Audio Engineering Society

Convention Paper 9740

Presented at the 142nd Convention
2017 May 20–23, Berlin, Germany

Parametric Joint Channel Coding of Immersive Audio

Heidi-Maria Lehtonen[1], Heiko Purnhagen[1], Lars Villemoes[1], Janusz Klejsa[1], and Stanislaw Gorlow[1]

[1]*Dolby Sweden AB, Gävlegatan 12 A, 113 30 Stockholm, Sweden*

Correspondence should be addressed to Heiko Purnhagen (heiko.purnhagen@dolby.com)

ABSTRACT

This paper presents a parametric joint channel coding scheme that enables the delivery of channel-based immersive audio content in formats such as 7.1.4, 5.1.4, or 5.1.2 at very low bit rates. It is based on a generalized approach for parametric spatial coding of groups of two, three, or more channels using a single downmix channel together with a compact parametrization that guarantees full covariance re-instatement in the decoder. By arranging the full-band channels of the immersive content into five groups, the content can be conveyed as a 5.1 downmix together with the parameters for each group. This coding scheme is implemented in the A-JCC tool of the AC-4 system recently standardized by ETSI, and listening test results illustrate its performance.

1 Introduction

Immersive audio experiences (3D audio) are a vital element of next-generation audio entertainment systems. Immersive audio content can be represented in different formats, and object-based as well as channel-based representations are widely adopted. While an object-based representation can combine intuitive content creation with optimal reproduction over a large range of playback configurations [1, 2, 3, 4], a channel-based representation of immersive audio can be seen as an evolution of established formats such as 5.1 surround [2, 5]. The focus of this paper is on channel-based immersive audio content and its delivery to consumer entertainment systems in a broadcast or streaming scenario. In such a scenario, transmission bandwidth limitations need to be taken into account, calling for a bitrate-efficient representation of the content.

To achieve this objective, parametric spatial coding techniques are studied. The fundamental idea of these techniques is to convey an N-channel signal by means of a reduced number $M < N$ of downmix signals together with parametric side information that enables the reconstruction of the N-channel signal in the decoder in a perceptually meaningful way [6, 7]. Such a system can be referred to as an N-M-N system, and the most prominent example is known as a Parametric Stereo system (2-1-2), where a 2-channel stereo signal is conveyed by means of a single mono downmix channel and parametric side information. In the decoder, the side information is used to control a time- and frequency-varying upmix process that reconstructs the 2-channel signal. This upmix process includes a decorrelator that enables to re-instate perceptually important cues like ambience or source width [8, 9]. A common approach is to control the upmix process such that it re-instates the time- and frequency-varying covariance matrix of the 2-channel signal that was observed in the encoder,

since this typically results in a perceptually meaningful reconstruction of spatial cues.

This paper discusses the use of parametric spatial coding techniques for channel-based immersive audio content with a large number of channels. It uses the 7.1.4 configuration with a 7.1 setup in the horizontal plane and 4 ceiling speakers as a prominent example, which is described in ITU-R BS.2051 [10] as Sound System G, except that the left and right "screen" channels were omitted here. This configuration comprises 11 full-band channels and a Low Frequency Effects (LFE) channel. In order to apply parametric spatial coding techniques to signals with more than two channels, different approaches can be used. The MPEG Surround system, for example, uses a tree-based parametrization approach that combines several 2-1-2 modules together with a 3-2-3 module in a tree-like structure [11, 12]. It allows to convey 5.1 surround signals using a 2-channel downmix, and supports also content with 7.1 or more channels using additional 2-1-2 modules. While this approach provides a parametrization requiring only a small amount of side information, it cannot ensure complete re-instatement of the covariance matrix. Another approach to handle a large number of channels using parametric spatial coding is employed in the Joint Object Coding system, where typically 11 to 15 channels (object signals) are conveyed using a downmix with 5 or 7 channels [3, 13]. It provides very flexible control of the upmix process including decorrelation, enabling partial or complete covariance re-instatement. This flexibility is advantageous when processing arbitrary object content. However, it requires more side information than less flexible schemes.

In particular at low target bit rates, it is desirable to use only a small amount of side information. Furthermore, also the number M of downmix channels should be chosen carefully, since a higher number of channels means that less bit rate is available per channel, which can result in a reduced quality of the decoded downmix channels that are then processed further by the upmix to reconstruct the N-channel output signal. One approach to convey a 7.1.4-channel signal is to form four groups with two full-band channels each, and use a parametric spatial coding module for each of these groups. This approach (which will be described in more detail in Sec. 4 and Fig. 3) results in a downmix with 7 full-band channels (one from each of the four groups, plus three unprocessed front channels) and the LFE. In this paper, we propose an alternative approach that requires only 5 full-band downmix channels, which can be beneficial at lower target bit rates. To achieve this, two groups with three full-band channels and two groups with two full-band channels are formed.

This paper is structured as follows. First, a generalized parametric spatial coding approach to convey $N \geq 2$ channels using a single downmix channel is introduced and a compact parametrization that enables full covariance re-instatement is described. Then, an Advanced Joint Channel Coding (A-JCC) paradigm is presented that utilizes this approach to process groups of two or three full-band channels and includes a mechanism to dynamically adapt the grouping to the properties of the content being encoded. Finally, experimental results are reported that compare the performance of this new paradigm using 5 downmix channels with the approach using 7 downmix channels for encoding of 7.1.4 content at target bit rates ranging from 128 to 384 kb/s, and conclusions are discussed.

2 Parametric spatial coding using a single downmix channel

This section presents a generalized parametric spatial coding approach to convey a group of $N \geq 2$ channels using a single downmix channel. A perceptually motivated separation of these signals into a set of non-uniform frequency bands together with temporal framing enables to compute and apply the processing steps discussed here in a time- and frequency-variant manner. The intersection of a frequency band and a temporal frame can be referred to as a time-frequency tile and the upmix parameters described below are computed for each tile. A common approach is to form frequency bands by applying a 64-band complex-valued pseudo-QMF analysis bank [14] to each of the signals, and then grouping the QMF bands into a set of typically 7 to 12 parameter bands according to a perceptual frequency scale. Temporal framing commonly uses overlapping analysis windows with a stride of typically 32 to 43 ms an corresponding temporal interpolation in the upmix process.

2.1 Synthesis model

Consider the case in which N audio signals x_n, $n = 1,\ldots,N$ are approximated by linear combinations of

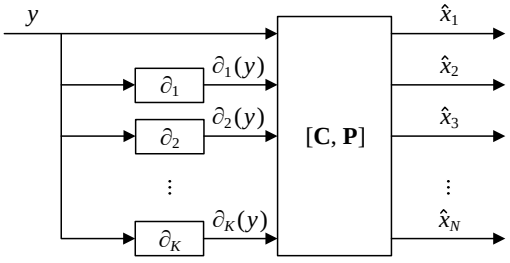

Fig. 1: Illustration of the upmix process from a downmix signal y to N output signals $\hat{x}$.

the downmix signal y and its decorrelated versions $\partial_k(y)$, $k = 1,\ldots,K$:

$$\hat{x} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_N \end{bmatrix} y + \begin{bmatrix} p_{11} & \cdots & p_{1,K} \\ p_{21} & \cdots & p_{2,K} \\ \vdots & \ddots & \vdots \\ p_{N,1} & \cdots & p_{N,K} \end{bmatrix} \begin{bmatrix} \partial_1(y) \\ \vdots \\ \partial_K(y) \end{bmatrix}, \tag{1}$$

where $\hat{x}$ consists of approximations of audio signals x_n and parameters c_n and $p_{n,k}$ are the dry and wet upmix parameters, respectively. Fig. 1 illustrates the block diagram of such a system.

For the sake of the analysis, all the signals y and $\partial_k(y)$ are assumed to be pairwise orthogonal and of equal 2-norm. With signals as row vectors, the matrix notation of (1) is

$$\hat{\mathbf{X}} = \mathbf{C}\mathbf{Y} + \mathbf{P}\partial(\mathbf{Y}) \tag{2}$$

where the dry upmix matrix $\mathbf{C}$ is of size $N \times 1$ and the wet upmix matrix $\mathbf{P}$ is of size $N \times K$. As both are assumed to have real entries we will consider only the real part of covariance matrices in the analysis that follows. Our notation for the sample covariance matrices is

$$\mathbf{R}_{\mathrm{uv}} = \mathrm{Re}(\mathbf{U}\mathbf{V}^*). \tag{3}$$

For instance, $\mathbf{R}_{\mathrm{yy}} = \|y\|^2$, since $\mathbf{Y}$ has a single row. From (2) and the assumptions on the decorrelators, it follows that

$$\mathbf{R}_{\hat{\mathrm{x}}\hat{\mathrm{x}}} = (\mathbf{C}\mathbf{C}^{\mathrm{T}} + \mathbf{P}\mathbf{P}^{\mathrm{T}})\|y\|^2. \tag{4}$$

The goal is now to choose the dry and wet coefficients in $\mathbf{C}$ and $\mathbf{P}$ such that the covariance in the reconstructed signals matches that of the original signals,

$$\mathbf{R}_{\hat{\mathrm{x}}\hat{\mathrm{x}}} = \mathbf{R}_{\mathrm{xx}}. \tag{5}$$

2.2 The cascaded approach

Our approach is to first find a dry upmix $\hat{\mathbf{X}}_0 = \mathbf{C}\mathbf{Y}$ which is optimal for waveform match in the least squares sense, by solving the normal equations

$$\mathbf{C}\|y\|^2 = \mathbf{R}_{\mathrm{xy}}. \tag{6}$$

From (6) it follows that $\mathrm{Re}((\hat{\mathbf{X}}_0 - \mathbf{X})\hat{\mathbf{X}}_0^*) = \mathbf{0}$ and it is easy to show by using this result that

$$\mathbf{R}_{\mathrm{xx}} = \mathbf{C}\mathbf{C}^{\mathrm{T}}\|y\|^2 + \Delta\mathbf{R}, \tag{7}$$

where $\Delta\mathbf{R}$ is the covariance of the approximation error $\hat{\mathbf{X}}_0 - \mathbf{X}$. Combining (4) and (7) shows that (5) holds if

$$\mathbf{P}\mathbf{P}^{\mathrm{T}}\|y\|^2 = \Delta\mathbf{R}. \tag{8}$$

The approach described above is cascaded in the sense that the target covariance is first approximated with the dry upmix, and the missing covariance is then compensated with the wet upmix. If $K = N$ and the downmix does not vanish, then (8) can be solved for a wet matrix $\mathbf{P}$ of size $N \times N$. As we shall see, an additional assumption leads to a more efficient parametrization.

2.3 The downmix model

Assume that the downmix is the sum of all N original audio signals x_n, $n = 1,\ldots,N$. Equivalently, the downmix process can be described by $\mathbf{Y} = \mathbf{D}\mathbf{X}$ with a downmix weight matrix $\mathbf{D} = [1,1,\ldots,1]$. If the column vector $\mathbf{C}$ is obtained by the least squares method, an application of $\mathbf{D}$ to both sides of (6) yields $\mathbf{D}\mathbf{C}\|y\|^2 = \|y\|^2$ so for non-degenerate downmixes we get

$$\mathbf{D}\mathbf{C} = \mathbf{I}, \tag{9}$$

or, equivalently,

$$c_1 + c_2 + \ldots + c_N = 1. \tag{10}$$

For the downmix of the approximation error, we get from (9) that

$$\mathbf{D}(\hat{\mathbf{X}}_0 - \mathbf{X}) = \mathbf{D}\mathbf{C}\mathbf{Y} - \mathbf{Y} = \mathbf{0}. \tag{11}$$

Hence, $\mathbf{D}\Delta\mathbf{R} = 0$ so the missing covariance has rank at most $N-1$ and we can factorize

$$\Delta\mathbf{R} = \mathbf{U}\mathbf{U}^{\mathrm{T}}. \tag{12}$$

where $\mathbf{U}$ is of size $N \times (N-1)$ with $\mathbf{DU} = \mathbf{0}$. By constructing $\mathbf{V}$ of size $N \times (N-1)$ with the space of vectors $\mathbf{v}$ with $\mathbf{Dv} = 0$ as columns, we can write

$$\mathbf{U} = \mathbf{VG}, \tag{13}$$

where $\mathbf{G}$ is of size $(N-1) \times (N-1)$. With the definition $\mathbf{R}_V = \mathbf{G}\mathbf{G}^{\mathrm{T}}$, the missing covariance can be expressed as

$$\mathbf{\Delta R} = \mathbf{V}\mathbf{R}_V\mathbf{V}^{\mathrm{T}}. \tag{14}$$

The condition (8) for covariance match can now be satisfied by putting

$$\mathbf{P} = \mathbf{VH}, \tag{15}$$

and choosing $\mathbf{H}$ of size $(N-1) \times (N-1)$ with

$$\mathbf{H}\mathbf{H}^{\mathrm{T}} = \frac{\mathbf{R}_V}{\|y\|^2}. \tag{16}$$

Thus, (5) can be achieved with $K = N-1$ decorrelators.

In order to get $\mathbf{R}_V$ from $\mathbf{\Delta R}$, one can use any $\mathbf{W}$ with $\mathbf{W}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$, such as the pseudo-inverse $\mathbf{W} = \mathbf{V}(\mathbf{V}^{\mathrm{T}}\mathbf{V})^{-1}$. Then one finds that $\mathbf{R}_V = \mathbf{W}^{\mathrm{T}}(\mathbf{\Delta R})\mathbf{W}$.

Note that (9) and (15) leads to $\mathbf{D}\hat{\mathbf{X}} = \mathbf{Y} = \mathbf{DX}$ in (2). This is a *downmix compatibility* condition, in the sense that the downmix of the upmix is equal to the downmix of the signals. It follows here from the use of a cascaded encoding approach, but it could also be a desirable feature in itself.

2.4 Reduction of parameter dimensionality

We have now seen that a full covariance match can be obtained by using the cascaded approach and that a downmix model makes this possible with $K = N-1$ decorrelators. The total number of parameters in (1) for this case is N^2. However, the number of dry parameters can be reduced from N to $N-1$ due to (10). For the wet parameters, once the choice of $\mathbf{V}$ is settled, there are many ways to parametrize the solutions to (16), such as Cholesky factorization leading to a lower triangular $\mathbf{H}$, positive square root giving a symmetric positive semi-definite $\mathbf{H}$, and polar, which writes $\mathbf{H} = \mathbf{O\Lambda}$ where $\mathbf{O}$ is orthogonal and $\mathbf{\Lambda}$ is diagonal. All of these require $(N-1)N/2$ parameters.

This approach reduces the total number of parameters from N^2 to $(N-1)(N/2+1)$. It should be pointed out that the general analysis of Sec. A.C.1 of [12] also implies that a covariance match (3) can be obtained by adding decorrelation described by a covariance matrix of size $N-1$, but no specific procedure to obtain this is described.

2.5 Examples of compact parametrization

Practical examples of compact parametrization for parametric spatial coding of groups of two, three, and four channels using a single downmix channel are provided below.

2.5.1 Group of $N = 2$ channels

This configuration is related to Parametric Stereo [9] and is used by the Advanced Coupling (A-CPL) tool defined in the AC-4 system [2, 14]. Here $\mathbf{D} = [1, 1]$ and there is only one solution for $\mathbf{V}$ up to scaling $\eta \neq 0$,

$$\mathbf{V} = \eta \begin{bmatrix} -1 \\ 1 \end{bmatrix}. \tag{17}$$

Since $N-1 = 1$, $\mathbf{R}_V$ and $\mathbf{H}$ are scalars and we can solve (16) with $\mathbf{H} = \sqrt{\mathbf{R}_V}/\|y\|$. All in all, the resulting upmix (1) can be parametrized by two parameters α, β with $\beta \geq 0$,

$$\hat{x} = \frac{1}{2}\begin{bmatrix} 1+\alpha \\ 1-\alpha \end{bmatrix} y + \frac{\beta}{2}\begin{bmatrix} -1 \\ 1 \end{bmatrix} \partial(y). \tag{18}$$

2.5.2 Group of $N = 3$ channels

In general, a matrix $\mathbf{V}$ with orthogonal columns is desirable, but in this case it would have coefficients which make it slightly difficult to achieve a zero decorrelator contribution to one of the outputs when the entries in $\mathbf{H}$ are quantized. Instead a better solution is to select

$$\mathbf{V} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 0 & -1 \\ -1 & 0 \end{bmatrix}. \tag{19}$$

Using the positive square root solution to (16), we transmit the elements h_{11}, h_{22}, and h_{12} of

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix}. \tag{20}$$

Since it is known that the matrix $\mathbf{H}$ is symmetric, the lower triangular elements are obtained from the upper triangular elements, i.e., $h_{21} = h_{12}$. For the dry part, the coefficients c_1 and c_2 are transmitted and c_3 is computed from (10).

2.5.3 Group of $N = 4$ channels

Here, the preferred choice is the orthogonal matrix

$$\mathbf{V} = \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \\ -1 & -1 & 1 \\ -1 & 1 & -1 \end{bmatrix}. \qquad (21)$$

In this case, the symmetric positive $\mathbf{H}$ can be described by its six upper triangular values, and the dry coefficients are defined by the first three values.

2.6 Comparison with tree structure

As noted in the Sec. 1, an alternative approach for parametric spatial coding of $N > 2$ channels using a single downmix channel is to employ a tree-based parametrization with several 2-1-2 modules, an approach for example utilized by the MPEG Surround system. For the case of $N = 3$ channels, this means that two 2-1-2 modules are used, with a total of four parameters, two for each module. To ensure full covariance re-instatement, a compact parametrization however requires five parameters, as shown in Sec. 2.5.2. In the general case, the tree-based approach requires $2(N-1)$ parameters, which is less than the $(N-1)(N/2+1)$ parameters required by the compact parametrization. This indicates that the cascaded approach is not able to ensure full covariance re-instatement.

3 Advanced Joint Channel Coding

This section presents a practical system to convey 7.1.4 channel-based immersive content at low target bit rates using only 5 full-band downmix channels. To achieve this, two groups with three full-band channels and two groups with two full-band channels are formed and processed using the approach presented in Sec. 2, while the last full-band channel remains unprocessed. We considered various different ways of grouping the original 11 full-band channels, and found that the two configurations shown in Fig. 2 are particularly interesting. These configurations are referred to as 5.1.0 and 3.1.2, respectively, indicating the format of the resulting 5-channel downmix. Both configurations are symmetric in the sense that they use the same two groups for the 5 channels on the left side as for those on the right side, while the center channel C remains unprocessed. They yield a total of five full-band downmix channels, which, together with the LFE channel, form a 5.1 downmix. In

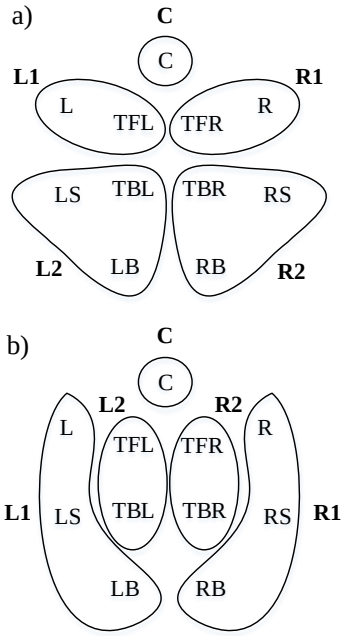

Fig. 2: Illustration of the two different downmix configurations used for 7.1.4 content in A-JCC. Panels a) and b) present the 5.1.0 and 3.1.2 downmixes, respectively. The channels of 7.1.4 content are L (left), R (right), C (center), LS (left surround), RS (right surround), LB (left back), RB (right back), TFL (top front left), TFR (top front right), TBL (top back left), and TBR (top back right), while the LFE (low frequency effects) is not shown here. The downmix channels L1, L2, C, R1, and R2 are denoted by bold labels.

the practial system presented here, a perceptual audio coding algorithm is used to convey the 5.1 downmix at a low bit rate, while the upmix parameters are included as side information in the bitstream. This paradigm is referred to as Advanced Joint Channel Coding (A-JCC), and was recently standardized by ETSI as part of the AC-4 system [15].

3.1 Content-adaptive downmix configuration

Initial experiments comparing the performance of the system for the two downmix configurations indicated a strong dependency on the properties of the immersive

7.1.4 content being encoded. For example for sections of the content where there is little activity in the ceiling channels, the 5.1.0 downmix configuration was often advantageous, while for sections of the content where sounds in the horizontal plane were very different from sounds in the ceiling channels, the 3.1.2 downmix configuration could provide a perceptually preferred 7.1.4 reconstruction.

To accommodate these observations, the selection of the downmix configuration was made content-adaptive, resulting in a scheme we also refer to as dynamic downmix. A simple approach to select the preferred downmix configuration for a given short temporal section (frame) of the content first computes the dry and wet upmix coefficients $\mathbf{C}$ and $\mathbf{P}$ for both downmix configurations. Then relative amount E of wet contributions to the total resulting upmix is computed as

$$E = \frac{E_{\mathbf{P}}}{E_{\mathbf{C}} + E_{\mathbf{P}}}, \tag{22}$$

where $E_{\mathbf{C}}$ and $E_{\mathbf{P}}$ denote the sum over all squared coefficients c_n^2 and $p_{n,k}^2$, respectively, when summed over all n, all k, all four downmix groups, and all frequency bands in the current frame. The downmix configuration that gives the lowest value of E is selected. The motivation behind this approach is to minimize the amount of wet (i.e., decorrelation-based) contributions to the upmix generated by the decoder, that is, to select the downmix that allows a reconstruction that is closer to the original signal when only the dry contributions are considered.

In order to avoid rapid switching from one downmix configuration to the other, the downmix decisions can be constrained to exhibit a certain degree of temporal continuity. In a practical scheme used for the experiments reported below, the transition from one downmix configuration to the other is only allowed if the new downmix configuration was selected for a certain number of consecutive frames. In order to align the downmix transition with respect to the audio content, such a scheme can require a corresponding amount of look-ahead on the encoder side.

3.2 Parameter quantization and coding

As indicated in Sec. 2, the upmix parameters are computed for each time-frequency tile using an appropriate time and frequency resolution. To convey these parameters as side information, they need to be quantized and coded. For groups comprising two channels, there are two parameters per time-frequency tile, the dry parameter α and the wet parameter β, as described in Sec. 2.5.1. These parameters are also used by A-CPL, and a perceptually motivated non-uniform quantization scheme is used, as described in [2]. For groups comprising three channels, there are a total of five parameters, two dry (c_1, c_2) and three wet (h_{11}, h_{12}, h_{22}), as described in Sec. 2.5.2. For these parameters, uniform quantization is used. After quantization of all parameters, time- or frequency-differential coding is applied, followed by Huffman coding.

3.3 Full decoding of 7.1.4

The decoder reconstructs upmix matrices $\mathbf{C}$ and $\mathbf{P}$ from the compact parametrization conveyed in the bitstream. As indicated in Sec. 2, temporal interpolation of the upmix matrix elements is used to ensure smooth transitions between frames. In order to enable smooth transitions between the two different downmix configurations, the upmix matrices for the different channel groups corresponding to the different downmix channels are used to construct two large but sparse upmix matrices $\mathbf{C}_{\text{full}}$ and $\mathbf{P}_{\text{full}}$ of size 11×5 and 11×6, respectively, that process all full-band channels simultaneously and utilize a total of 6 decorrelators. Considering the 5 channels on one side (e.g., left) in Fig. 2, a total of three decorrelators are required for the wet upmix contributions. Each decorrelator comprises an initial delay followed by an IIR all-pass filter and a "ducker" module that improves performance for transient signals [14]. Different IIR filter coefficients are used in the three decorrelators to ensure mutual decorrelation as assumed in Sec. 2.1. The assignment of decorrelators to downmix channels depends on the downmix configuration. While two of the decorrelators on each side can always be fed with the first and second downmix signal on that side, respectively, the third decorrelator is either fed by the first or the second downmix signal, depending on the downmix configuration. To ensure smooth transitions between downmix configurations, a cross-fade is applied to these decorrelator feeds.

3.4 Efficient core decoding to 5.1.2

In the case where the playback system has fewer channels than the original immersive content, it is possible to reduce the computational complexity of the decoder. This process is called core decoding, and is achieved

by a modified upmix process that requires a smaller number of decorrelators than what would be needed for the full decoding of all channels. Furthermore, the modified upmix process does not require any subsequent downmix and directly generates the signals for the available playback channel configuration. As specified in [15], the output configuration from A-JCC core decoding is 5.1.2, which is also illustrated in Fig. 3. Note that this configuration is the same as the downmix configuration of A-CPL, as explained in Sec. 4.

Let us take the downmix channel L1 from the 3.1.2 downmix shown in Fig. 2 b) as an example: L1 = L + LS + LB, where L, LS, and LB are the original channels. According to (2), full decoding would reconstruct the signals $\widehat{\mathrm{L}}$, $\widehat{\mathrm{LS}}$, and $\widehat{\mathrm{LB}}$ using two decorrelators as

$$\begin{bmatrix} \widehat{\mathrm{L}} \\ \widehat{\mathrm{LS}} \\ \widehat{\mathrm{LB}} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} \mathrm{L1} + \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \\ p_{31} & p_{32} \end{bmatrix} \begin{bmatrix} \partial_1(\mathrm{L1}) \\ \partial_2(\mathrm{L1}) \end{bmatrix}. \quad (23)$$

According to (10) $c_1 + c_2 + c_3 = 1$, and given the construction of $\mathbf{V}$ and (15), the columns of $\mathbf{P}$ sum up to zero, i.e., $p_{1,k} + p_{2,k} + p_{3,k} = 0$ for $k = 1, 2$. Thus the signals $\widehat{\mathrm{L}}$ and $\widehat{\mathrm{LS}} + \widehat{\mathrm{LB}}$ can be reconstructed as

$$\begin{bmatrix} \widehat{\mathrm{L}} \\ \widehat{\mathrm{LS}} + \widehat{\mathrm{LB}} \end{bmatrix} = \begin{bmatrix} c_1 \\ 1 - c_1 \end{bmatrix} \mathrm{L1} + \begin{bmatrix} p_{11} & p_{12} \\ -p_{11} & -p_{12} \end{bmatrix} \begin{bmatrix} \partial_1(\mathrm{L1}) \\ \partial_2(\mathrm{L1}) \end{bmatrix}. \quad (24)$$

Furthermore, by expressing $p_1^2 = p_{11}^2 + p_{12}^2$ and replacing the two decorrelators ∂_1 and ∂_2 by a single decorrelator ∂_1, (24) can be approximated as

$$\begin{bmatrix} \widetilde{\mathrm{L}} \\ \widetilde{\mathrm{LS}} + \widetilde{\mathrm{LB}} \end{bmatrix} = \begin{bmatrix} c_1 \\ 1 - c_1 \end{bmatrix} \mathrm{L1} + \begin{bmatrix} p_1 \\ -p_1 \end{bmatrix} \partial_1(\mathrm{L1}). \quad (25)$$

Note that the two reconstructed channels $\widetilde{\mathrm{L}}$ and $\widetilde{\mathrm{LS}} + \widetilde{\mathrm{LB}}$ are not exactly the same as $\widehat{\mathrm{L}}$ and $\widehat{\mathrm{LS}} + \widehat{\mathrm{LB}}$ in (24), since only one decorrelator is used. Nonetheless, the covariance of the two channels is fully re-instated.

4 Experimental results

This section compares the performance of different approaches to convey 7.1.4 content using parametric spatial coding techniques using either 7 or 5 full-band downmix channels for a range of different bit rates. The approach using 5 full-band downmix channels, referred to as Advanced Joint Channel Coding (A-JCC), was described in detail in Sec. 3. The approach using 7 full-band downmix channels, referred to as Advanced Coupling (A-CPL), uses only four groups of two channels each and was already outlined in Sec. 1. The specific grouping used by A-CPL is shown in Fig. 3, which results in a static (i.e., not content-adaptive) 5.1.2 downmix. Like A-JCC, also the A-CPL approach was recently standardized by ETSI as part of the AC-4 system [15].

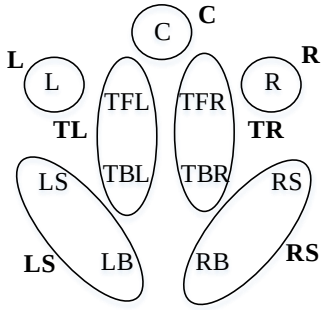

Fig. 3: Illustration of the 5.1.2 downmix configuration used for 7.1.4 content in A-CPL. The 7 full-band channels of this 5.1.2 downmix are denoted by bold labels and correspond also to the 5.1.2 output from A-JCC core decoding process. The LFE is not shown here.

In can be noted that both A-JCC and A-CPL also support encoding of content with fewer than 7.1.4 channels, such as 7.1.2, 5.1.4, and 5.1.2. This is achieved by simply mapping such content to 7.1.4, leaving the remaining two or four channels silent, and signal the chosen mapping in the bit stream.

4.1 Test setup

To assess the rate-distortion performance of A-JCC and compare it to the performance of A-CPL, a formal listening test according to the MUSHRA methodology [16] was conducted. The two systems under test, A-JCC and A-CPL, were operated at three different bit rates, namely 128 kb/s, 256 kb/s, and 384 kb/s. Each of

the systems was individually tuned for each operation point, to take the different values for the average bit rate available per downmix channel into account. This tuning includes the choice of an appropriate cross-over frequency between waveform coding and parametric bandwidth extension (Advanced Spectral Extension (A-SPX), see [2, 14]) for the encoded downmix channels. As required by the MUSHRA methodology, also a hidden reference and two low-pass anchors (3.5 kHz and 7 kHz bandwidth) were included. A total of 12 critical test items with channel-based immersive content in 7.1.4 format were used. The items were obtained by rendering object-based immersive content (in Dolby Atmos format), and are described in Tab. 1. The typical duration of each of the test items was 10 s.

4.2 Listening test results

The listening test results for 10 expert subjects that passed pre- and post-screening are shown in Fig. 4, indicating the mean scores and 95% confidence intervals for each of the 12 items, and when pooled over all items. At 128 kb/s, an analysis of the MUSHRA score differences between A-JCC and A-CPL for all items and subjects shows that A-JCC performs significantly better than A-CPL. Also at 256 kb/s, the mean score for A-JCC is better than for A-CPL, although this difference is not statistically significant. At 384 kb/s, however, the mean score for A-JCC is worse than for A-CPL, but also this difference is not statistically significant.

While the average bit rate required for the side information conveying all the upmix parameter for A-JCC and A-CPL is almost the same (7.4 kb/s and 7.6 kb/s, respectively, for the configuration used in this test), A-JCC operates with only 5 downmix channels compared to 7 channels for A-CPL. Considering also the bit rate needed for the LFE and bitstream framing, this means that at a target bit rate of 128 kb/s, there are in average approximately 24 kb/s and 17 kb/s available to encode each of the downmix channels for A-JCC and A-CPL, respectively. The test results show that at this low target bit rate, it is clearly advantageous to use more extensive parametric spatial coding (i.e., A-JCC instead of A-CPL), since the overall performance benefits from better quality of the decoded downmix channels possible when only fewer downmix channels need to be conveyed. Towards higher target bit rates, however, the performance of A-JCC saturates earlier than that of A-CPL, which can be clearly seen from the quality-rate curves for both A-JCC and A-CPL in Fig. 5, showing the mean score over all items as a function of the bit rate.

From the point of view of the AC-4 system, the performance of the upper hull of the two rate-quality curves in Fig. 5 is achieved by using A-JCC at lower rates, and switching to A-CPL for higher rates.

5 Conclusions

This paper presented a joint channel coding paradigm that enables the delivery of channel-based immersive audio content at low bit rates. This paradigm is used by the A-JCC tool in the recently standardized AC-4 system [15]. By enabling the joint parametric spatial coding of groups of more than two channels over a single downmix channel, it allows to reduce the number required downmix channels to convey 7.1.4 content compared to an alternative approach (A-CPL) that only uses groups of two channels. Listening test results show that this approach results in a significantly increased coding efficiency at a low target bit rate of 128 kb/s and is also beneficial at higher target bit rates like 256 kb/s.

References

[1] Riedmiller, J., Mehta, S., Tsingos, N., and Boon, P., "Immersive and Personalized Audio: A Practical System for Enabling Interchange, Distribution, and Delivery of Next-Generation Audio Experiences," *Motion Imaging Journal, SMPTE*, 124(5), pp. 1–23, 2015, ISSN 1545-0279, doi: 10.5594/j18578.

[2] Kjörling, K., Rödén, J., Wolters, M., Riedmiller, J., Biswas, A., Ekstrand, P., Gröschel, A., Hedelin, P., Hirvonen, T., Hörich, H., Klejsa, J., Koppens, J., Krauss, K., Lehtonen, H.-M., Linzmeier, K., Muesch, H., Mundt, H., Norcross, S., Popp, J., Purnhagen, H., Samuelsson, J., Schug, M., Sehlström, L., Thesing, R., Villemoes, L., and Vinton, M., "AC-4 — The Next Generation Audio Codec," in *Audio Engineering Society Convention 140*, 2016.

[3] Purnhagen, H., Hirvonen, T., Villemoes, L., Samuelsson, J., and Klejsa, J., "Immersive Audio Delivery Using Joint Object Coding," in *Audio Engineering Society Convention 140*, 2016.

Item #	Description
1	Forest ambience with numerous flapping wings sound effects.
2	Live concert with harmonica and applauding audience.
3	Ambient music and sound of ocean waves rolling over.
4	Fixed and panned clock chimes, mechanical sounds, gears, and bells with strong transients.
5	Panned creature dialog with strong cave reverberation. Subtle running water sounds.
6	Electronic music with panned percussive elements, cheering crowd, and applause ambience.
7	Orchestra and immersive sound effects.
8	Orchestra, rain sounds, and immersive sound effects.
9	Strong thunderclap and beginning rainfall.
10	Music with panned percussive elements and strong bass.
11	Rainfall with thunder rumble, wind noise, and music.
12	Intense immersive sound effects.

Table 1: Description of the 12 critical test items in the MUSHRA listening test.

[4] Dolby Laboratories, "Dolby Atmos," 2017, available: `http://www.dolby.com/us/en/brands/dolby-atmos.html`.

[5] Riedmiller, J. et al., "Delivering Scalable Audio Experiences using AC-4," *IEEE Transactions on Broadcasting*, 63(1), pp. 179–201, 2017.

[6] Faller, C. and Baumgarte, F., "Binaural cue coding: A novel and efficient representation of spatial audio," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 2002.

[7] Breebaart, J., Disch, S., Faller, C., Herre, J., Hilpert, J., Kjörling, K., Myburg, F., Purnhagen, H., and Schuijers, E., "The Reference Model Architecture for MPEG Spatial Audio Coding," in *Audio Engineering Society Convention 118*, 2005.

[8] Purnhagen, H., Engdegård, J., Rödén, J., and Liljeryd, L., "Synthetic Ambience in Parametric Stereo Coding," in *Audio Engineering Society Convention 116*, 2004.

[9] Purnhagen, H., "Low Complexity Parametric Stereo Coding in MPEG-4," in *Proc. Digital Audio Effects Workshop (DAFX)*, 2004.

[10] "Advanced sound system for programme production," Recommendation ITU-R BS.2051-0, 2014.

[11] Herre, J., Kjörling, K., Breebaart, J., Faller, C., Disch, S., Purnhagen, H., Koppens, J., Hilpert, J., Rödén, J., Oomen, W., Linzmeier, K., and Chong, K. S., "MPEG Surround — The ISO/MPEG Standard for Efficient and Compatible Multichannel Audio Coding," *J. Audio Eng. Soc*, 56(11), pp. 932–955, 2008.

[12] Hotho, G., Villemoes, L., and Breebaart, J., "A Backward-Compatible Multichannel Audio Codec," *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(1), pp. 83–93, 2008.

[13] Villemoes, L., Hirvonen, T., and Purnhagen, H., "Decorrelation for Audio Object Coding," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 2017.

[14] "Digital Audio Compression (AC-4) Standard; Part 1: Channel based coding," ETSI TS 103 190-1 V1.2.1, 2015.

[15] "Digital Audio Compression (AC-4) Standard; Part 2: Immersive and personalized audio," ETSI TS 103 190-2 V1.1.1, 2015.

[16] "Method for the subjective assessment of intermediate quality levels of coding systems," Recommendation ITU-R BS.1534-3, 2015.

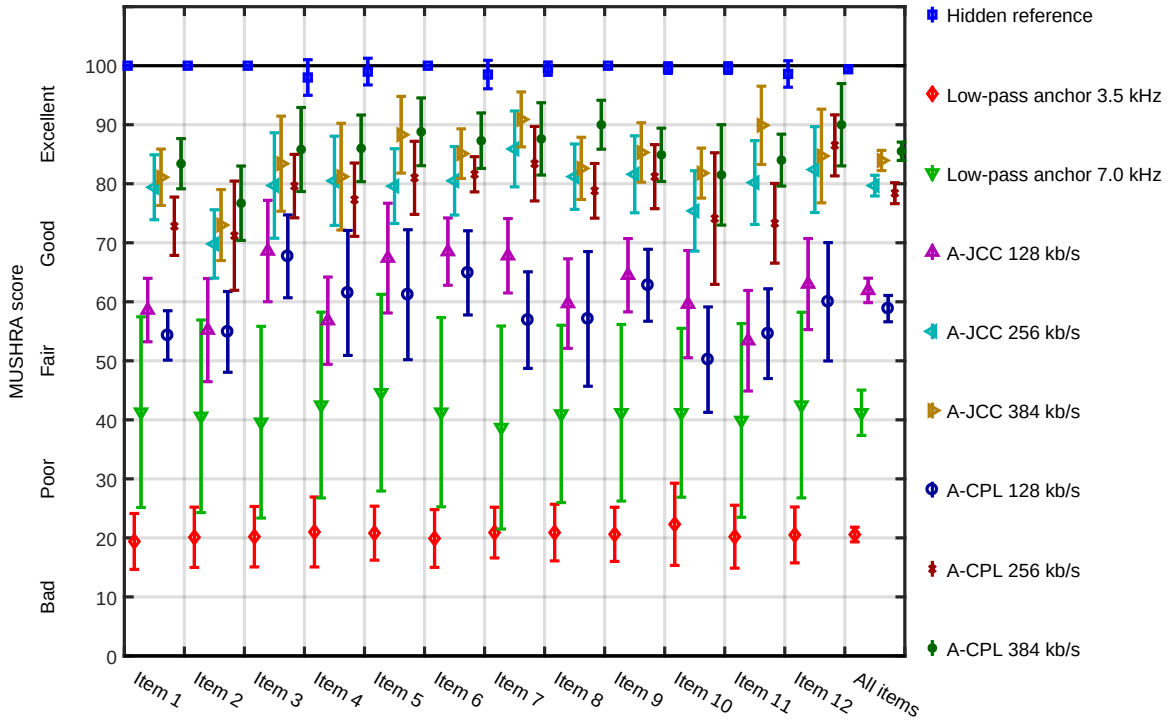

Fig. 4: MUSHRA listening test results (mean scores with 95% confidence intervals) for 10 expert listeners after post-screening for A-JCC and A-CPL at bit rates of 128, 256, and 384 kb/s encoding 12 critical test items of 7.1.4 channel-based immersive content.

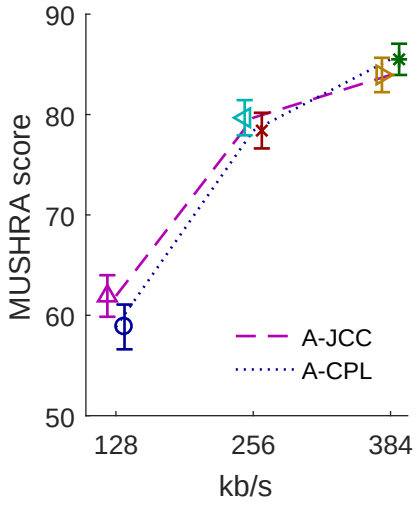

Fig. 5: Quality-rate curves for A-JCC and A-CPL derived from Fig. 4.