Audio Engineering Society

Convention Paper

Presented at the 120th Convention
2006 May 20–23 Paris, France

Parametric Joint-Coding of Audio Sources

Christof Faller[1]

[1] *Audiovisual Communications Laboratory, EPFL Lausanne, Switzerland*

Correspondence should be addressed to C. Faller (christof.faller@epfl.ch)

ABSTRACT

The following coding scenario is addressed: A number of audio source signals need to be transmitted or stored for the purpose of mixing stereo, multi-channel surround, wavefield synthesis, or binaural signals after decoding the source signals. The proposed technique offers significant coding gain when jointly coding the source signals, compared to separately coding them, even when no redundancy is present between the source signals. This is possible by considering statistical properties of source signals, properties of mixing techniques, and spatial perception. The sum of the source signals is transmitted plus the statistical properties which determine the spatial cues at the mixer output. Informal subjective evaluation indicates that the proposed scheme achieves high audio quality.

1. INTRODUCTION

In this paper, coding of a plurality of audio sources for the purpose of mixing after decoding is addressed. "Object based" audio systems require storage/transmission of the audio sources such that they can be mixed at the decoder side as desired. Also wave field synthesis systems are often driven with audio source signals. ISO/IEC MPEG-4 [1, 2, 3] addresses a general object-based coding scenario. It defines the scene description (= mixing parameters) and uses for each ("natural") source signal a separate mono audio coder. However, when a complex scene with many sources is to be coded the bitrate becomes high since the bitrate scales with the number of sources. An object based audio system is illustrated in Figure 1. A number of audio source signals are coded and stored/transmitted. The receiver mixes the decoded audio source signals to generate stereo [4, 5], surround [6, 5], wavefield synthesis [7, 8, 9], or binaural signals [10, 11].

It would be desirable to have an efficient coding paradigm for audio sources that will be mixed after decoding. However, from an information theoretic point of view, there is no additional coding gain when jointly coding independent sources compared to independently coding them. For example, given a number of independent instrument signals the best one can do with conventional wisdom is to apply to each instrument signal one coder (e.g. a perceptual audio coder such as AAC [12], AC-3 [13], ATRAC

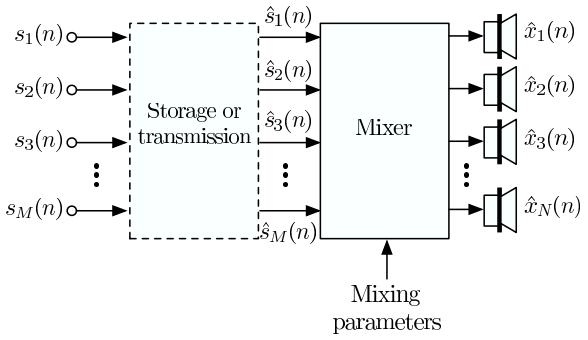

Fig. 1: The general problem that is addressed: Coding of a number of source signals for the purpose of mixing stereo, multi-channel, wavefield synthesis, or binaural audio signals with the decoded source signals.

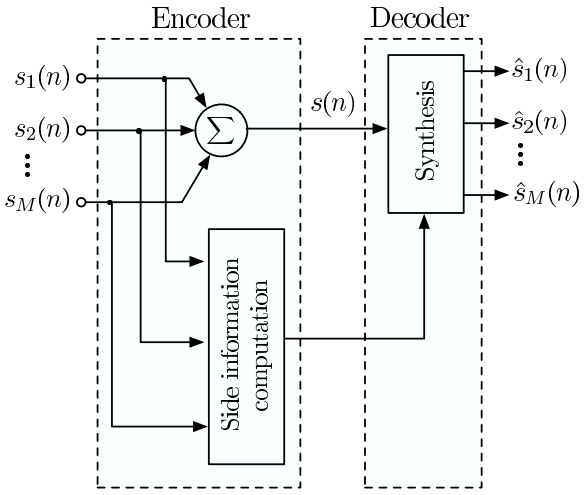

Fig. 2: A number of source signals are stored or transmitted as sum signal plus side information. The side information represents statistical properties of the source signals. At the receiver the source signals are recovered.

[14], MP3 [15], or PAC [16]).

Nevertheless, it is shown in this paper that joint-coding can be significantly more efficient than independent coding of the sources for a specific application scenario. Sources are joint-coded for the purpose of mixing after decoding. In this case, considering properties of the source signals, the mixing process, and spatial perception it is possible to significantly reduce the bitrate. The sources are represented as a single mono sum signal plus about 3kb/s side information per source. Conventional coding would require for 10 sources about $10 \cdot 80 = 800$ kb/s. The proposed technique requires only $80 + 10 \cdot 3 = 110$ kb/s and thus is significantly more efficient than conventional coding. (It was assumed that 80 kb/s are needed for high quality coding of a mono audio signal).

Previously, we addressed a special case of the described coding problem with a scheme denoted Binaural Cue Coding (BCC) for Flexible Rendering [17, 18, 19]. By storing/transmitting only the sum of the given source signals plus low bitrate side information, low bitrate is achieved. However, the source signals can not be recovered at the decoder and the scheme was limited to stereo and multi-channel surround signal generation. Also, only simplistic mixing was used, based on amplitude and delay panning. Thus, the direction of sources could be controlled but no other auditory spatial image attributes. Another limitation of this scheme was its limited audio quality. Especially, a decrease in audio quality as the number of sources is increased.

The proposed scheme for joint-coding of audio source signals is illustrated in Figure 2. Similarly to BCC for Flexible Rendering it is based on only transmitting[1] the sum of the audio source signals. But it overcomes the limitation that only simplistic mixers can be used and the quality does not anymore degrade as the number of source signals is increased. In addition to the sum signal also side information is transmitted.

The side information represents the statistical properties of the source signals which are the most important factors determining the perceptual spatial cues of the mixer output signals. It will be shown that these properties are temporally evolving spectral envelopes and auto-correlation functions. About 3 kb/s of side information is needed per source signal. A conventional audio or speech coder is used to efficiently represent the sum signal.

At the receiver, the source signals are recovered such that the before mentioned statistical properties approximate the corresponding properties of the original source signals. A stereo, surround, wavefield

[1] In the following *transmitting* a number of source signals always means either transmitting them or storing them on a medium.

synthesis, or binaural mixer is applied after decoding of the source signals to generate the output signal.

Conceptually, the aim of the proposed scheme is not to recover the *clean source signals* and it is not intended that one listens to these source signals separately prior to mixing. The goal is that the *mixed output signal* perceptually approximates the reference signal (the signal generated with the same mixer supplied with the original source signals).

The paper is organized as follows. Section 2 explains the psychoacoustic assumptions which are made. The proposed scheme for joint-coding of independent source signals is described in Section 3. Section 4 describes how the proposed scheme is implemented without explicit decoding of the source signals by use of a BCC decoder [17, 18, 20], parametric stereo decoder [21], or MPEG Surround decoder [22, 23] as a decoder/mixer. Section 5 describes how the proposed scheme is implemented using an FFT-based or QMF filterbank. Discussion of the proposed scheme and informal subjective evaluation are described in Sections 6 and 7, respectively. Conclusions are presented in Section 8.

2. PSYCHOACOUSTIC ASSUMPTIONS

For headphone playback, the psychoacoustic assumption we are making is that the interaural cues (time difference, level difference, coherence) represent the relevant attributes of the auditory spatial image that is perceived when listening to a stereo signal. This assumption implies that in principle it is possible to reduce a stereo signal to a single mono channel plus information about the binaural cues and recover a binaural signal which is perceptually equivalent to the original signal. In order to capture all binaural information processed by the auditory system, the binaural cues are considered in frequency bands mimicking the frequency decomposition of the auditory periphery and with a suitable time resolution.

For loudspeaker playback, it is assumed that the inter-channel cues represent all attributes of the auditory spatial image perceived by a listener. Similarly as the binaural cues for headphone playback, the inter-channel cues are considered in frequency bands mimicking the frequency decomposition of the auditory periphery and with a suitable time resolution. This assumption implies that in principle it is possible to reduce a multi-channel audio signal to a single mono channel plus information about the inter-channel cues and recover a multi-channel signal which is perceptually equivalent to the original signal.

The use of different inter-channel cues for representation of auditory spatial image properties can be motivated as follows. Summing localization [10] implies that perceptually relevant audio channel differences for a loudspeaker signal channel pair are the *inter-channel time difference* (ICTD) and *inter-channel level difference* (ICLD). ICTD and ICLD can be related to the perceived direction of auditory events [10, 24, 25]. Other auditory spatial image attributes, such as apparent source width [26] and listener envelopment [27], can be related to the *interaural cross-correlation coefficient* (IACC) [28, 26]. For loudspeaker pairs in the front or back of a listener, IACC is often directly related to the *inter-channel coherence* (ICC) [29] which is thus considered as a third audio channel difference measure. The considered inter-channel cues (ICTD, ICLD, ICC) are similar measures as the binaural cues, but considered between audio signal channels as opposed to ear entrance signal channels. For headphone playback the inter-channel cues are (ideally) identical to the binaural cues. Thus, in the following we are limiting the discussion to the inter-channel cues.

The made assumptions may at first sight seem a bit far fetched. In terms of source localization a recently proposed auditory model [30] supports the assumption that at least with respect to source localization in principle it would be enough to synthesize the inter-channel cues since source localization according to this model depends on these (even in reverberant multi-source scenarios). On the practical side, the surprisingly good performance achieved by Binaural Cue Coding (BCC) [17, 18, 20] and other schemes for parametric stereo [21] and multi-channel audio coding [22, 23] implies that the inter-channel cues capture a wide range of auditory spatial image attributes. The role ICTD, ICLD, and ICC may play in determining various attributes of the auditory spatial image is discussed in [20] (Chapter 3.3.3).

3. JOINT-CODING OF AUDIO SOURCE SIGNALS

As mentioned, the proposed scheme for joint-coding of audio source signals, shown in Figure 2, is based on only transmitting the sum of the audio source signals,

$$s(n) = \sum_{i=1}^{M} s_i(n)\,, \tag{1}$$

where M is the number of source signals and $s_i(n)$ are the source signals.

In addition to the sum signal, side information is transmitted. As mentioned in the previous section, the psychoacoustic assumption we are making, is, that the perceived auditory spatial image is largely determined by the ICTD, ICLD, and ICC. Therefore, as opposed to requiring "clean" source signals $s_i(n)$ as mixer input in Figure 1, we just require signals $\hat{s}_i(n)$ with the property that they result in similar ICTD, ICLD, and ICC at the mixer output as for the case of supplying the real source signals $s_i(n)$ to the mixer. There are three goals for the generation of $\hat{s}_i(n)$:

- If $\hat{s}_i(n)$ are supplied to a mixer, the mixer output channels will have approximately the same spatial cues (ICLD, ICTD, ICC) as if $s_i(n)$ were supplied to the mixer.
- $\hat{s}_i(n)$ are to be generated with as little as possible information about the original source signals $s_i(n)$ (because the goal is to have low bitrate side information).
- $\hat{s}_i(n)$ are generated from the transmitted sum signal $s(n)$ such that a minimum amount of signal distortion is introduced.

Without loss of generality, for deriving the proposed scheme, we are considering a stereo mixer. A further simplification over the general case is that only amplitude and delay panning are applied for mixing. If the discrete source signals were available at the decoder, a stereo signal would be mixed as shown in Figure 3, i.e.

$$\begin{aligned} x_1(n) &= \sum_{i=1}^{M} a_i s_i(n - c_i) \\ x_2(n) &= \sum_{i}^{M} b_i s_i(n - d_i)\,, \end{aligned} \tag{2}$$

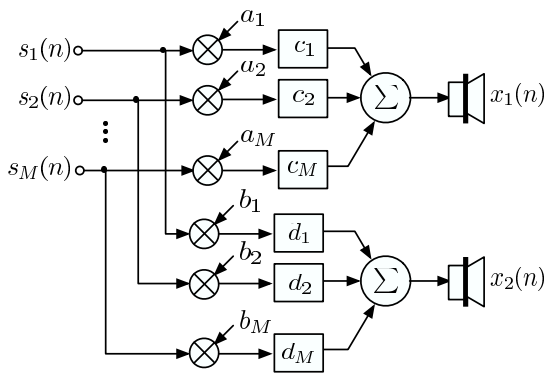

Fig. 3: A mixer for generating stereo signals given a number of source signals.

where a_i, b_i, c_i, and d_i are the mixing parameters.

Usually a mixer contains for each source user controls for gain and amplitude panning (pan pot). For the sake of generality we are also considering delay panning. Given for each source with index i the gain G_i in dB, pan pot position ΔL_i (expressed as level difference in dB), and the delay pan pot position τ_i in samples, the mixing parameters (2) can be computed:

$$\begin{aligned} a_i &= \frac{10^{G_i/20}}{\sqrt{1 + 10^{\Delta L_i/10}}} \\ b_i &= 10^{(G_i + \Delta L_i)/20} a_i \\ c_i &= \max\{-\tau_i, 0\} \\ d_i &= \max\{\tau_i, 0\}\,. \end{aligned} \tag{3}$$

In the following, we are computing ICTD, ICLD, and ICC of the stereo mixer output as a function of the input source signals $s_i(n)$. The obtained expressions will give indication which source signal properties determine the mixer output ICTD, ICLD, and ICC (together with the mixing parameters). $\hat{s}_i(n)$ are then generated such that the identified source signal properties approximate the corresponding properties of the original source signals.

3.1. ICTD, ICLD, and ICC of the mixer output

The cues are estimated in subbands and as a function of time. Psychoacoustics suggests that spatial perception is most likely based on a critical band representation of the acoustic input signal [10]. This

frequency resolution is considered by using an invertible filterbank with subbands with bandwidths equal or proportional to the critical bandwidth of the auditory system [31, 32].

In the following it is assumed that the source signals $s_i(n)$ are zero mean and mutually independent. A pair of subband signals of the mixer output (2) is denoted $\tilde{x}_1(n)$ and $\tilde{x}_2(n)$. Note that for simplicity of notation we are using the same time index n for time-domain and subband-domain signals. Also, no subband index is used and the described analysis/processing is applied to subbands at each frequency independently. The subband power of the two mixer output signals is

$$\begin{aligned} \mathrm{E}\{\tilde{x}_1^2(n)\} &= \sum_{i=1}^{M} a_i^2 \mathrm{E}\{\tilde{s}_i^2(n))\} \\ \mathrm{E}\{\tilde{x}_2^2(n)\} &= \sum_{i=1}^{M} b_i^2 \mathrm{E}\{\tilde{s}_i^2(n))\}\,, \end{aligned} \tag{4}$$

where $\tilde{s}_i(n)$ is one subband signal of source $s_i(n)$ and $\mathrm{E}\{.\}$ denotes short-time mean, e.g.

$$\mathrm{E}\{\tilde{s}_i^2(n)\} = \frac{1}{K} \sum_{n-K/2}^{n+K/2-1} \tilde{s}_i^2(n)\,, \tag{5}$$

where K determines the length of the moving average. Note that the subband power values $\mathrm{E}\{\tilde{s}_i^2(n)\}$ represent for each source signal the spectral envelope as a function of time. The time span considered for the averaging (5) determines the time resolution at which the inter-channel cues are considered.

The ICLD, $\Delta L(n)$, is

$$\Delta L(n) = 10 \log_{10} \frac{\sum_{i=1}^{M} b_i^2 \mathrm{E}\{\tilde{s}_i^2(n))\}}{\sum_{i=1}^{M} a_i^2 \mathrm{E}\{\tilde{s}_i^2(n))\}}\,. \tag{6}$$

For estimating ICTD and ICC the normalized cross-correlation function [33],

$$\Phi(n,d) = \frac{\mathrm{E}\{\tilde{x}_1(n)\tilde{x}_2(n+d)\}}{\sqrt{\mathrm{E}\{\tilde{x}_1^2(n)\}\mathrm{E}\{\tilde{x}_2^2(n+d)\}}}\,, \tag{7}$$

is estimated. The ICC, $c(n)$, is computed according to

$$c(n) = \max_d |\Phi(n,d)|\,. \tag{8}$$

For the computation of the ICTD, $\tau(n)$, the location of the highest peak on the delay axis is computed,

$$\tau(n) = \arg\max_d \Phi(n,d)\,. \tag{9}$$

Now the question is, how can the normalized cross-correlation function be computed as a function of the mixing parameters. Together with (2), (7) can be written as

$$\Phi(n,d) = \frac{\sum_{i=1}^{M} \mathrm{E}\{a_i b_i \tilde{s}_i(n-c_i)\tilde{s}_i(n-d_i+d)\}}{\sqrt{\mathrm{E}\{\sum_{i=1}^{M} a_i^2 s_i^2(n-c_i)\}\mathrm{E}\{\sum_{i=1}^{M} b_i^2 s_i^2(n-d_i)\}}}\,, \tag{10}$$

which is equivalent to

$$\Phi(n,d) = \frac{\sum_{i=1}^{M} a_i b_i \mathrm{E}\{\tilde{s}_i^2(n)\}\Phi_i(n,d-\tau_i)}{\sqrt{(\sum_{i=1}^{M} a_i^2 \mathrm{E}\{\tilde{s}_i^2(n)\})(\sum_{i=1}^{M} b_i^2 \mathrm{E}\{\tilde{s}_i^2(n)\})}}\,, \tag{11}$$

where the normalized auto-correlation function $\Phi_i(n,e)$ is

$$\Phi_i(n,e) = \frac{\mathrm{E}\{s_i(n)s_i(n+e)\}}{\mathrm{E}\{s_i^2(n)\}}\,, \tag{12}$$

and $\tau_i = d_i - c_i$. Note that for computing (11) given (10) it has been assumed that the signals are wide sense stationary within the considered range of delays, i.e.

$$\begin{aligned} \mathrm{E}\{\tilde{s}_i^2(n)\} &= \mathrm{E}\{\tilde{s}_i^2(n-c_i)\} \\ \mathrm{E}\{\tilde{s}_i^2(n)\} &= \mathrm{E}\{\tilde{s}_i^2(n-d_i)\} \\ \mathrm{E}\{\tilde{s}_i(n)\tilde{s}_i(n+c_i-d_i+d)\} &= \mathrm{E}\{\tilde{s}_i(n-c_i)\tilde{s}_i(n-d_i+d)\}\,. \end{aligned}$$

A numerical example for two source signals, illustrating the dependence between ICTD, ICLD, and ICC and the source subband power, is shown in Figure 4. The top, middle, and bottom panel of Figure 4 show $\Delta L(n)$, $\tau(n)$, and $c(n)$, respectively, as a function of the ratio of the subband power of the two sources, $a = \mathrm{E}\{\tilde{s}_1^2(n)\}/(\mathrm{E}\{\tilde{s}_1^2(n)\} + \mathrm{E}\{\tilde{s}_2^2(n)\})$, for different mixing parameters (3) ΔL_1, ΔL_2, τ_1, and τ_2 (with $G_i = 1$).

The top panel of Figure 4 indicates that when only one source has power in the subband ($a = 0$ or $a = 1$), then the mixer ICLD, $\Delta L(n)$ (6), is equal to the amplitude panning parameter ΔL_i (3) of the dominant source. When the power in the subbands fades from one source to the other, i.e. when a changes from zero to one, the mixer output level difference fades from the amplitude panning parameter

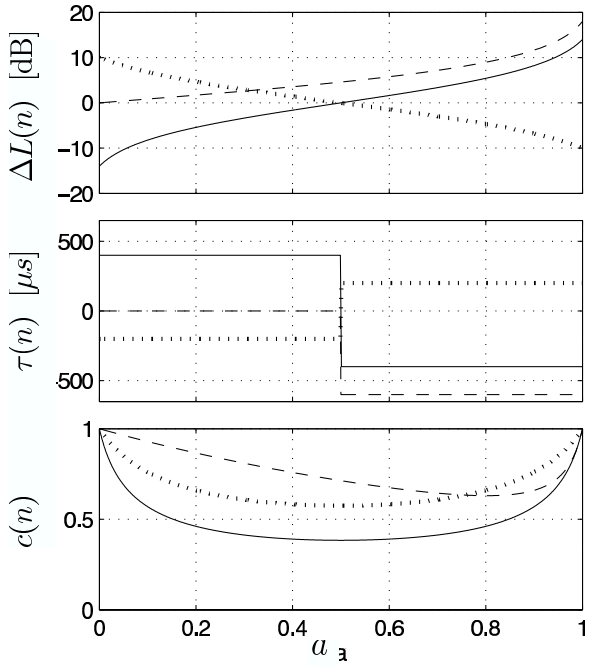

Fig. 4: $\Delta L(n)$ (top), $\tau(n)$ (middle), and $c(n)$ (bottom) for a critical band as a function of $a = \mathrm{E}\{\tilde{s}_1^2(n)\}/(\mathrm{E}\{\tilde{s}_1^2(n)\}+\mathrm{E}\{\tilde{s}_2^2(n)\})$. The mixer parameters (3) are: $\Delta L_1 = 14$ dB, $\Delta L_2 = -14$ dB, $\tau_1 = -400$ μs, $\tau_2 = 400$ μs (solid); $\Delta L_1 = 18$ dB, $\Delta L_2 = 0$ dB, $\tau_1 = -600$ μs, $\tau_2 = 0$ μs (dashed); $\Delta L_1 = -10$ dB, $\Delta L_2 = 10$ dB, $\tau_1 = 200$ μs, $\tau_2 = -200$ μs (dotted). The source gain has always been chosen to be $G_i = 0$ dB.

of one source to the amplitude panning parameter of the other source.

The middle panel of Figure 4 indicates that when only one source has power in the subband ($a = 0$ or $a = 1$), then the mixer ICTD, $\tau(n)$ (9), is equal to the delay panning parameter τ_i (3) of the dominant source. As opposed to the mixer output level difference, the mixer output time difference is determined by the delay panning parameter of the source which has more power in the subband, as indicated by the hard switch of $\tau(n)$ at $a = 0.5$.

The bottom panel of Figure 4 indicates that when only one source has power in the subband ($a = 0$ or $a = 1$), then the mixer output coherence, $c(n)$ (8), is equal to one. Mixer output coherence decreases when more than one source has power in the subband.

3.2. Necessary side information

The previously derived expressions for the inter-channel cues occurring when mixing the (original) source signals indicate what information other than the mixing parameters determine the inter-channel cues of the mixer output signal. The ICLD (6) depends on the mixing parameters (a_i, b_i, c_i, d_i) and on the short-time subband power of the sources, $\mathrm{E}\{\tilde{s}_i^2(n)\}$ (5). The normalized subband cross-correlation function $\Phi(n, d)$ (11), that is needed for ICTD (9) and ICC (8) computation, depends on $\mathrm{E}\{\tilde{s}_i^2(n)\}$ and additionally on the normalized subband auto-correlation function, $\Phi_i(n, e)$ (12), for each source signal.

For simplicity of synthesis and for reducing the amount of side information, only $\mathrm{E}\{\tilde{s}_i^2(n)\}$ is considered and it is assumed that the synthesized sources have the correct $\Phi_i(n, e)$ without explicitly applying processing for synthesizing it.

In order to further reduce the amount of side information, the relative dynamic range of the source signals is limited. At each time, for each subband, the power of the strongest source is selected. We found it sufficient to lower bound the corresponding subband power of all the other sources at a value 24 dB lower than the strongest subband power. Thus the dynamic range of the quantizer can be limited to 24 dB.

The power of the sources with indices $2 \le i \le M$ relative to the power of the first source is transmitted as side information,

$$\Delta\tilde{p}_i(n) = 10 \log_{10} \frac{\mathrm{E}\{\tilde{s}_i^2(n)\}}{\mathrm{E}\{\tilde{s}_1^2(n)\}} \,. \qquad (13)$$

Note that dynamic range limiting as described previously is carried out prior to (13), avoiding numerical problems when $\mathrm{E}\{\tilde{s}_1^2(n)\}$ vanishes. For a sampling frequency of 44.1 kHz we use 20 subbands and transmit for each subband $\Delta\tilde{p}_i(n)$ ($2 \le i \le M$) about every 12 ms. The relative power values are quantized with a scheme similar to the ICLD quantizer described in [34], resulting in a bitrate of approximately $3(M-1)$ kb/s.

As opposed to transmitting the subband power values $\mathrm{E}\{\tilde{s}_i^2(n)\}$, other information representing the spectral envelopes of the source signals could be transmitted. For example, linear predictive coding

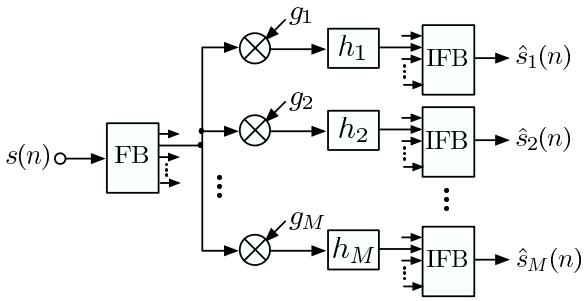

Fig. 5: The process for generation of $\hat{s}_i(n)$. The sum signal is converted to the subband domain. The subbands are scaled such that the subband power is approximately the same as the subband power of the original source signals. Filtering is applied to the scaled subbands for de-correlation. The shown processing is carried out independently for each subband. FB is a filterbank with subbands with bandwidths motivated by perception. IFB is the corresponding inverse filterbank.

(LPC) parameters [35, 36] could be transmitted, or corresponding other parameters such as lattice filter parameters or line spectral pair (LSP) parameters [37].

3.3. Reconstructing the sources

Figure 5 illustrates the process that is used to recreate the source signals, given the sum signal (1). This process is part of the "Synthesis" block in Figure 2. The individual source signals are recovered by scaling each subband of the sum signal with $g_i(n)$ and by applying a de-correlation filter with impulse response $h_i(n)$,

$$\begin{aligned}\hat{\tilde{s}}_i(n) &= h_i(n) \star (g_i(n)\tilde{s}(n)) \\ &= h_i(n) \star \left(\sqrt{\frac{\mathrm{E}\{\tilde{s}_i^2(n)\}}{\mathrm{E}\{\tilde{s}^2(n)\}}}\tilde{s}(n) \right) , \quad (14)\end{aligned}$$

where $\star$ is the linear convolution operator and $\mathrm{E}\{\tilde{s}_i^2(n)\}$ is computed with the side information by

$$\mathrm{E}\{\tilde{s}_i^2(n)\} = \begin{cases} 1/\sqrt{1+\sum_{i=2}^{M} 10^{\frac{\Delta\tilde{p}_i(n)}{10}}}, & \text{for } i=1 \\ 10^{\frac{\Delta\tilde{p}_i(n)}{10}}\mathrm{E}\{\tilde{s}_1^2(n)\}, & \text{otherwise}. \end{cases} \quad (15)$$

As de-correlation filters $h_i(n)$, complementary comb filters [38], allpass filters [39, 40], delays [41], or filters with random impulse responses [42, 20] may be used. The goal for the de-correlation process is to reduce correlation between the signals while not modifying how the individual waveforms are perceived. Different de-correlation techniques cause different artifacts. Complementary comb filters cause coloration. All the described techniques are spreading the energy of transients in time causing artifacts such as "pre-echoes". Given their potential for artifacts, de-correlation techniques should be applied as little as possible.

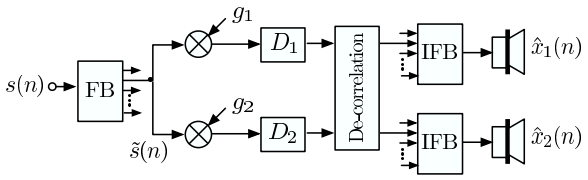

Fig. 6: A mixer for generating directly stereo signals given the sum of a number of source signals without explicit computation of the source signals. Gain factors, delays, and de-correlation are applied independently in subbands.

When applying no de-correlation processing ($h_i(n) = \delta(n)$ in (14)) good audio quality can also be achieved. It is a compromise between artifacts introduced by the de-correlation processing and artifacts due to the fact that the source signals $\hat{s}_i(n)$ are correlated. When no de-correlation processing is used the resulting auditory spatial image may suffer from instability [20]. But the mixer may introduce itself some de-correlation when reverberators or other effects are used and thus there is less need for de-correlation processing.

4. USING SPATIAL AUDIO DECODERS AS MIXERS

Mixing is directly applied to the transmitted sum signal (1) without explicit computation of $\hat{s}_i(n)$. A BCC synthesis scheme is used for this purpose. In the following, we are considering the stereo case, but all the described principles can be applied for generation of multi-channel audio signals as well. (Similarly, a parametric stereo [21] or MPEG Surround [22, 23] decoder can be used as mixer).

A stereo BCC synthesis scheme, applied for processing the sum signal (1), is shown in Figure 6. Desired would be that the BCC synthesis scheme generates a

signal that is perceived similarly as the output signal of a mixer as shown in Figure 3. This is so, when ICTD, ICLD, and ICC between the BCC synthesis scheme output channels are similar as the corresponding cues appearing between the mixer output (2) signal channels.

The same side information as for the previously described more general scheme is used, allowing the decoder to compute the short-time subband power values $\mathrm{E}\{\tilde{s}_i^2(n)\}$ of the sources. Given $\mathrm{E}\{\tilde{s}_i^2(n)\}$, the gain factors g_1 and g_2 in Figure 6 are computed as

$$\begin{aligned} g_1(n) &= \sqrt{\frac{\sum_{i=1}^{M} a_i^2 \mathrm{E}\{\tilde{s}_i^2(n)\}}{\mathrm{E}\{\tilde{s}^2(n)\}}} \\ g_2(n) &= \sqrt{\frac{\sum_{i=1}^{M} b_i^2 \mathrm{E}\{\tilde{s}_i^2(n)\}}{\mathrm{E}\{\tilde{s}^2(n)\}}}, \end{aligned} \qquad (16)$$

such that the output subband power and ICLD (6) are the same as for the mixer in Figure 3.

The ICTD $\tau(n)$ is computed according to (9), determining the delays D_1 and D_2 in Figure 6,

$$\begin{aligned} D_1(n) &= \max\{-\tau(n), 0\} \\ D_2(n) &= \max\{\tau(n), 0\}. \end{aligned} \qquad (17)$$

The ICC $c(n)$ is computed according to (8) determining the de-correlation processing in Figure 6. De-correlation processing (ICC synthesis) is described in [18, 43, 40, 42, 20]. The advantages of applying de-correlation processing to the mixer output channels compared to applying it for generating independent $\hat{s}_i(n)$ are:

1. Usually the number of source signals M is larger than the number of audio output channels N. Thus, the number of independent audio channels that need to be generated is smaller when de-correlating the N output channels as opposed to de-correlating the M source signals.

2. Often the N audio output channels are correlated (ICC > 0) and less de-correlation processing can be applied than would be needed for generating independent M or N channels.

Due to less de-correlation processing better audio quality is expected.

Best audio quality is expected when the mixer parameters are constrained such that $a_i^2 + b_i^2 = 1$, i.e. $G_i = 0$ dB. In this case, the subband power of each source in the transmitted sum signal (1) is the same as the power of the same source in the mixed decoder output signal. The decoder output signal (Figure 6) is the same as if the mixer output signal (Figure 3) were encoded and decoded by a BCC encoder/decoder in this case. Thus, also similar audio quality can be expected.

Also, in this case the decoder can not only determine the direction at which each source is to appear but also the gain of each source can be varied. The gain is increased by choosing $a_i^2 + b_i^2 > 1$ ($G_i > 0$ dB) and decreased by choosing $a_i^2 + b_i^2 < 1$ ($G_i < 0$ dB) in (16).

5. COMPUTATIONALLY EFFICIENT IMPLEMENTATION

As opposed to directly using a filterbank mimicking the frequency resolution of the auditory system, the proposed scheme can be implemented using a discrete short time Fourier transform (STFT) using an efficient fast Fourier transform (FFT) algorithm. Alternatively, other computationally efficient uniform complex filterbanks, such as a quadrature mirror filterbank (QMF) may be used. In order to mimic the non-uniform frequency resolution of the auditory system, the frequency coefficients are "grouped" such that each group (denoted "partition" in [18]) corresponds to an auditory critical band. The previously described processing is then applied to each group of coefficients. Details on this type of processing can be found in [18]. Also, it is described in [18] how to compute ICLD, ICTD, and ICC in the STFT domain.

Figure 7 shows our real-time joint-source coding decoder implementation including a mixer allowing to control amplitude panning and gain for each source. Mixing of stereo, 4-channel, 5.1 surround, and 8-channel audio signals is supported. Note that the same bitstream supports all these formats since the output signal format depends on the mixer and not on the joint-source coding scheme. The decoder and corresponding encoder are implemented using a STFT transform with similar parameters as used in [18]).

Fig. 7: The graphical user interface of our real-time decoder implementation.

6. DISCUSSION

The proposed scheme was motivated and derived using a stereo mixer with amplitude panning, delay panning, and gains for each source. For a multi-channel mixer, the inter-channel cues of the mixer output relate to the same source properties as for the stereo case. Thus, the proposed scheme is also applicable for multi-channel mixers.

Note that mechanically the side information, relative power in subbands between the sources (13), is similar to the ICLD used by a BCC (or parametric multi-channel audio coding) scheme. However, the meaning of these parameters is different. In the case of the proposed joint-source coding scheme, the parameters are relative subband power values for each source. In the case of BCC, the parameters are the inter-channel cues between audio channels important for spatial perception.

Also, the joint-coding synthesis, shown in Figure 5, at first sight looks similar to a BCC synthesis scheme. The differences are that the filters have not the purpose of reproducing an ICC parameter related to a multi-channel signal but merely the purpose of mimicking the mutual (time invariant) independence of the output source signals. Another important difference to BCC is that the joint-source coding output signals are not intended for listening but post-mixing is required.

7. INFORMAL SUBJECTIVE EVALUATION

The audio quality of the proposed scheme is clearly significantly better than for the previously proposed scheme BCC for Flexible Rendering. The hard decision "frequency mask" used in BCC for Flexible Rendering causes more problems the higher the number of sources is. But even for only two sources some artifacts are already noticeable when the sources are concurrently active and are overlapping in time and frequency.

We also experimented using a more complex mixing process, not only using panning but also effects such as reverberation. Informal listening revealed that the proposed scheme performs equally well for more complicated mixing processes as it does for a simple mixing process.

As also argued in Section 4, the audio quality of the proposed scheme is comparable to the audio quality achieved by BCC and other parametric multi-channel audio coding techniques for source gains $G_i = 0$ dB. For small source gain variations the quality is hardly impaired, while when large source gain variations are used quality degrades by a certain degree.

8. CONCLUSIONS

A scheme for joint-coding of independent audio source signals was proposed. By considering that the sources are mixed before listening to them, significant coding gain improvement can be achieved compared to the case of independently coding the sources. Only the information of the individual sources is coded which is relevant for perception of the sources after mixing. The waveform of the sum signal is used as a basis for reconstructing each of the sources.

By transmitting only the sum signal and low bitrate side information a plurality of sources is coded extremely efficiently, while maintaining at the decoder side the flexibility to mix the sources to stereo, multi-channel surround, wavefield synthesis, or binaural audio signals.

Alternatively, it is shown how to generate a mix of the sources without explicit decoding of the source signals by use of a BCC or parametric multi-channel

audio decoder. This is done by computation of the BCC parameters as a function of the side information and mixing parameters.

The audio quality of the proposed scheme is about the same as the audio quality achieved by a BCC or parametric multi-channel audio coding scheme, if the source gains are not modified. The more the source gains are modified the more potentially the audio quality degrades.

9. REFERENCES

[1] ISO/IEC, *MPEG-4 Overview*, ISO/IEC, March 2002, JTC1/SC29/WG11 N4668.

[2] Eric D. Scheirer, "Structured audio and effects processing in the MPEG-4 multimedia standard," *Multimedia Systems*, vol. 7, no. 1, pp. 11–22, 1999.

[3] ISO/IEC JTC1/SC29/WG11, *Subpart 5: MPEG-4 Structured Audio*, Oct. 1998, Final Commitee Draft FCD 14496-3: Coding of Audiovisual Objects, Part 3: Audio.

[4] A. Blumlein, "Improvements in and relating to sound transmission, sound recording and sound reproduction systems," *British Patent Specification 394325*, 1931, Reprinted in *Stereophonic Techniques*, Aud. Eng. Soc., New York, 1986.

[5] F. Rumsey, *Spatial Audio*, Focal Press, Music Technology Series, 2001.

[6] Rec. ITU-R BS.775, *Multi-Channel Stereophonic Sound System with or without Accompanying Picture*, ITU, 1993, http://www.itu.org.

[7] A. J. Berkhout, D. de Vries, and P. Vogel, "Wave front synthesis: a new direction in electroacoustics," in *Preprint 93th Conv. Aud. Eng. Soc.*, Oct. 1992.

[8] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *J. Acoust. Soc. Am.*, vol. 93, no. 5, pp. 2764–2778, May 1993.

[9] E. N. G. Verheijen, *Sound Reproduction by Wave Field Synthesis*, Ph.D. thesis, Delft University of Technology, 1997.

[10] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, The MIT Press, Cambridge, Massachusetts, USA, revised edition, 1997.

[11] D. R. Begault, *3-D Sound for Virtual Reality and Multimedia*, Academic Press, Cambridge, MA, 1994.

[12] ISO/IEC, *Generic coding of moving pictures and associated audio information – Part 7: Advanced Audio Coding*, ISO/IEC 13818-7 International Standard, 1997, JTC1/SC29/WG11.

[13] L. D. Fielder, M. Bosi, G. Davidson, M. Davis, C. Todd, and S. Vernon, "AC-2 and AC-3: Low-complexity transform-based audio coding," in *Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds., pp. 54–72. Audio Engineering Society Inc., 1996.

[14] K. Tsutsui, H. Suzuki, O. Shimoyoshi, M. Sonohara, K. Akagiri, and R. M. Heddle, "ATRAC: Adaptive transform acoustic coding for MiniDisc," in *Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds., pp. 95–101. Audio Engineering Society Inc., 1996.

[15] ISO/IEC, *Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s – Part 3: Audio*, ISO/IEC 11172-3 International Standard, 1993, JTC1/SC29/WG11.

[16] D. Sinha, J. D. Johnston, S. Dorward, and S. Quackenbush, "The perceptual audio coder (PAC)," in *The Digital Signal Processing Handbook*, V. Madisetti and D. B. Williams, Eds., chapter 42. CRC Press, IEEE Press, Boca Raton, Florida, 1997.

[17] C. Faller and F. Baumgarte, "Efficient representation of spatial audio using perceptual parametrization," in *Proc. IEEE Workshop on Appl. of Sig. Proc. to Audio and Acoust.*, Oct. 2001, pp. 199–202.

[18] C. Faller and F. Baumgarte, "Binaural Cue Coding - Part II: Schemes and applications," *IEEE Trans. on Speech and Audio Proc.*, vol. 11, no. 6, pp. 520–531, Nov. 2003.

[19] C. Faller and F. Baumgarte, "Binaural Cue Coding applied to audio compression with flexible rendering," in *Preprint 113th Conv. Aud. Eng. Soc.*, Oct. 2002.

[20] C. Faller, *Parametric Coding of Spatial Audio*, Ph.D. thesis, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland, July 2004, Thesis No. 3062, http://library.epfl.ch/theses/?nr=3062.

[21] E. Schuijers, W. Oomen, A. C. den Brinker, and A. J. Gerrits, "Advances parametric coding for high-quality audio," in *Proc. MPCA*, Nov. 2002.

[22] J. Herre, C. Faller, S. Disch, C. Ertel, J. Hilpert, A. Hoelzer, K. Linzmeier, C. Spenger, and P. Kroon, "Spatial Audio Coding: Next-generation efficient and compatible coding of multi-channel audio," in *Preprint 117th Conv. Aud. Eng. Soc.*, October 2004.

[23] 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 *Preprint 119th Conv. Aud. Eng. Soc.*, Oct. 2005.

[24] G. Theile and G. Plenge, "Localization of lateral phantom sources," *J. Audio Eng. Soc.*, vol. 25, no. 4, pp. 196–200, 1977.

[25] V. Pulkki, "Localization of amplitude-panned sources II: Two- and three-dimensional panning," *J. Audio Eng. Soc.*, vol. 49, no. 9, pp. 753–757, 2001.

[26] T. Okano, L. L. Beranek, and T. Hidaka, "Relations among interaural cross-correlation coefficient ($IACC_E$), lateral fraction (LF_E), and apparent source width (asw) in concernt halls," *J. Acoust. Soc. Am.*, vol. 104, no. 1, pp. 255–265, July 1998.

[27] M. Morimoto and Z. Maekawa, "Auditory spaciousness and envelopment," in *Proc. 13th Int. Congr. on Acoustics*, Belgrade, 1989, vol. 2, pp. 215–218.

[28] J. S. Bradley, "Comparison of concert hall measurements of spatial impression," *J. Acoust. Soc. Am.*, vol. 96, no. 6, pp. 3525–3535, 1994.

[29] K. Kurozumi and K. Ohgushi, "The relationship between the cross-correlation coefficient of two-channel acoustic signals and sound image quality), and apparent source width (asw) in concernt halls," *J. Acoust. Soc. Am.*, vol. 74, no. 6, pp. 1726–1733, Dec. 1983.

[30] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Am.*, vol. 116, no. 5, pp. 3075–3089, Nov. 2004.

[31] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, Springer, New York, 1999.

[32] 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.

[33] Y. L. Lee, *Statistical Theory of Communication*, John Wiley, New York, 1960.

[34] C. Faller and F. Baumgarte, "Binaural Cue Coding applied to stereo and multi-channel audio compression," in *Preprint 112th Conv. Aud. Eng. Soc.*, May 2002.

[35] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Am.*, vol. 50, pp. 637–655, 1971.

[36] W. B. Kleijn and K. K. Paliwal, *An Introduction to Speech Coding*, Elsevier, Amsterdam, 1995.

[37] F. K. Soong and B.-H. Juang, "Line spectrum pair (LSP) and speech data compression," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 1984, pp. 1.10.1–1.10.4.

[38] H. Lauridsen, "Nogle forsog med forskellige former rumakustic gengivelske," *Ingenioren*, vol. 47, pp. 906, 1954, (cited in Schroeder 1961 and Blauert 1997).

[39] M. R. Schroeder, "Improved quasi-stereophony and "colorless" artificial reverberation," *J. Acoust. Soc. Am.*, vol. 33, pp. 1061–1064, 1961.

[40] J. Engdegard, H. Purnhagen, J. Roden, and L. Liljeryd, "Synthetic ambience in parametric stereo coding," in *Preprint 117th Conv. Aud. Eng. Soc.*, May 2004.

[41] M. Bouŕi and C. Kyirakakis, "Audio signal decorrelation based on a critical band approach," in *Preprint 117th Conv. Aud. Eng. Soc.*, Oct. 2004, p. preprint 6291.

[42] C. Faller, "Parametric multi-channel audio coding: Synthesis of coherence cues," *IEEE Trans. on Speech and Audio Proc.*, vol. 14, no. 1, pp. 299–310, Jan. 2006.

[43] E. Schuijers, W. Oomen, B. den Brinker, and J. Breebaart, "Advances in parametric coding for high-quality audio," in *Preprint 114th Conv. Aud. Eng. Soc.*, Mar. 2003.