Audio Engineering Society
Convention Paper
Presented at the 116th Convention
2004 May 8–11 Berlin, Germany

High-quality parametric spatial audio coding at low bit rates

Jeroen Breebaart[1], Steven van de Par[1], Armin Kohlrausch,[1,2], Erik Schuijers[3]

[1]*Philips Research Laboratories Eindhoven*

[2]*Eindhoven University of Technology*

[3]*Philips Digital Systems Laboratories Eindhoven*

Correspondence should be addressed to Jeroen Breebaart (jeroen.breebaart@philips.com)

ABSTRACT

Recently, so-called binaural cue coding schemes have been introduced. These audio coding schemes transmit two perceptually relevant sound localization cues (i.e., level and time differences between the input channels), combined with a mono audio signal. Although these schemes are able to reconstruct the locations of various sound sources quite effectively, other aspects of the spatial ambience (such as the spatial diffuseness of reverberation) cannot be captured in this way. In this paper, we will present an extension to these spatial coding schemes, which comprises a spatial sound-field parameter that is able to capture ambience properties. Experiments show that the combination of three spatial parameters enables highly efficient, high-quality stereo audio representations.

1. INTRODUCTION

Efficient coding of wideband audio has gained large interest during the last decades. With the increasing popularity of mobile applications, internet and wireless communication protocols, the demand for more efficient coding systems is still sustaining. A large variety of different coding strategies and algorithms has been proposed and several of them have been incorporated in standardization bodies [1, 2]. These coding strategies reduce the required bit rate by exploiting two main principles for bit-rate reduction. The first principle is the fact that signals may exhibit redundant information. A signal may be partly predictable from its past, or the signal can be described more efficiently using a suitable set of signal functions. For example, a single sinusoid can be described by its successive time-domain samples, but a more efficient description would be to transmit its amplitude, frequency and starting phase. This source of bit-rate reduction is often referred to as 'signal redundancy'. The second principle (or source) for bit-rate reduction is the exploitation of 'perceptual irrelevancy'. Signal properties that are irrelevant from a perceptual point of view can be discarded without loss in perceptual quality. In particular, a significant amount of bit-rate reduction in current state-of-the-art audio coders is obtained by exploiting auditory masking.

If audio material consists of more than one channel, redundancies and irrelevancies can not only be exploited within each channel, but also across channels. For example, a technique referred to as 'mid-side coding' exploits the common part of a stereophonic input signal by encoding the sum and difference signals of the two input signals rather than the input signals themselves [3]. If the two input signals are sufficiently correlated, sum/difference coding requires less bits than dual-mono coding. However, some investigations have shown that the amount of mutual information in the signals for a sum/difference transform is usually rather low [4]. One possible explanation for this finding is related to the (limited) signal model, which assumes that both input signals have a common component. To be more specific, the cross-correlation coefficient (or the value of the cross-correlation function at lag zero) of the two input signals must be significantly different from zero in order to obtain a bit rate reduction. If the two input signals are (nearly) identical but have a relative time delay, the cross-correlation coefficient will (in general) be very low, despite the fact that there exists significant signal redundancy between the input signals. Such a relative time delay may result from the usage of a stereo microphone setup during the recording stage or may result from effect processors that insert delays. In this case, the cross-correlation function shows a clear maximum at a certain, non-zero delay. The maximum value of the cross-correlation as a function of the relative delay is also known as 'coherence'. Coherent signals can in principle be modeled using more advanced signal models, for example using cross-channel prediction schemes. However, studies indicate only limited success in exploiting coherence using such techniques [5, 6]. These results indicate that exploiting cross-channel *redundancies*, even if the signal model is able to capture relative time delays, does not lead to a large coding gain.

Besides cross-channel redundancies, cross-channel perceptual irrelevancies may also be exploited. For example, it is well known that for high frequencies (typically above 2 kHz), the human auditory system is not sensitive to fine-structure phase differences between the left and right signals in a stereo recording [7, 8]. This phenomenon is exploited by a technique referred to as 'intensity stereo' [9, 10]. Using this technique, a single audio signal is transmitted for the high-frequency range, combined with time- and frequency-dependent scale factors to encode level differences. More recently, so-called binaural-cue coding (BCC) schemes have been described that aim at modeling the most relevant sound source localization cues, while discarding all other spatial attributes [11, 12, 13]. These BCC schemes can be seen as an extension to intensity stereo. For the full frequency range, only a single audio channel is transmitted, combined with time and frequency-dependent differences in level and arrival time between the input channels. Although the BCC schemes are able to capture a large part of the spatial properties of a sound field, they suffer from narrowing of the stereo image and spatial instabilities [14, 15], suggesting that these techniques are mostly advantageous at low bit rates [16]. A solution that we suggest to reduce the narrowing stereo image artefact is to transmit the inter-channel coherence as a third parameter (cf. [17]).

In this paper, a parametric description of the spatial soundfield will be presented which is based on the three spatial properties described above (i.e., level differences, time differences, and the coherence). The analysis, encoding and synthesis of these parameters is heavily based on binaural psychoacoustics. Furthermore, it will be

shown that the presented parameterization of the stereo image is able to achieve a significant bit-rate reduction if it is applied as part of a stereo audio coder. The amount of spatial information that is extracted and transmitted is scalable; at low parameter rates (typically in the order of 1 to 3 kbits/s), the coder is able to represent the most important stereo cues. For these bit rates, sound source positioning is expected to be similar as for BCC schemes, but without spatial narrowing artefacts, given the fact that the perceived 'width' is captured in the coherence parameter. This coder configuration targets at stereo coding at extremely low bitrates (typically less than 32 kbits/s), where 'transparent' coding is virtually impossible. But besides the gain obtained at such low bitrates, it will also be demonstrated that if the spatial parameter bit rate is increased up to about 8 kbits/s, a spatial parameterization system is obtained with a quality which is equivalent to current high-quality stereo audio coders (such as MPEG-1 layer 3 at a bit rate of 128 kbits/s). The bit-rate scalability options and the fact that a high-quality stereo image can be obtained enables integration of parametric stereo in state-of-the-art transform-based [18, 19] and parametric [17] mono audio coders for a wide quality/bitrate range.

The paper outline is as follows. First the psychoacoustic background of the parametric stereo coder is discussed. Section 3 discusses the general structure of the coder. In Section 4 the encoder is described in detail. In section 5, the corresponding decoder is outlined. In section 6, results from a listening test are discussed, followed by a last concluding section.

2. PSYCHOACOUSTIC BACKGROUND

In 1907, Lord Rayleigh formulated the duplex theory [20], which states that sound source localization is facilitated by interaural intensity differences (IIDs) at high frequencies and by interaural time differences (ITDs) at low frequencies. This theory was (in part) based on the observation that at low frequencies, IIDs between the eardrums do not occur due to the fact that the signal wavelength is much larger than the size of the head, and hence the acoustical shadow of the head is virtually absent. According to Lord Rayleigh, this had the consequence that human listeners can only use ITD cues for sound source localization at low frequencies. Since then, a large amount of research has been conducted to investigate the human sensitivity to both IIDs and ITDs as a function of various stimulus parameters. One of the striking findings is that although it seems that IID cues are virtually absent at low frequencies for free-field listening conditions, humans are nevertheless very sensitive to IID and ITD cues at low *and* high frequencies. Stimuli with specified values of the ITD and IID can be presented over headphones, resulting in a lateralization of the sound source which depends on the magnitude of the ITD or IID [21, 22, 23]. The usual result of such laboratory headphone-based experiments is that the source images are located inside the head, and are lateralized along the axis connecting the left and the right ears. In order to change the position of the lateralized image, the binaural cues have to change by a certain minimum amount. This observation reflects the limited spatial resolution of the human auditory system. For IID cues, the resolution is between 0.5 and 1 dB for a reference IID of 0 dB and is relatively independent of frequency and stimulus level [24, 25, 26, 27]. If the reference IID increases, IID thresholds increase also. For reference IIDs of 9 dB, the IID threshold is about 1.2 dB, and for a reference IID of 15 dB, the IID threshold amounts between 1.5 and 2 dB [28, 29, 30].

The sensitivity to changes in ITDs strongly depends on frequency. For frequencies below 1000 Hz, this sensitivity can be described as a constant interaural phase difference (IPD) sensitivity of about 0.05 rad [7, 31, 25, 32]. At higher frequencies, the binaural auditory system is not able to detect time differences in the fine-structure waveforms. However, time differences in the envelopes can be detected quite accurately [33, 34]. The reference ITD has some effect on the ITD thresholds: large ITDs in the reference condition tend to decrease sensitivity to changes in the ITDs [24, 35]. There is almost no effect of stimulus level on ITD sensitivity [8].

If multiple sound sources at different spatial positions are presented simultaneously or sequentially, or with the occurence of head movements, the localization cues vary across frequency and time. Extensive psychophysical research (cf. [36, 37, 38]) and efforts to model the binaural auditory system (cf. [39, 40, 41, 42, 43]) have suggested that the human auditory system extracts IID and ITD cues as a function of time and frequency. To be more specific, there is considerable evidence that the binaural auditory system renders its binaural cues in a set of frequency bands, without having the possibility to acquire these properties at a finer frequency resolution. This spectral resolution of the binaural auditory system can be described by a filter bank with filter bandwidths that follow the ERB (Equivalent Rectangular Bandwidth)

scale [44, 45, 46].

The limited temporal resolution at which the auditory system can track binaural localization cues is often referred to as 'binaural sluggishness', and the associated time constants are between 30 and 100 milliseconds [47, 38]. Although the auditory system is not able to *follow* IIDs and ITDs that vary quickly over time, this does not mean that listeners are not able to detect the *presence* of quickly varying cues. Slowly-varying IIDs and/or ITDs result in a movement of the perceived sound source location, while fast changes in binaural cues lead to a percept of 'spatial diffuseness', or a reduced 'compactness' [48]. Despite the fact that the perceived 'quality' of the presented stimulus depends on the movement speed of the binaural cues, it has been shown that the detectability of IIDs and ITDs is practically *independent* of the variation speed [49]. The sensitivity of human listeners to time-varying changes in binaural cues can be described by sensitivity to changes in the maximum of the cross-correlation function (e.g., the *coherence*) of the incoming waveforms [50, 51, 52, 53]. There is considerable evidence that the sensitivity to changes in the coherence is the basis of the phenomenon of the Binaural Masking Level Difference (BMLD) [54, 55, 56]. Moreover, the sensitivity to quasi-static ITDs can also be described by the (changes in the) cross-correlation function [41, 57, 42].

There is, however, one important exception in which the auditory system does not seem to integrate spatial information across time. In reverberant rooms, the perceived location of a sound source is dominated by the first 2 milliseconds of the onset of the sound source, while the remaining signal is largely discarded in terms of spatial cues. This phenomenon is referred to as 'the law of the first wavefront' or 'precedence effect' [58, 59, 60, 61].

The sensitivity to changes in the coherence strongly depends on the reference coherence. For a reference coherence of +1 (for example when two signals are identical), changes of about 0.002 can be perceived, while for a reference coherence around 0, the change in coherence must be about 100 times larger to be perceptible [62, 63, 64, 65]. The sensitivity to interaural coherence is practically independent of stimulus level, as long as the stimulus is sufficiently above the absolute threshold [66]. At high frequencies, the *envelope* coherence seems to be the relevant descriptor of the spatial diffuseness [67, 53].

Recently, it has been demonstrated that the concept of 'spatial diffuseness' mostly depends on the coherence value itself and is relatively unaffected by the temporal fine-structure details of the coherence within the temporal integration time of the binaural auditory system. For example, van de Par *et al.* [68] measured the detection and discrimination thresholds for two types of test signals presented in an interaurally in-phase masker. The masker consisted of a broadband noise with a certain bandwidth. Two different test signal types were used: the first test signal type was a band-pass noise, while the second test signal type consisted of a harmonic tone complex having the same bandwidth and signal level as the band-pass noise. In different experiments, the test signal was presented interaurally in-phase (to facilitate monaural sensitivity measurements) as well as interaurally out-of-phase (to investigate the binaural sensitivity). Their results can be summarized as follows:

- In all conditions, an interaurally out-of-phase test signal resulted in considerably lower detection thresholds than the in-phase signal, indicating a binaural masking level difference (BMLD) of up to 15 dB.
- Subjects had large difficulty to discriminate between an interaurally out-of-phase noise and an interaurally out-of-phase harmonic tone complex. This result suggests that although human listeners may be very sensitive to the *presence* of an out-of-phase signal (i.e., a change in the coherence), they are not able to *indicate* the nature of the test signal (i.e., a noise sequence or a harmonic tone complex).

These findings suggest that it is sufficient from a perceptual point of view to describe a stereo audio signal with 1) a mono (or dominant) signal with a certain spatial location which is parameterized by time and frequency dependent IID and ITD cues, and 2) a side (or residual) signal which induces a certain (in)coherence upon the output signals, of which the fine-structure details are relatively unimportant. More specifically, it seems that the side signal can be replaced by another side signal, as long as their spectro-temporal envelopes are equal in terms of the time and frequency resolution of the human auditory system. This observation forms the basis for the parametric coding scheme as presented below.

3. CODER OVERVIEW

The generic structure of the parametric stereo encoder is shown in Fig. 1. The two input channels are fed to a stage

that extracts spatial parameters and generates a mono downmix of the two input channels. The spatial parameters are subsequently quantized and encoded, while the mono downmix is encoded using an arbitrary mono audio coder. The resulting mono bit stream is combined with the encoded spatial parameters to form the output bit stream. If the spatial parameters are transmitted in the ancillary part of the bitstream, ensuring backwards (mono) compabitility.

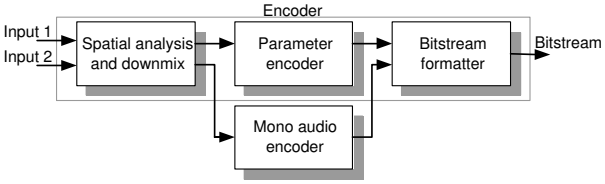

Fig. 1: *Structure of the parametric stereo encoder. The two input signals are first processed by a parameter extraction and downmix stage. The parameters are subsequently quantized and encoded, while the mono downmix can be encoded using an arbitrary mono audio coder. The mono bit stream and spatial parameters are subsequently combined into a single output bit stream.*

The parametric stereo decoder basically performs the reverse process, as shown in Fig. 2. The spatial parameters are separated from the incoming bit stream and decoded. The mono bit stream is decoded using a mono audio decoder. The decoded audio signal is fed into the spatial synthesis stage, which reinstates the spatial image, resulting in a two-channel output.

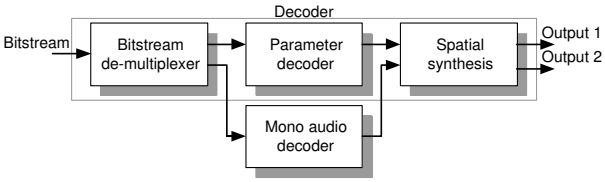

Fig. 2: *Structure of the parametric decoder. The demultiplexer splits mono and spatial parameter information. The mono audio signal is decoded and fed into the spatial synthesis stage, which reinstates the spatial cues based on the decoded spatial parameters.*

The spatial parameters are estimated (at the encoder side) and reinstated (at the decoder side) as a function of time and frequency. Therefore, both the encoder and decoder require a transform or filterbank that generates individual time/frequency tiles. The frequency resolution of this stage should be non-uniform according to the frequency resolution of the human auditory system. Furthermore, the temporal resolution should generally be fairly low (in the order of tens of milliseconds) reflecting the concept of binaural sluggishness, except in the case of transients, where the precedence effect dictates a time resolution of only a few milliseconds. Furthermore, the transform or filter bank should be oversampled, since time- and frequency-dependent changes will be made to the signals which would lead to audible aliasing distortion in a critically-sampled system. Finally, a complex-valued transform or filter bank is preferred to enable easy estimation and modification of (inter-channel) phase information. A process that meets these requirements is a variable segmentation process with temporally overlapping segments, followed by forward and inverse FFTs. Complex-modulated filter banks can be employed as a low-complexity alternative [18, 19].

4. ENCODER

The spatial analysis and downmix stage of the encoder is shown in more detail in Fig. 3. The two input signals $x_1[n]$ and $x_2[n]$ (with n the sample index) are first segmented using variable segmentation to account for the precedence effect in the case of a transient. Subsequently, each windowed segment is transformed to the frequency domain using a Fast Fourier Transform (FFT). The frequency-domain input signals $X_1[k], X_2[k]$, with k the FFT bin index ($k = [0..N/2]$) and N the FFT length in samples are divided into 34 subbands b by grouping of the FFT bins. The frequency bands are formed in such a way that each band with center frequency f (in Hz) has a bandwidth, BW (in Hz), which is approximately equal to the Equivalent Rectangular Bandwidth (ERB) scale [46], following:

$$BW = 24.7\left(0.00437 f + 1\right). \tag{1}$$

The groups of FFT bins are subsequently used to extract spatial parameters and to generate a mono downmix signal. The mono signal is transformed to the time domain using an inverse FFT, followed by windowing and overlap-add. More information with respect to the segmentation and frequency separation processes can be found in [69].

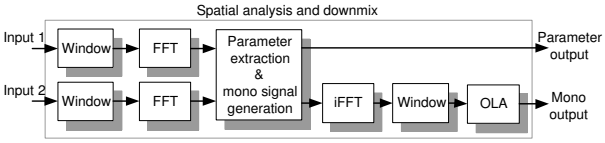

Fig. 3: *Spatial analysis and downmix stage of the encoder.*

4.1. Parameter extraction

For each frequency band b, three spatial parameters are computed. The first parameter is the interchannel intensity difference (IID[b]), defined as the logarithm of the power ratio of corresponding subbands from the input signals:

$$\mathrm{IID}[b] = 10\log_{10}\frac{\sum_{k=k_b}^{k_{b+1}-1} X_1[k]X_1^*[k]}{\sum_{k=k_b}^{k_{b+1}-1} X_2[k]X_2^*[k]}, \qquad (2)$$

where * denotes complex conjugation. The second parameter is the relative phase rotation. The phase rotation aims at optimal (in terms of correlation) phase alignment between the two signals. This parameter is denoted by the interchannel phase difference (IPD[b]) and is obtained as follows:

$$\mathrm{IPD}[b] = \angle\left(\sum_{k=k_b}^{k_{b+1}-1} X_1[k]X_2^*[k]\right). \qquad (3)$$

IPD parameters are only transmitted for frequency bands up to about 2 kHz, given the fact that human listeners are insensitive to interaural fine-structure phase differences above that frequency.

The third parameter is the interchannel coherence (IC[b]), which is, in our context, defined as the normalized cross-correlation coefficient after phase alignment according to the IPD. The coherence is derived from the cross-spectrum in the following way:

$$\mathrm{IC}[b] = \frac{\left|\sum_{k=k_b}^{k_{b+1}-1} X_1[k]X_2^*[k]\right|}{\sqrt{\left(\sum_{k=k_b}^{k_{b+1}-1} X_1[k]X_1^*[k]\right)\left(\sum_{k=k_b}^{k_{b+1}-1} X_2[k]X_2^*[k]\right)}}. \qquad (4)$$

4.2. Downmix and parameter quantization

A suitable mono signal $S[k]$ is obtained by a linear combination of the input signals $X_1[k]$ and $X_2[k]$:

$$S[k] = w_1X_1[k] + w_2X_2[k], \qquad (5)$$

where w_1 and w_2 are weights that determine the relative amount of X_1 and X_2 in the mono output signal. For $w_1 = w_2 = 0.5$, the output will consist of the average of the two input signals.

After the mono signal is generated, the last parameter that has to be extracted is computed. The IPD parameter as described above specifies the *relative* phase difference between the stereo input signals (and hence the stereo output signals at the decoder side). Thus the IPD does not indicate how the decoder should distribute these phase differences across the output channels. To signal the actual distribution of phase modifications to yield a relative IPD, an overall phase difference (OPD) is computed and transmitted. To be more specific, the decoder applies a phase modification equal to the OPD to compute the first output signal, and applies a phase modification of the OPD minus the IPD to obtain the second output signal. Given this specification, the OPD is computed as the average phase difference between $X_1[k]$ and $S[k]$, according to

$$\mathrm{OPD}[b] = \angle\left(\sum_{k=k_b}^{k=k_{b+1}-1} X_1[k]S^*[k]\right). \qquad (6)$$

Subsequently, the mono signal $S[k]$ is transformed to the time domain using an inverse FFT. Finally, a synthesis window is applied to each segment followed by overlap-add, resulting in the desired mono output signal. The IID, IPD, OPD and IC parameters are quantized according to perceptual criteria. The quantization process aims at introducing quantization errors which are just inaudible. See [69] for more information on quantization and

coding. The resulting entropy per symbol, using modulo-differential coding, averaged across a large set of audio material, and the resulting contribution to the overall bit rate are given in Table 1. The values are determined for a parameter update rate of 23 ms.

The total estimated parameter bit rate for the configuration as described above, excluding bit-stream overhead and averaged across a large amount of representative stereo material, amounts to 7.7 kbits/s. However, by reducing the number of frequency bands, the parameter update rate and the quantization accuracy, the parameter bitrate can be scaled down to approximately 1 kbits/s (see [69]).

5. DECODER

The spatial synthesis part of the decoder receives a mono input signal $s[n]$ and has to generate two output signals $y_1[n]$ and $y_2[n]$. These two output signals should obey the transmitted spatial parameters. A more detailed overview of the spatial synthesis stage is shown in Fig. 4.

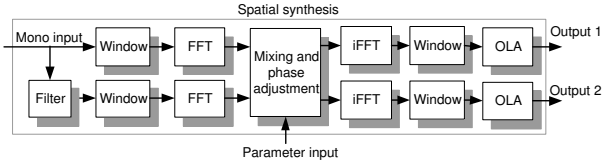

Fig. 4: *Spatial synthesis stage of the decoder.*

In order to generate two output signals with a variable (i.e., parameter dependent) coherence, a second signal has to be generated which has a similar spectral-temporal envelope as the mono input signal, but is incoherent (i.e., orthogonal) from a fine-structure waveform point of view. This incoherent signal, $s_d[n]$ is obtained by convolving the mono input signal $s[n]$ with an allpass decorrelation filter $h_d[n]$. A very cost-effective all-pass filter is obtained by a simple delay. However, since the filtered and original signal will be mixed in the final output, a fixed delay will result in harmonic comb-filter effects in the output. To prevent harmonically-related comb-filter peaks and troughs in the output, a frequency-dependent delay is used as decorrelation filter. The decorrelation filter consists of a single period of a positive Schroeder-phase complex [70] of length $N_s = 640$ (i.e., with a fundamental frequency of f_s/N_s). The filter's impulse response $h_d[n]$ for $0 \leq n \leq N_s - 1$ is given by:

$$h_d[n] = \sum_{k=0}^{N_s/2} \frac{2}{N_s} \cos\left(\frac{2\pi kn}{N_s} + \frac{2\pi k(k-1)}{N_s}\right). \quad (7)$$

Subsequently, the segmentation, windowing and transform operations that are performed in the decoder match those that were applied in the encoder, resulting in the frequency-domain representations $S[k]$ and $S_d[k]$, for the mono input signal $s[n]$ and its decorrelated version $s_d[n]$, respectively. The next step consists of computing linear combinations of the two input signals to arrive at the two frequency-domain output signals $Y_1[k]$ and $Y_2[k]$. The mixing process, which is performed on a subband basis, is described by the following matrix multiplication. For each subband b, we have

$$\begin{bmatrix} Y_1[k] \\ Y_2[k] \end{bmatrix} = \mathbf{P}[b]\mathbf{R}[b]\mathbf{G}[b] \begin{bmatrix} S[k] \\ S_d[k] \end{bmatrix}. \quad (8)$$

The diagonal matrix $\mathbf{G}$ enables scaling of the two orthogonal signals $S[k]$ and $S_d[k]$. The matrix $\mathbf{R}$ is a rotation in the two-dimensional signal space, i.e., $\mathbf{R}^{-1} = \mathbf{R}^T$, and the diagonal matrix $\mathbf{P}$ enables modification of the complex-phase relationships between the output signals, hence $|p_{ij}| = 1$ for $i = j$ and 0 otherwise.

It can be shown that the following solution for $\mathbf{P}$, $\mathbf{R}$ and $\mathbf{G}$ results in output signals that match the spatial parameters:

$$\mathbf{P}[b] = \begin{bmatrix} e^{j\mathrm{OPD}[b]} & 0 \\ 0 & e^{j\mathrm{OPD}[b]-j\mathrm{IPD}[b]} \end{bmatrix}, \quad (9)$$

$$\mathbf{R}[b] = \begin{bmatrix} \cos(\alpha[b]) & -\sin(\alpha[b]) \\ \sin(\alpha[b]) & \cos(\alpha[b]) \end{bmatrix}, \quad (10)$$

$$\mathbf{G}[b] = \sqrt{2} \begin{bmatrix} \cos(\gamma[b]) & 0 \\ 0 & \sin(\gamma[b]) \end{bmatrix}, \quad (11)$$

with $\alpha[b]$ being a rotation angle in the two-dimensional signal space, which is given by:

$$\alpha[b] = 0.5 \arctan\left(\frac{2c[b](\mathrm{IC})[b]}{c[b]^2 - 1}\right), \quad (12)$$

and $\gamma[b]$ a parameter for relative scaling of S and S_d:

Parameter	Bits/symbol	symbols/second	bit rate (bits/s)
IID	1.94	1464	2840
IPD	1.58	732	1157
OPD	1.31	732	959
IC	1.88	1464	2752
Total			7708

Table 1: *Entropy per parameter symbol, number of symbols per second and bit rate for spatial parameters averaged across a large set of excerpts.*

$$\gamma[b] = \arctan\sqrt{\frac{1-\sqrt{\mu[b]}}{1+\sqrt{\mu[b]}}}, \tag{13}$$

with

$$\mu[b] = 1 + \frac{4\mathrm{IC}^2[b]-4}{(c[b]+1/c[b])^2}, \tag{14}$$

and $c[b]$ the RMS ratio of the two subband output signals:

$$c[b] = 10^{\mathrm{IID}[b]/20}. \tag{15}$$

The mixing process is visualized in Fig. 5. The two orthogonal input signals S and S_d are first scaled and subsequently rotated by an angle α to obtain the two output signals Y_1 and Y_2.

Finally, the frames are transformed to the time domain and windowed (using equal synthesis windows as in the encoder), and combined using overlap-add to result in continuous time-domain output signals.

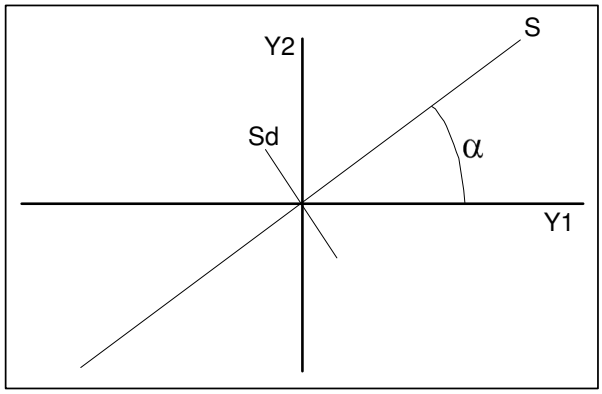

Fig. 5: *Graphic representation of the mixing process which is performed in the decoder. The two output signals Y_1 and Y_2 consist of scaled and subsequently rotated versions of the two orthogonal signals S and S_d.*

6. PERCEPTUAL EVALUATION

To evaluate the parametric stereo coder, a listening test has been conducted. Nine well-trained listeners participated in the experiment. In a double-blind MUSHRA test [71], 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 labels. All excerpts were presented over Stax Lambda Pro headphones. The processed items included:

1. Encoding and decoding using a state-of-the-art MPEG-1 layer 3 (MP3) coder at a bit rate of 128 kbit/s stereo and using its highest possible quality settings.
2. Encoding and decoding using the FFT-based parametric stereo coder as described above without mono coder (i.e., assuming transparent mono coding) operating at 8 kbits/s.
3. Encoding and decoding using the FFT-based parametric stereo coder without mono coder operating at a bit rate of 5 kbits/s (using 20 analysis frequency bands instead of 34).
4. The original as hidden reference.

The 13 test excerpts are listed in Table 2. All items are stereo, 16 bits resolution per sample, at a sampling frequency of 44.1 kHz.

The subjects could listen to each excerpt as often as they liked and could switch in real time between the four versions of each item. From a large set of test items, 13

Item index	Name	Origin/artist
1	Starship Trooper	Yes
2	Day tripper	the Beatles
3	Eye in the sky	Alan Parsons
4	Harpsichord	MPEG si01
5	Castanets	MPEG si02
6	Pitch pipe	MPEG si03
7	Glockenspiel	MPEG sm02
8	Plucked string	MPEG sm03
9	Yours is no disgrace	Yes
10	Man in the long black coat	Bob Dylan
11	Vogue	Madonna
12	Applause	SQAM disc
13	Two voices	left = MPEG es03 = english female right = MPEG es02 = german male

Table 2: *Description of test material.*

excerpts that showed to be the most critical items for parametric stereo were selected. These items had a duration of about 10 seconds and contained a large variety of audio classes. The average scores of all subjects are shown in Fig. 6. The top panel shows mean opinion scores (MOS) for 8 kbit/s parametric stereo (black bars) and MP3 at 128 kbit/s (white bars) as a function of the test item. The right-most bars indicate the mean across all test excerpts. Most excerpts show very similar scores, except for excerpts 4, 8, 10 and 13. Excerpts 4 ('Harpsichord') and 8 ('Plucked string') show a significantly higher quality for parametric stereo, while excerpts 10 ('Man in the long black coat') and 13 ('Two voices') have higher scores for MP3. On average, both coders have equal scores.

The middle panel shows results for the parametric stereo coder working at 5 kbit/s (black bars) and 8 kbit/s (white bars). In most cases, the 8 kbit/s coder has a higher quality than the 5 kbit/s coder, except for excerpts 5 ('Castanets') and 7 ('Glockenspiel'). On average, the quality of the 5 kbit/s coder is only marginally lower than for 8 kbit/s, which demonstrates the shallow bit-rate/quality slope for the parametric stereo coder.

The bottom panel shows 128 kbit/s MP3 (white bars) against the hidden reference (black bars). As expected, the hidden reference scores are close to 100. For fragments 7 ('Glockenspiel') and 10 ('Man in the long black coat'), the hidden reference scores lower than MP3 at 128 kbits/s, which indicates transparent coding.

It is important to note that the results described here were obtained for headphone listening conditions. We have found that headphone listening conditions are much more critical for parametric stereo than playback using loudspeakers. In fact, a listening test has shown that on average, the difference in MOS between headphones and loudspeaker playback is 17 points in favor of loudspeaker playback using an 8-kbits/s parameter bitstream. This means that the perceptual quality for loudspeaker playback has an average MOS of over 90, indicating excellent perceptual quality. The difference between these playback conditions is most probably the result of the combination of an unnaturally large channel separation which is obtained using headphones on the one hand, and crosstalk resulting from the downmix procedure on the other hand. It seems that the amount of crosstalk that is inherently introduced by transmission of a single audio channel only is less than the amount of crosstalk that occurs in free-field listening conditions. A consequence of this observation is that a comparison of the present coder with BCC schemes is rather difficult, since the BCC algorithms were all tested under sub-critical conditions using loudspeaker playback (cf. [14, 13, 15, 12, 16]).

7. CONCLUSIONS

We have described a parametric stereo coder which enables stereo coding using a mono audio channel and spatial parameters. Depending on the desired spatial quality, the spatial parameters require between 1 and 8 kbits/s.

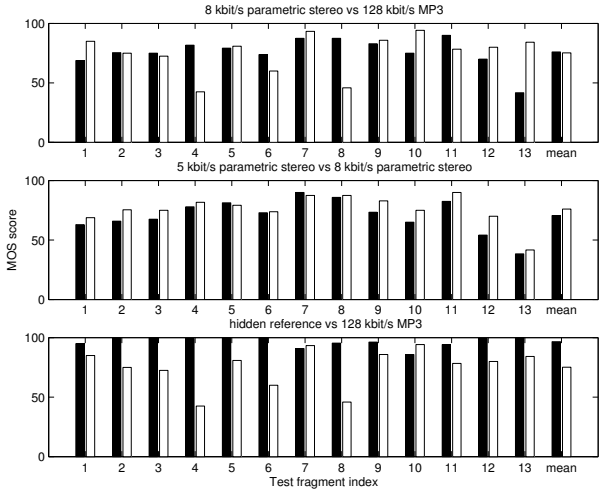

Fig. 6: *Mean opinion scores (MOS) averaged across listeners as a function of test item and various coder configurations (see text). The upper panel shows the results for 8 kbits/s parametric stereo (black bars) against stereo MP3 at 128 kbits/s (white bars). The middle panel shows the results for 5 kbits/s parametric stereo (black bars) vs. 8 kbits/s parametric stereo (white bars). The lower panel shows the hidden reference (black bars) vs. MP3 at 128 kbits/s (white bars).*

It has been demonstrated that for headphone playback, a spatial parameter bit stream of 8 kbits/s is sufficient to reach a quality level that is comparable to popular coding techniques currently on the market (i.e., MPEG-1 layer 3). Furthermore, it has been shown that a reduction in bit rate from 8 to 5 kbits/s results, on average, in only a minor quality degradation.

8. ACKNOWLEDGMENTS

We would like to thank our colleagues at Philips Digital Systems Labs for carrying out the listening tests. We would like to thank our colleagues Michel van Loon and Gerard Hotho for their useful suggestions for improving the manuscript.

9. REFERENCES

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

[2] K. Brandenburg. Mp3 and aac explained. In *Proceedings of the 17th International AES Conference, Florence, Italy*, 1999.

[3] J. D. Johnston and A. J. Ferreira. Sum-difference stereo transform coding. In *Proc. ICASSP*, San-Francisco, 1992.

[4] R. G. van der Waal and R. N. J. Veldhuis. Subband coding of stereophonic digital audio signals. In *Proc. ICASSP*, Toronto, 1991.

[5] S. S. Kuo and J. D. Johnston. A study of why cross channel prediction is not applicable to perceptual audio coding. *IEEE signal processing letters*, 8:245–247, 2001.

[6] T. Liebchen. Lossless audio coding using adaptive multichannel prediction. In *113th AES convention, Los Angeles, USA*, 2002.

[7] R. G. Klumpp and H. R. Eady. Some measurements of interaural time difference thresholds. *J. Acoust. Soc. Am.*, 28:859–860, 1956.

[8] J. Zwislocki and R. S. Feldman. Just noticeable differences in dichotic phase. *J. Acoust. Soc. Am.*, 28:860–864, 1956.

[9] J. D. Johnston and K. Brandenburg. Wideband coding - perceptual considerations for speech and music. In S. Furui and M.M. Sondhi, editors, *Advances in speech signal processing*, chapter 4, pages 109–140. Marcel Dekker, Inc., New York, Basel, Honkong, 1992.

[10] J. Herre, K. Brandenburg, and D. Lederer. Intensity stereo coding. In *Preprint 3799, 96th AES convention*, 1994.

[11] C. Faller and F. Baumgarte. Efficient representation of spatial audio using perceptual parameterization. In *WASPAA, workshop on applications of signal processing on audio and acoustics*, 2001.

[12] C. Faller and F. Baumgarte. Binaural cue coding: A novel and efficient representation of spatial audio. In *Proc. ICASSP*, 2002.

[13] F. Baumgarte and C. Faller. Design and evaluation of binaural cue coding schemes. In *Proceedings of the 113th AES convention*, 2002.

[14] F. Baumgarte and C. Faller. Why binaural cue coding is better than intensity stereo coding. In *Preprint 5575, 112th AES convention, Munich (D)*, 2002.

[15] F. Baumgarte and C. Faller. Estimation of auditory spatial cues for binaural cue coding. In *Proc. ICASSP*, 2002.

[16] C. Faller and F. Baumgarte. Binaural cue coding applied to stereo and multi-channel audio compression. In *Preprint 5574, 112th AES convention, Munich (D)*, 2002.

[17] E. Schuijers, W. Oomen, B. den Brinker, and J. Breebaart. Advances in parametric coding for high-quality audio. In *Preprint 5852, 114th AES convention, Amsterdam, The Netherlands*, 2003.

[18] E. Schuijers, J. Breebaart, H. Purnhagen, and J. Engdegård. Low complexity parametric stereo coding. In *Proc. 116th AES convention, Berlin, Germany*, 2004.

[19] J. Engdegård, H. Purnhagen, J. Rödén, and L. Liljeryd. Synthetic ambience in parametric stereo coding. In *Proc. 116th AES convention, Berlin, Germany*, 2004.

[20] Strutt (Lord Rayleigh). On our perception of sound direction. *Philos. Mag.*, 113:214–232, 1907.

[21] B. Sayers. Acoustic image lateralization judgments with binaural tones. *J. Acoust. Soc. Am.*, 36:923–926, 1964.

[22] E. R. Hafter and S. C. Carrier. Masking-level differences obtained with pulsed tonal maskers. *J. Acoust. Soc. Am.*, 47:1041–1047, 1970.

[23] W. A. Yost. Lateral position of sinusoids presented with interaural intensive and temporal differences. *J. Acoust. Soc. Am.*, 70:397–409, 1981.

[24] R.M. Hershkowitz and N.I. Durlach. Interaural time and amplitude jnds for a 500-hz tone. *J. Acoust. Soc. Am.*, 46:1464–1467, 1969.

[25] D. McFadden, L. A. Jeffress, and H. L. Ermey. Difference in interaural phase and level in detection and lateralization: 250 Hz. *J. Acoust. Soc. Am.*, 50:1484–1493, 1971.

[26] W. A. Yost. Weber's fraction for the intensity of pure tones presented binaurally. *Percept. Psychophys.*, 11:61–64, 1972.

[27] D. W. Grantham. Interaural intensity discrimination: insensitivity at 1000 Hz. *J. Acoust. Soc. Am.*, 75:1191–1194, 1984.

[28] A.W. Mills. Lateralization of high-frequency tones. *J. Acoust. Soc. Am.*, 32:132–134, 1960.

[29] R.C. Rowland Jr and J.V. Tobias. Interaural intensity difference limen. *J. Speech Hear. Res.*, 10:733–744, 1967.

[30] W. A. Yost and E. R. Hafter. *Lateralization*. Yost & Gourevich, 1991.

[31] L. A. Jeffress and D. McFadden. Differences of interaural phase and level in detection and lateralization. *J. Acoust. Soc. Am.*, 49:1169–1179, 1971.

[32] W. A. Yost, D. W. Nielsen, D. C. Tanis, and B. Bergert. Tone-on-tone binaural masking with an antiphasic masker. *Percept. Psychophys.*, 15:233–237, 1974.

[33] S. van de Par and A. Kohlrausch. A new approach to comparing binaural masking level differences at low and high frequencies. *J. Acoust. Soc. Am.*, 101:1671–1680, 1997.

[34] L. R. Bernstein and C. Trahiotis. The effects of signal duration on NoSo and NoSπ thresholds at 500 Hz and 4 kHz. *J. Acoust. Soc. Am.*, 105:1776–1783, 1999.

[35] W. A. Yost. Discrimination of interaural phase differences. *J. Acoust. Soc. Am.*, 55:1299–1303, 1974.

[36] B. Kollmeier and I. Holube. Auditory filter bandwidths in binaural and monaural listening conditions. *J. Acoust. Soc. Am.*, 92:1889–1901, 1992.

[37] Marcel van der Heijden and C. Trahiotis. Binaural detection as a function of interaural correlation and bandwidth of masking noise: Implications for estimates of spectral resolution. *J. Acoust. Soc. Am.*, 103:1609–1614, 1998.

[38] I. Holube, M. Kinkel, and B. Kollmeier. Binaural and monaural auditory filter bandwidths and time constants in probe tone detection experiments. *J. Acoust. Soc. Am.*, 104:2412–2425, 1998.

[39] H. S. Colburn and N. I. Durlach. Models of binaural interaction. In E.C. Carterette and M.P. Friedman, editors, *Handbook of Perception*, volume IV: Hearing, pages 467–518. Academic Press, New York, San Francisco, London, 1978.

[40] W. Lindemann. Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals. *J. Acoust. Soc. Am.*, 80:1608–1622, 1986.

[41] R. M. Stern, A. S. Zeiberg, and C. Trahiotis. Lateralization of complex binaural stimuli: A weighted-image model. *J. Acoust. Soc. Am.*, 84:156–165, 1988.

[42] W. Gaik. Combined evaluation of interaural time and intensity differences: psychoacoustic results and computer modeling. *J. Acoust. Soc. Am.*, 94:98–110, 1993.

[43] J. Breebaart, S. van de Par, and A. Kohlrausch. Binaural processing model based on contralateral inhibition. I. Model setup. *J. Acoust. Soc. Am.*, 110:1074–1088, 2001.

[44] J. W. Hall and M. A. Fernandes. The role of monaural frequency selectivity in binaural analysis. *J. Acoust. Soc. Am.*, 76:435–439, 1984.

[45] A. Kohlrausch. Auditory filter shape derived from binaural masking experiments. *J. Acoust. Soc. Am.*, 84:573–583, 1988.

[46] B. R. Glasberg and B. C. J. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47:103–138, 1990.

[47] B. Kollmeier and R. H. Gilkey. Binaural forward and backward masking: evidence for sluggishness in binaural detection. *J. Acoust. Soc. Am.*, 87:1709–1719, 1990.

[48] J. Blauert. *Spatial hearing: the psychophysics of human sound localization*. the MIT Press, Cambridge, Massachusetts, 1997.

[49] J. Breebaart, S. van de Par, and A. Kohlrausch. The contribution of static and dynamically varying ITDs and IIDs to binaural detection. *J. Acoust. Soc. Am.*, 106:979–992, 1999.

[50] L. A. Jeffress. A place theory of sound localization. *J. Comp. Physiol. Psych.*, 41:35–39, 1948.

[51] H. S. Colburn. Theory of binaural interaction based on auditory-nerve data. II. Detection of tones in noise. *J. Acoust. Soc. Am.*, 61:525–533, 1977.

[52] R. M. Stern and G. D. Shear. Lateralization and detection of low-frequency binaural stimuli: Effects of distribution of internal delay. *J. Acoust. Soc. Am.*, 100:2278–2288, 1996.

[53] L. R. Bernstein and C. Trahiotis. The normalized correlation: accounting for binaural detection across center frequency. *J. Acoust. Soc. Am.*, 100:3774–3787, 1996.

[54] N. I. Durlach. Equalization and cancellation theory of binaural masking-level differences. *J. Acoust. Soc. Am.*, 35:1206–1218, 1963.

[55] D. M. Green. Signal-detection analysis of equalization and cancellation model. *J. Acoust. Soc. Am.*, 40:833–838, 1966.

[56] M. van der Heijden and C. Trahiotis. Masking with interaurally delayed stimuli: the use of internal delays in binaural detection. *J. Acoust. Soc. Am.*, 105:388–399, 1999.

[57] T.M. Shackleton, R. Meddis, and M.J. Hewitt. Across frequency integration in a model of lateralization. *J. Acoust. Soc. Am.*, 92:2276–2279, 1992.

[58] H. Wallach, E. B. Newman, and M. R. Rosenzweig. The precedence effect in sound localization. *AM. J. Psychol.*, 62:315–336, 1949.

[59] P. M. Zurek. The precedence effect and its possible role in the avoidance of interaural ambiguities. *J. Acoust. Soc. Am.*, 67:952–964, 1980.

[60] B. G. Shinn-Cunningham, P. M. Zurek, and N. I. Durlach. Adjustment and discrimination measurements of the precedence effect. *J. Acoust. Soc. Am.*, 93:2923–2932, 1993.

[61] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman. The precedence effect. *J. Acoust. Soc. Am.*, 106:1633–1654, 1999.

[62] D.E. Robinson and L.A. Jeffress. Effect of varying the interaural noise correlation on the detectability of tonal signals. *J. Acoust. Soc. Am.*, 35:1947–1952, 1963.

[63] T. L. Langford and L. A. Jeffress. Effect of noise crosscorrelation on binaural signal detection. *J. Acoust. Soc. Am.*, 36:1455–1458, 1964.

[64] K. J. Gabriel and H. S. Colburn. Interaural correlation discrimination: I. Bandwidth and level dependence. *J. Acoust. Soc. Am.*, 69:1394–1401, 1981.

[65] J. F. Culling, H. S. Colburn, and M. Spurchise. Interaural correlation sensitivity. *J. Acoust. Soc. Am.*, 110:1020–1029, 2001.

[66] J. W. Hall and A. D. G. Harvey. NoSo and NoSπ thresholds as a function of masker level for narrowband and wideband masking noise. *J. Acoust. Soc. Am.*, 76:1699–1703, 1984.

[67] L. R. Bernstein and C. Trahiotis. Discrimination of interaural envelope correlation and its relation to binaural unmasking at high frequencies. *J. Acoust. Soc. Am.*, 91:306–316, 1992.

[68] S. van de Par, A. Kohlrausch, J. Breebaart, and M. McKinney. Discrimination of different temporal envelope structures of diotic and dichotic target signals within diotic wide-band noise. In D. Pressnitzer, A. de Cheveigné, S. McAdams, and L. Collet, editors, *Auditory signal processing: physiology, psychoacoustics, and models*, volume Proc. 13th int. symposium on hearing. Springer Verlag, New York, 2004.

[69] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers. Parametric coding of stereo audio. *EURASIP J. on Applied Signal Processing*, XX:Under review, 2004.

[70] M. R. Schroeder. Synthesis of low-peak-factor signals and binary sequences with low autocorrelation. *IEEE Transact. Inf. Theor.*, 16:85–89, 1970.

[71] G. Stoll and F. Kozamernik. EBU listening tests on internet audio codecs. In *EBU Technical Review no 283*, 2000.