Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2008, Article ID 531693, 9 pages
doi:10.1155/2008/531693

Research Article

Multichannel Coding of Applause Signals

Gerard Hotho, Steven van de Par, and Jeroen Breebaart

Digital Signal Processing Group, Philips Research, High Tech Campus 36, 5656 AE Eindhoven, The Netherlands

Correspondence should be addressed to Steven van de Par, steven.van.de.par@philips.com

Received 21 December 2006; Revised 23 May 2007; Accepted 26 July 2007

Recommended by Antonio Ortega

We develop a parametric multichannel audio codec dedicated to coding signals consisting of a dense series of transient-type events. These signals of which applause is a typical example are known to be problematic for such audio codecs. The codec design is based on preservation of both timbre and transient-type event density. It combines a very low complexity and a low parameter bit rate (0.2 kbps). In a formal listening test, we compared the proposed codec to the recently standardised MPEG Surround multichannel codec, with an associated parameter bit rate of 9 kbps. We found the new codec to have a significantly higher audio quality than the MPEG Surround codec for the two multichannel applause signals under test. Though this seems promising, the technique presented is not fully mature, for example, because issues related to integration of the proposed codec in the MPEG Surround codec were not addressed.

1. INTRODUCTION

Audio compression algorithms for wideband 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]. In subsequent years, the MPEG-2 and MPEG-4 standards were developed. They achieve a higher compression efficiency and include multichannel audio support. At the same time, content creators shifted from stereo to multichannel audio formats such as those available on SACD and DVD video, to increase the realism and improve the consumer experience. However, despite the popularity increase of multichannel audio, most broadcast services still operate in traditional stereo. This is mainly due to bandwidth and compatibility constraints. In conventional audio transmission systems, the required bandwidth grows approximately linearly with the number of audio channels. As such, a 5.1-channel audio broadcast requires almost three times as much bandwidth as a conventional stereo broadcast. In many cases, this increased bandwidth is undesirable and sometimes even not allowed.

Even if this increased bandwidth was available, the large installed base of stereo-only receivers poses another challenge to any attempt to upgrade a stereo service to a multichannel service. Backward compatibility with existing equipment is a prerequisite for market acceptance of an upgrade from stereo to multichannel in a broadcast environment.

Recently, the so-called *spatial* audio coders have been introduced, which solve the problem of bandwidth constraints and backward compatibility for digital broadcasting services [2–8]. Whereas conventional audio coders have very limited possibilities to exploit perceptual irrelevancy and signal redundancy between the various channels, spatial audio coders generate a down mix from multichannel content that can be encoded and transmitted using an existing mono or stereo service. The topology of such a coding scheme is illustrated in Figure 1, where the spatial audio encoder receives an N-channel input signal that is transformed into an M-channel down mix (with $M < N$). The degraded spatial impression resulting from the down-mix process is compensated for by a small amount of side information that captures the perceptually relevant aspects of the original multichannel content. These so-called "spatial parameters" are stored in the ancillary data part of, for example, a legacy mono ($M = 1$) or stereo ($M = 2$) coder. (Alternatively, the spatial parameters can be added to the bit stream as an extra layer. However, in this way backward compatibility is lost because the resulting bit stream cannot be decoded by a legacy decoder.) Backward compatability is ensured because a user that has access to a legacy decoder only can still listen to the M-channel down

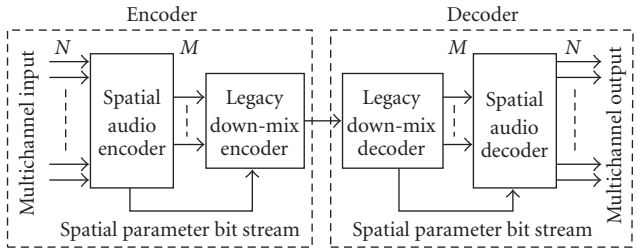

Figure 1: Encoder and decoder configurations for a spatial audio coder. An N-channel input signal is down-mixed to an M-channel signal that is encoded by a legacy encoder. Spatial parameters are embedded in the ancillary data part.

mix. Since the spatial parameters require a very low-bit rate, the reduction of N input channels to M down-mix channels results in a reduction of the bit rate. At the decoder side, a dedicated spatial audio decoder interprets the spatial parameter bit stream and up-mixes the down mix into an N channel representation by reinstating the appropriate perceptually relevant aspects.

Spatial audio coders effectively exploit known limitations of the human hearing system with respect to sound localisation and sound source separation abilities. It is well known that the human auditory system bases its estimates of sound source location on two interaural "cues": interaural level differences (ILDs) and interaural time differences (ITDs) (see, e.g., [9]). The perception of "spaciousness" or sound source "compactness" is closely related to the interaural coherence (IC) [10]. These cues are exactly the properties that spatial audio coders analyse at the encoder side and reinstate at the decoder side. Time and intentensity differences between channels are estimated, encoded, and reinstated at the decoder side, per frequency band in line with human spectral resolution. In a similar way, across-channel coherence is estimated, encoded, and reinstated at the decoder side by mixing a so-called decorrelated signal to the output signals. This decorrelated signal is a similarly sounding filtered version of the decoded down-mix signal.

Although the above-mentioned principles of spatial audio coding are powerful for a wide range of signal types, one signal type has been shown to be problematic: it is a signal consisting of a series of transient-type events that both occur at a rate faster than the frame update rate and are more or less randomly distributed in time and space. Examples are applause, rainfall, and crackling sounds. In this paper, it is explained that a frequency specific processing of these transient-type signals leads to perceptual degradation, and a coder is presented that operates strictly in the time domain. The coder is evaluated in a formal listening experiment.

2. PROBLEMS AND SOLUTIONS

A particular example of the signal type that we focus on in this paper is a multichannel *applause* signal. The transient-type events are the hand clap sounds that create the applause signal. Due to the specific nature of this signal type, the following complications arise when it is coded with a standard spatial audio coder.

(I) The first complication is that in the encoder the interchannel level difference parameters are measured *per frequency band*. In the decoder, these interchannel level differences are recreated by properly modifying copies of the down-mix signal. Since the level difference parameters typically vary across frequency, the amplitude spectrum of the down-mix signal is modified, causing temporal smearing of the transient-type clapping events.

(II) A second complication results from the so-called *decorrelators* [5, 11, 12], which are used to reinstate interchannel coherence. These decorrelators consist of the combination of delays and all-pass filters. However, the employed all-pass filters have a highly nonlinear phase characteristic. This causes the clapping events to be smeared in time leading to a clearly noticeable change of the timbre.

(III) The third complication arises due to the *down-mixing operation* that is performed in the encoder. In this down-mixing operation, different input channels are summed. Because the clapping events of the individual input channels are independent, this summing operation will lead to an increased clap density of the down-mix signals. For the same reason, a summing of signals in the *up-mixing procedure* leads to an increased clap density.

(IV) A fourth complication arises due to the *limited update rate* of the spatial parameters across time in the coder framework. Although changes in spatial parameters can only be sampled by the auditory system on a relatively coarse time scale, the global spatial percept depends to some extent on the rate of change of these spatial parameters [13]. In an applause signal, the rate of change of the spatial parameters is determined by the clap density. Thus in order to faithfully represent this dynamically changing spatial pattern, each hand clap would need to be labelled with one set of spatial parameters. However, this is difficult to implement in practice and would lead to a too high parameter bit rate.

To tackle the above-mentioned problems, we proceed as follows. Both in the encoder and the decoder, the applause signal is treated without applying any spectral filtering to it to ensure that the temporal transient structure is not affected in any negative way (Problems (I) and (II)). In order to avoid an increase in clap density a new down- and up-mix method is employed. Each down-mix signal consists of a weighted sum of a *limited* selection of the original input signals to avoid that at this level the clap density increases too much. This solution also holds for the up-mixing procedure (Problem (III)). In order to create decorrelated signals at the decoder, short portions of the down-mix signals are redistributed in random temporal order (Problem (II)). Different redistributions enable different decorrelated signals. In this way, it is possible to ensure that the different output signals are mutually uncorrelated. Therefore, each clap in one channel will

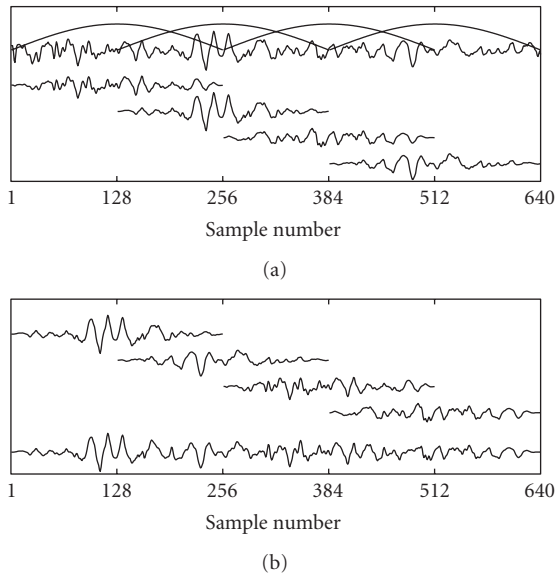

FIGURE 2: Decomposition of a segment into subsegments (a) and composition of the reordered subsegments into the decorrelated signal (b).

be independent of events in other channels. In this way, the problem associated with a limited update rate of the spatial parameters is avoided (Problem (IV)).

3. A DECORRELATOR FOR APPLAUSE SIGNALS

In a spatial audio decoder, the number of down-mix signals has to be extended to the number of channels of the original multichannel input signal. An important element in this channel extension process is the so-called *decorrelator* [5]. A decorrelator generates an output (or decorrelated) signal that has a similar timbre as the decorrelator input signal, but is uncorrelated with it. Typical decorrelation schemes consist of the combination of delays and all-pass filters [5]. For most signal types, this leads to desired decorrelated signals. However, when applying applause signals to these decorrelators, the timbre of the applause signals is significantly altered due to temporal smearing of the transient-type events. Therefore, we introduce in this section a decorrelator that preserves the timbre of applause signals. In Section 3.1, we present the new decorrelator. Comments with respect to using this new decorrelator in a multichannel coder are given in Section 3.2.

3.1. Decorrelator description

We start by giving the implementation of the new decorrelator. The time domain input signal is segmented into segments of length K. Such a segment is divided into L subsegments that are 50% overlapping. Each subsegment is windowed with a square-root Hanning (or sine) window. This process is depicted in Figure 2(a) for $K = 640$ and $L = 4$. In the top part of Figure 2(a), the data segment is shown along with the four square-root Hanning windows. In the lower part of Figure 2(a), the data of the windowed subsegments is shown. In the next step, the order of the subsegments within the segment is changed. Finally, the data of the subsegments is merged using overlap add. This process is shown in Figure 2(b). The subsegment window is chosen such that the decorrelated signal has the same energy as the input signal, using the assumption that signals in consecutive subsegments after reordering are uncorrelated. Therefore, it should be prevented that two consecutive subsegments before reordering are consecutive subsegments after reordering, because in that case *plain* Hanning windows preserve the signal energy.

A decorrelated signal should fulfill two requirements. Firstly, its timbre should match, as closely as possible, that of the original signal. Secondly, when playing the original and the decorrelated applause signal on different channels of headphones, the spatial image of the resulting stereo signal should sound as wide as that for two independent applause signals. Having two variables to tune, the subsegment length, L, and the segment length, K, we found a properly wide stereo image and an only marginally altered timbre at a sampling frequency of 44 100 Hz for $L = 16$ and $K = 2048 + 2048/L$ (hence using square-root Hanning windows of length 256 with 50% overlap).

3.2. Using the decorrelators in a multichannel coder

A multichannel audio decoder typically needs several independent decorrelators. It is possible for our system to make several decorrelators by appropriately choosing different reordering operations. However, the number of *independent* decorrelators is limited for the fixed choice of K and L resulting in high-quality decorrelated signals, because mutually independent reordering operations have to be selected. This limitation can be overcome by adding different delays to the different decorrelated signals. Another advantage of adding a delay is that shifts of events forward in time, occurring due to the reordering operation, can be avoided. The drawback of a larger delay is that more memory is required.

The drawback of the described decorrelation procedure is that applying it to nonapplause-like input signals can lead to severe artefacts due to the reordering operation. Therefore, the decorrelator should be applied to applause signals only.

The complexity of the dedicated applause decorrelator is low due to the fact that neither a frequency domain transformation nor filtering is applied.

4. CODER STRUCTURE

The coder structure will be explained in two phases. First, in Section 4.1, we give an overview of the *generic* applause coder structure. In Section 4.2, we present the structure of the more specific 5-2-5 multichannel applause coder. To conclude, in Section 4.3, we shortly discuss aspects related to integrating the applause coder into a spatial audio coder.

The design of the applause coder is based on the following assumptions. Firstly, the low-frequency effects (LFEs) channel of the multichannel applause signal is not used. Secondly, the encoder down-mix output signals are identical to

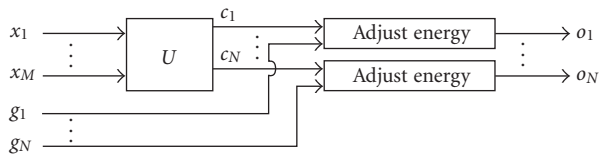

Figure 3: Generic decoder structure of the applause coder.

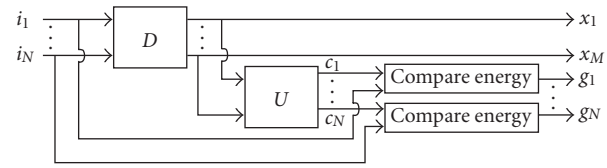

Figure 4: Generic encoder structure of the applause coder.

the decoder input signals, that is, the down-mix signals are not coded with a (legacy) audio coder. Thirdly, the multichannel input signal contains only applause signals that are mutually (highly) uncorrelated, which is a valid assumption for most applause signals. However, if the applause signals were correlated, it would be possible to compute correlation parameters at the encoder. At the decoder, we would generate uncorrelated output signals that are subsequently mixed using these correlation parameters to obtain their desired mutual correlation. A drawback of this mixing operation is that it leads to an increase of the clap density as mentioned in Problem (III) in Section 2.

4.1. Generic coder structure

In this subsection, the generic structure of the spatial audio encoder and decoder of Figure 1 is presented. The structure is generic in the sense that it holds for all positive integer values of M and N, $M < N$. The generic encoder, given in Figure 4, down-mixes N input signals i_n to M down-mix signals x_m in unit D and extracts spatial parameters g_n. In order to be able to generate decoder output signals with the same energy as the encoder input signals, these parameters are computed by first up mixing the down-mix signals to N up-mix signals c_n in unit U, and then comparing the energy of these up-mix signals to the energy of the original input signals. The generic decoder, shown in Figure 3, up mixes the M down-mix signals to N up-mix signals, where decorrelators are employed to enable the increase in the number of signals. Finally, the energy of the up-mix signals is matched to that of the original multichannel input signals using the encoder parameters. This process is explained in more detail in the next sections, where we start with the description of the decoder, because the encoder structure is based on the analysis-by-synthesis principle.

4.1.1. Decoder

The generic structure of the N-M-N applause decoder is shown in Figure 3. In the expression N-M-N, the first N refers to the number of input channels, M refers to the number of down-mix channels and the second N refers to the number of output channels. Let x_m, with $1 \le m \le M$, denote the discrete time domain waveform of the mth down-mix channel. These down-mix channels are segmented in segments of K samples (segmentation not shown in the figure). The qth segment (or frame) $\mathbf{x}_{m,q}$, with $1 \le m \le M$, is a $K \times 1$ vector containing the time samples $x_{m,q}[(q-1)V+1], x_{m,q}[(q-1)V+2], \ldots, x_{m,q}[(q-1)V+K]$, where V denotes the coder update interval. The frame index q is dropped henceforth for ease of notation. The M down-mix segments are up mixed in the unit labelled U resulting in N up-mix segments $\mathbf{c}_n$, with $1 \le n \le N$, given by

$$\mathbf{C} = \mathbf{X}_d \mathbf{U}, \tag{1}$$

where $\mathbf{C}$ is a $K \times N$ matrix containing the N up-mix segments, $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_N]$, and $\mathbf{X}_d$ is a $K \times (M+W)$ matrix containing the M down-mix segments and W decorrelated signals, $\mathbf{X}_d = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M, \mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_W]$, where $\mathbf{d}_w$ denotes the wth decorrelated signal. Furthermore, $\mathbf{U}$ is the $(M+W) \times N$ up-mix matrix that is fixed and a priori known.

The up-mix segments are then fed to the units "Adjust energy" together with the coder parameters g_n, with $1 \le n \le N$. Applying these parameters to the up-mix segments results in N output segments $\mathbf{o}_n$, with $1 \le n \le N$, for which holds that

$$\mathbf{o}_n = g_n \mathbf{c}_n, \quad 1 \le n \le N. \tag{2}$$

The coder parameters are computed in the encoder such that the decoder output segments have the same energy as the associated encoder input segments.

Relating the generic decoder description to the problem statement of Section 2, we make the following remarks. By taking the decorrelators of Section 3, we counter Problem (II). When combining (1) and (2), we observe that the output signals are linear combinations of the down-mix signals and decorrelated signals derived thereof, which solves for Problems (I) and (IV). Finally, by keeping matrix $\mathbf{U}$ sparse, we tackle the up-mix part of Problem (III).

4.1.2. Encoder

The generic encoder structure of the N-M-N applause coder is shown in Figure 4. Let i_n denote the discrete time domain waveform of the nth input channel, with $1 \le n \le N$. These input channels are segmented (not shown in the figure), resulting in the segments $\mathbf{i}_n$, $1 \le n \le N$. Down-mixing of these input segments in the unit labelled D results in M down-mix segments $\mathbf{x}_m$, with $1 \le m \le M$. This is expressed by

$$\mathbf{X} = \mathbf{J}\,\mathbf{D}, \tag{3}$$

where $\mathbf{X}$ is a $K \times M$ matrix containing the M down-mix segments, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M]$, $\mathbf{J}$ is a $K \times N$ matrix containing the N input segments, $\mathbf{J} = [\mathbf{i}_1, \mathbf{i}_2, \ldots, \mathbf{i}_N]$, and $\mathbf{D}$ is the $N \times M$ down-mix matrix.

Next, in order to compute the coder parameters g_n, the encoder down-mix segments are first transformed to N up-mix segments $\mathbf{c}_n$, with $1 \leq n \leq N$, in the unit labelled U. This unit is identical to the decoder up-mixing unit, so that its operation is expressed by (1).

After having up-mixed the down-mix signals, the next step is to compute the coder parameters. From (2), it follows that by computing

$$g_n = \frac{||\mathbf{i}_n||}{||\mathbf{c}_n||}, \quad 1 \leq n \leq N, \text{ with } \|\mathbf{a}\|^2 \equiv \mathbf{a}^H \cdot \mathbf{a}, \tag{4}$$

the decoder output segments have the same energy as the encoder input segments. Because the parameters represent RMS ratios, they can be quantised like the ILD parameters of the MPEG Surround (MPS) coder [5].

Relating the generic encoder description to the problem statement of Section 2, we make the following remarks. By keeping the down-mix matrix $\mathbf{D}$ sparse, we counter the down-mix part of Problem (III). Moreover, because no filtering is applied, we counter Problem (I).

We conclude by remarking that the computational complexity of both the encoder and decoder is low (no frequency domain transformations, no filtering, and low complexity decorrelators).

4.2. The structure of the 5-2-5 coder

The highest quality of the MPS coder is attained for the 5-2-5 structure (hence $N = 5$ and $M = 2$). Therefore, we want to compare our coder to that configuration. However, the settings of the generic coder structure of Section 4.1, $\mathbf{D}$, $\mathbf{U}$, and the decorrelated signals still have to be determined. We present the coder implementation that achieved a high audio quality in informal listening experiments for the 5 multichannel applause signals at our disposal. For each of these signals, its five input channels sounded similarly (and were uncorrelated).

4.2.1. Encoder

An overview of an encoder implementation of the 5-2-5 applause coder is given in Figure 5. We have five input channels: left front (l_f), left surround (l_s), right front (r_f), right surround (r_s), and centre (c). These channels are segmented (not shown in the figure). In the down-mixing procedure, the left ($\mathbf{l}_{\text{dmx}}$) and right ($\mathbf{r}_{\text{dmx}}$) down-mix segments are simply the left and right front segments, respectively, so $\mathbf{l}_{\text{dmx}} = \mathbf{l}_f$, $\mathbf{r}_{\text{dmx}} = \mathbf{r}_f$. Because we assume that the five input channels sound sufficiently similar, both surround channels and the centre channel can be left out of the down mix. If, for example, the front and surround channels contain differently sounding applause signals, a different down-mixing procedure is necessary. In this case, one front and one back channel should be used as down-mix signals. It will be clear that when more than two input channels sound clearly different, a two channel down mix as we propose cannot be used to reproduce all differently sounding input channels.

For determining the coder parameters, we first remark that the encoder up-mixing procedure does not use decorrelators. This is done to simplify the encoder scheme and has no negative impact because the down-mix signals are uncorrelated. The latter follows from the down-mix signal equations, the assumption that the input signals are uncorrelated, and the assumption that the coder parameters are determined on a frame-by-frame basis, so that the reordering operation within a frame does not influence the energy measured in that frame (and hence neither the coder parameters). When the temporal resolution of the coder parameters needs to be higher, this simplification of the encoder up-mixing procedure cannot be applied. In the up-mixing procedure, we first apply a delay δ to the down-mix segments $\mathbf{l}_{\text{dmx}}$ and $\mathbf{r}_{\text{dmx}}$, which amounts to 896 samples at a sampling frequency of 44 100 Hz (about 20 milliseconds). The down-mix segments are delayed, because in the decoder we want to ensure that all decorrelators delay their input events. This is achieved, as mentioned in Section 3.2, by properly choosing a joint delay and reordering per decorrelator. Let $\mathbf{l}_{\text{dmx}}(\delta)$ denote the delayed left down-mix signal. The parameter p_4 is now given by

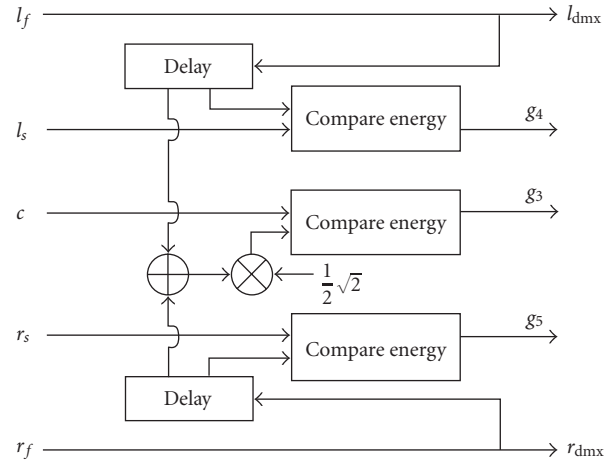

Figure 5: Encoder implementation of the 5-2-5 applause coder.

$$p_4 = \frac{||\mathbf{l}_s||}{||\mathbf{l}_{\text{dmx}}(\delta)||}. \tag{5}$$

Analogously we compute

$$\begin{aligned} p_3 &= \frac{\|\mathbf{c}\|}{(1/2)\sqrt{2}(||\mathbf{l}_{\text{dmx}}(\delta)|| + ||\mathbf{r}_{\text{dmx}}(\delta)||)}, \\ p_5 &= \frac{||\mathbf{r}_s||}{||\mathbf{r}_{\text{dmx}}(\delta)||}. \end{aligned} \tag{6}$$

The parameters p_3, p_4, and p_5 are low-pass filtered as follows:

$$g_{n,q} = \frac{1}{4}p_{n,q} + \frac{3}{4}g_{n,q-1}, \quad n = 3, 4, 5, \tag{7}$$

where q denotes the frame number. The time constant of the low-pass filter amounts to 161 milliseconds. Low-pass filtering is performed to obtain more stable output signals l_s, c, and r_s. Next, consecutive down-mix segments, $\mathbf{l}_{\text{dmx}}$ and $\mathbf{r}_{\text{dmx}}$,

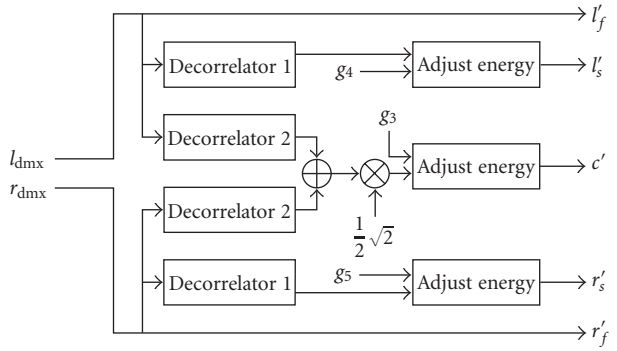

Figure 6: Decoder implementation of the 5-2-5 applause coder.

are combined using overlap add (not shown), resulting in the two output signals l_{dmx} and r_{dmx}.

Relating the encoder down-mixing operation to that of the generic encoder as given by (3), we have

$$\mathbf{X} = \begin{bmatrix} \mathbf{l}_{\text{dmx}} & \mathbf{r}_{\text{dmx}} \end{bmatrix}, \qquad \mathbf{D} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}^T, \tag{8}$$
$$\mathbf{J} = \begin{bmatrix} \mathbf{l}_f & \mathbf{r}_f & \mathbf{c} & \mathbf{l}_s & \mathbf{r}_s \end{bmatrix}.$$

4.2.2. Decoder

An overview of the decoder implementation of the 5-2-5 applause coder is given in Figure 6. The two time domain down-mix segments, $\mathbf{l}_{\text{dmx}}$ and $\mathbf{r}_{\text{dmx}}$, are obtained by segmenting the time domain down-mix signals l_{dmx} and r_{dmx} (not shown). In order to obtain the output segments $\mathbf{l}'_f$ and $\mathbf{r}'_f$, these down-mix segments are simply fed through. Next, both $\mathbf{l}_{\text{dmx}}$ and $\mathbf{r}_{\text{dmx}}$ are applied to two independent decorrelators. Because l_{dmx} and r_{dmx} are expected to be uncorrelated, the decorrelators applied to $\mathbf{l}_{\text{dmx}}$ can be identical to the decorrelators applied to $\mathbf{r}_{\text{dmx}}$. In the decorrelators, the 20-millisecond delay is applied before the reordering operation. After having made the decorrelated signals, the energy of $\mathbf{c}'$, $\mathbf{l}'_s$, and $\mathbf{r}'_s$ is adjusted in the blocks "Adjust energy" using the coder parameters. Finally, consecutive segments are combined using overlap add (not shown) resulting in five time domain output signals l'_f, r'_f, c', l'_s, and r'_s.

Relating this decoder to the generic decoder of Section 4.1, we have

$$\mathbf{X}_d = \begin{bmatrix} \mathbf{l}_{\text{dmx}} & \mathbf{r}_{\text{dmx}} & \mathbf{d}_1 & \mathbf{d}_2 \end{bmatrix}, \qquad \mathbf{U} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2}\sqrt{2} & 1 & 0 \\ 0 & 0 & \frac{1}{2}\sqrt{2} & 0 & 1 \end{bmatrix}, \tag{9}$$

where $\mathbf{d}_1$ and $\mathbf{d}_2$ are decorrelated signals (Section 3), whose associated reordering operations are determined by the permutations

$$\pi_1 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 \\ 3 & 7 & 15 & 2 & 1 & 14 & 6 & 4 & 10 & 5 & 11 & 9 & 8 & 13 & 16 & 12 \end{pmatrix},$$
$$\pi_2 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 \\ 5 & 1 & 4 & 6 & 3 & 9 & 2 & 8 & 15 & 12 & 7 & 13 & 16 & 11 & 10 & 14 \end{pmatrix}, \tag{10}$$

respectively. The permutations were hand-picked, whilst keeping in mind to generate uncorrelated signals, and the quality of the resulting decorrelated signals was checked using the assessment criteria mentioned in Section 3.1.

To complete the decoder description the following data are given: the coder update interval is 2048 samples and the overlap between segments amounts to 128 samples (hence the segment length K equals 2048 + 128). The overlapping begin and end parts of each segment are windowed using a half-sided plain Hanning window. For the decorrelators, we have $L = 16$, and each subsegment contains 256 samples, has 50% overlap, and is square-root Hanning windowed.

4.3. Integration in a spatial audio coder

The most basic way to integrate the applause coder in a spatial audio coder is to put the two coders in parallel, and switch between them, depending on the input signal being applause or not. To this end an applause classifier needs to be developed. It was mentioned in Section 3.2 that applying the decorrelators of the applause coder to nonapplause signals can lead to severe artefacts. Therefore, when tuning the classifier it should be taken into account that correctly classifying nonapplause signals is more important than correctly classifying applause signals. Another issue is to avoid artefacts when switching between the two coders. Though this problem is not addressed in this paper, it cannot be solved straightforwardly.

5. SUBJECTIVE EVALUATION

In this section, we compare the 5-2-5 applause coder to that specific configuration of the 5-2-5 MPS coder that achieves the highest audio quality for applause signals, by means of a listening test. In Section 5.1, we describe the conditions of the listening test. Next, in Section 5.2, we present the listening test results. A discussion is held in Section 5.3.

5.1. Method and stimuli

The list of coders used in the test is given in Table 1. Three alternative configurations were evaluated. Configuration (1) is the so-called guided envelope shaping mode of the MPS coder that performs best as to audio quality for applause signals and uses about 9 kbps for parameters [14]. For coding the stereo down-mix signal, a state-of-the-art AAC encoder is used that operates at a bit rate of 160 kbps. This bit rate is commonly used for high-quality stereo transmission. The coder of the stereo down mix is henceforth referred to as

Table 1: Coders under test.

5-2-5 Configuration	Coder	Core bit rate [kbps]	Parameter bit rate [kbps]
(1)	AAC stereo + MPS	160	9
(2)	AAC stereo + AC	160	0.2
(3)	None + AC	n/a	0.2

core coder. Configuration (2) is the applause coder (AC) as described in Section 4.2. The parameter bit rate amounts to 0.2 kbps. The core coder is identical to that used for the first configuration. Configuration (3) is similar to the second configuration, except that no core coder is employed. This configuration is used to gain insight in the effect on the audio quality of perceptually coding the stereo down-mix signal. The MPS coder without core coder was not evaluated in the test, because informal listening tests showed a performance similar to that of the MPS coder with core coder. This can be understood as follows. The MPS encoder substantially degrades the temporal structure of the applause signal and the additional degradation of the temporal structure by the core coder is perceived to be small. Moreover, the small artefacts introduced by the core coder are insignificant relative to the artefacts introduced by the MPS *decoder*. For the applause coder, however, the core coder does introduce clearly perceivable artefacts because the encoder preserves the temporal structure of the original applause signal.

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 artefacts. In a double-blind MUSHRA test [15], 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 5 intervals, labelled "bad," "poor," "fair," "good," and "excellent." A hidden reference and a low-pass filtered anchor (cut-off 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 D-to-A converter. Discrete pre amplifiers (array obsydian A-1) and power amplifiers (array quartz M-1) were used to feed a 5.1 loudspeaker setup employing B&W Nautilus 800 speakers in a dedicated listening room according to ITU recommendation [16]. The two test items used are part of the MPEG call for proposals (CfP) on spatial audio coding [17], being labelled "BBC applause" and "ARL applause." These two items are applause signals that contain no shouting or whistling (i.e., human utterances) and can be described as quite regular. For each item, all input channels sound quite similar and the clap density of the first item is significantly higher than that of the second. All input and output items were sampled at 44.1 kHz.

5.2. Results

The subjective listening test results are shown in Figure 7(a). The horizontal axis shows the two 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, labelled with "Mean" on the horizontal axis, 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 it was detected by the listeners. The 3.5 kHz low-pass filtered anchor received lowest scores of about 20. For the encoded items, the MPS coder (upward triangles) scored lowest. The applause coder (AC) + AAC (downward triangles) scores about 9 points higher in the mean, while the applause coder alone (diamonds) is again about 9 points better in the mean. The core coder appears to have a large influence on the quality of the applause coder for the "BBC applause" item.

Because the 95% confidence intervals are overlapping in the left panel, a pairwise two-tailed t-test was done to determine whether the differences between the MPS coder and the applause coder with AAC are statistically significant. For this purpose, in the right panel of Figure 7, difference scores are shown between the MPS coder on the one hand and the applause coder on the other hand. As can be seen, the confidence intervals for the mean scores are above the zero line, indicating that for both versions of the applause coder (with and without AAC) the difference with the MPS coder is statistically significant ($P < .05$) in favour of the applause coder.

The feedback of the listeners revealed a slight preference for the MPS coder as to preservation of the spatial image. At the same time, the applause coder was perceived to be much better in preserving the timbre of the original signal. This indicates that depending on whether the emphasis was put on correct representation of the spatial image or the timbre, the results of the individual listeners varied between being comparable for both coders, or being clearly in favour of the applause coder.

5.3. Discussion

The applause coder was found to have a significantly better audio quality than the best MPS coder for the two applause signals tested, whilst it employs a significantly lower parameter bit rate. This result indicates the added value of the applause coder. The increase in quality is due to a better

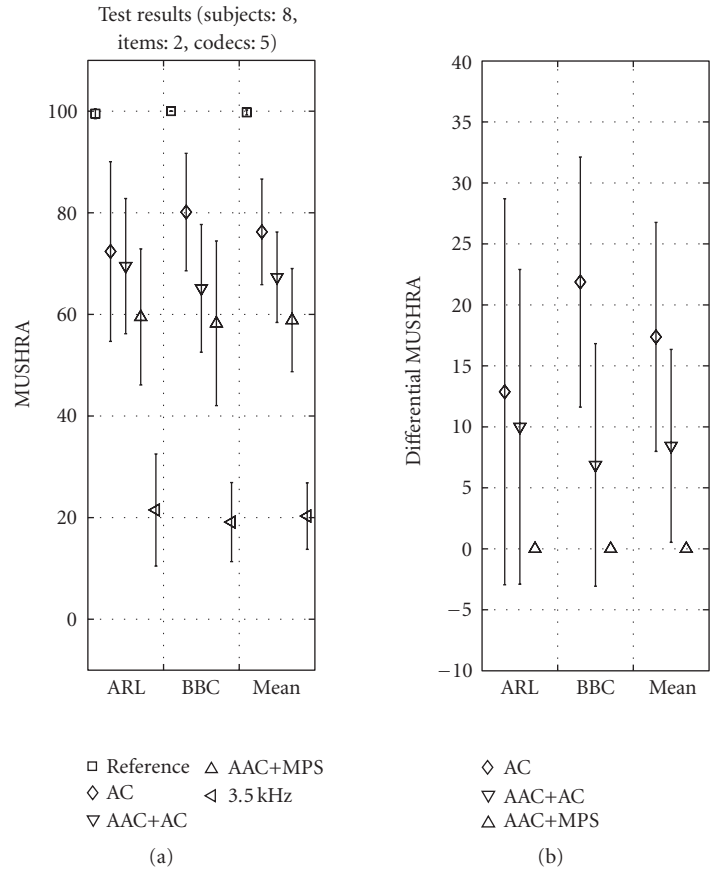

Figure 7: Subjective listening test results. (a) shows MUSHRA scores for the applause coder alone (diamonds), applause coder plus AAC core coder (downward trianges), and MPS coder (upward triangles). In addition, the 3.5 kHz low-pass filtered anchor (leftwards triangles) and hidden reference (squares) are shown. In (b), difference scores are shown relative to the MPS coder.

preservation of the timbre of the original multichannel signal and the avoidance of an increase in clap density. This is achieved by not applying any spectral filtering in the coder framework, having a new type of decorrelators and having a sparse down- and up-mix matrix. At the same time, the coder structure is of very low complexity. This is due to the fact that frequency specific manipulations are avoided and basic decorrelators are used in the proposed applause coder. However, as mentioned in Section 4.3, integration of the applause coder in the structure of the MPS coder is not a straightforward task. Another issue is the fact that we focused on applause signals with similarly sounding channels. However, we briefly saw that the down-mixing procedure depends on the number and positions of the differently sounding channels. Therefore, when coding the more general applause signal, an *adaptive* down-mixing (and up-mixing) procedure might be required. Finally, it should be noted that in the listening test there was dissension among the listeners, related to putting the emphasis on correct representation of the spatial image or the timbre, so that 8 listeners might be a too small number for a truly representative listening test.

In the listening test, we observed that the MUSHRA score dropped by 15 points for the BBC item when applying a core coder to the stereo down-mix of the applause coder. This shows that for this specific signal, the state-of-the-art AAC coder, operating at 80 kbps per channel, is not close to transparency when used in the environment of the multichannel applause coder.

The proposed coder was only evaluated for applause signals. However, we expect the coder to achieve a good audio quality as well for other signals consisting of frequently occurring transients, like rainfall and crackling sounds. This is due to the fact that the problem-solution approach of Section 2 focuses on this kind of signals. Moreover, we expect the proposed coder to achieve high-audio quality for both coloured- and white-noise (like) input signals. Also for these types of signals, the new decorrelators produce high-quality decorrelated signals being uncorrelated with their input signals, yet having a very similar timbre. Therefore, the coder output signals will be independent coloured- and white-noise signals, respectively, as desired. At the same time, fluctuations of the temporal envelope can be captured by the coder parameters.

6. CONCLUSIONS

In this paper, we describe a multichannel audio codec dedicated to the coding of applause signals. It is based on timbre

and clap density preservation. The codec combines a very low complexity and a low parameter bit rate (0.2 kbps). When comparing the audio quality of this codec to the best MPEG Surround multichannel codec for applause signals, with an associated parameter bit rate of about 9 kbps, we found the proposed codec to perform significantly better for the two applause signals under test. Though this seems promising, the technique presented is not fully mature, for example, because issues related to integration of the proposed codec in the MPEG Surround codec were not addressed. Moreover, we mainly focussed on a solution for applause signals with similarly sounding channels and we have not evaluated other types of applause signals.

ACKNOWLEDGMENTS

The authors wish to thank the reviewers and their colleagues Bert den Brinker and Arno van Leest for their useful remarks and suggestions to earlier versions of the manuscript.

REFERENCES

[1] K. Brandenburg, G. Stoll, Y.-F. Dehery, J. D. Johnston, L. Kerkhof, and E. F. Schroder, "ISO-MPEG-1 audio: a generic standard for coding of high-quality digital audio," *Journal of the Audio Engineering Society*, vol. 42, no. 10, pp. 780–792, 1994.

[2] J. Herre, H. Purnhagen, J. Breebaart, et al., "The reference model architecture of MPEG spatial audio coding," in *Proceedings of 118th Audio Engineering Society Convention*, pp. 1–13, Barcelona, Spain, May 2005.

[3] J. Breebaart, J. Herre, C. Faller, et al., "MPEG spatial audio coding/MPEG surround: overview and current status," in *Proceedings of 119th Audio Engineering Society Convention*, New York, NY, USA, October 2005.

[4] L. Villemoes, J. Herre, J. Breebaart, et al., "MPEG surround: the forthcoming ISO standard for spatial audio coding," in *Proceedings of the 28th Audio Engineering Society International Conference*, pp. 213–230, Pitea, Sweden, June 2006.

[5] 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 multi-channel audio compression," *Journal of Audio Engineering Society*, vol. 55, pp. 331–351, 2007.

[6] A. Seefelt, M. S. Vinton, and C. Q. Robinson, "New techniques in spatial audio coding," in *Proceedings of 119th Audio Engineering Society Convention*, New York, NY, USA, October 2005.

[7] F. Baumgarte and C. Faller, "Binaural cue coding—part I: psychoacoustic fundamentals and design principles," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 509–519, 2003.

[8] C. Faller and F. Baumgarte, "Binaural cue coding—part II: schemes and applications," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 520–531, 2003.

[9] W. A. Yost, "Lateral position of sinusoids presented with interaural intensive and temporal differences," *Journal of the Acoustical Society of America*, vol. 70, no. 2, pp. 397–409, 1981.

[10] D. W. Grantham, "Spatial hearing and related phenomena," in *Handbook of Perception and Cognition: Hearing*, B. C. J. Moore, Ed., pp. 297–345, Academic Press, London, UK, 2nd edition, 1995.

[11] H. Purnhagen, "Low complexity parametric stereo coding in MPEG-4," in *Proceedings of the 7th International Conference on Digital Audio Effects (DAFx '04)*, Naples, Italy, October 2004.

[12] C. Faller, "Parametric multichannel audio coding: synthesis of coherence cues," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 299–310, 2006.

[13] I. Pollack, "Temporal switching between binaural information sources," *Journal of the Acoustical Society of America*, vol. 63, no. 2, pp. 550–558, 1978.

[14] ISO/IEC FDIS 23003-1:2006(E), "MPEG audio technologies—part 1: MPEG surround," 2004.

[15] ITU-R Recommend. BS.1534, "Method for the subjective assessment of intermediate quality level of coding systems (MUSHRA)," 2001.

[16] ITU-R Recommend. BS.1116-1, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," 1997.

[17] Audio Subgroup, "Call for proposals on spatial audio coding," ISO/IEC JTC1/SC29/WG11 N6455, 2004.