A Backward-Compatible Multichannel Audio Codec

Gerard Hotho, Lars F. Villemoes, *Member, IEEE*, and Jeroen Breebaart

***Abstract*—We propose in this paper a backward-compatible multichannel audio codec. This codec represents a multichannel audio input signal by a down mix and parametric data. In order to enable backward compatibility, it is necessary to have the possibility of exerting control over the down-mixing procedure. At the same time, in order to achieve a high coding efficiency, both signal and perceptual redundancies should be exploited. In this paper, we describe a codec that unifies the above-mentioned conditions: backward compatibility and exploitation of both signal and perceptual redundancies. The codec combines a high audio quality and a low parameter bit rate. Moreover, its design is flexible, examples of which are the scalability of the audio quality to (in principle) transparency and the possibility to preserve the correlation structure of the original input signals by using synthetic signals. A stereo backward compatible version of the proposed codec is used as a component of the recently standardized MPEG Surround multichannel audio codec.**

***Index Terms*—Audio coding, Auditory system, codecs, digital audio broadcasting, estimation, prediction, redundancy, signal processing.**

I. Introduction

AUDIO compression algorithms for wide-band audio have been a continuous topic of research and development during the last decades. Initially, research in this area focused predominantly on efficient transmission of mono or stereo content, which led to the well-known MPEG-1 standard [1], [2]. This standard comprises several "layers" that have different complexity/efficiency tradeoffs and enables a broad range of applications, such as audio storage on digital compact cassettes (DCC), digital broadcasting of audio, efficient storage and playback of music from flash memory (so-called "MP3-players"), and online download services. Several years later, the MPEG-2 standard extended MPEG-1 with multichannel capabilities and more advanced compression tools (AAC, cf. [3]).

The MPEG-1 and 2 compression algorithms typically employ three sources for bit-rate reduction. First, they exploit the phenomenon of *auditory masking*. The accuracy of the signal representation can be adjusted individually in various time/frequency tiles. The resulting quantization noise that is introduced is kept below the masked threshold. Second, there is a limited repertoire to exploit *cross-channel redundancies*. For stereo material, quantization noise can be introduced in each channel independently [4], or on a mid/side projection [5], [6]. The latter is especially beneficial if the two channels are highly correlated. Third, further *redundancies* are exploited using entropy coding of the remaining signal components after the mid/side projection and signal quantization.

MPEG-4 extended the predominant signal-domain repertoire for bit-rate reduction with *parametric* techniques. For example, a fully parametric audio coder was introduced that decomposes an audio signal into sinusoidal components, transients, and noise [7], [8]. Also, hybrid techniques were introduced that combine filter-bank or transform-domain compression with parametric representations. One such method is known as "spectral band replication" (SBR), which regenerates high-frequency content using a parameter-guided copy from the low-frequency components that are coded using filter-bank or transform coders [9]–[11]. Another well-known example of hybrid techniques is "parametric stereo" (PS), also known as "binaural cue coding" (BCC). This method parameterizes the perceptually-relevant spatial aspects of a stereo recording [12]–[14]. As such, this method is very effective in exploiting *perceptual irrelevancies* between audio channels. The resulting parameters are combined with a mono down mix of the stereo signal pair. This mono down mix can subsequently be encoded with any existing mono compression algorithm. The combination of AAC as band-limited, mono coder, with SBR and PS is standardized as high-efficiency AAC version 2 (HE-AAC v2) [15].

Recent trends in audio recording and reproduction demonstrate a shift from stereo to multichannel audio. This shift poses new challenges to exploit perceptual irrelevancies and cross-channel redundancies. Methods to exploit cross-channel redundancies in a multichannel setting are not so widespread. Some conventional audio coders such as MPEG-4 AAC can use mid/side projections on channel pairs. More advanced, experimental proposals incorporate multidimensional principle component analysis (PCA) to exploit cross-signal redundancies [16], [17].

Parametric techniques to exploit irrelevancies have also been proposed for surround material. So-called "spatial audio coding" techniques extend the scope of parametric techniques to multichannel audio by encoding level differences and correlation coefficients between various channels, accompanied by a mono down mix [18], [19].

One interesting application of spatial audio coding techniques is the extension of existing stereo services to multichannel audio. In such a scenario, parametric side information can be transmitted along with a backward-compatible *stereo* down mix. The transmission of parametric side information has several important advantages when compared to matrix-surround systems [20]. In matrix-surround systems, the transmitted down mix is created such that surround channels cause the down-mix channels to be out of phase. A matrix-surround decoder detects

Manuscript received January 15, 2007; revised August 24, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. George Tzanetakis.

G. Hotho and J. Breebaart are with Philips Research Laboratories, 5656 AA (WO 02), Eindhoven, The Netherlands (e-mail: gerard.hotho@hotmail.com; jeroen.breebaart@philips.com).

L. F. Villemoes is with Coding Technologies, SE-113 30, Stockholm, Sweden (e-mail: lars.villemoes@codingtechnologies.com).

Digital Object Identifier 10.1109/TASL.2007.910768

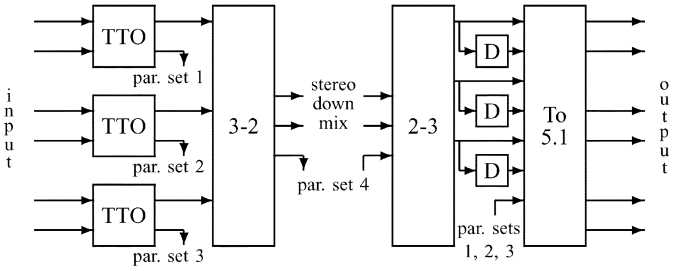

Fig. 1. Generic coder structure of the MPEG Surround 5.1-2-5.1 coder including a 3-2 encoder and a 2-3 decoder element.

these properties to steer the down mix to the front or surround channels. This method does not require any additional side information to be transmitted and can also be used in analog systems. However, the quality of the multichannel reconstruction has been shown to be rather limited [21], [22].

There have been proposals to extend a stereo service to multichannel audio based on a parametric approach. For example, Faller [23] proposed to extend the BCC approach (describing level differences, time differences, and coherence values between certain audio channels) to a stereo down mix. In essence, it aims at (partial) reconstruction of those statistical properties of multichannel audio signals that are most relevant from a perceptual point of view. While such a parametric representation results in a very high compression efficiency, it also has two drawbacks. The first drawback is that it does not provide any means to specifically exploit signal *redundancies* in its parameterization. Second, it is often observed that parametric methods provide unsurpassed compression efficiency at low bit rates, but often fail to reach very high quality levels (perceptual transparency) due to limitations of the underlying parametric model.

The approach described in the current paper aims at extending the fully parametric approach with dedicated methods to exploit both perceptual irrelevancy as well as signal redundancy (inevitably introduced by a down-mix process where at least one audio channel is present in at least two down-mix channels) and to provide means to overcome quality limitations of a parametric method. Examples of the latter are the possibility to regenerate the correlation structure of the original input signals at the output by adding so-called decorrelated signals and the scalability of the coder to (in principle) transparency by making use of residual signals. A so-called 3-2-3 version of the proposed approach is part of the current ISO-MPEG standard for multichannel audio, called "MPEG Surround" [24], [21], [25]. This standard comprises a decoding module that converts a stereo down-mix signal to a three-channel configuration based on transmitted parameters and exploits both cross-channel signal redundancies as well as perceptual irrelevancies. Moreover, this module has different modes to adapt the processing to the extent to which the waveform is preserved by the audio coder employed to code the stereo down mix.

The incorporation of the 3-2-3 module in the stereo backward compatible MPEG Surround coder, which is henceforth referred to as the MPS 5.1-2-5.1 coder, is shown in Fig. 1. The six (5.1) input channels of the encoder (left panel) are first pairwise combined using two-to-one (TTO) encoder elements, resulting in three intermediate signals and three parameter sets (one set for each TTO element). The three intermediate signals are subsequently processed by a 3-2 encoder element that generates two down-mix signals and a fourth parameter set.

The decoder process (shown in the right panel of Fig. 1) performs the inverse process of the encoder. The two input signals and appropriate parameters are first processed by a 2-3 decoder that generates three intermediate signals. These three intermediate signals and decorrelated versions thereof (generated by decorrelator blocks "D") subsequently serve as input to the block "To 5.1," that generates six (5.1) output channels.

In this paper, we give a detailed description of the codec's encoding and decoding blocks. First, in Section II, we describe the prediction mode of the general codec, with special focus on its 3-2-3 version. In the next section, we discuss the energy mode of the 3-2-3 version. Subsequently, in Section IV the codec is evaluated by means of a subjective listening test. Finally, in Section V, conclusions are drawn.

II. Prediction Coder

In this section, we first treat the general $N-M-N$ coder. This means that we consider a coder that represents N input channels by M down-mix channels and parametric data. Because $N-M$ channels are discarded, information is lost and perfect reconstruction is impossible. In order to get the best possible reconstruction (in the sense of least square errors) of the N input channels at the decoder using only M channels, principal component analysis (PCA) [26] should be used. A drawback of PCA is the fact that no control can be exerted over the perceptual quality of the M down-mix channels, which are not fixed, but input signal dependent. In the case of two down-mix channels, or $M=2$, this means that a good quality of the stereo image of the two down-mix channels is not guaranteed when employing PCA. When imposing a fixed down mix on the M down-mix channels, for $M=2$, a good quality of the stereo image of the two down-mix channels can be obtained. As opposed to PCA, whose N channels are orthogonal so that the $N-M$ discarded channels cannot be predicted using the M down-mix channels, now the $N-M$ channels can—to some extent—be predicted from the M down-mix channels. It is this predictability that can be exploited at the decoder, by sending the appropriate prediction parameters.

A. Coder Using a Fixed Down-Mix Matrix

1) The $N-M-N$ Coder: In this section, we explain an optimal $N-M-N$ coder that uses a fixed (hence, input signal-independent) down mix. The coder structure is shown in Fig. 2. We see N time-domain input signals, denoted by $s_1, s_2, \ldots, s_N$. These signals are segmented resulting in the signal segments $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N$ (not shown in the figure). Next, these segments are decomposed into time/frequency tiles using an analysis filter bank, resulting in the signals $\mathbf{x}_{1,1}, \ldots, \mathbf{x}_{1,K}, \ldots, \mathbf{x}_{N,1}, \ldots, \mathbf{x}_{N,K}$, where $\mathbf{x}_{n,k}$ denotes the kth frequency tile, or parameter band, of the signal segment $\mathbf{s}_n$. For ease of notation, the index k is henceforth omitted. For each time/frequency tile, the encoder generates M down-mix signals, $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M$, and parametric data. The down-mix signals are transformed back to the time-domain using a

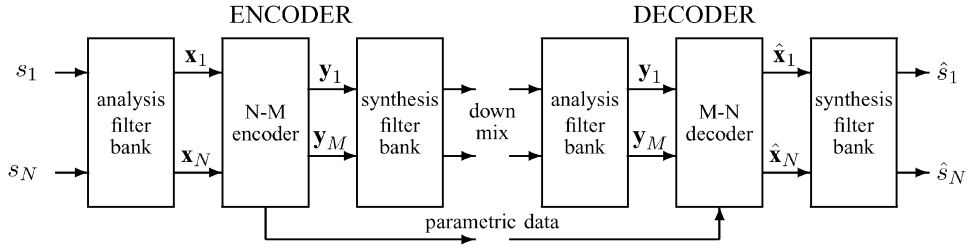

Fig. 2. Generic coder structure of the $N - M - N$ coder.

synthesis filter bank. These signals are sent along with the parametric data to the decoder. At the decoder, the M down-mix signals are decomposed into time/frequency tiles. Next, the decoder generates for each time/frequency tile N output signals, $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_N$, using the M down-mix signals and the parametric data. These signals are converted to the time-domain by means of a synthesis filter bank, resulting in the output signals $\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_N$. This process is described in more detail in the following.

The $L \times 1$ input signals segments $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are obtained by applying an analysis filter bank to the $K \cdot L \times 1$ input signal segments $\mathbf{s}_1, \ldots, \mathbf{s}_N$. This filter bank should mimic the temporal and spectral resolution of the human listener. This is realized by a linear filter bank and grouping of the resulting frequency bands into nonlinearly spaced parameter bands that mimic critical bands [27]. Moreover, because we employ time-variant signal processing (especially at the decoder side), we use an oversampled signal representation in order to reduce aliasing artefacts that would result from a critically sampled filter bank. Finally, because we perform signal prediction at the decoder on the basis of the input signals of the encoder, we use a (near) perfect reconstruction filter bank. For more details of the filter bank, the reader is referred to [22]. Down mixing of the N input signals (i.e., time/frequency tiles) to the M down-mix signals is described by

$$\mathbf{Y} = \mathbf{X}\mathbf{D} \tag{1}$$

where $\mathbf{Y}$ denotes the $L \times M$ matrix containing the M down-mix signals, $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M]$, $\mathbf{X}$ denotes the $L \times N$ matrix containing the N input signals, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$, and $\mathbf{D}$ is a fixed $N \times M$ down-mix matrix.

The M down-mix signals can be extended with $N - M$ channels, denoted by $\mathbf{y}_{M+1}, \mathbf{y}_{M+2}, \ldots, \mathbf{y}_N$ such that

$$\mathbf{Y}_N = \mathbf{X}\mathbf{D}_N \tag{2}$$

where the N columns of $\mathbf{Y}_N$ correspond to the N down-mix signals; hence, $\mathbf{Y}_N = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N]$, and $\mathbf{D}_N$ represents a fixed $N \times N$ mixing matrix. In this case, perfect reconstruction is possible at the decoder in the case that matrix $\mathbf{D}_N$ is nonsingular, by computing

$$\mathbf{X} = \mathbf{Y}_N \mathbf{D}_N^{-1} \tag{3}$$

when the mixing matrix $\mathbf{D}_N$ is known at the decoder. Because in our case only M down-mix signals are available at the decoder, a different approach is required. At the encoder, the $N - M$ discarded channels $\mathbf{y}_{M+1}, \mathbf{y}_{M+2}, \ldots, \mathbf{y}_N$ can be predicted as linear combinations of the M transmitted down-mix channels. This is described by the following equation:

$$\hat{\mathbf{Y}}_{N-M} = \mathbf{Y}\tilde{\mathbf{C}} \tag{4}$$

where $\hat{\mathbf{Y}}_{N-M}$ is the $L \times (N - M)$ matrix containing the approximations of the $N - M$ segments, $\hat{\mathbf{Y}}_{N-M} = [\hat{\mathbf{y}}_{M+1}, \hat{\mathbf{y}}_{M+2}, \ldots, \hat{\mathbf{y}}_N]$, and $\tilde{\mathbf{C}}$ is the $M \times (N - M)$ matrix containing the $M \cdot (N - M)$ prediction parameters.

For choosing these prediction parameters of $\tilde{\mathbf{C}}$ various optimization criteria are possible. We choose a least squares approach described by the problem

$$\begin{aligned} \min_{\hat{\mathbf{Y}}_{N-M}} & \operatorname{trace}\{(\mathbf{Y}_{N-M} - \hat{\mathbf{Y}}_{N-M})^H (\mathbf{Y}_{N-M} - \hat{\mathbf{Y}}_{N-M})\} \\ &= \min_{\tilde{\mathbf{C}}} \operatorname{trace}\{(\mathbf{Y}_{N-M} - \mathbf{Y}\tilde{\mathbf{C}})^H (\mathbf{Y}_{N-M} - \mathbf{Y}\tilde{\mathbf{C}})\} \end{aligned} \tag{5}$$

where the columns of the matrix $\mathbf{Y}_{N-M}$ contain the $N-M$ discarded signals $\mathbf{y}_i$; hence, $\mathbf{Y}_{N-M} = [\mathbf{y}_{M+1}, \mathbf{y}_{M+2}, \ldots, \mathbf{y}_N]$. The error measure of (5) is the square of the Hilbert–Schmidt norm of the error matrix [28], and it is a sum of contributions from each column of $\tilde{\mathbf{C}}$. Hence, the problem can be solved by independently solving a least squares problem for each column of $\tilde{\mathbf{C}}$. The combined solution to this problem in terms of $\hat{\mathbf{Y}}_{N-M}$ is the orthogonal projection of the columns of $\mathbf{Y}_{N-M}$ on the vector space spanned by the columns of $\mathbf{Y}$, which, for the case that $\mathbf{Y}^H\mathbf{Y}$ is nonsingular, is expressed by

$$\hat{\mathbf{Y}}_{N-M} = \mathbf{Y}\tilde{\mathbf{C}} = \mathbf{Y}(\mathbf{Y}^H\mathbf{Y})^{-1}\mathbf{Y}^H\mathbf{Y}_{N-M} \tag{6}$$

so that we find for $\tilde{\mathbf{C}}$

$$\tilde{\mathbf{C}} = (\mathbf{Y}^H\mathbf{Y})^{-1}\mathbf{Y}^H\mathbf{Y}_{N-M}. \tag{7}$$

The down-mix signals $\mathbf{y}_1, \ldots, \mathbf{y}_M$ are converted to time-domain signals using a synthesis filter bank. These time-domain down-mix signals are sent, along with the parametric data contained in $\tilde{\mathbf{C}}$, to the decoder.

At the decoder, the time-domain encoder output signals are assumed to be identical to the time-domain decoder input signal. The decoder time-domain input signals are converted into time/frequency tiles using an analysis filter bank, which is identical to the encoder analysis filter bank. This results in M down-mix signals $\mathbf{y}_1, \ldots, \mathbf{y}_M$ (assuming a perfectly reconstructing filter bank). Subsequently, the $N - M$ discarded signals contained in $\mathbf{Y}_{N-M}$ are predicted using the coder parameters $\tilde{\mathbf{C}}$ as expressed by (6). The output signals that are contained in the columns of the $L \times N$ matrix $\hat{\mathbf{X}}$ are computed as

$$\hat{\mathbf{X}} = \hat{\mathbf{Y}}_N \mathbf{D}_N^{-1} \tag{8}$$

where the $L \times N$ matrix $\hat{\mathbf{Y}}_N$ contains the M down-mix signals and the predictions of the $N-M$ discarded signals, $\hat{\mathbf{Y}}_N = [\mathbf{y}_1, \ldots, \mathbf{y}_M, \hat{\mathbf{y}}_{M+1}, \ldots, \hat{\mathbf{y}}_N]$. It is assumed that the mixing matrix $\mathbf{D}_N$ is *a priori* known at the decoder. In order to obtain signals that cover the entire frequency band, (8) is evaluated for all parameter bands. These signals are combined using a synthesis filter bank, resulting in the time domain signal segments, $\hat{\mathbf{s}}_1, \hat{\mathbf{s}}_2, \ldots, \hat{\mathbf{s}}_N$. The time domain signals $\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_N$ are obtained by concatenating consecutive associated time domain segments.

2) 3-2-3 Coder: In this section, the $N-M-N$ coder of the previous section is elaborated for $N = 3$ and $M = 2$; hence, in the form in which it is used in the MPS 5.1-2-5.1 multichannel coder.

The 3-2-3 encoder has three input channels, left, l, right, r, and center, c. We start by premultiplying the center channel as follows:

$$c \mapsto \frac{1}{2}\sqrt{2}c. \tag{9}$$

The output channels are the two down-mix channels, left, l_0 and right, r_0. Extending these two down-mix channels with a third channel, referred to as c_0, $\mathbf{X}$ and $\mathbf{Y}_N$ of (2) are given by $\mathbf{X} = [\mathbf{l}, \mathbf{r}, \mathbf{c}]$ and $\mathbf{Y}_N = [\mathbf{l}_0, \mathbf{r}_0, \mathbf{c}_0]$. The three channels of $\mathbf{Y}_N$ are given by

$$\begin{aligned} \mathbf{l}_0 &= \mathbf{l} + \mathbf{c}, \\ \mathbf{r}_0 &= \mathbf{r} + \mathbf{c}, \\ \mathbf{c}_0 &= \mathbf{l} + \mathbf{r} - \mathbf{c} \end{aligned} \tag{10}$$

where the specific choice for $\mathbf{l}_0$, and $\mathbf{r}_0$ is driven by the demand for a good quality of their stereo image. The premultiplication of channel c, as expressed by (9), was performed in order to create a phantom c channel with an energy similar to that of the original c channel. Furthermore, the third down-mix channel $\mathbf{c}_0$ is chosen such that its down-mix weight-vector is orthogonal to those of $\mathbf{l}_0$ and $\mathbf{r}_0$.

Parameter matrix $\tilde{\mathbf{C}} = [\tilde{C}_{1,1}, \tilde{C}_{2,1}]^T$, whose elements are the two prediction coefficients for predicting the center channel, is found after some algebra using (7)

$$\begin{aligned} \tilde{C}_{1,1} &= \frac{< \mathbf{l}_0, \mathbf{c}_0 >^* \, ||\mathbf{r}_0||^2 - < \mathbf{r}_0, \mathbf{c}_0 >^* < \mathbf{l}_0, \mathbf{r}_0 >^*}{||\mathbf{l}_0||^2 ||\mathbf{r}_0||^2 - | < \mathbf{l}_0, \mathbf{r}_0 > |^2}, \\ \tilde{C}_{2,1} &= \frac{< \mathbf{r}_0, \mathbf{c}_0 >^* \, ||\mathbf{l}_0||^2 - < \mathbf{l}_0, \mathbf{c}_0 >^* < \mathbf{l}_0, \mathbf{r}_0 >}{||\mathbf{l}_0||^2 ||\mathbf{r}_0||^2 - | < \mathbf{l}_0, \mathbf{r}_0 > |^2} \end{aligned} \tag{11}$$

with

$$\begin{aligned} < \mathbf{a}, \mathbf{b} > &\equiv \sum_k a[k] b^*[k] \\ ||\mathbf{a}||^2 &\equiv \sum_k |a[k]|^2. \end{aligned} \tag{12}$$

In the case of $\mathbf{r}_0 \approx \beta \mathbf{l}_0$, (11) describing the variables $\tilde{C}_{1,1}$ and $\tilde{C}_{2,1}$ becomes ill conditioned because then the denominators approach zero. Now, the two-channel (both $\mathbf{l}_0$ and $\mathbf{r}_0$) optimization problem for $\tilde{C}_{1,1}$ and $\tilde{C}_{2,1}$ that can be written as

$$\min_{\hat{\mathbf{c}}_0} ||\mathbf{c}_0 - \hat{\mathbf{c}}_0||^2 = \min_{\tilde{C}_{1,1}, \tilde{C}_{2,1}} ||\mathbf{c}_0 - \tilde{C}_{1,1}\mathbf{l}_0 - \tilde{C}_{2,1}\mathbf{r}_0||^2 \tag{13}$$

is reduced to two single-channel problems

$$\min_{\tilde{C}'_{1,1}} ||\mathbf{c}_0 - \tilde{C}'_{1,1}\mathbf{l}_0||^2 \text{ and } \min_{\tilde{C}'_{2,1}} ||\mathbf{c}_0 - \tilde{C}'_{2,1}\mathbf{r}_0||^2 \tag{14}$$

for which the following solutions are found:

$$\begin{aligned} \tilde{C}'_{1,1} &= \frac{< \mathbf{l}_0, \mathbf{c}_0 >^*}{||\mathbf{l}_0||^2} \\ \tilde{C}'_{2,1} &= \frac{< \mathbf{r}_0, \mathbf{c}_0 >^*}{||\mathbf{r}_0||^2}. \end{aligned} \tag{15}$$

Having two sets of parameters, one for the single-channel and one for the two-channel problem, we next investigate how these sets are related. Because the single-channel problem is a special case of the two-channel problem, it is possible to return to a single parameter set using this relation. A single parameter set is beneficial in terms of coder efficiency. Rewriting the two-channel problem for the case of $\mathbf{r}_0 \approx \beta \mathbf{l}_0$

$$\hat{\mathbf{c}}_0 = \tilde{C}_{1,1}\mathbf{l}_0 + \tilde{C}_{2,1}\mathbf{r}_0 \approx (\tilde{C}_{1,1} + \beta \tilde{C}_{2,1})\mathbf{l}_0 \tag{16}$$

and observing the descriptions of the two single-channel problems

$$\begin{aligned} \hat{\mathbf{c}}_0 &= \tilde{C}'_{1,1}\mathbf{l}_0, \\ \hat{\mathbf{c}}_0 &= \tilde{C}'_{2,1}\mathbf{r}_0 \approx \beta \tilde{C}'_{2,1}\mathbf{l}_0 \end{aligned} \tag{17}$$

we see that parameters of the single-channel problem relate to the parameters of the two-channel problem in the following way:

$$\tilde{C}'_{1,1} \approx \beta \tilde{C}'_{2,1} \approx \tilde{C}_{1,1} + \beta \tilde{C}_{2,1}. \tag{18}$$

We fix the relations between the two single-channel problem parameters and the two two-channel problem parameters as follows:

$$\tilde{C}_{1,1} = \frac{1}{2}\tilde{C}'_{1,1} \text{ and } \tilde{C}_{2,1} = \frac{1}{2}\tilde{C}'_{2,1}. \tag{19}$$

Having a single set of parameters for both the single-channel and the two-channel problems, we need to obtain a gradual transition between the single-channel solutions and the two-channel solutions. To this end, the following expressions are used for computing the variables that are actually transmitted to the decoder:

$$\tilde{C}_{i,1} = (1 - s^\eta)\tilde{C}^{(2)}_{i,1} + s^\eta \tilde{C}^{(1)}_{i,1}, i = 1, 2 \tag{20}$$

where $\tilde{C}^{(2)}_{i,1}$ are the solutions for the case that $\mathbf{r}_0 \not\approx \beta \mathbf{l}_0$, as given by (11), $\tilde{C}^{(1)}_{i,1}$ are the solutions for the case that $\mathbf{r}_0 \approx \beta \mathbf{l}_0$, as given by (15) and (19) and s is the measure of similarity between $\mathbf{l}_0$ and $\mathbf{r}_0$, which is given by

$$s = \frac{| < \mathbf{l}_0, \mathbf{r}_0 > |^2}{||\mathbf{l}_0||^2 ||\mathbf{r}_0||^2}. \tag{21}$$

The value of s lies in between 0 (when there is no correlation between $\mathbf{l}_0$ and $\mathbf{r}_0$) and 1 (when $\mathbf{r}_0 = \beta \mathbf{l}_0$). The value of η was determined on the basis of the need for a smooth, yet swift,

transition between the two solutions. Comparing several values in an informal listening experiment yielded a value of 8.

At the decoder, we approximate the output signals using (8), which can be written as

$$\begin{aligned}\hat{\mathbf{X}} &= \hat{\mathbf{Y}}_N \mathbf{D}_N^{-1} \\ &= [\mathbf{l}_0, \mathbf{r}_0, \tilde{C}_{1,1}\mathbf{l}_0 + \tilde{C}_{2,1}\mathbf{r}_0]\frac{1}{3}\begin{bmatrix} 2 & -1 & 1 \\ -1 & 2 & 1 \\ 1 & 1 & -1 \end{bmatrix} \\ &= \frac{1}{3}[(\tilde{C}_{1,1}+2)\mathbf{l}_0 + (\tilde{C}_{2,1}-1)\mathbf{r}_0 \\ &\quad \times (\tilde{C}_{1,1}-1)\mathbf{l}_0 + (\tilde{C}_{2,1}+2)\mathbf{r}_0 \\ &\quad \times (1-\tilde{C}_{1,1})\mathbf{l}_0 + (1-\tilde{C}_{2,1})\mathbf{r}_0]. \end{aligned} \tag{22}$$

Finally, the premultiplication of the center channel as expressed by (9) is corrected for.

B. Residual Signals and Energy Preservation

Residual signals are those signals that make a perfect reconstruction of the input signals by the decoder possible, in absence of signal quantization, ignoring windowing effects and assuming perfectly reconstructing filter banks. For the coder that was described in the previous section, the residual signals, contained in the $L \times (N-M)$ residual matrix $\mathbf{Y}_r$, are the difference between the discarded $N-M$ down-mix signals and the predictions thereof, which is expressed by

$$\mathbf{Y}_r = \mathbf{Y}_{N-M} - \hat{\mathbf{Y}}_{N-M}. \tag{23}$$

It is possible to send the residual signals parameterized to the decoder, so that the input signals can, in principle, be perfectly reconstructed. To allow for perfect reconstruction, it is necessary to compute at the encoder the discarded signals contained in $\hat{\mathbf{Y}}_{N-M}$ using quantized parameters $\tilde{\mathbf{C}}$. Sometimes the available bit rate is too limited to send the full-band residual signals to the decoder. In that case, it is beneficial to transmit only the low-frequency part of these residual signals, as this results in the largest quality improvement. If no bit-rate is available for transmitting the residual signals, an alternative procedure can be followed.

For this alternative procedure, we first restate the geometrical interpretation that the prediction signals are the orthogonal projection of the discarded signals on the vector space spanned by the down-mix signals, as expressed by (6). Therefore, the residual signals as defined by (23) are orthogonal to (or uncorrelated with) the down-mix signals. From this it follows that the prediction signals have at most the same amount of energy as the discarded signals themselves (if the prediction is perfect while ignoring signal quantization). In all other cases, an energy loss is associated with the prediction signals. This energy loss can be compensated by using an energy preservation parameter γ which is computed at the encoder as follows:

$$\gamma = \sqrt{\frac{\sum_{i=1}^{N} \|\hat{\mathbf{x}}_i\|^2}{\sum_{i=1}^{N} \|\mathbf{x}_i\|^2}} \tag{24}$$

where $\hat{\mathbf{x}}_i$ denotes the ith column of the matrix containing the decoder output signals $\hat{\mathbf{X}}$, that for the 3-2-3 coder can be computed using (22). Obviously, $\gamma \leq 1$. When at the decoder, the output signals are scaled with γ^{-1}; hence

$$\hat{\mathbf{X}} \mapsto \gamma^{-1}\hat{\mathbf{X}} \tag{25}$$

the summed energy of the scaled output signals matches the summed energy of the input signals. To prevent multiplication of the output signals with too large an amplification factor, the value of γ, is limited from below as follows:

$$\gamma \mapsto \max\left(\gamma, \frac{5}{6}\right). \tag{26}$$

The value of $(5)/(6)$ was experimentally established.

C. Correlation Reproduction

Residual signals are used with the goal of reconstructing the waveforms of the original input signals. Without residual signals a transmitted energy preservation parameter γ enables a reconstruction of the correct total energy. The intermediate solution to be described here will result in a reconstruction of the correlation structure of the original input signals by means of replacing the residual signals with so-called decorrelation signals [22]. An important consequence is that, apart from deficiencies due to imperfect decorrelators, any linear combination of the output channels will have the correct power. This method also extends the paradigm of parametric stereo coding [12]–[14] to the $N-M-N$ coder in a natural way.

1) N-M-N Coder: As we saw in Section II-B, the down-mix signals $\mathbf{Y}$ are orthogonal to the residual signals $\mathbf{Y}_r$, or

$$\mathbf{Y}^H\mathbf{Y}_r = \mathbf{0}. \tag{27}$$

It follows that in order to reproduce the original signal correlation, or rather its sample covariance structure $\mathbf{X}^H\mathbf{X}$, it suffices to replace the residual signal (or prediction error signal) matrix $\mathbf{Y}_r$ with a synthetic $L \times (N-M)$ signal matrix $\mathbf{S}$ satisfying

$$\mathbf{Y}^H\mathbf{S} = \mathbf{0} \text{ and } \mathbf{S}^H\mathbf{S} = \mathbf{Y}_r^H\mathbf{Y}_r \tag{28}$$

which we will verify next.

Assume (28) holds and consider the enhanced predicted signal

$$\tilde{\mathbf{Y}}_{N-M} = \hat{\mathbf{Y}}_{N-M} + \mathbf{S}. \tag{29}$$

The corresponding enhanced extended down-mix signal is $\tilde{\mathbf{Y}}_N = [\mathbf{Y}, \tilde{\mathbf{Y}}_{N-M}]$. For the sample covariance of the enhanced extended down-mix signals, we find by applying block notation

$$\tilde{\mathbf{Y}}_N^H\tilde{\mathbf{Y}}_N = \begin{bmatrix} \mathbf{Y}^H \\ \hat{\mathbf{Y}}_{N-M}^H + \mathbf{S}^H \end{bmatrix}[\mathbf{Y} \quad \hat{\mathbf{Y}}_{N-M} + \mathbf{S}] \tag{30}$$

which equals

$$\begin{bmatrix} \mathbf{Y}^H\mathbf{Y} & \mathbf{Y}^H(\hat{\mathbf{Y}}_{N-M}+\mathbf{S}) \\ \left(\hat{\mathbf{Y}}_{N-M}^H + \mathbf{S}^H\right)\mathbf{Y} & \left(\hat{\mathbf{Y}}_{N-M}^H + \mathbf{S}^H\right)(\hat{\mathbf{Y}}_{N-M}+\mathbf{S}) \end{bmatrix}. \tag{31}$$

In Section II-A1, we saw that the predicted discarded signals $\hat{\mathbf{Y}}_{N-M}$ are the result of an orthogonal projection on the vector space spanned by the columns of $\mathbf{Y}$. Therefore, using (28), we find that $\hat{\mathbf{Y}}_{N-M}^H\mathbf{S} = \mathbf{0}$. Using this last result and (28), we

find that $\mathbf{Y}^H(\hat{\mathbf{Y}}_{N-M}+\mathbf{S}) = \mathbf{Y}^H\mathbf{Y}_{N-M}$, and $(\hat{\mathbf{Y}}^H_{N-M}+\mathbf{S}^H)(\hat{\mathbf{Y}}_{N-M}+\mathbf{S}) = \mathbf{Y}^H_{N-M}\mathbf{Y}_{N-M}$. Substitution of these results in (30) shows that

$$\tilde{\mathbf{Y}}^H_N\tilde{\mathbf{Y}}_N = \mathbf{Y}^H_N\mathbf{Y}_N. \tag{32}$$

The enhanced output signal matrix is $\tilde{\mathbf{X}} = \tilde{\mathbf{Y}}_N\mathbf{D}_N^{-1}$. Therefore, using (3) and (32), we find that the sample covariance of the enhanced output signals $\tilde{\mathbf{X}}^H\tilde{\mathbf{X}}$ equals the sample covariance of the output signals $\mathbf{X}^H\mathbf{X}$, which was to be proven.

In practice, the $N-M$ synthetic signal columns of $\mathbf{S}$ are obtained by first filtering of the down-mix signal rows of $\mathbf{Y}$, or of the decoded predicted signal rows of $\hat{\mathbf{X}}$, with a set of decorrelation filters in order to obtain $N-M$ mutually orthogonal decorrelation signals. A suitable linear combination of those signals is then constructed in order to meet the correlation structure specification given by the second part of (28). Parameters describing the correlation matrix $\mathbf{Y}^H_r\mathbf{Y}_r$ have to be transmitted in addition to the prediction parameters.

2) 3-2-3 Coder: For the 3-2-3 coder ($N=3$ and $M=2$) that is used in the MPS 5.1-2-5.1 coder, the theory of Section II-C1 becomes simpler. Since $N-M=1$, (29) turns into a vector equation. The enhanced predicted signal $\tilde{\mathbf{c}}_0$ is the sum of the synthetic signal $\mathbf{s}$ and the predicted signal $\hat{\mathbf{c}}_0$, or

$$\tilde{\mathbf{c}}_0 = \hat{\mathbf{c}}_0 + \mathbf{s}. \tag{33}$$

By taking the synthetic signal to be a decorrelation signal, we comply with the first condition for the synthetic signal expressed in (28). The second condition in this equation is complied with when the energy of the synthetic signal $\mathbf{s}$ equals the residual signal (or prediction error signal) energy

$$||\mathbf{s}||^2 = ||\mathbf{c}_0 - \hat{\mathbf{c}}_0||^2. \tag{24}$$

We assume that a decorrelator $\mathbf{d}\{\cdot\}$ both preserves the energy of its input signal and produces an output signal that is uncorrelated with (or orthogonal to) its input signal. At the decoder, one could generate $\mathbf{s}$ by feeding a combination of the down-mix channels or the predicted channels to a decorrelator and applying a gain adjustment in order to fulfill (34). For example, with $\mathbf{s} = g_1\,\mathbf{d}\{\hat{\mathbf{c}}_0\}$, the value of the gain adjustment factor g_1 can be derived from the predicted signal quotient $\kappa_0 = ||\hat{\mathbf{c}}_0||/||\mathbf{c}_0||$, via $g_1 = \sqrt{1-\kappa_0^2}/\kappa_0$. The advantage of using such a *relative* parameter κ_0, which should be transmitted to the decoder, is that no energy measurement is necessary in the decoder. Moreover, its range $0 \le \kappa_0 \le 1$ allows for efficient quantization in the encoder. However, instead of introducing a new parameter κ_0, a reuse of the transmitted energy preservation parameter γ can be enabled by using a sum of decorrelated versions of all three predicted output channels

$$\mathbf{s} = g_2(\mathbf{d}_1\{\hat{\mathbf{x}}_1\} + \mathbf{d}_2\{\hat{\mathbf{x}}_2\} + \mathbf{d}_3\{\hat{\mathbf{x}}_3\}) \tag{35}$$

assuming we have three mutually orthogonal decorrelators $\mathbf{d}_1, \mathbf{d}_2$, and $\mathbf{d}_3$. With this assumption, it follows that

$$||\mathbf{s}||^2 = g_2^2\sum_{i=1}^{3}||\hat{\mathbf{x}}_i||^2. \tag{36}$$

To see how the decorrelator gain g_2 can be adjusted to meet the requirement of (34), based entirely on the energy preservation parameter γ, the starting point is the observation that the residual matrix satisfies $\hat{\mathbf{Y}}^H_N\mathbf{Y}_r = \mathbf{0}$. This follows from $\hat{\mathbf{Y}}_N = [\mathbf{Y}, \hat{\mathbf{Y}}_{N-M}]$, (4) and (27). Hence, we have

$$\mathbf{Y}^H_N\mathbf{Y}_N = \hat{\mathbf{Y}}^H_N\hat{\mathbf{Y}}_N + [\mathbf{0} \quad \mathbf{Y}_r]^H[\mathbf{0} \quad \mathbf{Y}_r]. \tag{37}$$

By postmultiplication with the inverse of the extended down-mix matrix $\mathbf{D}_N^{-1}$ and premultiplication with its adjoint, it follows that

$$\mathbf{X}^H\mathbf{X} = \hat{\mathbf{X}}^H\hat{\mathbf{X}} + \left([\mathbf{0} \quad \mathbf{Y}_r]\mathbf{D}_N^{-1}\right)^H \times \left([\mathbf{0} \quad \mathbf{Y}_r]\mathbf{D}_N^{-1}\right) \tag{38}$$

and by inserting $\mathbf{Y}_r = \mathbf{c}_0 - \hat{\mathbf{c}}_0$, using the expression for $\mathbf{D}_N^{-1}$ from (22), and taking matrix traces, we find that

$$\sum_{i=1}^{3}||\mathbf{x}_i||^2 = \sum_{i=1}^{3}||\hat{\mathbf{x}}_i||^2 + \frac{1}{3}||\mathbf{c}_0 - \hat{\mathbf{c}}_0||^2. \tag{39}$$

By definition of the energy preservation parameter γ, it holds that

$$\sum_{i=1}^{3}||\mathbf{x}_i||^2 = \gamma^{-2}\sum_{i=1}^{3}||\hat{\mathbf{x}}_i||^2. \tag{40}$$

Combining this with (36) and (39) leads to

$$||\mathbf{s}||^2 = \frac{g_2^2\gamma^2}{3(1-\gamma^2)}||\mathbf{c}_0 - \hat{\mathbf{c}}_0||^2 \tag{41}$$

and a comparison with (34) gives the appropriate gain adjustment factor g_2 in (35)

$$g_2 = \sqrt{3}\frac{\sqrt{1-\gamma^2}}{\gamma}. \tag{42}$$

Experimentally, it was found that always adding decorrelation according to the above rule leads to a clear improvement of audio quality in terms of wideness and image stability for many excerpts. On the other hand, especially in cases where the original multichannel signal has a dominant and dry center component, the added decorrelation signal can be perceived as an artefact. Fortunately, since the decorrelator contribution can be shut off by setting $\gamma = 1$ in (42), an optimal decision is, in principle, enabled at the encoding stage. This, however, was not further investigated.

D. Parameters of the 3-2-3 Coder

1) Real Versus Complex Prediction Parameters: The prediction parameters of the $N-M-N$ coder, as expressed by (7), are complex. Because real parameters are cheaper in terms of bit rate, it is investigated if they suffice. For real parameters, (7) changes to

$$\tilde{\mathbf{C}} = \Re\{(\mathbf{Y}^H\mathbf{Y})^{-1}\mathbf{Y}^H\mathbf{Y}_{N-M}\}. \tag{43}$$

For the 3-2-3 coder, the complex parameters are given by (11), (15), and (19). By replacing the terms $< \mathbf{a}, \mathbf{b} >$, as defined by

(12), by their real counterpart, $\Re\{< \mathbf{a}, \mathbf{b} >\}$, in the equations for the complex parameters, we find expressions for the real parameters.

A comparison between real and complex parameters was done using an informal listening test on various excerpts. Besides the fact that no large differences were found between real and complex parameters, different preferences also were found for different excerpts. It was decided to use real parameters, because they are cheaper in terms of bit-rate. In this way, a problem associated with using complex parameters is avoided: the problem of matching the phases of the signals of consecutive segments. Although this problem is solved for the parametric stereo coder by means of the so-called OPD parameter [14], the solution for the multichannel coder cannot straightforwardly be derived thereof.

2) Parameter Quantization: In order to obtain a low bit-rate, the coder parameters, $\gamma, \tilde{C}_{1,1}$ and $\tilde{C}_{2,1}$, need to be quantized. The parameter γ is quantized like the interchannel coherence (ICC) parameter [14] of the parametric stereo coder. Basically, this quantization scheme uses six discrete values in the interval $[0, 1]$, where the quantization step size decreases as the discrete level 1 is approached. The distribution of both (real) parameters, $\tilde{C}_{1,1}$ and $\tilde{C}_{2,1}$, is quite similar in 96 different 5.1-channel excerpts. We found a minimum value of -2 and a maximum value of 3 for either parameter to be a sufficient margin. In between the maximum and the minimum value, we quantize using a fixed step size of 0.1. This step size was chosen on the basis of informal listening experiments. We found, both for $\tilde{C}_{1,1}$ and $\tilde{C}_{2,1}$, an estimated bit rate of about 2.1 kb/s, based on 44 100/2048 updates per second and 28 parameter bands, when using the parameter coding scheme of the MPS coder. Coding of the ICC parameter resulting from one single TTO element requires about 0.8 kb/s in the same setting.

III. 3-2-3 Energy-Based Coder

The transmitted parameters in the so-called "energy mode" of the MPS 5.1-2-5.1 coder convey information regarding the energy distribution of the original three input channels, left, right, and center. This type of information is more absolute and robust than the prediction parameters of the previous section, which are defined relative to a down mix. The energy mode parameters can be used in situations where the encoding and decoding of the down mix by the henceforth-called core coder alters the signal waveforms to such an extent that it leads to problems for the prediction mode. For example, the HE-AAC coder, where SBR is used [9]–[11], completely modifies the waveform in the high-frequency range. When using this coder as a core coder, it is possible to use the prediction mode in the lower frequency range, where no SBR is used, and the energy mode in the high-frequency range where the original waveform is completely lost due to SBR.

A. Plain Energy Mode

In this section, we describe the energy mode for the case that the waveform of the down-mix signals is completely lost. In this case, it is usually not appropriate to use an up-mix matrix that predicts the left and right signals from *both* down-mix signals.

For both the left and the right signal, we aim for energy preservation, as expressed by

$$||\mathbf{Y}\mathbf{u}_1|| = ||\mathbf{l}||, \qquad ||\mathbf{Y}\mathbf{u}_2|| = ||\mathbf{r}|| \tag{44}$$

where $\mathbf{Y} = [\mathbf{l}_0, \mathbf{r}_0]$ contains the two *original* down-mix signals, and $\mathbf{u}_i$ denotes the ith column of the up-mix matrix $\mathbf{U}$. Excluding cross-terms from $\mathbf{l}$ to $\mathbf{r}_0$ and from $\mathbf{r}$ to $\mathbf{l}_0$, we find straightforwardly

$$[\mathbf{u}_1, \mathbf{u}_2] = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix} = \begin{bmatrix} \sqrt{\frac{L}{L_0}} & 0 \\ 0 & \sqrt{\frac{R}{R_0}} \end{bmatrix} \tag{45}$$

where L, R, L_0 and R_0 denote the energies of the left and right input signal and the left and right original down-mix signal, respectively. In the case that the energies of the original down-mix signals is preserved by the core coder, the output signals will be endowed with the same energies as the input channels.

For the center signal, we do not necessarily aim for energy preservation, but mix the estimations on the basis of the left and the right down-mix signal as follows:

$$\mathbf{u}_3 = \begin{bmatrix} U_{13} \\ U_{23} \end{bmatrix} = \begin{bmatrix} \frac{1}{2}\sqrt{\frac{C}{L_0}} \\ \frac{1}{2}\sqrt{\frac{C}{R_0}} \end{bmatrix} \tag{46}$$

where C denotes the energy of the center input signal. From this equation, we see that mixing is performed such that the contribution of the largest center to down-mix signal energy ratio is weighted more heavily. With this choice for the center signal up-mixing procedure, the synthesized energy of the center signal, denoted by C', becomes

$$C' = \left(\frac{1}{2} + \frac{C}{2\sqrt{L_0 R_0}}\right) C \tag{47}$$

in the case that both the three input signals are uncorrelated and the correlation structure of the down-mix signals is preserved by the core coder. This implies that for a strong center signal, the energy reconstruction is close to perfect, as desired.

B. Energy Mode With Center Cancellation

In this section, we describe the energy mode for the case that at least part of the waveform of the down-mix signals is preserved, but the prediction mode is unsuited to handle them. One can think of situations where intricate phase relations between the input channels leads to a suboptimal real valued prediction, or where the down-mix modifications are subtle but strong enough to destabilize the decoder prediction. In such a case, it can be beneficial to use all terms of the energy-based up-mix matrix in order to regain the multichannel signal wideness.

The derivation is based on the model that the three original channels $\mathbf{X} = [\mathbf{l}, \mathbf{r}, \mathbf{c}]$ are uncorrelated. Although this assumption does not seem a realistic one, the method described here was found to give good results in practice. Furthermore, the up-mix matrix $\mathbf{U}$ is defined for each channel by the principle of best waveform match subject to correct energy reproduction

$$\min_{\mathbf{u}_i} \{||\mathbf{x}_i - \mathbf{Y}\mathbf{u}_i||\} \quad \text{subject to} \quad ||\mathbf{Y}\mathbf{u}_i|| = ||\mathbf{x}_i||, \qquad i = 1, 2, 3. \tag{48}$$

Let $\mathbf{Y}\check{\mathbf{u}}_i$ be the orthogonal projection of $\mathbf{x}_i$ onto the span of the down-mix vectors. This is the solution to the unconstrained part of the problem (48). Then, we have

$$\|\mathbf{x}_i - \mathbf{Y}\mathbf{u}_i\|^2 = \|\mathbf{x}_i - \mathbf{Y}\tilde{\mathbf{u}}_i\|^2 + \|\mathbf{Y}\tilde{\mathbf{u}}_i - \mathbf{Y}\mathbf{u}_i\|^2, \quad i = 1,2,3. \tag{49}$$

It follows that the constrained problem is solved by post normalization of the unconstrained projection

$$\mathbf{u}_i = \kappa_i \tilde{\mathbf{u}}_i \text{ where } \kappa_i = \frac{\|\mathbf{x}_i\|}{\|\mathbf{Y}\tilde{\mathbf{u}}_i\|}, \qquad i = 1,2,3. \tag{50}$$

The unconstrained projections are simultaneously found for all channels in the special case that these channels are mutually uncorrelated, as this is the underlying assumption for the energy mode. The resulting up-mix matrix of the unconstrained projection $\tilde{\mathbf{U}}$ is given by

$$\begin{aligned}\tilde{\mathbf{U}} &= \begin{bmatrix} L+C & C \\ C & R+C \end{bmatrix}^{-1} \begin{bmatrix} L & 0 & C \\ 0 & R & C \end{bmatrix} \\ &= \frac{1}{LC+RC+LR} \\ &\quad \times \begin{bmatrix} LC+LR & -RC & RC \\ -LC & RC+LR & LC \end{bmatrix}. \end{aligned} \tag{51}$$

The up-mix matrix $\mathbf{U}$ results from combining (50) and (51) and can be expressed as

$$\mathbf{U} = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ u_{21} & u_{22} & u_{23} \end{bmatrix} \tag{52}$$

with

$$u_{11} = \sqrt{\frac{LC+LR}{LC+RC+LR}}, \tag{53}$$

$$u_{12} = -\sqrt{\frac{C}{C+L}}\sqrt{\frac{RC}{LC+RC+LR}} \tag{54}$$

$$u_{13} = \sqrt{\frac{R}{R+L}}\sqrt{\frac{RC}{LC+RC+LR}} \tag{55}$$

$$u_{21} = -\sqrt{\frac{C}{C+R}}\sqrt{\frac{LC}{LC+RC+LR}} \tag{56}$$

$$u_{22} = \sqrt{\frac{RC+LR}{LC+RC+LR}} \tag{57}$$

$$u_{23} = \sqrt{\frac{L}{L+R}}\sqrt{\frac{LC}{LC+RC+LR}}. \tag{58}$$

From this equation, we see that the left output channel $\hat{\mathbf{l}}$ is given by

$$\begin{aligned}\hat{\mathbf{l}} &= U_{11}\mathbf{l}_0' - \sqrt{\frac{C}{C+R}}\sqrt{\frac{LC}{LC+RC+LR}}\mathbf{r}_0' \\ &= U_{11}\mathbf{l}_0' - \left(\sqrt{\frac{LC}{LC+RC+LR}}\right)\hat{\mathbf{c}}_{\mathbf{r}_0} \end{aligned} \tag{59}$$

where $\mathbf{l}_0'$ and $\mathbf{r}_0'$ indicate the left and right down-mix signal after coding with the core coder, respectively, and $\hat{\mathbf{c}}_{\mathbf{r}_0}$ denotes the estimation of the center channel from the right down-mix channel in the plain energy mode. Because the left output channel equals a weighted left down-mix channel minus a weighted estimate of the center channel, and this similarly holds for the right channel, this mode of operation is referred to as energy mode with center cancellation.

The dynamic upmixing method proposed in [23] also consists of subtracting an estimated center channel from the down-mix channels, but the weights are derived with a focus on the energy reconstruction of the center channel. Moreover, as it will be described in the next subsection, the current method relies entirely on transmitted parameters, whereas the BCC system of [23] requires energy and correlation measurements on the decoded down-mix channels.

C. Coder Parameters

In this section, we first describe the parameters of the energy mode. Then, we describe their quantization. It turns out that all energy up-mix weights can be expressed as smooth functions of two energy ratios, q_1 and q_2, that are given by

$$q_1 = \frac{L+R}{C}, \qquad q_2 = \frac{L}{R}. \tag{60}$$

With these two energy ratios, the up-mix matrix of the plain energy mode can be written as

$$\mathbf{U} = \begin{bmatrix} \sqrt{\frac{q_1 q_2}{q_1 q_2 + q_2 + 1}} & 0 & \frac{1}{2}\sqrt{\frac{q_2+1}{q_1 q_2 + q_2 + 1}} \\ 0 & \sqrt{\frac{q_1}{q_1+q_2+1}} & \frac{1}{2}\sqrt{\frac{q_2+1}{q_1+q_2+1}} \end{bmatrix}. \tag{61}$$

For the up-mixing procedure of the energy mode with center cancellation, we choose for the center channel to use the plain energy mode. This choice is made to limit the decoder complexity, as informal listening revealed only subtle differences between the two methods for this channel. The up-mix matrix of the energy mode using center cancellation is now given by

$$\mathbf{U} = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ u_{21} & u_{22} & u_{23} \end{bmatrix} \tag{62}$$

with

$$u_{11} = \sqrt{\frac{q_1 q_2 + q_2(1+q_2)}{q_1 q_2 + (q_2+1)^2}} \tag{63}$$

$$u_{12} = -\sqrt{\frac{q_2(q_2+1)^2}{(q_1+q_2+1)(q_1 q_2 + (q_2+1)^2)}} \tag{64}$$

$$u_{13} = \frac{1}{2}\sqrt{\frac{q_2+1}{q_1 q_2 + q_2 + 1}} \tag{65}$$

$$u_{21} = -\sqrt{\frac{(q_2+1)^2}{(q_1 q_2 + q_2 + 1)(q_1 q_2 + (q_2+1)^2)}} \tag{66}$$

$$u_{22} = \sqrt{\frac{q_1 q_2 + q_2 + 1}{q_1 q_2 + (q_2+1)^2}} \tag{67}$$

$$u_{23} = \frac{1}{2}\sqrt{\frac{q_2+1}{q_1+q_2+1}}. \tag{68}$$

Because the parameters q_1 and q_2 represent energy ratios, they can be straightforwardly quantized like the interchannel intensity difference (IID) parameters [14] of the parametric stereo coder. For 44 100/2048 updates per second and 28

parameter bands, the estimated bit-rate of each of these parameters amounts to about 1.7 kb/s, when using the MPS parameter coding scheme.

IV. Subjective Evaluation

A. Method and Stimuli

The objective of the listening test is in the first place to investigate the effect of the different modes of the 3-2-3 coder on the perceived audio quality. At the same time, we want to gain insight in the quality loss that is induced by the 3-2-3 coder in the 5.1-2-5.1 MPS coder, of which it is a module. Therefore, the stereo down-mix signal is not coded. Two alternative configurations were evaluated. Configuration (1) is the 3-2-3 coder using the prediction mode. The average parameter bit rate amounts to 5.0 kb/s. Configuration (2) is the 3-2-3 coder using the plain energy mode, with an associated average parameter bit rate of 3.7 kb/s. For both configurations, the standard MPS 5.1-2-5.1 coder configuration was chosen, which includes 28 parameter bands and an update interval of 2048 time samples at a sampling frequency of 44 100 Hz. The two configurations were chosen because they are expected to represent the two extremes as to coder quality. Moreover, the plain energy up-mix can be seen as a representative of a conventional up-mixing procedure, in so far that it does not exploit signal redundancies (i.e., signal predictability). It was an issue how to represent the three channels spatially in the listening test. In order to gain insight in the worst case operation of the 3-2-3 coder, we investigated two (extreme) spatial settings. In the first setting, the left and right channel were played at the loudspeaker position of the left front and right front channel of the standard 5.1 loudspeaker setting, respectively. In the second setting, the surround loudspeakers were used instead of the front loudspeakers. The center channel was played in both cases at the position of the center channel of the standard 5.1 loudspeaker setting. By means of an informal listening experiment, we found the "surround" setting to be the most critical. Therefore, this setting was used in the formal listening experiment.

Eight listeners participated in the experiment. All listeners had significant experience in evaluating audio coders and were specifically instructed to evaluate both the spatial audio quality as well as any other noticeable artifacts. In a double-blind MUSHRA test [29], the listeners had to rate the perceived quality of several processed items against the original (i.e., 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 of 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 item. The experiment was controlled from a PC and audio was played with an RME Digi 96/24 sound card using ADAT digital out. Digital-to-analog conversion was provided by an RME ADI-8 DS 8-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, of which only the center, left surround, and right surround speaker played content, employing B&W Nautilus 800 speakers in a dedicated listening room according to ITU recommendation [30].

TABLE I
Test Items

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

A total of 11 three-channel excerpts were selected that are listed in Table I. These excerpts were based on the 5.1 multichannel excerpts used in the MPEG Call for Proposals (CfP) on spatial audio coding [31]. The left channel was obtained from the 5.1 multichannel signal by summing the left front and left surround channel, where the surround channel was attenuated by $(1)/(2)\sqrt{2}$. Similarly, the right channel was obtained from the right channels of the 5.1 multichannel signal. Finally, the center channel was identical to the center channel of the 5.1 multichannel signal. The items range from pathological signals (designed to be critical items for the technology at hand) to movie sound and multichannel productions. All input and output items were sampled at 44 100 Hz.

B. Results

The subjective listening test results are shown in Fig. 3. The horizontal axis shows the 11 excerpts under test, the vertical axis the mean MUSHRA score averaged across listeners. Moreover, the mean MUSHRA score averaged across listeners and items is shown labeled with "Mean," indicating the mean coder performance. Furthermore, different symbols indicate different configurations, and the error bars denote 95% confidence intervals of the means.

As can be seen, the hidden reference scores are essentially 100 indicating that the results of the listeners are reliable. The 3.5-kHz low-pass filtered anchor received lowest scores between 13 and 21. For the encoded items, the plain energy mode (downward triangles) scores lowest, with about 84 in the mean. The prediction mode (diamonds) scores about 91 in the mean. Because the 95% confidence intervals of the mean scores of the prediction and plain energy mode are not overlapping, the prediction mode performs better than the plain energy mode as to audio quality. Looking at the scores of the individual items, we find them to be consistently high for the prediction mode (MUSHRA score above 87), except for the "BBC applause" item. This is partly explained by the fact that the three input channels of this item are both uncorrelated and

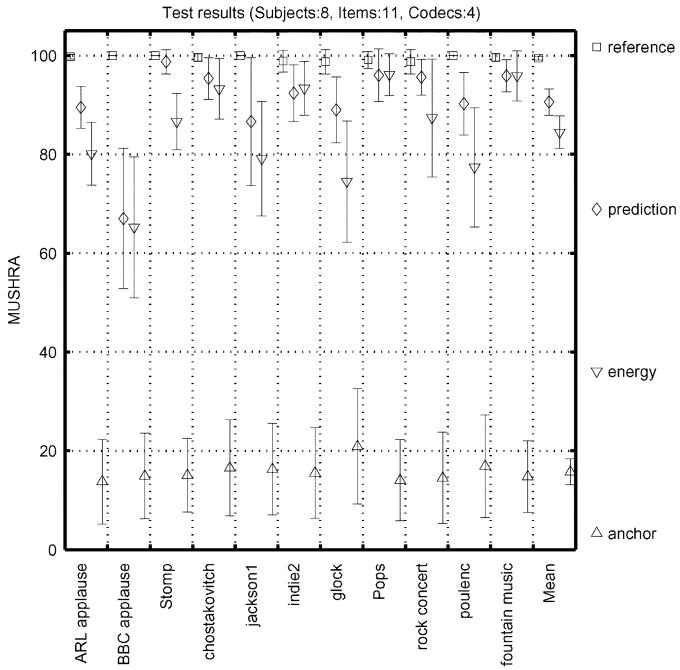

Fig. 3. Subjective listening test results. The mean MUSHRA scores are shown for the 3-2-3 prediction coder (diamonds) and 3-2-3 energy mode (downward triangles). In addition, the 3.5-kHz low-pass filtered anchor (upward triangles) and hidden reference (squares) are shown.

spectrally overlapping, so that a 3 to 2 down-mix operation cannot be undone by the decoder.

We see in Fig. 3 that the 95% confidence intervals of the prediction and plain energy mode are overlapping for all but one of the individual items. Therefore, a pair-wise two-tailed t-test was done to determine whether the differences between the two modes are statistically significant for the individual items. For this purpose, we investigated the difference score of the two modes. For the "ARL applause," "Stomp," "jackson1," "glock," and "poulenc" items, we found the differences to be statistically significant $(p < 0.05)$ in favor of the prediction mode. This is almost half of the items (5 out of 11).

The feedback of the listeners revealed for some items a change of the timbre of the center channel and/or spatial image. The first was most pronounced for the "jackson1" item, the latter for the "poulenc" item.

C. Discussion

We find the audio quality of the prediction mode to be high (MUSHRA scores above 87 for the individual items), except for one applause item (MUSHRA score of 67). Moreover, the prediction mode of the 3-2-3 coder was found have a significant better audio quality than the plain energy mode, at the expense of a slight increase in parameter bit rate (1.3 kb/s). This result indicates the added value of exploiting channel predictability in the up-mix procedure of the 3-2-3 coder. Yet, for both coders, the associated parameter bit rate is low as compared to the bit rate required for coding a stereo signal by a state-of-the-art stereo coder.

The relatively low MUSHRA score of the applause item does not come as a surprise, because this type of signal is known to be problematic in audio coding. We further investigated this applause item in an informal listening test. In this test, we compared the audio quality of the output signals of the 3-2-3 prediction coder to that of the original three multichannel input signals. We also compared the quality of the output signals of the MPS 5.1-2-5.1 coder to that of the original five multichannel input signals. We found the 3-2-3 output signals to be of higher quality, because the timbre of the three multichannel input signals was quite well preserved, whereas the timbre of the five multichannel input signals was significantly changed by the 5.1-2-5.1 MPS coder.

The results of the listening test show that the plain energy mode should not be used when the waveform of the stereo down-mix is preserved by the core coder. However, informal listening experiments demonstrated the benefit of employing the plain energy mode instead of the prediction mode whenever the core coder does not preserve the waveform. When taking the HE-AAC codec, that uses SBR in the high-frequency range, as the core codec, we found the prediction mode to have serious leaking problems, unlike the plain energy mode.

V. Conclusion

We describe in this paper a multichannel audio codec that exploits both signal redundancies (i.e., predictabilities) and perceptual redundancies, while it employs a fixed down-mixing procedure. The latter enables control over the down-mixing procedure, which is necessary for backward compatibility. A subjective listening test reveals a high audio quality and the benefit of making use of signal redundancies for the 3-2-3 system. Moreover, it has a low parameter bit rate (5.0 kb/s) and its design is flexible, examples of which are the scalability of the audio quality to (in principle) transparency , the option to adapt the processing to properties of the codec that is applied to code the stereo down-mix signal and the possibility to preserve the correlation structure of the original input signals by using synthetic signals. The 3-2-3 system of the proposed codec is used as a component of the recently standardized MPEG Surround multichannel audio codec.

Acknowledgment

The authors would like to thank both the reviewers and their colleagues B. den Brinker, E. Sarroukh, and S. van de Par for their useful remarks and suggestions on earlier versions of the manuscript.

References

[1] K. Brandenburg and G. Stoll, "ISO-MPEG-1 audio: A generic standard for coding of high-quality digital audio," *J. Audio Eng. Soc.*, vol. 42, pp. 780–792, 1994.

[2] H. G. Musmann, "Genesis of the MP3 audio coding standard," *IEEE Trans. Consumer Electron.*, vol. 52, no. 3, pp. 1043–1049, Aug. 2006.

[3] K. Brandenburg, "MP3 and AAC explained," in *Proc. 17th Int. AES Conf.*, Florence, Italy, 1999, pp. 99–100.

[4] 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.

[5] R. G. van der Waal and R. N. J. Veldhuis, "Subband coding of stereophonic digital audio signals," in *Proc. ICASSP*, Toronto, QC, Canada, 1991, pp. 3601–3604.

[6] J. D. Johnston and A. J. Ferreira, "Sum-difference stereo transform coding," in *Proc. ICASSP*, San Francisco, CA, 1992, pp. 569–572.

[7] A. C. den Brinker, E. G. P. Schuijers, and A. W. J. Oomen, "Parametric coding for high-quality audio," in *Proc. 112th AES Convention*, Munich, Germany, 2002, preprint 5554.
[8] E. Schuijers, W. Oomen, B. den Brinker, and J. Breebaart, "Advances in parametric coding for high-quality audio," in *Proc. 114th AES Convention*, Amsterdam, The Netherlands, 2003, preprint 5852.
[9] P. Ekstrand, "Bandwidth extension of audio signals by spectral band replication," in *Proc. 1st IEEE Benelux Workshop Model-Based Process. Coding of Audio (MPCA-2002)*, Leuven, Belgium, Nov. 2002, pp. 53–58.
[10] M. Dietz, L. Liljeryd, K. Kjörling, and O. Kunz, "Spectral band replication, a novel approach in audio coding," in *Proc. 112th AES Conv.*, Munich, Germany, 2002, preprint 5553.
[11] O. Kunz, "Enhancing MPEG-4 AAC by spectral band replication," in *Proc. Tech. Sessions Workshop Exhibition MPEG-4 (WEMP4)*, San Jose, CA, 2002, pp. 41–44.
[12] F. Baumgarte and C. Faller, "Why binaural cue coding is better than intensity stereo coding," in *Proc. 112th AES Conv.*, Munich, Germany, 2002, preprint 5575.
[13] F. Baumgarte and C. Faller, "Binaural cue coding—Part I: Psychoacoustic fundamentals and design principles," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 509–519, Nov. 2003.
[14] 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.
[15] E. Schuijers, J. Breebaart, H. Purnhagen, and J. Engdegard, "Low complexity parametric stereo coding," in *Proc. 116th AES Conv.*, Berlin, Germany, 2004, preprint 5852.
[16] H. P. Kramer and M. V. Mathews, "A linear coding for transmitting a set of correlated signals," *IRE Trans. Inf. Theory*, vol. 23, pp. 41–46, Sep. 1956.
[17] D. T. Yang, C. Kyriakakis, and C. C. Jay Kuo, "High-fidelity multichannel audio coding with Karhunen–Loève transform," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 4, pp. 365–380, Jul. 2003.
[18] C. Faller and F. Baumgarte, "Binaural cue coding applied to stereo and multichannel audio compression," in *112th AES Conv.*, Munich, Germany, 2002, preprint 5574.
[19] J. Breebaart and C. Faller, *Spatial Audio Processing: MPEG Surround and Other Applications*. New York: Wiley, 2007.
[20] J. M. Eargle, "Multichannel stereo matrix systems: An overview," *J. Audio Eng. Soc.*, vol. 19, no. 7, pp. 552–559, Jul. 1971.
[21] 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 Conf.*, Pitea, Sweden, 2006, pp. 213–230.
[22] J. Breebaart, G. Hotho, J. Koppens, E. Schuijers, W. Oomen, and S. van de Par, "MPEG Surround: The ISO/MPEG standard for efficient and backward compatible multichannel audio compression," *J. Audio Eng. Soc.*, vol. 55, pp. 331–351, 2007.
[23] C. Faller, "Coding of spatial audio compatible with different playback formats," in *Proc. 117th Conv. Aud. Eng. Soc.*, Oct. 2004, paper 6187.
[24] 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," in *Proc. 119th AES Conv.*, New York, 2005, paper 6599.
[25] *ISO IEC. MPEG Audio Technologies—Part 1: MPEG Surround*, ISO/IEC FDIS 23003-1:2006(E), 2004.
[26] T. W. Lee, *Independent Component Analysis: Theory and Applications.*. New York: Kluwer, 1998.
[27] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, pp. 103–138, 1990.
[28] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.
[29] *Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems (MUSHRA)*, , 2001, ITU-R., ITU-R Rec. BS.1534.
[30] *Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems*, , 1997, ITU-R, ITU-R Rec. BS.1116-1.
[31] *Call for Proposals on Spatial Audio Coding.*, ISO/IEC JTC1/SC29/WG11 N6455, 2004, ISO IEC.

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

He is currently with Philips Research Laboratories, Eindhoven, The Netherlands. As a Researcher, he is very much inspired by the ideas of J. Goethe and R. Steiner. Professionally, he has worked for ten years on digital signal processing topics, 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.

Lars F. Villemoes (M'06) was born in Frederiksberg, Denmark, in 1965. He received the M.Sc. degree in engineering and the Ph.D. degree in mathematics from the Technical University of Denmark, Lyngby. in 1989 and 1992, respectively, and the TeknD. and the Swedish Docent degrees in mathematics from the Royal Institute of Technology, Stockholm, Sweden, in 1995 and 2001, respectively.

From 1995 to 1997, as a Postdoctoral Researcher, he visited the Department of Mathematics, Yale University, New Haven, CT, and the Signal Processing Group, Department of Signals, Systems, and Sensors, Royal Institute of Technology. From 1997 to 2001, he was a Research Associate in wavelet theory in the Department of Mathematics, Royal Institute of Technology. Since 2001, he has been with Coding Technologies, Stockholm, where he is currently Senior Research Advisor. His main research interests include applied harmonic analysis and audio coding.

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

Currently, he is a Researcher in the Digital Signal Processing Group, 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. He has 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.