EURASIP Journal on Applied Signal Processing 2005:9, 1305–1322

Parametric Coding of Stereo Audio

Jeroen Breebaart
Digital Signal Processing Group, Philips Research Laboratories, 5656 AA Eindhoven, The Netherlands
Email: jeroen.breebaart@philips.com

Steven van de Par
Digital Signal Processing Group, Philips Research Laboratories, 5656 AA Eindhoven, The Netherlands
Email: steven.van.de.par@philips.com

Armin Kohlrausch
Digital Signal Processing Group, Philips Research Laboratories, 5656 Eindhoven, The Netherlands
Department of Technology Management, Eindhoven University of Technology, 5656 AA Eindhoven, The Netherlands
Email: armin.kohlrausch@philips.com

Erik Schuijers
Philips Digital Systems Laboratories, 5616 LW Eindhoven, The Netherlands
Email: erik.schuijers@philips.com

Received 27 January 2004; Revised 22 July 2004

Parametric-stereo coding is a technique to efficiently code a stereo audio signal as a monaural signal plus a small amount of parametric overhead to describe the stereo image. The stereo properties are analyzed, encoded, and reinstated in a decoder according to spatial psychoacoustical principles. The monaural signal can be encoded using any (conventional) audio coder. Experiments show that the parameterized description of spatial properties enables a highly efficient, high-quality stereo audio representation.

Keywords and phrases: parametric stereo, audio coding, perceptual audio coding, stereo coding.

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 international standards [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 a 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.

Basically, two different coding approaches can be distinguished that aim at bit-rate reduction. The first approach, often referred to as "waveform coding," describes the actual waveform (in frequency subbands or transform-based) with a limited (sample) accuracy. By ensuring that the quantization noise that is inherently introduced is kept below the masking curve (both across time and frequency), the concept of auditory masking (e.g., perceptual intrachannel irrelevancy) is effectively exploited.

The second coding approach relies on parametric descriptions of the audio signal. Such methods decompose the audio signal in several "objects," such as transients, sinusoids, and noise (cf. [3, 4]). Each object is subsequently parameterized and its parameters are transmitted. The decoder at the receiving end resynthesizes the objects according to the transmitted parameters. Although it is difficult to obtain transparent audio quality using such coding methods, parametric coders often perform better than waveform or transform coders (i.e., with a higher perceptual quality) at extremely low bit rates (typically up to about 32 kbps).

Recently, hybrid forms of waveform coders and parametric coders have been developed. For example, spectral band replication (SBR) techniques are proposed as a parametric coding extension for high-frequency content combined with a waveform or transform coder operating at a limited bandwidth [5, 6]. These techniques reduce the bit rate of waveform or transform coders by reducing the signal bandwidth that is sent to the encoder, combined with a small amount of parametric overhead. This parametric overhead describes how the high-frequency part, which is not encoded by the waveform coder, can be resynthesized from the low-frequency part.

The techniques described up to this point aim at encoding a single audio channel. In the case of a multichannel signal, these methods have to be performed for each channel individually. Therefore, adding more independent audio channels will result in a linear increase of the total required bit rate. It is often suggested that for multichannel material, cross-channel redundancies can be exploited to increase the coding efficiency. 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 [7]. If the two input signals are sufficiently correlated, sum/difference coding requires less bits than dual-mono coding. However, some investigations have suggested that the amount of mutual information in the signals for such a transform is rather low [8].

One possible explanation for this finding is related to the (limited) signal model. 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 apply (relative) delays to the input signals. In this case, the cross-correlation function shows a clear maximum at a certain nonzero 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 [9, 10]. 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.

The second source for bit-rate reduction in multichannel audio relates to cross-channel perceptual *irrelevancies*. 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 [11, 12]. This phenomenon is exploited by a technique referred to as "intensity stereo" [13, 14]. 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, the so-called binaural-cue coding (BCC) schemes have been described that initially aimed at modeling the most relevant sound-source localization cues [15, 16, 17], while discarding other spatial attributes such as the ambiance level and room size. BCC schemes can be seen as an extension of intensity stereo in terms of bandwidth and parameters. 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 the majority of the sound localization cues, they suffer from narrowing of the stereo image and spatial instabilities [18, 19], suggesting that these techniques are mostly advantageous at low bit rates [20]. A solution that was suggested to reduce the narrowing stereo image artifact is to transmit the interchannel coherence as a third parameter [4]. Informal listening results in [21, 22] claim improvements in spatial image width and stability.

In this paper, a parametric description of the spatial sound field 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 largely based on binaural psychoacoustics. The amount of spatial information is extracted and parameterized in a scalable fashion. At low parameter rates (typically in the order of 1 to 3 kbps), the coder is able to represent the spatial sound field in an extremely compact way. It will be shown that this configuration is very suitable for low-bit-rate audio coding applications. It will also be demonstrated that, in contrast to statements on BCC schemes [20, 21], if the spatial parameters bit rate is increased to about 8 kbps, the underlying spatial model is able to encode and recreate a spatial image which has a subjective quality which is equivalent to the quality of current high-quality stereo audio coders (such as MPEG-1 layer 3 at a bit rate of 128 kbps/s). Inspection of the coding scheme proposed here and BCC schemes reveals (at least) three important differences that all contribute to quality improvements:

(1) dynamic window switching (see Section 5.1);
(2) different methods of decorrelation synthesis (see Section 6);
(3) the necessity of encoding interchannel time or phase differences, even for loudspeaker playback conditions (see Section 3.1).

Finally, the bit-rate scalability options and the fact that a high-quality stereo image can be obtained enable integration of parametric stereo in state-of-the-art transform-based [23, 24] and parametric [4] mono audio coders for a wide quality/bit-rate range.

The paper outline is as follows. First the psychoacoustic background of the parametric-stereo coder is discussed. Section 4 discusses the general structure of the coder. In Section 5, an FFT-based encoder is described. In Section 6, an FFT-based decoder is outlined. In Section 7, an alternative decoder based on a filter bank is given. In Section 8, results from listening tests are discussed, followed by a concluding section.

2. PSYCHOACOUSTIC BACKGROUND

In 1907, Lord Rayleigh formulated the duplex theory [25], 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, frequency-independent 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 as well as the IID [26, 27, 28]. 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. The reason for the fact that these stimuli are not perceived externalized is that the single frequency-independent IID or ITD is a poor representation of the acoustic signals at the listener's eardrums in free-field listening conditions. The waveforms of sounds are filtered by the acoustical transmission path between the source and the listener's eardrums, which includes room reflections and pinna filtering, resulting in an intricate frequency dependence of the ITD and IID [29]. Moreover, if multiple sound sources with different spectral properties exist at different spatial locations, the spatial cues of the signals arriving at the eardrums will show a frequency dependence which is even more complex because they are constituted by (weighted) combinations of the spatial cues of the individual's sound sources.

Extensive psychophysical research (cf. [30, 31, 32]) and efforts to model the binaural auditory system (cf. [33, 34, 35, 36, 37]) have suggested that the human auditory system extracts spatial 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 [38, 39, 40].

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 [32, 41]. 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" [42]. 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 [43]. 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 [44, 45, 46, 47]. There is a considerable evidence that the sensitivity to changes in the coherence is the basis of the phenomenon of the binaural masking level difference (BMLD) [48, 49]. Moreover, the sensitivity to quasistatic ITDs can also be described by (changes in) the cross-correlation function [35, 36, 50].

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. [51] measured the detectability and discriminability of interaurally out-of-phase test signals presented in an interaurally in-phase masker. The subjects were perfectly able to *detect* the presence of the out-of-phase test signal, but they had great difficulty in *discriminating* different test signal types (i.e., noise versus harmonic tone complexes).

Besides the limited spectral and temporal resolution that seems to underly the extraction of spatial sound-field properties, it has also been shown that the auditory system exhibits a limited *spatial* resolution. The spatial parameters have to change by a certain minimum amount before subjects are able to detect the change. For IIDs, 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 [52, 53, 54, 55]. 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 [56, 57, 58].

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 [11, 53, 59, 60]. 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 [52, 61]. There is almost no effect of stimulus level on ITD sensitivity [12]. 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 [62, 63]. Despite this high-frequency sensitivity, ITD-based sound-source localization is dominated by low-frequency cues [64, 65].

The sensitivity to changes in the coherence strongly depends on the reference coherence. For a reference coherence of +1, 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 [66, 67, 68, 69]. The sensitivity to interaural coherence is practically independent of stimulus level, as long as the stimulus is sufficiently above the absolute threshold [70]. At high frequencies, the *envelope* coherence seems to be the relevant descriptor of the spatial diffuseness [47, 71].

The threshold values described above are typical for spatial properties that exist during a prolonged time (i.e., 300 to 400 milliseconds). If the duration is smaller, thresholds generally increase. For example, if the duration of the IID and ITD in a stimulus is decreased from 310 to 17 milliseconds, the thresholds may increase by up to a factor of 4 [72]. Interaural coherence sensitivity also strongly depends on the duration [73, 74, 75]. It is often assumed that the increased sensitivity for longer durations results from temporal integration properties of the auditory system. 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” [76, 77, 78, 79].

In summary, it seems that the auditory system performs a frequency separation and temporal averaging process in its determination of IIDs, ITDs, and the coherence. This estimation process leads to the concept of a certain sound-source location as a function of frequency and time, while the variability of the localization cues leads to a certain degree of “diffuseness,” or spatial “widening,” with hardly any interaction between diffuseness and location [72]. Furthermore, these cues are rendered with a limited (spatial) resolution. These observations form the basis of the parametric stereo coder as described in the following sections. The general idea is to encode all (monaurally) relevant sound sources using a *single* audio channel, combined with a parameterization of the spatial sound stage. The parameterized sound stage consists of IID, ITD, and coherence parameters as a function of frequency and time. The update rate, frequency resolution, and quantization of these parameters is determined by the human sensitivity to (changes in) these parameters.

3. CODING ISSUES

3.1. Headphones versus loudspeaker rendering

The psychoacoustic background as discussed in Section 2 is based on spatial cues at the level of the listener’s eardrums. In the case of headphone rendering, the spatial cues which are presented to the human hearing system (i.e., the *interaural* cues ILD, ITD, and coherence) are virtually the same as the spatial cues in the original stereo signal (*interchannel* cues). For loudspeaker playback, however, the complex acoustical transmission paths between loudspeakers and eardrums (as described in Section 2) may cause significant changes in the spatial cues. It is therefore highly unlikely that the spatial cues of the original stereo signal (e.g., the interchannel cues) and the spatial cues at the level of the listener’s eardrums (interaural cues) are even comparable in the case of loudspeaker playback. In fact, it has been suggested that the acoustical transmission path effectively converts certain spatial cues (for example interchannel intensity differences) to other cues at the level of the eardrums (e.g., interaural time differences) [80, 81]. However, this effect of the transmission path is not necessarily problematic for parametric-stereo coding. As long as the *interaural* cues are *the same* for original material and material which has been processed by a parametric-stereo coder, the listener should have a similar percept of the spatial sound field. Although a detailed analysis of this problem is beyond the scope of this paper, we state that given certain restrictions on the acoustical transmission path, it can be shown that the interaural spatial cues are indeed comparable for original and decoded signal, provided that *all three* interchannel parameters are encoded and reconstructed correctly. Moreover, well-known algorithms that aim at widening of the perceived sound stage for loudspeaker playback (so-called crosstalk-cancellation algorithms, which are used frequently in commercial recordings) heavily rely on correct interchannel phase relationships (cf. [82]). These observations are in contrast to statements by others (cf. [18, 21, 22]) that interchannel time or phase differences are irrelevant for loudspeaker playback.

Supported by the observations given above, we will refer to ILD, ITD, and coherence as interchannel parameters. If all three interchannel parameters are reconstructed correctly, we assume that the interaural parameters of original and decoded signals are very similar as well (but different from the interchannel parameters).

3.2. Mono coding effects

As discussed in Section 1, bit-rate reduction in conventional lossy audio coders is obtained predominantly by exploiting the phenomenon of masking. Therefore, lossy audio coders rely on accurate and reliable masking models, which are often applied to individual channel signals in the case of a stereo or multichannel signal. For a parametric-stereo extended audio coder, however, the masking model is applied only once on a certain combination of the two input signals. This scheme has two implications with respect to masking phenomena.

The first implication relates to spatial unmasking of quantization noise. In stereo waveform or transform coders,

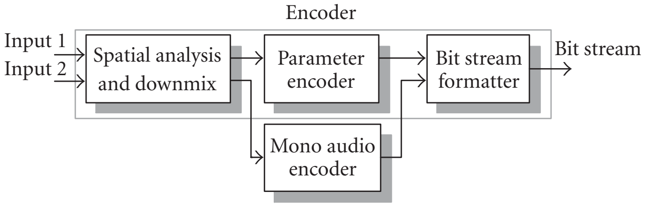

Figure 1: Structure of the parametric-stereo encoder. The two input signals are first processed by a parameter extraction and downmix stage. The parameters are subsequently quantized and encoded, while the mono downmix can be encoded using an arbitrary mono audio coder. The mono bit stream and spatial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

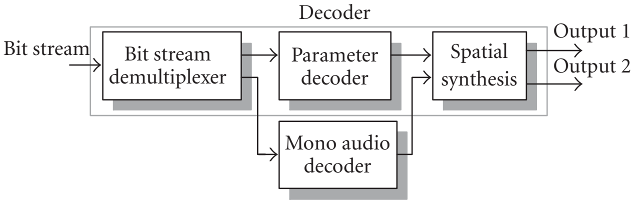

Figure 2: Structure of the parametric-stereo decoder. The demultiplexer splits mono and spatial parameter information. The mono audio signal is decoded and fed into the spatial synthesis stage, which reinstates the spatial cues based on the decoded spatial parameters.

individual quantizers are applied on the two input signals or on linear combinations of the input signals. As a consequence, the injected quantization noise may exhibit different spatial properties than the audio signal itself. Due to binaural unmasking, the quantization noise may thus become audible, even if it is inaudible if presented monaurally. For tonal material, this unmasking effect (or BMLD, quantified as threshold difference between a binaural condition and a monaural reference condition) has shown to be relatively small (about 3 dB, see [83, 84]). However, we expect that for broadband maskers, the unmasking effect is much more prominent. If one assumes an interaurally in-phase noise as a masker, and a quantization noise which is either inter-aurally in-phase or interaurally uncorrelated, BMLDs are reported of 6 dB [85]. More recent data revealed BMLDs of 13 dB for this condition, based on a sensitivity of changes in the correlation of 0.045 [86]. To prevent these spatial unmasking effects of quantization noise, conventional stereo coders often apply some sort of spatial unmasking protection algorithm.

For a parametric stereo coder, on the other hand, there is only one waveform or transform quantizer, working on the mono (downmix) signal. In the stereo reconstruction phase, both the quantization noise and the audio signal present in each frequency band will obey the same spatial properties. Since a difference in spatial characteristics of quantization noise and audio signal is a prerequisite for spatial unmasking, this effect is less likely to occur for parametric-stereo enhanced coders than for conventional stereo coders.

4. CODER IMPLEMENTATION

The generic structure of the parametric-stereo encoder is shown in Figure 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.

The parametric-stereo decoder basically performs the reverse process, as shown in Figure 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.

Since the spatial parameters are estimated (at the encoder side) and applied (at the decoder side) as a function of time and frequency, both the encoder and decoder require a transform or filter bank that generates individual time/frequency tiles. The frequency resolution of this stage should be nonuniform 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 (cross-channel) phase- or time-difference 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 [23, 24].

5. FFT-BASED ENCODER

The spatial analysis and downmix stage of the encoder is shown in more detail in Figure 3. The two input signals are first segmented by an analysis windowing process. Subsequently, each windowed segment is transformed to the frequency domain using a fast fourier transform (FFT). The transformed segments are 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 synthesis windowing and overlap-add (OLA).

5.1. Segmentation

The encoder receives a stereo input signal pair $x_1[n]$, $x_2[n]$ with a sampling rate f_s. The input signals are segmented

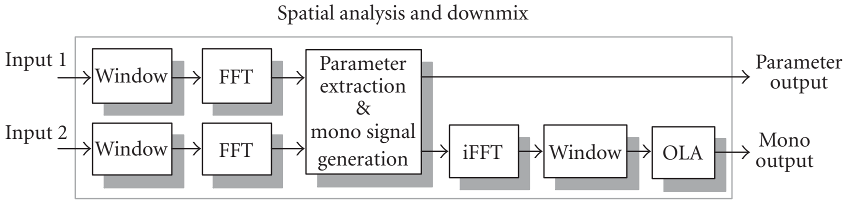
Spatial analysis and downmix
Input 1
Input 2
Window
FFT
Window
FFT
Parameter extraction & mono signal generation
iFFT
Window
OLA
Parameter output
Mono output

Figure 3: Spatial analysis and downmix stage of the encoder.

using overlapping frames of total length N with a (fixed) hop size of N_h samples. If no transients are detected, the analysis window length and the window hop size (or parameter update rate) should match the lower bound of the measured time constants of the binaural auditory system. In the following, a parameter update interval of approximately 23 milliseconds is used. Each segment is windowed using overlapping analysis windows and subsequently transformed to the frequency domain using an FFT. Dynamic window switching is used in the case of transients. The purpose of window switching is twofold: firstly, to account for the precedence effect, which dictates that only the first 2 milliseconds of a transient in a reverberant environment determine its perceived location; secondly, to prevent pre-echos resulting from the frequency-dependent processing which is applied in otherwise relatively long segments. The window switching procedure, of which the essence is demonstrated in Figure 4, is controlled by a transient detector.

If a transient is detected at a certain temporal position, a stop window of variable length is applied which just stops before the transient. The transient itself is captured using a very short window (in the order of a few milliseconds). A start window of variable length is subsequently applied to ensure segmentation at the same temporal grid as before the transient.

5.2. Frequency separation

Each segment is transformed to the frequency domain using an FFT of length N ($N = 4096$ for a sampling rate f_s of 44.1 kHz). The frequency-domain signals $X_1[k]$, $X_2[k]$ ($k = [0, 1, \dots, N/2]$) are divided into nonoverlapping subbands by grouping of FFT bins. The frequency bands are formed in such a way that each band has a bandwidth, BW (in Hz), which is approximately equal to the equivalent rectangular bandwidth (ERB) [40], following

$$BW = 24.7(0.00437f + 1), \tag{1}$$

with f the (center) frequency given in Hz. This process results in $B = 34$ frequency bands with FFT start indices k_b of subband b ($b = [0, 1, \dots, B-1]$). The center frequencies of each analysis band vary between 28.7 Hz ($b = 0$) to 18.1 kHz ($b = 33$).

5.3. 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:

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

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:

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

Using the IPD as specified in (3), (relative) delays between the input signals which are represented as a constant phase difference in each analysis frequency band, hence result in a fractional delay. Thus, *within* each analysis band, the constant *slope* of phase with frequency is modeled by a constant phase difference per band, which is a somewhat limited model for the delay. On the other hand, constant phase differences across the input signals are described accurately, which is in turn not possible if an ITD parameter (i.e., a parameterized slope of phase with frequency) would have been used. An advantage of using IPDs over ITDs is that the estimation of ITDs requires accurate unwrapping of bin-by-bin phase differences within each analysis frequency band, which can be prone to errors. Thus, usage of IPDs circumvents this potential problem at the cost of a possibly limited model for ITDs.

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:

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

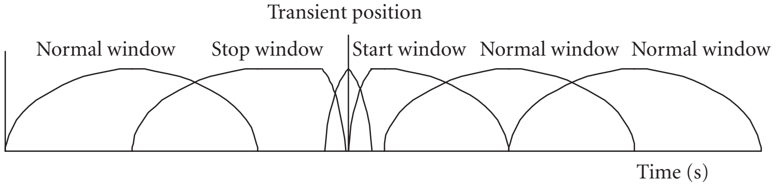

Figure 4: Schematic presentation of dynamic window switching in case of a transient. A stop window is placed just before the detected transient position. The transient itself is captured using a short window.

5.4. *Downmix*

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

$$S[k] = w_1 X_1[k] + w_2 X_2[k], \tag{5}$$

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 example, if $w_1 = w_2 = 0.5$, the output will consist of the average of the two input signals. A downmix that is created using fixed weights however bears the risk that the power of the downmix signal strongly depends on the cross-correlation of the two input signals. To circumvent signal loss and signal coloration due to time- and frequency-dependent cross-correlations, the weights w_1 and w_2 are (1) complex-valued, to prevent phase cancellation, and (2) varying in magnitude, to ensure overall power preservation. Specific details of the downmix procedure are however beyond the scope of this paper.

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 signal (at the encoder) and the stereo output signals (at the decoder). Hence the IPD does not indicate how the decoder should distribute these phase differences across the output channels. In other words, an IPD parameter alone does not indicate whether a first signal is lagging the second signal, or vice versa. Thus, it is generally impossible to reconstruct the absolute phase for the stereo signal pair using only the relative phase difference. Absolute phase reconstruction is required to prevent signal cancellation in the applied overlap-add procedure in both the encoder as well as the decoder (see below). To signal the actual distribution of phase modifications, 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 $S[k]$, following

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

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.

5.5. *Parameter quantization and coding*

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. For the IID, this constraint requires a nonlinear quantizer, or nonlinearly spaced IID values given the fact that the sensitivity for changes in IID depends on the reference IID. The vector **IIDs** contains the possible discrete IID values that are available for the quantizer. Each element in **IIDs** represents a single quantization level for the IID parameter and is indicated by $\text{IID}_q[i]$ ($i = [0, \dots, 30]$):

$$\begin{aligned}\mathbf{IIDs} &= [\text{IID}_q[0], \text{IID}_q[1], \text{IID}_q[30]] \\ &= [-50, -45, -40, -35, -30, -25, -22, \dots, \\ &\quad -19, -16, -13, -10, -8, -6, -4, -2, 0, \dots, \\ &\quad 2, 4, 6, 8, 10, 13, 16, 19, 22, 25, 30, 35, 40, 45, 50].\end{aligned} \tag{7}$$

The IID index for subband b, $\text{IDX}_{\text{IID}}[b]$, is then equal to

$$\text{IDX}_{\text{IID}}[b] = \arg\left(\min_i |\text{IID}[b] - \text{IID}_q[i]|\right). \tag{8}$$

For the IPD parameter, the vector **IPDs** represents the available quantized IPD values:

$$\begin{aligned}\mathbf{IPDs} &= [\text{IPD}_q[0], \text{IPD}_q[1], \dots, \text{IPD}_q[7]] \\ &= \left[0, \frac{\pi}{4}, \frac{2\pi}{4}, \frac{3\pi}{4}, \frac{4\pi}{4}, \frac{5\pi}{4}, \frac{6\pi}{4}, \frac{7\pi}{4}\right].\end{aligned} \tag{9}$$

This repertoire is in line with the finding that the human sensitivity to changes in timing differences at low frequencies can be described by a constant phase difference sensitivity. The IPD index for subband b, $\text{IDX}_{\text{IPD}}[b]$, is given by

$$\text{IDX}_{\text{IPD}}[b] = \text{mod}\left(\left\lfloor \frac{4\text{IPD}[b]}{\pi} + \frac{1}{2} \right\rfloor, \Lambda_{\mathbf{IPDs}}\right), \tag{10}$$

where mod$(\cdot)$ means the modulo operator, $\lfloor\cdot\rfloor$ the floor function, and $\Lambda_{\mathbf{IPDs}}$ the cardinality of the set of possible quantized IPD values (i.e., the number of elements in **IPDs**). The OPD is quantized using the same quantizer, resulting in $\mathrm{IDX}_{\mathrm{OPD}}[b]$ according to

$$\mathrm{IDX}_{\mathrm{OPD}}[b] = \mathrm{mod}\left(\left\lfloor \frac{4\mathrm{OPD}[b]}{\pi} + \frac{1}{2}\right\rfloor, \Lambda_{\mathbf{IPDs}}\right). \tag{11}$$

Finally, the repertoire for IC, represented in the vector **ICs**, is given by (see also (21))

$$\begin{aligned}\mathbf{ICs} &= \left[\mathrm{IC}_q[0], \mathrm{IC}_q[1], \ldots, \mathrm{IC}_q[7]\right] \\ &= [1, 0.937, 0.84118, 0.60092, 0.36764, 0, -0.589, -1].\end{aligned} \tag{12}$$

This repertoire is based on just-noticeable differences in correlation reported by [69]. The coherence index $\mathrm{IDX}_{\mathrm{IC}}[b]$ for subband b is determined by

$$\mathrm{IDX}_{\mathrm{IC}}[b] = \arg\left(\min_i \left|\mathrm{IC}[b] - \mathrm{IC}_q[i]\right|\right). \tag{13}$$

The IPD and OPD indices are not transmitted for subbands $b > 17$ (approximately 2 kHz), given the fact that the human auditory system is insensitive to fine-structure phase differences at high frequencies. ITDs present in the high-frequency envelopes are supposed to be represented by the time-varying nature of IID parameters (hence discarding ITDs presented in envelopes that fluctuate faster than the parameter update rate).

Thus, for each frame, 34 indices for the IID and IC have to be transmitted, and 17 indices for the IPD and OPD. All parameters are transmitted differentially across time. In principle, differential coding of indices Λ ($\lambda = \{0, \ldots, \Lambda - 1\}$) requires $2\Lambda - 1$ codewords $\lambda_d = \{-\Lambda + 1, \ldots, 0, \ldots, \Lambda - 1\}$. Assuming that each differential index λ_d has a probability of occurrence $p(\lambda_d)$, the entropy $H(p)$ (in bits/symbol) of this distribution is given by

$$H(p) = \sum_{\lambda_d=-\Lambda+1}^{\lambda=\Lambda-1} -p(\lambda_d)\log_2\left(p(\lambda_d)\right). \tag{14}$$

Given the fact that the cardinality of each parameter Λ is known by the decoder, each differential index λ_d can also be modulo-encoded by λ_{mod}, which is given by

$$\lambda_{\mathrm{mod}} = \mathrm{mod}\left(\lambda_d, \Lambda\right). \tag{15}$$

The decoder can simply retain the transmitted index λ recursively following

$$\lambda[q] = \mathrm{mod}\left(\lambda_{\mathrm{mod}}[q] + \lambda[q-1], \Lambda\right), \tag{16}$$

with q the frame number of the current frame. The entropy for λ_{mod}, $H(p_{\mathrm{mod}})$, is given by

Table 1: Entropy per parameter symbol, number of symbols per second, and bit rate for spatial parameters.

Parameter	Bits/symbol	Symbols/s	Bit rate (bps)
IID	1.94	1464	2840
IPD	1.58	732	1157
OPD	1.31	732	959
IC	1.88	1464	2752
Total	—	—	7708

$$H(p_{\mathrm{mod}}) = \sum_{\lambda_{\mathrm{mod}}=0}^{\Lambda-1} -p_{\mathrm{mod}}(\lambda_{\mathrm{mod}})\log_2\left(p_{\mathrm{mod}}(\lambda_{\mathrm{mod}})\right). \tag{17}$$

Given that

$$\begin{aligned} p_{\mathrm{mod}}(0) &= p(0), \\ p_{\mathrm{mod}}(z) &= p(z) + p(z-\Lambda) \quad \text{for } z = \{1, \ldots, \Lambda - 1\}, \end{aligned} \tag{18}$$

it follows that the difference in entropy between differential and modulo-differential coding, $H(p) - H(p_{\mathrm{mod}})$, equals

$$\begin{aligned} H(p) &- H(p_{\mathrm{mod}}) \\ &= \sum_{\lambda_d=1}^{\lambda_d=\Lambda-1} p(\lambda_d)\log_2\frac{p(\lambda_d) + p(\lambda_d - \Lambda)}{p(\lambda_d)} \\ &\quad + \sum_{\lambda_d=1}^{\lambda_d=\Lambda-1} p(\lambda_d - \Lambda)\log_2\frac{p(\lambda_d) + p(\lambda_d - \Lambda)}{p(\lambda_d - \Lambda)}. \end{aligned} \tag{19}$$

For nonnegative probabilities $p(\cdot)$, it follows that

$$H(p) - H(p_{\mathrm{mod}}) \geq 0. \tag{20}$$

In other words, modulo-differential coding results in an entropy which is equal to or smaller than the entropy obtained for non modulo-differential coding. However, the bit-rate gains for modulo time-differential coding compared to time-differential coding are relatively small: about 15% for the IPD and OPD parameters, and virtually no gain for the IID and IC parameters. The entropy per symbol, using modulo-differential coding, and the resulting contribution to the overall bit rate are given in Table 1. These numbers were obtained by analysis of 80 different audio recordings representing a large variety of material.

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 kbps. If further parameter bit-rate reduction is required, the following changes can be made.

(i) Reduction of the number of frequency bands (e.g., using 20 instead of 34). The parameter bit rate increases approximately linearly with the number of bands. This results in a bit rate of approximately 4.5 kbps for the 20-band case, assuming an update rate of 23 milliseconds and including

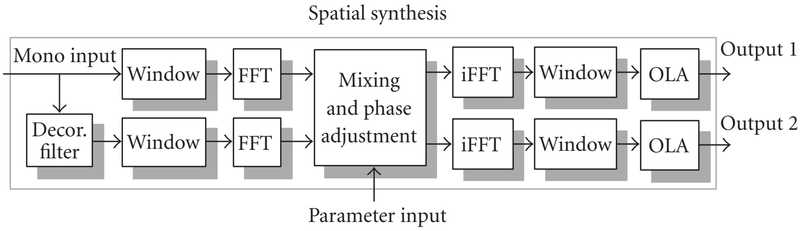

Figure 5: Spatial synthesis stage of the decoder.

transmission of IPD and OPD parameters. Informal listening experiments showed that lowering the number of frequency bands below 10 results in severe degradation of the perceived spatial quality.

(ii) No transmission of IPD and OPD parameters. As described above, the coherence is a measure of the difference between the input signals which cannot be accounted for by (subband) phase and level differences. A lower bit rate is obtained if the applied signal model does not incorporate phase differences. In that case, the normalized cross-correlation is the relevant measure of differences between the input signals that cannot be accounted for by level differences. In other words, phase or time differences between the input signals are modeled as (additional) changes in the coherence. The estimated coherence value (which is in fact the normalized cross-correlation) is then derived from the cross-spectrum following

$$\mathrm{IC}[b] = \frac{\mathrm{Re}\left\{\sum_{k=k_b}^{k_{b+1}-1} X_1[k]X_2^*[k]\right\}}{\sqrt{\left(\sum_{k=k_b}^{k_{b+1}-1} X_1[k]X_1^*[k]\right)\left(\sum_{k=k_b}^{k_{b+1}-1} X_2[k]X_2^*[k]\right)}}. \tag{21}$$

The associated bit-rate reduction amounts to approximately 27% compared to parameter sets which do include the IPD and OPD values.

(iii) Increasing the quantization errors of the parameters. The bit-rate reduction is only marginal, given the fact that the distribution of time-differential parameters is very peaky.

(iv) Decreasing the parameter update rate. The bit rate scales approximately linear with the update rate.

In summary, the parameter bit rate can be scaled between approximately 8 kbps for maximum quality (using 34 analysis bands, an update rate of 23 milliseconds, and transmitting all relevant parameters) to about 1.5 kbps (using 20 analysis frequency bands, an update rate of 46 milliseconds, and no transmission of IPD and OPD parameters).

6. FFT-BASED 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 Figure 5.

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 from a fine-structure waveform point of view. This incoherent (or orthogonal) 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 decorrelation allpass filter is obtained by a simple delay. The combination of a delay and a (fixed) mixing matrix to produce two signals with a certain spatial diffuseness is known as a Lauridsen decorrelator [87]. The decorrelation is produced by complementary comb-filter peaks and troughs in the two output signals. This approach works well provided that the delay is sufficiently long to result in multiple comb-filter peaks and troughs in each auditory filter. Due to the fact that the auditory filter bandwidth is larger at higher frequencies, the delay is preferably frequency dependent, being shorter at higher frequencies. A frequency-dependent delay has the additional advantage that it does not result in harmonic comb-filter effects in the output. A suitable decorrelation filter consists of a single period of a positive Schroeder-phase complex [88] of length $N_s = 640$ (i.e., with a fundamental frequency of f_s/N_s). The Schroeder-phase complex exhibits low autocorrelation at nonzero lags and its impulse response $h_d[n]$ for $0 \leq n \leq N_s - 1$ is given by

$$h_d[n] = \sum_{k=0}^{N_s/2} \frac{2}{N_s} \cos\left(\frac{2\pi kn}{N_s} + \frac{2\pi k(k-1)}{N_s}\right). \tag{22}$$

Subsequently, the segmentation, windowing, and transform operations that are performed are equal to those performed 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 dynamic mixing process, which is performed on a subband basis, is described by the matrix multiplication R_B. For each subband b (i.e., $k_b \leq k < k_{b+1}$), we have

$$\begin{bmatrix} Y_1[k] \\ Y_2[k] \end{bmatrix} = \mathbf{R}_B \begin{bmatrix} S[k] \\ S_d[k] \end{bmatrix}, \tag{23}$$

with

$$\mathbf{R}_B[b] = \sqrt{2}\mathbf{P}[b]\mathbf{A}[b]\mathbf{V}[b]. \tag{24}$$

The diagonal matrix $\mathbf{V}$ enables real-valued (relative) scaling of the two orthogonal signals $S[k]$ and $S_d[k]$. The matrix $\mathbf{A}$ is a real-valued rotation in the two-dimensional signal space, that is, $\mathbf{A}^{-1} = \mathbf{A}^T$, 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. The nonzero entries in the matrices $\mathbf{P}$, $\mathbf{A}$, and $\mathbf{V}$ are determined by the following constraints.

(1) The power ratio of the two output signals must obey the transmitted IID parameter.
(2) The coherence of the two output signals must obey the transmitted IC parameter.
(3) The average energy of the two output signals must be equal to the energy of the mono input signal.
(4) The total amount of $S[k]$ present in the two output signals should be maximum (i.e., v_{11} should be maximum).
(5) The average phase difference between the output signals must be equal to the transmitted IPD value.
(6) The average phase difference between $S[k]$ and $Y_1[k]$ should be equal to the OPD value.

The solution for the matrix $\mathbf{P}$ is given by

$$\mathbf{P}[b] = \begin{bmatrix} e^{j\mathrm{OPD}[b]} & 0 \\ 0 & e^{j\mathrm{OPD}[b]-j\mathrm{IPD}[b]} \end{bmatrix}. \tag{25}$$

The matrices $\mathbf{A}$ and $\mathbf{V}$ can be interpreted as the eigenvector, eigenvalue decomposition of the covariance matrix of the (desired) output signals, assuming (optimum) phase alignment ($\mathbf{P}$) prior to correlation. The solution for the eigenvectors and eigenvalues (maximizing the first eigenvalue v_{11}) results from a singular value decomposition (SVD) of the covariance matrix. The matrices $\mathbf{A}$ and $\mathbf{V}$ are given by (see [89] for more details)

$$\begin{aligned} \mathbf{A}[b] &= \begin{bmatrix} \cos(\alpha[b]) & -\sin(\alpha[b]) \\ \sin(\alpha[b]) & \cos(\alpha[b]) \end{bmatrix}, \\ \mathbf{V}[b] &= \begin{bmatrix} \cos(\gamma[b]) & 0 \\ 0 & \sin(\gamma[b]) \end{bmatrix}, \end{aligned} \tag{26}$$

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

$$\alpha[b] = \begin{cases} \dfrac{\pi}{4} & \text{for } (\mathrm{IC}[b], c[b]) = (0,1), \\ \bmod\left(\dfrac{1}{2}\arctan\left(\dfrac{2c[b]\mathrm{IC}[b]}{c[b]^2-1}\right), \dfrac{\pi}{2}\right) & \\ & \text{otherwise}, \end{cases} \tag{27}$$

and $\gamma[b]$ a parameter for relative scaling of S and S_d (i.e., the relation between the eigenvalues of the desired covariance matrix):

$$\gamma[b] = \arctan\sqrt{\frac{1-\sqrt{\mu[b]}}{1+\sqrt{\mu[b]}}}, \tag{28}$$

with

$$\mu[b] = 1 + \frac{4\mathrm{IC}^2[b] - 4}{\left(c[b] + 1/c[b]\right)^2}, \tag{29}$$

and $c[b]$ the square root of the power ratio of the two subband output signals:

$$c[b] = 10^{\mathrm{IID}[b]/20}. \tag{30}$$

It should be noted that a two-dimensional eigenvector problem has in principle four possible solutions: each eigenvector, which is represented as columns in the matrix $\mathbf{A}$, may be multiplied with a factor -1. The modulo operator in (27) ensures that the first eigenvector is always positioned in the first quadrant. However, this technique only works under the constraint of IC > 0, which is guaranteed if phase alignment is applied. If no IPD/OPD parameters are transmitted, however, the IC parameters may become negative, which requires a different solution for the matrix $\mathbf{R}$. A convenient solution is obtained if we maximize $S[k]$ in the sum of the output signals (i.e., $Y_1[k] + Y_2[k]$). This results in the mixing matrix $\mathbf{R}_A[b]$:

$$\mathbf{R}_A[b] = \begin{bmatrix} c_1\cos(\nu[b]+\mu[b]) & c_1\sin(\nu[b]+\mu[b]) \\ c_2\cos(\nu[b]-\mu[b]) & c_2\sin(\nu[b]-\mu[b]) \end{bmatrix}, \tag{31}$$

with

$$\begin{aligned} c_1[b] &= \sqrt{\frac{2c^2[b]}{1+c^2[b]}}, \\ c_2[b] &= \sqrt{\frac{2}{1+c^2[b]}}, \\ \mu[b] &= \frac{1}{2}\arccos(\mathrm{IC}[b]), \\ \nu[b] &= \frac{\mu[b](c_2[b]-c_1[b])}{\sqrt{2}}. \end{aligned} \tag{32}$$

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

7. QMF-BASED DECODER

The FFT-based decoder as described in the previous section requires a relatively long FFT length to provide sufficient frequency resolution at low frequencies. As a result, the resolution at high frequencies is unnecessarily high, and consequently the memory requirements of an FFT-based decoder are larger than necessary. To reduce the frequency resolution at high frequencies while still maintaining the required resolution at low frequencies, a hybrid complex filter bank is used. To be more specific, a hybrid complex-modulated quadrature mirror filter bank (QMF) is used which is an extension to the filter bank as used in spectral band replication (SBR) techniques [5, 6, 90]. The outline of the QMF-based parametric-stereo decoder is shown in Figure 6.

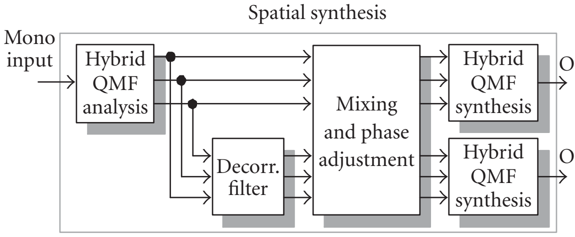

Figure 6: Structure of the QMF-based decoder. The signal is first fed through a hybrid QMF analysis filter bank. The filter-bank output and a decorrelated version of each filter-bank signal are subsequently fed into the mixing and phase-adjustment stage. Finally, two hybrid QMF banks generate the two output signals.

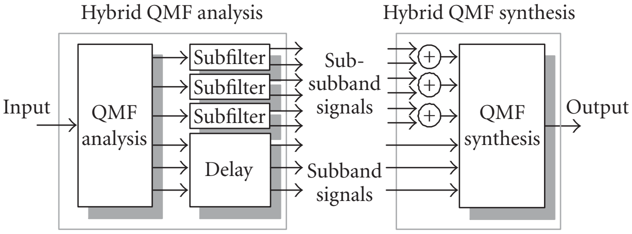

Figure 7: Structure of the hybrid QMF analysis and synthesis filter banks.

The input signal is first processed by the hybrid QMF analysis filter bank. A copy of each filter-bank output is processed by a decorrelation filter. This filter has the same purpose as the decorrelation filter in the FFT-based decoder; it generates a decorrelated version of the input signal in the QMF domain. Subsequently, both the QMF output and its decorrelated version are fed into the mixing and phase-adjustment stage. This stage generates two hybrid QMF-domain output signals with spatial parameters that match the transmitted parameters. Finally, the output signals are fed through a pair of hybrid QMF synthesis filter banks to result in the final output signals.

The hybrid QMF analysis filter bank consists of a cascade of two filter banks. The structure is shown in Figure 7.

The first filter bank is compatible with the filter bank as used in SBR algorithms. The subband signals which are generated by this filter bank are obtained by convolving the input signal with a set of analysis filter impulse responses $h_k[n]$ given by

$$h_k[n] = p_0[n] \exp\left\{j\frac{\pi}{4K}(2k+1)(2n-1)\right\}, \qquad (33)$$

with $p_0[n]$, for $n = 0,\dots,N_q-1$, the prototype window of the filter, $K = 64$ the number of output channels, k the subband index ($k = 0,\dots,K-1$), and $N_q = 640$ the filter length. The filtered outputs are subsequently down sampled by a factor K, to result in a set of down-sampled QMF outputs (or subband signals) $S_k[q]$:[1]

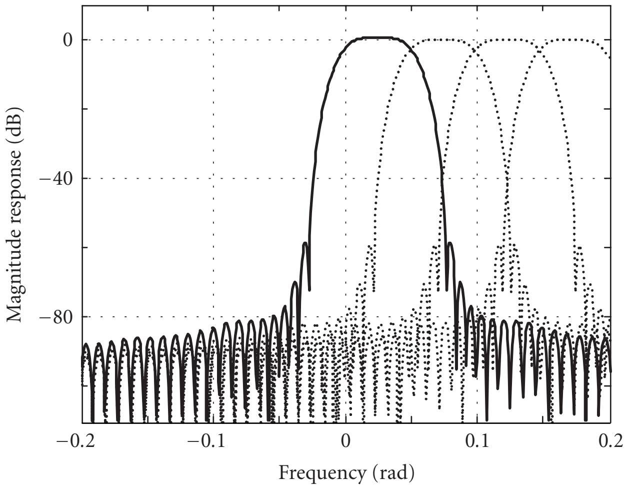

Figure 8: Magnitude responses of the first 4 of the 64-band SBR complex-exponential modulated analysis filter bank. The magnitude for $k = 0$ is highlighted.

$$S_k[q] = (s * h_k)[Kq]. \qquad (34)$$

The magnitude responses of the first 4 frequency bands ($k = 0,\dots,3$) of the QMF analysis bank are illustrated in Figure 8.

The down-sampled subband signals $S_k[q]$ of the lowest QMF subbands are subsequently fed through a second complex-modulated filter bank (sub-filter bank) to further enhance the frequency resolution; the remaining subband signals are delayed to compensate for the delay which is introduced by the sub-filter bank. The output of the hybrid (i.e., combined) filter bank is denoted by $S_{k,m}[q]$, with k the subband index of the initial QMF bank, and m the filter index of the sub-filter bank. To allow easy identification of the two filter banks and their outputs, the index k of the first filter bank will be denoted "subband index," and the index m of the subfilter bank is denoted "sub-subband index." The sub-filter bank has a filter order of $N_s = 12$, and an impulse response $G_{k,m}[q]$ given by

$$G_{k,m}[q] = g_k[q] \exp\left\{j\frac{2\pi}{M_k}\left(m+\frac{1}{2}\right)\left(q-\frac{N_s}{2}\right)\right\}, \qquad (35)$$

with $g_k[q]$ the prototype window associated with QMF band k, q the sample index, and M_k the number of sub-subbands in QMF subband k ($m = 0,\dots,M_k-1$). Table 2 gives the number of sub-subbands M_k as a function of the QMF band k, for both the 34 and 20 analysis-band configurations. As an example, the magnitude response of the 4-band sub-filter

[1] The equations given here are purely analytical; in practice the computational efficiency of the filter bank can be increased using decomposition methods.

Table 2: Specification of M_k for the first 5 QMF subbands.

QMF subband (k)	M_k ($B = 34$)	M_k ($B = 20$)
0	12	8
1	8	4
2	4	4
3	4	1
4	4	1

bank ($M_k = 4$) is given in Figure 9. Obviously, due to the limited prototype length ($N_s = 12$), the stop-band attenuation is only in the order of 20 dB.

As a result of this hybrid QMF filter-bank structure, 91 (for $B = 34$) or 77 ($B = 20$) down-sampled filter outputs $S_{k,m}[q]$ and their filtered (decorrelated) counterparts $S_{k,m,d}[q]$ are available for further processing. The decorrelation filter can be implemented in various ways. An elegant method comprises a reverberator [24]; a low-complexity alternative consists of a (frequency-dependent) delay T_k of which the delay time depends on the QMF subband index k.

The next stage of the QMF-based spatial synthesis stage performs a mixing and phase-adjustment process. For each sub-subband signal pair $S_{k,m}[q]$, $S_{k,m,d}[q]$, an output signal pair $Y_{k,m,1}[q]$, $Y_{k,m,2}[q]$ is generated by

$$\begin{bmatrix} Y_{k,m,1}[q] \\ Y_{k,m,2}[q] \end{bmatrix} = \mathbf{R}_{k,m} \begin{bmatrix} S_{k,m}[q] \\ S_{k,m,d}[q] \end{bmatrix}. \tag{36}$$

The mixing matrix $\mathbf{R}_{k,m}$ is determined as follows. Each quartet of the parameters IID, IPD, OPD, and IC for a single parameter subband b represents a certain frequency range and a certain moment in time. The frequency range depends on the specification of the encoder analysis frequency bands (i.e., the grouping of FFT bins), while the position in time depends on the encoder time-domain segmentation. If the encoder is designed properly, the time/frequency localization of each parameter quartet coincides with a certain sample index in a sub-subband or set of sub-subbands in the QMF domain. For that particular QMF sample index, the mixing matrices are exactly the same as their FFT-based counterparts (as specified by (25)–(32)). For QMF sample indices in between, the mixing matrices are interpolated linearly (i.e., its real and imaginary parts are interpolated individually).

The mixing process is followed by a pair of hybrid QMF synthesis filter banks (one for each output channel), which also consist of two stages. The first stage comprises summation of the sub-subbands m which stem from the same subband k:

$$\begin{aligned} Y_{k,1}[q] &= \sum_{m=0}^{M_k-1} Y_{k,m,1}[q], \\ Y_{k,2}[q] &= \sum_{m=0}^{M_k-1} Y_{k,m,2}[q]. \end{aligned} \tag{37}$$

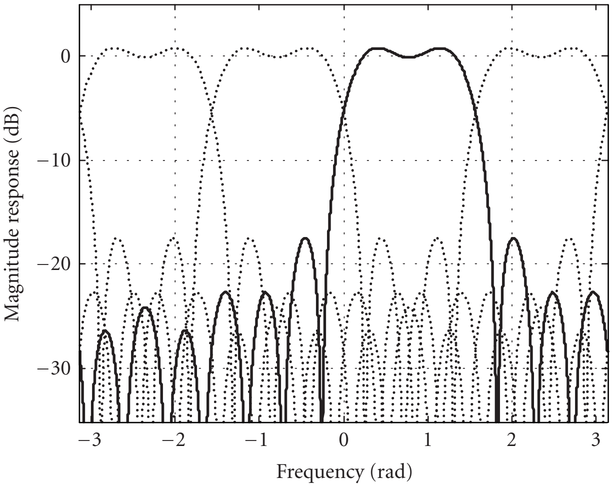

Figure 9: Magnitude response of the 4-band sub-filter bank. The response for $m = 0$ is highlighted.

Finally, upsampling and convolution with synthesis filters (which are similar to the QMF analysis filters as specified by (33)) results in the final stereo output signal.

The fact that the same filter-bank structure is used for both PS and SBR enables an easy and low-cost integration of SBR and parametric stereo in a single decoder structure (cf. [23, 24, 91, 92]). This combination is known as enhanced aacPlus and is under consideration for standardization in MPEG-4 as the HE-AAC/PS profile [93]. The structure of the decoder is shown in Figure 10. The incoming bit stream is demultiplexed into a band-limited AAC bit stream, SBR parameters, and parametric-stereo parameters. The AAC bit stream is decoded by an AAC decoder and fed into a 32-band QMF analysis bank. The output of this filter bank is processed by the SBR stage and by the sub-filter bank as described in Section 7. The resulting full-bandwidth mono signal is converted to stereo by the PS stage, which performs decorrelation and mixing. Finally, two hybrid QMF synthesis banks result in the final output signals. More details on enhanced aacPlus can be found in [23, 92].

8. PERCEPTUAL EVALUATION

To evaluate the parametric-stereo coder, two listening tests were conducted. The first test aims at establishing the *maximum perceptual quality* that can be obtained given the underlying spatial model. Other authors have argued that parametric-stereo coding techniques are only advantageous in the low-bit-rate range, since near transparency could not be achieved [20, 21, 22]. Therefore, this experiment is useful for two reasons: firstly, to verify statements by others on the maximum quality that can be obtained using parametric stereo, secondly, if parametric stereo is included in an audio coder, the maximum overall bit rate at which parametric stereo still leads to a coding gain compared to conventional stereo techniques is in part dependent on the quality limitations induced by the parametric-stereo algorithm

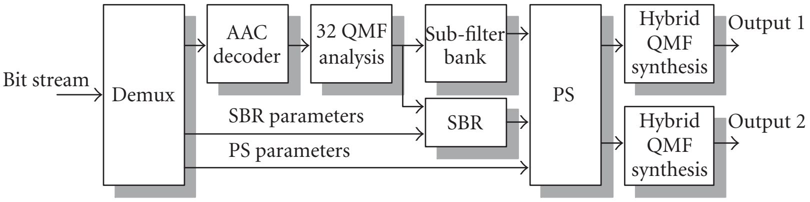

Figure 10: Structure of enhanced aacPlus.

only. To exclude quality limitations induced by other coding processes besides parametric stereo, this experiment was performed without a mono coder. The second listening test was performed to derive the actual coding gain of parametric stereo in a complete coder. For this purpose, a comparison was made between a state-of-the-art stereo coder (i.e., aacPlus) and the same coder extended with parametric stereo (e.g., enhanced aacPlus) as described in Section 7.

8.1. Listening test I

Nine listeners participated in this experiment. All listeners had experience in evaluating audio codecs and were specifically instructed to evaluate both the spatial audio quality as well as other noticeable artifacts. In a double-blind MUSHRA test [94], 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 anchors. 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 kbps 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 kbps;
(3) encoding and decoding using the FFT-based parametric-stereo coder without mono coder operating at a bit rate of 5 kbps (using 20 analysis frequency bands instead of 34);
(4) the original as hidden reference.

The 13 test excerpts are listed in Table 3. All items are stereo, 16-bit 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. The 13 selected items showed to be the most critical items from an 80-item test set for either parametric stereo or MP3 during development and in-between evaluations of the algorithms described in this paper. The 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 Figure 11. The top panel shows mean MUSHRA scores for 8 kbps parametric stereo (black bars) and MP3 at 128 kbps (white bars) as a function of the test item. The rightmost 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. These items contain many tonal components, a property that is typically problematic for waveform coders due to the large audibility of quantization noise for such material. On the other hand, excerpts 10 (“Man in the long black coat”) and 13 (“Two voices”) have higher scores for MP3. Item 13 exhibits an (unnaturally) large amount of channel separation, which is partially lost after parametric-stereo decoding. On average, both coders have equal scores.

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

The bottom panel shows 128 kbps 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 kbps, 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 MUSHRA scores between headphones and loudspeaker playback amounts to 17 points in favor of loudspeaker playback for an 8 kbps FFT-based encoder/decoder. 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 interchannel crosstalk that is inherently introduced by transmission of a single audio channel only is less than the amount of interaural 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 subcritical conditions using loudspeaker playback (cf. [16, 17, 18, 19, 20]).

Table 3: Description of test material.

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 disk

13	Two voices	Left = MPEG es03 = English female
		Right = MPEG es02 = German male

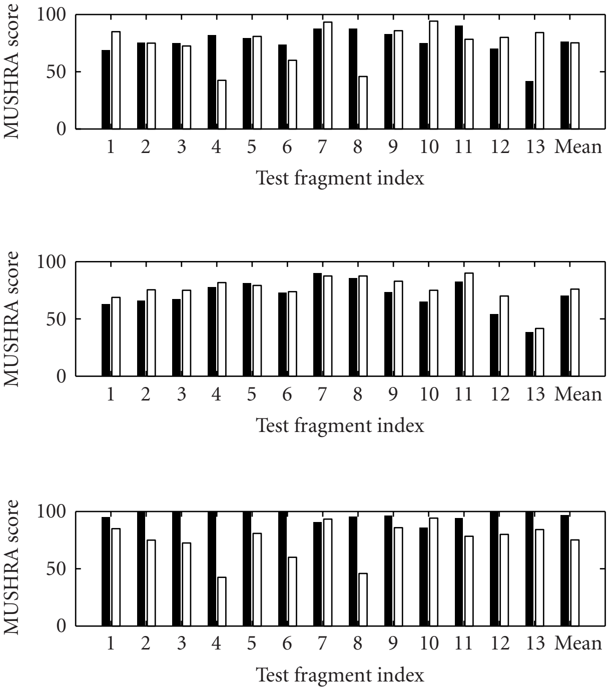

Figure 11: MUSHRA scores averaged across listeners as a function of test item and various coder configurations (see text). The upper panel shows the results for 8 kbps parametric stereo (black bars) against stereo MP3 at 128 kbps (white bars). The middle panel shows the results for 5 kbps parametric stereo (black bars) versus 8 kbps parametric stereo (white bars). The lower panel shows the hidden reference (black bars) versus MP3 at 128 kbps (white bars).

8.2. Listening test II

This test also employed MUSHRA [94] methodology and included 10 items which were selected for the MPEG-4 HE-AAC stereo verification test [95]. The following versions of each item were included in the test:

(1) the original as hidden reference;
(2) a first lowpass filtered anchor (3.5 kHz bandwidth);
(3) a second lowpass filtered anchor (7 kHz bandwidth);
(4) aacPlus (HE-AAC) encoded at a bitrate of 24 kbps;
(5) aacPlus (HE-AAC) encoded at a bit rate of 32 kbps;
(6) enhanced aacPlus (HE-AAC/PS) encoded at a total bit rate of 24 kbps. Twenty analysis bands were used, and no IPD or OPD parameters were transmitted. The average parameter update rate amounted to 46 milliseconds. For each frame, the required number of bits for the stereo parameters was calculated. The remaining number of bits was available for the mono coder (HE-AAC).

Two different test sites participated in the test, with 8 and 10 experienced subjects per site, respectively. All excerpts were presented over headphones. The results per site, averaged across excerpts, are given in Figure 12.

At both test sites, it was found that aacPlus with parametric stereo (enhanced aacPlus) at 24 kbps achieves a respectable average subjective quality of around 70 on a MUSHRA scale. Moreover, at 24 kbps, the subjective quality of enhanced aacPlus is equal to aacPlus at 32 kbps and significantly better than aacPlus at 24 kbps. These results indicate a coding gain for enhanced aacPlus of 25% over stereo aacPlus.

9. 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 kbps. It has been demonstrated that for headphone playback, a spatial parameter bit stream of 5 to 8 kbps 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 state-of-the-art coder such as aacPlus benefits from a significant reduction in bit rate without subjective quality loss if enhanced with parametric stereo.

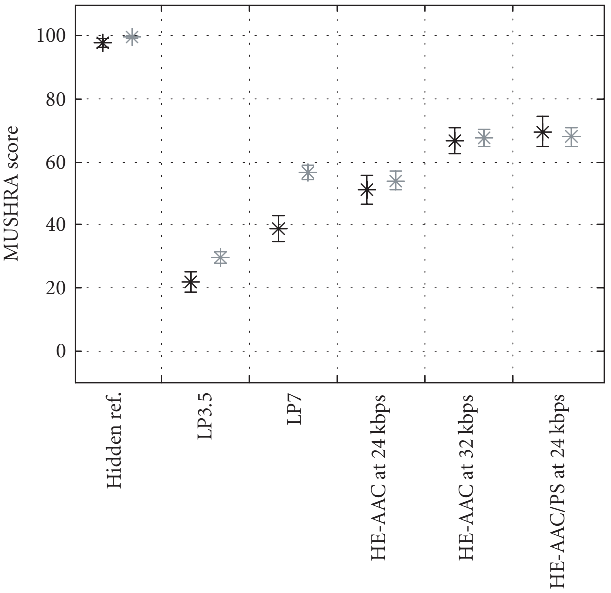

FIGURE 12: MUSHRA listening test results for two sites (black and gray symbols) showing mean grading and 95% confidence interval.

ACKNOWLEDGMENTS

We would like to thank the colleagues at Coding Technologies for their fruitful cooperation in the development of HE-AAC/PS and especially Heiko Purnagen and Jonas Engdegård for their valuable input for this paper. We would also like to thank our colleagues Rob Sluijter, Michel van Loon, and Gerard Hotho for their helpful comments on earlier versions of this manuscript. Furthermore, we would like to thank our colleagues at the Philips Digital Systems Lab for carrying out the listening tests. Finally, we would like to thank the anonymous reviewers for their thorough review and helpful suggestions to improve the manuscript.

REFERENCES

[1] K. Brandenburg and G. Stoll, "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] K. Brandenburg, "MP3 and AAC explained," in *Proc. 17th International AES Conference*, Florence, Italy, September 1999.

[3] A. C. den Brinker, E. G. P. Schuijers, and A. W. J. Oomen, "Parametric coding for high-quality audio," in *Proc. 112th AES Convention*, Munich, Germany, May 2002, preprint 5554.

[4] E. Schuijers, W. Oomen, B. den Brinker, and J. Breebaart, "Advances in parametric coding for high-quality audio," in *Proc. 114th AES Convention*, Amsterdam, The Netherlands, March 2003, preprint 5852.

[5] O. Kunz, "Enhancing MPEG-4 AAC by spectral band replication," in *Technical Sessions Proceedings of Workshop and Exhibition on MPEG-4 (WEMP4)*, pp. 41–44, San Jose, Calif, USA, June 2002.

[6] M. Dietz, L. Liljeryd, K. Kjörling, and O. Kunz, "Spectral band replication, a novel approach in audio coding," in *Proc. 112th AES Convention*, Munich, Germany, May 2002, preprint 5553.

[7] J. D. Johnston and A. J. Ferreira, "Sum-difference stereo transform coding", in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '92)*, vol. 2, pp. 569–572, San Francisco, Calif, USA, March 1992.

[8] R. G. van der Waal and R. N. J. Veldhuis, "Subband coding of stereophonic digital audio signals," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '91)*, Toronto, Ontario, Canada, April 1991.

[9] S.-S. Kuo and J. D. Johnston, "A study of why cross channel prediction is not applicable to perceptual audio coding," *IEEE Signal Processing Lett.*, vol. 8, no. 9, pp. 245–247, 2001.

[10] T. Liebchen, "Lossless audio coding using adaptive multichannel prediction," in *Proc. 113th AES Convention*, Los Angeles, Calif, USA, October 2002, preprint 5680.

[11] R. G. Klumpp and H. R. Eady, "Some measurements of interaural time difference thresholds," *Journal of the Acoustical Society of America*, vol. 28, pp. 859–860, 1956.

[12] J. Zwislocki and R. S. Feldman, "Just noticeable differences in dichotic phase," *Journal of the Acoustical Society of America*, vol. 28, pp. 860–864, 1956.

[13] J. D. Johnston and K. Brandenburg, "Wideband coding—Perceptual considerations for speech and music," in *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds., chapter 4, pp. 109–140, Marcel Dekker, New York, NY, USA, 1992.

[14] J. Herre, K. Brandenburg, and D. Lederer, "Intensity stereo coding," in *Proc. 96th AES Convention*, Amsterdam, The Netherlands, February–March 1994, preprint 3799.

[15] C. Faller and F. Baumgarte, "Efficient representation of spatial audio using perceptual parameterization," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '01)*, pp. 199–202, New Platz, NY, USA, October 2001.

[16] C. Faller and F. Baumgarte, "Binaural cue coding: a novel and efficient representation of spatial audio," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '02)*, vol. 2, pp. 1841–1844, Orlando, Fla, USA, May 2002.

[17] F. Baumgarte and C. Faller, "Design and evaluation of binaural cue coding schemes," in *Proc. 113th AES Convention*, Los Angeles, Calif, USA, October 2002, preprint 5706.

[18] F. Baumgarte and C. Faller, "Why binaural cue coding is better than intensity stereo coding," in *Proc. 112th AES Convention*, Munich, Germany, May 2002, preprint 5575.

[19] F. Baumgarte and C. Faller, "Estimation of auditory spatial cues for binaural cue coding," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '02)*, vol. 2, pp. 1801–1804, Orlando, Fla, USA, May 2002.

[20] C. Faller and F. Baumgarte, "Binaural cue coding applied to stereo and multi-channel audio compression," in *Proc. 112th AES Convention*, Munich, Germany, May 2002, preprint 5574.

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

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

[23] E. Schuijers, J. Breebaart, H. Purnhagen, and J. Engdegård, "Low complexity parametric stereo coding," in *Proc. 116th AES Convention*, Berlin, Germany, May 2004, preprint 6073.

[24] H. Purnhagen, J. Engdegård, J. Rödén, and L. Liljeryd, "Synthetic ambience in parametric stereo coding," in *Proc. 116th AES Convention*, Berlin, Germany, May 2004, preprint 6074.

[25] J. W. Strutt (Lord Rayleigh), "On our perception of sound direction," *Philosophical Magazine*, vol. 13, pp. 214–232, 1907.

[26] B. Sayers, "Acoustic image lateralization judgments with binaural tones," *Journal of the Acoustical Society of America*, vol. 36, pp. 923–926, 1964.

[27] E. R. Hafter and S. C. Carrier, "Masking-level differences obtained with pulsed tonal maskers," *Journal of the Acoustical Society of America*, vol. 47, pp. 1041–1047, 1970.

[28] 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.

[29] F. L. Wightman and D. J. Kistler, "Headphone simulation of free-field listening. I. Stimulus synthesis," *Journal of the Acoustical Society of America*, vol. 85, no. 2, pp. 858–867, 1989.

[30] B. Kollmeier and I Holube, "Auditory filter bandwidths in binaural and monaural listening conditions," *Journal of the Acoustical Society of America*, vol. 92, no. 4, pp. 1889–1901, 1992.

[31] M. 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," *Journal of the Acoustical Society of America*, vol. 103, no. 3, pp. 1609–1614, 1998.

[32] I. Holube, M. Kinkel, and B. Kollmeier, "Binaural and monaural auditory filter bandwidths and time constants in probe tone detection experiments," *Journal of the Acoustical Society of America*, vol. 104, no. 4, pp. 2412–2425, 1998.

[33] H. S. Colburn and N. I. Durlach, "Models of binaural interaction," in *Handbook of Perception*, E. C. Carterette and M. P. Friedman, Eds., vol. IV, pp. 467–518, Academic Press, New York, NY, USA, 1978.

[34] W. Lindemann, "Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals," *Journal of the Acoustical Society of America*, vol. 80, no. 6, pp. 1608–1622, 1986.

[35] R. M. Stern, A. S. Zeiberg, and C. Trahiotis, "Lateralization of complex binaural stimuli: A weighted-image model," *Journal of the Acoustical Society of America*, vol. 84, no. 1, pp. 156–165, 1988.

[36] W. Gaik, "Combined evaluation of interaural time and intensity differences: psychoacoustic results and computer modeling," *Journal of the Acoustical Society of America*, vol. 94, no. 1, pp. 98–110, 1993.

[37] J. Breebaart, S. van de Par, and A. Kohlrausch, "Binaural processing model based on contralateral inhibition. I. Model structure," *Journal of the Acoustical Society of America*, vol. 110, no. 2, pp. 1074–1088, 2001.

[38] J. W. Hall and M. A. Fernandes, "The role of monaural frequency selectivity in binaural analysis," *Journal of the Acoustical Society of America*, vol. 76, no. 2, pp. 435–439, 1984.

[39] A. Kohlrausch, "Auditory filter shape derived from binaural masking experiments," *Journal of the Acoustical Society of America*, vol. 84, no. 2, pp. 573–583, 1988.

[40] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1-2, pp. 103–138, 1990.

[41] B. Kollmeier and R. H. Gilkey, "Binaural forward and backward masking: evidence for sluggishness in binaural detection," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1709–1719, 1990.

[42] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, MIT Press, Cambridge, Mass, USA, 1997.

[43] J. Breebaart, S. van de Par, and A. Kohlrausch, "The contribution of static and dynamically varying ITDs and IIDs to binaural detection," *Journal of the Acoustical Society of America*, vol. 106, no. 2, pp. 979–992, 1999.

[44] L. A. Jeffress, "A place theory of sound localization," *Journal of Comparative and Physiological Psychology*, vol. 41, pp. 35–39, 1948.

[45] H. S. Colburn, "Theory of binaural interaction based on auditory-nerve data. II. Detection of tones in noise," *Journal of the Acoustical Society of America*, vol. 61, no. 2, pp. 525–533, 1977.

[46] R. M. Stern and G. D. Shear, "Lateralization and detection of low-frequency binaural stimuli: Effects of distribution of internal delay," *Journal of the Acoustical Society of America*, vol. 100, no. 4, pp. 2278–2288, 1996.

[47] L. R. Bernstein and C. Trahiotis, "The normalized correlation: accounting for binaural detection across center frequency," *Journal of the Acoustical Society of America*, vol. 100, no. 6, pp. 3774–3784, 1996.

[48] N. I. Durlach, "Equalization and cancellation theory of binaural masking-level differences," *Journal of the Acoustical Society of America*, vol. 35, pp. 1206–1218, 1963.

[49] D. M. Green, "Signal-detection analysis of equalization and cancellation model," *Journal of the Acoustical Society of America*, vol. 40, pp. 833–838, 1966.

[50] T. M. Shackleton, R. Meddis, and M. J. Hewitt, "Across frequency integration in a model of lateralization," *Journal of the Acoustical Society of America*, vol. 91, no. 4, pp. 2276–2279, 1992.

[51] 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 *Auditory Signal Processing: Physiology, Psychoacoustics, and Models*, D. Pressnitzer, A. de Cheveigné, S. McAdams, and L. Collet, Eds., Springer, New York, NY, USA, November 2004.

[52] R. M. Hershkowitz and N. I. Durlach, "Interaural time and amplitude jnds for a 500-Hz tone," *Journal of the Acoustical Society of America*, vol. 46, pp. 1464–1467, 1969.

[53] D. McFadden, L. A. Jeffress, and H. L. Ermey, "Difference in interaural phase and level in detection and lateralization: 250 Hz," *Journal of the Acoustical Society of America*, vol. 50, pp. 1484–1493, 1971.

[54] W. A. Yost, "Weber's fraction for the intensity of pure tones presented binaurally," *Perception and Psychophysics*, vol. 11, pp. 61–64, 1972.

[55] D. W. Grantham, "Interaural intensity discrimination: insensitivity at 1000 Hz," *Journal of the Acoustical Society of America*, vol. 75, no. 4, pp. 1191–1194, 1984.

[56] A. W. Mills, "Lateralization of high-frequency tones," *Journal of the Acoustical Society of America*, vol. 32, pp. 132–134, 1960.

[57] R. C. Rowland Jr. and J. V. Tobias, "Interaural intensity difference limen," *Journal of Speech and Hearing Research*, vol. 10, pp. 733–744, 1967.

[58] W. A. Yost and E. R. Hafter, "Lateralization," in *Directional Hearing*, W. A. Yost and G. Gourevitch, Eds., pp. 49–84, Springer, New York, NY, USA, 1987.

[59] L. A. Jeffress and D. McFadden, "Differences of interaural phase and level in detection and lateralization," *Journal of the Acoustical Society of America*, vol. 49, pp. 1169–1179, 1971.

[60] W. A. Yost, D. W. Nielsen, D. C. Tanis, and B. Bergert, "Tone-on-tone binaural masking with an antiphasic masker," *Perception and Psychophysics*, vol. 15, pp. 233–237, 1974.

[61] W. A. Yost, "Discrimination of interaural phase differences," *Journal of the Acoustical Society of America*, vol. 55, pp. 1299–1303, 1974.

[62] S. van de Par and A. Kohlrausch, "A new approach to comparing binaural masking level differences at low and high frequencies," *Journal of the Acoustical Society of America*, vol. 101, no. 3, pp. 1671–1680, 1997.

[63] L. R. Bernstein and C. Trahiotis, "The effects of signal duration on NoSo and NoSπ thresholds at 500 Hz and 4 kHz," *Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1776–1783, 1999.

[64] F. A. Bilsen and J. Raatgever, "Spectral dominance in binaural hearing," *Acustica*, vol. 28, pp. 131–132, 1973.

[65] F. A. Bilsen and J. Raatgever, "Spectral dominance in binaural lateralization," *Acustica*, vol. 28, pp. 131–132, 1977.

[66] D. E. Robinson and L. A. Jeffress, "Effect of varying the interaural noise correlation on the detectability of tonal signals," *Journal of the Acoustical Society of America*, vol. 35, pp. 1947–1952, 1963.

[67] T. L. Langford and L. A. Jeffress, "Effect of noise crosscorrelation on binaural signal detection," *Journal of the Acoustical Society of America*, vol. 36, pp. 1455–1458, 1964.

[68] K. J. Gabriel and H. S. Colburn, "Interaural correlation discrimination: I. Bandwidth and level dependence," *Journal of the Acoustical Society of America*, vol. 69, no. 5, pp. 1394–1401, 1981.

[69] J. F. Culling, H. S. Colburn, and M. Spurchise, "Interaural correlation sensitivity," *Journal of the Acoustical Society of America*, vol. 110, no. 2, pp. 1020–1029, 2001.

[70] J. W. Hall and A. D. G. Harvey, "NoSo and NoSπ thresholds as a function of masker level for narrow-band and wideband masking noise," *Journal of the Acoustical Society of America*, vol. 76, no. 6, pp. 1699–1703, 1984.

[71] L. R. Bernstein and C. Trahiotis, "Discrimination of interaural envelope correlation and its relation to binaural unmasking at high frequencies," *Journal of the Acoustical Society of America*, vol. 91, no. 1, pp. 306–316, 1992.

[72] L. R. Bernstein and C. Trahiotis, "The effects of randomizing values of interaural disparities on binaural detection and on discrimination of interaural correlation," *Journal of the Acoustical Society of America*, vol. 102, no. 2, pp. 1113–1120, 1997.

[73] U. T. Zwicker and E. Zwicker, "Binaural masking-level difference as a function of masker and test-signal duration," *Hearing Research*, vol. 13, no. 3, pp. 215–219, 1984.

[74] R. H. Wilson and C. G. Fowler, "Effects of signal duration on the 500-Hz masking-level difference," *Scandinavian Audiology*, vol. 15, no. 4, pp. 209–215, 1986.

[75] R. H. Wilson and R. A. Fugleberg, "Influence of signal duration on the masking-level difference," *Journal of Speech and Hearing Research*, vol. 30, no. 3, pp. 330–334, 1987.

[76] H. Wallach, E. B. Newman, and M. R. Rosenzweig, "The precedence effect in sound localization," *American Journal of Psychology*, vol. 62, pp. 315–336, 1949.

[77] P. M. Zurek, "The precedence effect and its possible role in the avoidance of interaural ambiguities," *Journal of the Acoustical Society of America*, vol. 67, no. 3, pp. 952–964, 1980.

[78] B. G. Shinn-Cunningham, P. M. Zurek, and N. I. Durlach, "Adjustment and discrimination measurements of the precedence effect," *Journal of the Acoustical Society of America*, vol. 93, no. 5, pp. 2923–2932, 1993.

[79] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *Journal of the Acoustical Society of America*, vol. 106, no. 4, pp. 1633–1654, 1999.

[80] S. P. Lipshitz, "Stereo microphone techniques; are the purists wrong?" *Journal of the Audio Engineering Society*, vol. 34, no. 9, pp. 716–744, 1986.

[81] V. Pulkki, M. Karjalainen, and J. Huopaniemi, "Analyzing virtual sound source attributes using a binaural auditory model," *Journal of the Audio Engineering Society*, vol. 47, no. 4, pp. 203–217, 1999.

[82] B. S. Atal and M. R. Schroeder, "Apparent sound source translator," US Patent 3,236,949, February 1966.

[83] A. J. M. Houtsma, C. Trahiotis, R. N. J. Veldhuis, and R. van der Waal, "Bit rate reduction and binaural masking release in digital coding of stereo sound," *Acustica/Acta Acustica*, vol. 82, pp. 908–909, 1996.

[84] A. J. M. Houtsma, C. Trahiotis, R. N. J. Veldhuis, and R. van der Waal, "Further bit rate reduction through binaural processing," *Acustica/Acta Acustica*, vol. 82, pp. 909–910, 1996.

[85] N. I. Durlach and H. S. Colburn, "Binaural phenomena," in *Handbook of Perception*, E. C. Carterette and M. P. Friedman, Eds., vol. IV, pp. 365–466, Academic Press, New York, NY, USA, 1978.

[86] S. E. Boehnke, S. E. Hall, and T. Marquardt, "Detection of static and dynamic changes in interaural correlation," *Journal of the Acoustical Society of America*, vol. 112, no. 4, pp. 1617–1626, 2002.

[87] H. Lauridsen, "Experiments concerning different kinds of room-acoustics recording," *Ingenioren*, vol. 47, 1954.

[88] M. R. Schroeder, "Synthesis of low-peak-factor signals and binary sequences with low autocorrelation," *IEEE Trans. Inform. Theory*, vol. 16, no. 1, pp. 85–89, 1970.

[89] R. Irwan and R. M. Aarts, "Two-to-five channel sound processing," *Journal of the Audio Engineering Society*, vol. 50, no. 11, pp. 914–926, 2002.

[90] M. Wolters, K. Kjörling, D. Homm, and H. Purnhagen, "A closer look into MPEG-4 high efficiency AAC," in *Proc. 115th AES Convention*, New York, NY, USA, October 2003, preprint 5871.

[91] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers, "High-quality parametric spatial audio coding at low bitrates," in *Proc. 116th AES Convention*, Berlin, Germany, May 2004, preprint 6072.

[92] H. Purnhagen, "Low complexity parametric stereo coding in MPEG-4," in *Proc. 7th International Conference on Digital Audio Effects (DAFx '04)*, Naples, Italy, October 2004, available: http://dafx04.na.infn.it/.

[93] ISO/IEC, "Coding of audio-visual objects—Part 3: Audio, AMENDMENT 1: Bandwidth Extension," ISO/IEC Int. Std. 14496-3:2001/Amd.1:2003, 2003.

[94] G. Stoll and F. Kozamernik, "EBU listening tests on internet audio codecs," in *EBU Technical Review*, no. 28, 2000.

[95] ISO/IEC JTC1/SC29/WG11, "Report on the Verification Tests of MPEG-4 High Efficiency AAC," ISO/IEC JTC1/SC29/WG11 N6009, October 2003.

Jeroen Breebaart was born in the Netherlands in 1970. He studied biomedical engineering at the Technical University Eindhoven. He received his Ph.D. degree in 2001 from the Institute for Perception Research (IPO) in the field of mathematical models of human spatial hearing. Currently, he is a researcher in the Digital Signal Processing Group, Philips Research Laboratories Eindhoven. His main fields of interest and expertise are spatial hearing, parametric stereo and multichannel audio coding, automatic audio content analysis, and audio signal processing tools. He published several papers on binaural detection, binaural modeling, and spatial audio coding. He also contributed to the development of parametric stereo coding algorithms as currently standardized in MPEG-4 and 3GPP.

Steven van de Par studied physics at the Technical University Eindhoven, and received his Ph.D. degree in 1998 from the Institute for Perception Research on a topic related to binaural hearing. As a Postdoc at the same institute, he studied auditory-visual interaction and he was a Guest Researcher at the University of Connecticut Health Centre. In the beginning of 2000, he joined Philips Research Laboratories in Eindhoven. Main fields of expertise are auditory and multisensory perception and low-bit-rate audio coding. He published various papers on binaural detection, auditory-visual synchrony perception, and audio-coding-related topics. He participated in several projects on low-bit-rate audio coding based on sinusoidal techniques and is presently participating in the EU project Adaptive Rate-Distortion Optimized audio codeR (ARDOR).

Armin Kohlrausch studied physics at the University of Göttingen, Germany, and specialized in acoustics. He received his M.S. degree in 1980 and his Ph.D. degree in 1984, both in perceptual aspects of sound. From 1985 until 1990, he worked at the Third Physical Institute, University of Göttingen, being responsible for research and teaching in the fields psychoacoustics and room acoustics. In 1991, he joined the Philips Research Laboratories in Eindhoven and worked in the Speech and Hearing Group of the Institute for Perception Research (IPO). Since 1998, he combines his work at Philips Research Laboratories with a Professor position for multisensory perception at the TU/e. In 2004, he was appointed a Research Fellow of Philips Research. He is a member of a great number of scientific societies, both in Europe and the US. Since 1998, he has been a Fellow of the Acoustical Society of America and serves currently as an Associate Editor for the Journal of the Acoustical Society of America, covering the areas of binaural and spatial hearing. His main scientific interest is in the experimental study and modelling of auditory and multisensory perception in humans and the transfer of this knowledge to industrial media applications.

Erik Schuijers was born in the Netherlands in 1976. He received the M.S. degree in electrical engineering from the Eindhoven University of Technology, the Netherlands, in 1999. Since 2000, he has joined the Sound Coding Group of Philips Digital Systems Laboratories in Eindhoven, the Netherlands. His main activity has been the research and development of the MPEG-4 parametric-audio and parametric-stereo coding tools. Currently Mr. Schuijers is contributing to the recent standardization of MPEG-4 spatial audio coding.