in 1999.

In 2000 he joined the Signal Processing Group of Philips Applied Technologies in Eindhoven. His main activity has been the research and development of the MPEG-4 parametric audio and parametric stereo coding tools. He has also been actively involved in the development of the MPEG Surround spatial audio coding system.

●

Werner Oomen received an Ingenieur degree in electronics from the University of Eindhoven, The Netherlands, in 1992.

He joined Philips Research Laboratories in Eindhoven, in the Digital Signal Processing Group in 1992, leading and contributing to a diversity of audio signal processing projects. His main activities are in the field of audio source coding algorithms. Since 1999 he has been with Philips Applied Technologies, Eindhoven, in the Digital Signal Processing Group, where he leads and contributes to different topics related to digital signal processing of audio signals.

Since 1995 Mr. Oomen has been involved with standardization bodies, primarily 3GPP and MPEG, where for the latter he has actively contributed to the standardization of MPEG2-AAC, MPEG4-WB CELP, parametric (stereo) coding, lossless coding of 1-bit oversamples audio, and MPEG Surround.

Steven van de Par studied physics at the Eindhoven University of Technology, The Netherlands, and received a Ph.D. degree in 1998 from the Institute for Perception Research on a topic related to binaural hearing. As a postdoctoral researcher at the same institute he studied auditory-visual interaction and he was a guest researcher at the University of Connecticut Health Center. In the beginning of 2000 he joined Philips Research in Eindhoven. His main fields of expertise are auditory and multisensory perception and low-bit-rate audio coding.

Dr. van de Par has published various papers on binaural detection, auditory-visual synchrony perception, and audio coding related topics. He participated in several projects on low-bit-rate audio coding, most recently in the EU project Adaptive Rate-Distortion Optimized Sound CodeR (ARDOR).