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

Synthetic Ambience in Parametric Stereo Coding

Jonas Engdegård[1], Heiko Purnhagen[1], Jonas Rödén[1], Lars Liljeryd[1]

[1]*Coding Technologies, Döbelnsgatan 64, 11352 Stockholm, Sweden*

Correspondence should be addressed to Jonas Engdegård (je@codingtechnologies.com)

ABSTRACT

Parametric stereo coding in combination with an efficient coder for the underlying monaural audio signal results in the most efficient coding scheme for stereo signals at very low bit rates available today. While techniques for lateral localization have been studied since early intensity stereo coding tools, synthesis of stereophonic ambience was only recently applied in parametric stereo coding systems. This paper studies different techniques for synthetic ambience generation in the context of parametric stereo coding systems and discusses their mono-compatibility. Implementations of these techniques in combination with mp3PRO and aacPlus are presented together with experimental results.

1 INTRODUCTION

Parametric stereo (PS) coding has recently proven to be a vital component in state-of-the-art audio coders, as it considerably enhances the performance for stereo coding at low and medium bit rates. Here, the term "parametric stereo" denotes a system that reconstructs a 2-channel stereo signal from a full bandwidth mono signal and so-called stereo parameters. In a complete PS coding system, the underlying mono signal is conveyed using a state-of-the-art audio coder, and the stereo parameters are added as side information to the bit stream.

Techniques for joint stereo coding [1, Section 11.2.8] have long been used in audio coding systems like those standardized in MPEG-1 [2] and MPEG-2 [3]. While

mid/side (M/S) coding still requires the transmission of two signals, intensity stereo (IS) can be seen as a parametric stereo technique. However, IS only permits to control lateral localization of a sound, using a parameter that acts similar to the "pan-pot" on a mixing desk, while it fails to re-create stereophonic ambience that might have been present in the original signal. Furthermore, IS is typically only used for high frequency range, combined with normal stereo coding for the remaining low frequency range.

Techniques to create a pseudo-stereo signal from a mono signal have long been known and can employ e.g. simple phase shifting techniques [4]. Also room acoustic simulations or (stereo) reverberation systems [5] can convert a mono into a stereo signal. Such techniques, however, significantly influence signal properties like the perceived room size. By integrating pseudo-stereo generation techniques in the decoder of a parametric stereo coding system, it is now possible to re-create also stereophonic ambience. Different from normal pseudo-stereo systems, the amount and nature of the ambience synthesized in the decoder is controlled by additional parameters that are estimated in the encoder and conveyed as side information. This approach makes it possible to reproduce the characteristics of the ambience present in the original signal. The mp3PRO coding system [6] released in 2001 was the first coding system to implement a very simple version of a synthetic ambience parametric stereo coding technique.

Since then, there has been much progress in parametric stereo coding in general and synthetic ambience generation for such systems in particular. Coding systems utilizing parametric stereo are now deployed in commercial systems, like Digital Radio Mondiale (DRM) [7], and the specification of an efficient low complexity parametric stereo coding tool [8] was recently finalized as a part of the MPEG-4 standard [9].

This paper illuminates aspects of how to reproduce, from a mono (downmixed) signal, an appropriate stereophonic ambience, that is, the uncorrelated components of the left and right channel. The need of such inter-channel decorrelation in PS systems has been acknowledged also by other recent publications, like [10]. However, appropriate methods for generating such a decorrelated "synthetic ambience" signal have received little attention until now. In this paper, we will provide a survey of applicable techniques and present the design of efficient synthetic ambience generators specifically developed for parametric stereo coding.

This paper will mainly focus on PS as an enhancement to aacPlus, the combination of Spectral Band Replication (SBR) and Advanced Audio Coding (AAC) that was standardized by MPEG in 2003 as "High Efficiency AAC" (HE-AAC) [11]. The combination of PS with aacPlus is referred to as "enhanced aacPlus." Specific technical advantages of this combination as well as further details on stereo parameters and integration aspects are presented in [8]. More details about the stereo parameter processing can be found in [12].

The structure of this paper is as follows. Section 2 outlines the main principles of PS coding. In Section 3, motivation of and background to the problem of synthetic ambience, i.e., stereo decorrelation, are given. This is followed by a detailed discussion of decorrelator design aspects. Section 4 presents existing PS systems and the integration of decorrelation tools in enhanced aacPlus. Finally, conclusions are drawn in Section 5.

2 THE PARAMETRIC STEREO CODING PARADIGM

2.1 Overview

PS coding is based on perceptual inter-channel redundancy and takes it much further than other joint stereo coding methods such as e.g. mid/side (M/S) coding [1, Section 11.2.8]. In M/S coding, the left and right channels are matrixed into mid and side channels, primarily to profit from the sometimes significantly lower side channel energy. Those large energy differences can often be observed for only certain frequency bands, which also can be exploited. Similar to the mid signal in M/S coding, also PS coding transmits a (downmixed) mono core signal that does not include stereo information. However, different from M/S coding, where also the side signal carrying the stereo information is transmitted, PS coding reduces this stereo information to a concise parameteric respresentation.

As an example of the quality gain possible by PS coding at a given bit rate, it can be compared to traditional left-right ("dual mono") stereo coding where both channels each get 50% of the available bit rate. If the stereo parameters in PS coding uses 10% of the total bit rate, while letting the mono core coder have the remaining 90%, the mono coder will thus have a 80% higher bit rate available compared to left-right stereo coding. Even

though M/S coding can be more efficient than left-right coding, also compared to M/S coding there is usually a considerable quality gain of the mono core signal when using PS. For higher bit rates though, when the quality of the mono coder approaches perceptual transparency, the overall quality of a PS-based coder can suffer more from imperfections in the PS stereo reconstruction than from coding artifacts in the mono core.

Figure 1 depicts the signal flow of a PS-based encoder from a top level view and shows how PS is combined with an arbitrary audio coder operating in mono. The mono compatibility is clearly illustrated here as the mono encoded core bit stream is not affected by the stereo parameter extraction and the PS bit stream. When decoding a bit stream containing PS information as in Figure 2, the mono core is decoded separately and independently from all parts of the PS information. Hence, any decoder not supporting PS will be able to sucessfully decode the mono bit stream. This behavior requires an appropriate bit stream syntax that allows such extension formats. This is the case with aacPlus [11].

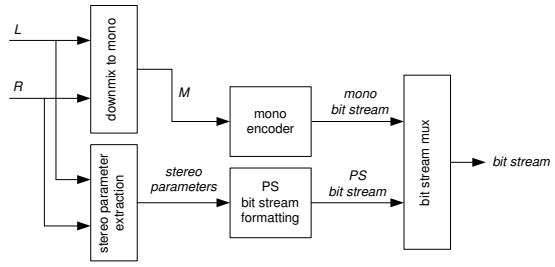

Figure 1: *PS integration into a mono encoder.*

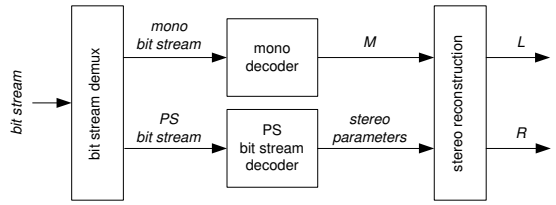

Figure 2: *PS integration into a mono decoder.*

2.2 QMF bank

Essential parts of the PS system are the time-to-frequency (t/f) and frequency-to-time (f/t) transforms required to enable frequency-selective processing. Experiments have shown that the complex-exponential modulated (Pseudo) Quadrature Mirror Filter (QMF) bank is very well suited for PS. A more detailed introduction of using the QMF bank as a flexible signal modifier can be found in [13]. The QMF bank used in aacPlus is a 64 channel complex-valued filter-bank with near alias-free behavior even when altering the gains of neighboring subbands excessively, which is a fundamental requirement for use as t/f transform in a PS system. Also the fact that the QMF is the native filter-bank in aacPlus is important. In the enhanced aacPlus decoder, the re-use of this aacPlus framework enables significant savings in computational complexity, which make it possible for a decoder utilizing PS to stay within the same complexity range as a normal stereo decoder.

Band	Frequency range	Bandwidth
0	0–86 Hz	86 Hz
1	86–172 Hz	86 Hz
2	172–258 Hz	86 Hz
3	258–345 Hz	86 Hz
4	345–517 Hz	172 Hz
5	517–689 Hz	172 Hz
6	689–861 Hz	172 Hz
7	861–1034 Hz	172 Hz
8–68	1034–22050 Hz	345 Hz

Table 1: *Frequency resolution of the hybrid filter-bank for baseline configuration. Figures are based on 44.1 kHz sampling rate.*

The 64 channel QMF bank offers linearly spaced frequency bands, hence the effective bandwidth for a sampling rate of 44.1 kHz is approximately 345 Hz per frequency band. This frequency resolution is far lower than suggested bandwidth scales, like the Equivalent Rectangular Bandwidth (ERB) scale, that are more relevant to auditory perception [8]. This motivates the introduction of a hybrid filter-bank structure, comprising additional sub-band filtering for the lower QMF bands. The complete hybrid filter-bank presented in [8] achieves effective bandwidths of approximately 86 Hz for the lowest bands up to the inherent 345 Hz for the upper bands. In Table 1, the frequency division of the baseline configuration hybrid filter-bank is shown, as defined for the MPEG PS tool [9]. In addition, the MPEG PS tool also provides an advanced hybrid filter-bank with an even higher frequency resolution offering in total 75 bands.

3 GENERATION OF SYNTHETIC AMBIENCE

3.1 Motivation

Typical parameters used for describing stereo image properties are inter-channel intensity difference (IID) and inter-channel time difference (ITD), where ITD also could be substituted by the inter-channel phase difference (IPD) for a multi-band representation. The basic principles for analysis and synthesis of those inter-channel properties are rather uncomplicated. For the special case of stereo signals only containing static stereo information in terms of full-bandwith IID and ITD, nearly perfect reconstruction of the stereo signal can be achieved, assuming that parameter quantization is fine enough.

There is reason to believe that these parameters are not sufficient to complete the parameterized representation of the spatial scene. This motivates the introduction of an inter-channel coherence (IC) parameter. Unlike the IID and ITD parameters, IC does not have the same obvious relation between analysis and synthesis. For the above mentioned special case with full-bandwith static stereo information but including IC only, reconstruction of the side signal (in an M/S sense) is generally not possible at all. However, from a human perception point of view, the impression of spatial width due to low inter-channel coherence can be approximated by decorrelation methods as discussed in this paper.

Various decorrelation methods are based on very simplified models of real signal situations. Consider the single source case as an acoustical model containing one source $x(t)$ and a coincident stereo microphone delivering the signal $l(t)$, $r(t)$ in a room or other acoustic environment:

$$l(t) = g_l x(t) + h_l(t) * x(t) \quad (1)$$

$$r(t) = g_r x(t) + h_r(t) * x(t) \quad (2)$$

where $*$ denotes convolution. Here, g_l and g_r describe the different amplitudes due to the directional characteristics of the stereo microphone. The functions $h_l(t)$ and $h_r(t)$ are the room impulse response functions describing all non-direct sound, i.e., both early and late reflections and the reverberation in the room.

From the microphone signals, a downmixed mono signal $m(t)$ can be derived e.g. as follows:

$$m(t) = (g_l + g_r)x(t) + (h_l(t) + h_r(t)) * x(t) \quad (3)$$

Given this mono signal $m(t)$, a stereo signal $l'(t)$, $r'(t)$ can be synthesized in order to approximate the original microphone signal:

$$l'(t) = g'_l m(t) + g'_{dl} h_d(t) * m(t) \quad (4)$$

$$r'(t) = g'_r m(t) - g'_{dr} h_d(t) * m(t) \quad (5)$$

Here, g'_l and g'_r are used to re-create the original level balance between the left and right channel as determined in the encoder. g'_{dl} and g'_{dr} control the level of the decorrelated signal in left and right channel, respectively. These gain values are determined by a decorrelation estimator in the encoder.

To reconstruct the stereo ambience, a decorrelation process using a filter with the impulse response $h_d(t)$ is applied to the downmixed signal $m(t)$. As can be seen in Equation 3, $m(t)$ already contains both the left and right part of the reverberated signal. To decompose those two signals would be fairly complicated and is out of scope for this paper. Instead the signal is decorrelated in a more general way and added to both synthesized channels but with opposite sign (Equations 4, 5). The signal $m(t)$, which is the input to the decorrelation process, of course also contains the direct source signal $x(t)$. Unfortunately, this is not desirable for ambience generation. However, in this context, it should be noted that both the direct signal gains g'_l, g'_r as well as the decorrelation gains g'_{dl}, g'_{dr} are time- and frequency-dependent. Hence, assuming that each region in the time-frequency plane is dominated either by the direct signal or by the decorrelated signal, appropriate control of the gain factors g' allows to minimize this effect.

If complex material containing multiple sources is considered, the simplified signal model above does not apply anymore. Multiple source materials including modern sound mixes, i.e., mixes featuring an artificial acoustical stage, do not have the same inherit separation of direct source and ambience. Instead of modeling such signals, it is suggested to study the inter-aural properties of human perception for arbitrary signals. It is believed that when too many independent sources are presented as binaural information in a limited frequency region, human perception cannot distinguish between the spatial properties of the different sources. Thereupon, a general stereo decorrelator without other a priori information could generate a similar spatial impression for such exposed frequency regions.

3.2 Known decorrelation solutions

There are a couple of methods available for creation of decorrelated signals. Ideally, a linear time invariant (LTI) function with all-pass frequency response is desired. One obvious method for achieving this is by using a constant delay, which is the approach for the pseudo-stereo tool in mp3PRO [6]. This pseudo-stereo tool, usually referred to as "low complexity stereo," is intended for very low bit rates and is a native part of the mp3PRO format. Similar techniques can be found in an AES paper as early as 1957 [4]. However, using a delay, or any other LTI all-pass functions, will result in non-all-pass response after adding the non-processed signal. In the case of a delay, the result will be a typical comb-filter. The comb-filter often gives an undesirable "metallic" sound that, even if the stereo widening effect can be efficient, reduces much naturalness of the original. In mp3PRO, above problem is addressed by using a filter to shape the spectrum of the side signal in a manner that the comb-filter effect will be perceptually less harmful. This filter has large attenuation for lower frequencies (under approximately 300 Hz) and also a negative slope for the higher frequencies. Since the level of the decorrelation signal is conveyed in the mp3PRO bit stream as only one parameter with no frequency selectivity, the filter is a necessary compromise. A major difficulty with this "one frequency band" decorrelation parameter is the estimation in the encoder. The typical case of such problem are speech-music transitions, where the music benefits from decorrelation while e.g. the overlayed speech clearly suffers from it. In the mp3PRO encoder, these cases are taken care of by a dedicated detector. For a multi-band system however, decorrelation parameter extraction could be trivial using straight-forward cross-correlation estimates.

In Figure 3 a typical implementation of a stereo decorrelator based on a simplified version of Equations 4, 5 is shown. Here, a pseudo-stereo signal is created by adding the original mono signal processed by the transfer function $H_d(z)$ to the original. In the figure, g' denotes the gain for the mono signal and g'_d denotes the gain of the decorrelated signal. This figure depicts the typical concept of the low complexity stereo tool in mp3PRO, where $H_d(z)$ corresponds to a delay and the gain, g'_d is controlled by a width parameter, which is conveyed in the bit stream. Assuming $g' = 1$, the figure also nicely shows the mono-compatibility of this approach.

For the single delay solution, an important design parameter is the delay length. A too short delay might lead to perceived smaller stereo width and worse comb-filter problems, while a too long delay introduces audible echoes in case of transients in the original signal. A possible alternative is to choose a delay length which satisfies typical auditory time constants at different frequency bands, hence a delay length that decreases with increasing frequency. If the delay decreases linearly with frequency, the resulting transfer function of such a delay would have the impulse response of a chirp signal [14]. The chirp approach certainly helps, since it does not introduce the in frequency linearly spaced notches and nodes as is characteristic for a comb-filter. However, it also introduces other artifacts e.g. due to the somewhat funny transient response which in extreme cases actually can be perceived as a chirp.

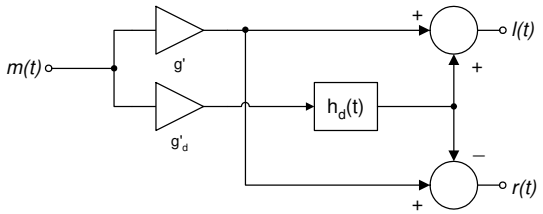

Figure 3: *The basic principle of a stereo decorrelator.*

For more advanced functions as decorrelation methods, one could study the art of simulating reverberation using LTI systems such as IIR or FIR filters [5]. Typical properties of a well-designed artificial reverberator are high resonance density with not too regular resonances/notches along the frequency axis, and a diffuse impulse response that is noise-like and with no repetitive pulses. These design aspects are indeed also very relevant for the design of a stereo decorrelator. More about artificial reverberators can be found in Sections 3.3 and 3.4.

The decorrelation methods discussed above all operate on time domain signals or on sub-samples produced by a very short t/f transform, as e.g. the 64 band QMF. If a long t/f transform is used, applying delay based decorrelation gets more difficult. An obvious, but surely not from a complexity point of view beneficial, solution is to perform decorrelation in the time domain and subsequently make a separate t/f transform for the decorrelated signal. It would however be considerably more efficient

to implement the decorrelator as a frequency domain process. Such an optimization was proposed in [15], but might have a slight effect on the perceptual quality.

Another example of frequency domain decorrelation is suggested by Faller and Baumgarte (2003) [16]. A random sequence was added to the IID values along the frequency axis, where different sequences were used for the different audio channels, i.e., left and right. It was possible to apply the random sequence and still preserve the main IID properties, as the frequency domain representation offered high frequency resolution. Overlapping DFTs with a size of 1024 were typically used which enables modifications of several DFT bins within one frequency band that corresponds to one stereo parameter, where such frequency bands normally have a size that is similar to critical bands as known from psychoacoustics.

One problem with frequency domain decorrelation by the random sequence modifications is the introduction of pre-echoes. Subjective tests have shown that for non-stationary signals, pre-echoes are by far more annoying than post-echoes, which is also well supported by established psycho acoustical principles. This problem could be reduced by dynamically adapting transform sizes to the signal characteristics in terms of transient content. However, switching transform sizes is always a hard (i.e., binary) decision that affects the full signal bandwidth and that can be difficult to accomplish in a robust manner.

3.3 Artificial reverberators

Throughout history of audio engineering, designing artificial reverberators has always been one of the most debated topics that rarely lacks attention. For true acoustical modeling and for purists, the actual impulse response of the room is desired in order to make a filter realization based on a long FIR filter. For any real-time application, this method is generally considered as extremely expensive, as the impulse response usually extends to several seconds. Therefore, practical artificial reverberation design usually requires an IIR filter based approach. Hence, it is not surprising that research in this field in the recent decades has to a large extent focused on how to get down complexity while preserving quality.

Elemental modeling of room acoustics uses sets of parallel connected delay loops, also known as comb-filters, where every loop unit could correspond to a one-dimensional resonance mode in the room. In such structure, the modal density of the reverberator will increase with the number of parallel comb-filter links added, assuming that the delay line lengths are wisely chosen. M. R. Schroeder (1961) [17] suggested using all-pass filters instead of comb-filters in order to achieve “colorless” artificial reverberation. Further research led to a variety of combinations of all-pass and comb-filters, where shorter all-pass filters often were used to improve echo density [5]. Modern research in this field tends to prefer feedback delay networks (FDN) as framework to describe such systems, where FDN allows a more generic notation for delay networks using matrix algebra.

The behavior of those reverberation simulating filter networks described here, has much in common with real room acoustics. Due to the delay feedback structure of such filters, a build-up phase and a decay phase can be clearly identified. During the build-up phase the echo density is constantly increasing as the output from the different delay loops combine and the all-pass filters start multiplying echoes. This can be a problem if high echo density is needed at an early stage. If an FIR filter is not an option e.g. as a complement, the delay line lengths of the IIR filter have to be scaled down. Normally, it is not trivial to do so, at least not while preserving quality. One obvious reason is the fact that delay line lengths preferably are chosen as mutually prime numbers. An efficient solution to this problem is discussed in Section 3.4.

A conclusion that is still valid today is the necessity to use a large number of comb-filters and/or all-pass filters to achieve satisfactory modal and echo density. This might appear a bit contradictory to the design criteria and constrains of the stereo decorrelator discussed in this paper. Yet, the primary goals are in common, as to achieve maximal modal and echo density by a minimal complex design.

3.4 The all-pass (IIR) approach

Due to the fact that design constrains, in terms of computational complexity and memory requirements, practically always are present, the IIR filter is the only reasonable approach if the desired reverberation time corresponds to a larger number of samples. The special case of realizing such an implementation within a filterbank gives certain advantages even though some disadvantages are obvious. The QMF bank used here is, because of its complex valued structure, oversampled by a factor two. This naturally implies a complexity increase of the same factor, both computational and memory wise.

However, complexity can be reduced by introducing a bandwidth limitation of the decorrelation process. This is easily achieved when operating within the QMF bank by just processing up to the bandwidth limitation, leaving the upper QMF bands unprocessed. This is further discussed in Section 3.6. One could easily imagine that e.g. discarding a few kHz in the upper region of a 44.1 or 48 kHz sampled signal would be a reasonable compromise also from a quality point of view. Furthermore, processing the different QMF frequency bands individually and totally independently, opens up many new, probably not much exploited, possibilities to perform the reverberation process frequency dependent. Applying a decay time that decreases with higher frequencies can be a vital part of a time domain reverberation system. The corresponding functionality in a QMF bank is implemented by the quite straight-forward method of using different decay coefficients in the different QMF frequency bands.

Choosing delay line lengths is one of the most crucial parts in reverberator design. Best results are usually obtained by using lengths that are large numbers and are mutually prime. This is a problem in the QMF domain because of insufficient time resolution. The sub-samples in the QMF domain are the result of 64 times subsampling, hence for a 44.1 kHz original signal, a sample period T of approximately 1.5 ms is obtained within the filter-bank. To access finer time resolution, delay by fractions of the sampling interval has to be considered. Fractional delay can easily be approximated by rotating the phase of the complex (approximately analytical) QMF sub-band signals by the angle that corresponds to the desirable fraction of a sample at the center frequency of the QMF band in question. The delay fraction, δ is defined as the fraction of a QMF sub-sample. Hence, if a delay time of τ is desired the delay fraction, δ is calculated as:

$$\delta = \frac{\tau}{T} \tag{6}$$

where T denotes the sample period of the QMF sub-band signals. Because of the cyclical properties of the phase rotation, δ should only be defined within the range $[-0.5, 0.5]$.

For a given δ, the phase rotation for the N QMF bands are below defined by the phase vector α:

$$\alpha_k = \delta\pi(k+0.5) \tag{7}$$

where k is the QMF band index, $k = 0, 1, ..., N-1$ and hence the complex samples x_k should be rotated according to:

$$x'_k = x_k e^{i\alpha_k} \tag{8}$$

This fractional delay implementation using the QMF bank only gives a rough approximation of a delayed signal. Admittedly, any fractional delayed signal is an approximation, since an exact version would always spread in time from $-\infty$ to ∞. This results in a somewhat time-smeared signal coming out from the fractional delay. Interestingly, this can be a positive effect for an artificial reverberator striving for a more diffuse time response.

In Figure 4, the resulting impulse response and spectrum are plotted for such a decorrelation process that uses a set of all-pass filter in the QMF bands. The responses of two alternative systems are shown, one with and one without fractional delays. In both time plots the impulse occurs at $t = 0$. In the time plot for the system without fractional delay, the individual echoes can clearly been identified. In the fractional delay case however, the impulse response appears to be more noise-like, which also is indicated by its power spectrum plot. It can be noted that the impulse response of the fractional delay system has the typical appearance of the impulse response of a reverberator, as it decays with increasing echo density. The non-fractional delay system achieves only a limited echo density since its echoes are quantized in time to QMF sub-samples. Apparently, the fractional delay works most efficiently as it significantly improves both time and frequency behavior in terms of echo and modal density, respectively. This quality improvement provided by the fractional delay has also been confirm by subjective listening impressions.

A widely used technique to increase perceived modal density is to use time-varying delay lengths. This can be implemented by letting a low frequency oscillator (LFO) control small delay length deviations by inter-sample interpolation. Even though the modulation depth is kept low to avoid audible vibrato effects, it is difficult to reduce it enough, while keeping its desirable smoothing properties. This makes it hard to apply it in the field of natural audio coding.

Conventional reverberation algorithms usually give an increasing impulse density over time. As the typical comb-filter delay lines often are in the range of 40–80 ms, it takes some time to build up a high degree of

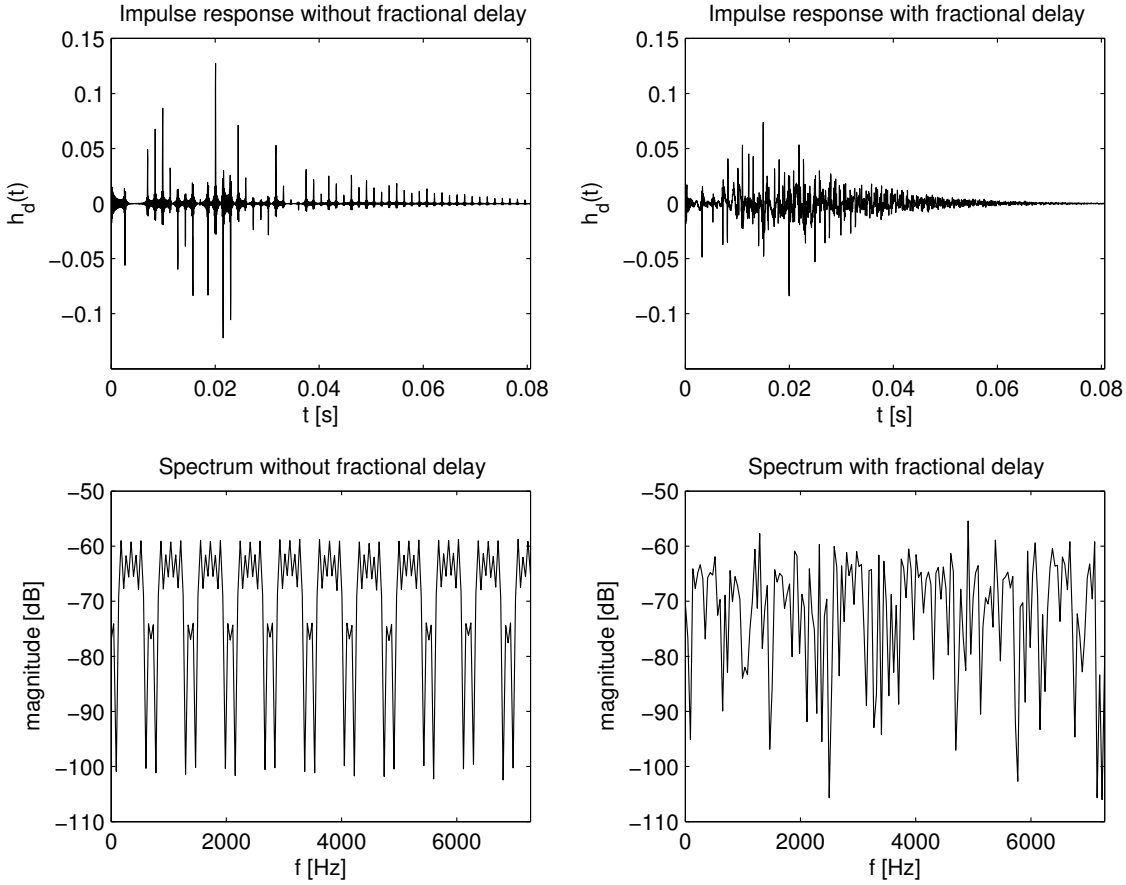

Figure 4: *Impulse response of short IIR reverberator, with and without fractional delay, and corresponding power spectra.*

diffusion (impulse density). As a decorrelator tool this is not fast enough. An immediate diffuse time response is preferred, which actually would speak for the FIR filter design. Empirical research has shown that a reverberation time (RT60) of about 80 ms or below, can be useful for a stereo decorrelator without perceptually adding or changing the perceived room size of the original signal, under the condition that this signal does not contain strong transients. RT60 denotes the time for a 60 dB decay of the reverberator response. By using three serially connected all-pass links with the transfer functions:

$$H_n(z) = \frac{g(n) + z^{-m(n)}}{1 + g(n) z^{-m(n)}} \tag{9}$$

a reverberator structure is obtained, reminding of the attempts to do very simple artificial reverberators in the early days. Here, the delay lengths, m (measured in QMF samples) for the three links $n = 0, 1, 2$ are:

$$m(0) = 3 \tag{10}$$

$$m(1) = 4 \tag{11}$$

$$m(2) = 5 \tag{12}$$

and the filter coefficients, $g(n)$ for the three links $n = 0, 1, 2$ are tuned to make all links have the same decay. As an additional pre-delay, the whole system is delayed by two QMF samples. This structure would hardly be useful for a conventional reverberator design, at least not if targeting longer decay times. Nevertheless, its impulse response reads some nicely located early echos, which

is desired. The complete reverberator, including fractional delay and a continuously changing decay over frequency, has a build-up phase of approximately 15–20 ms followed by a decay of about 0.75 dB/ms.

3.5 Improving transient behavior

A major problem when incorporating delays or all-pass filters that include long decays into a stereo decorrelator, is the performance at transients due to the risk of generating audible post-echoes or unnatural colorization. Two well-known solutions to this problem are a) to use a sufficiently short delay or b) to adapt the delay time to the transient conditions by explicit signaling such cases. The first method mentioned is simple but compromises quality for the general, non-transient case. The latter is more sophisticated and can take advantage of having a decorrelator that is optimized for several signal cases and by that, adapts itself. On the other hand, to signal transients explicitely might increase bit rate and keeping signaling overhead low often suggests using hard decisions, whereas transients can appear in various magnitudes.

A third and novel approach, very advantageous in this context, is to detect transients on the decoder side, i.e., within the PS synthesis, and thereafter reduce the level of the decorrelated signal using soft decisions. A certain consequence of such scheme is the inability to decorrelate the actual transients. However, a delay based decorrelation method tends to decorrelate the echo after the transient rather than the actual transient, anyway. It is preferred to let the transient reduction scheme operate in different independent frequency bands, in order to deal with transient-like signals within complex audio material. As the transient reduction process removes reverberation only from transients, this solution also addresses the problem with the decorrelator enlarging the perceived room size. This is evident since transient-like signals to a larger extent makes the reverberator perceptible.

The transient reduction works in an efficient and robust manner, when it attenuates post-echoes that potentially degrade quality. A corresponding approach built upon a frequency domain decorrelator would not be easily implemented. Hence, a frequency domain decorrelator that handles transients in a proper way would be hard to design without incorporating adaptive window switching, which in turn affects the full signal spectrum and offers no frequency selectivity. Again, the time domain approach employing the QMF bank proves to have advantages well suited for PS.

Figure 5 shows the behavior of the transient reducing process. The original input signal used in this experiment is a dirac-like impulse run through a conventional artificial reverberator. Typical settings were used for the reverberator and a decay time of approximately 2.5 seconds (RT60) was chosen. Also shown are the processed signals containing the sum of the original and the decorrelated signals (i.e., synthetic ambience), with and without the transient reducer. It can be observed from the figure that the impulse response of the decorrelator is much attenuated when applying the transient reduction process. Hence, the reverberation characteristics (e.g. percieved room size) of the original signal remains nearly unchanged.

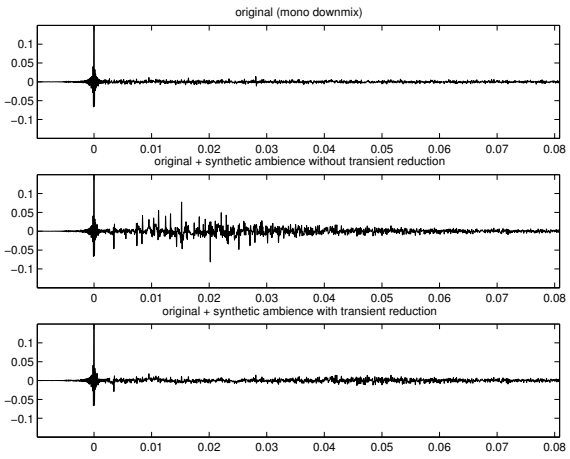

Figure 5: *Impulse fed to a conventional synthetic reverberator. Top: original mono downmixed signal (no decorrelation). Center: original signal + short IIR reverberator decorrelation without transient reduction. Bottom: original signal + short IIR reverberator decorrelation with transient reduction.*

3.6 Complexity aspects

For complexity reasons, one would prefer to keep the reverberation bandwidth, i.e., the audio bandwidth of the decorrelated signal, as low as possible. A rather successful compromise is to combine an all-pass reverberation structure with a simple delay for the higher frequencies. This saves both computational complexity as well as static memory buffers, as the delay lengths in the upper frequency region do not have to be very long and require quite few calculations compared to the reverberation algorithm. It has been proven by experiments that the unpleasant comb-filter effect and other draw-

backs with the delay methods are much less significant for higher frequencies, something that is exploited here.

An alternative complexity reduction method is to discard any decorrelation above a certain frequency. Such bandwidth limitation can be motivated by the minor importance of decorrelation at high frequencies compared to the lower. This is used in the PS system defined in Digital Radio Mondial (DRM) [7], where decorrelation only operates up to about 8.5 kHz. Higher up, only intensity stereo is applied by means of IID processing.

4 PARAMETRIC STEREO CODING UTILIZING SYNTHETIC AMBIENCE

4.1 Early systems

Parametric stereo today has only a few industrial applications but is expected to play an increasing role in the near future for low bit rate audio coding. Some systems existing today are the enhanced aacPlus as defined in MPEG, aacPlus in DRM and the quite minimalistic version of parametric stereo as used in mp3PRO.

Probably the first commercial system including a very simple parametric stereo tool was mp3PRO. For low bit rates, mp3PRO allows a pseudo-stereo mode which excites the mono decoded signal into a wide stereo signal. The stereo width is controlled by only two bits per frame or even less at unchanged stereo width, which corresponds to a maximum side information rate of about 77 bit/s for 44.1 kHz sampling rate. This quite limited stereo tool obviously only provides a stereo output with equal signal power in left and right channel and hence would not qualify as a tool for spatial localization or true reconstruction of the stereo scene. Still its extremely low complexity and bit rate overhead have made it hard to replace.

The parametric stereo included in DRM [7] is a scaled down version of the PS tool recently defined in the MPEG-4 standard [18], [19], [9]. It utilizes the same QMF bank since it is already included in the mono core coder (aacPlus) but lacks of the additional low frequency filtering which increases low frequency resolution. Also the bit stream and parameterization are simplified and a bit less flexible. These simplifications relaxes the PS synthesis scheme complexity wise to some extent. DRM was the first commercial broadcasting service using a parametric stereo coder that includes both intensity and decorrelation parameters.

4.2 Integration with aacPlus

The integration of PS into aacPlus is a most beneficial concept due to the already present QMF bank. At mono decoding, aacPlus provides a complete QMF spectrum immediately prior to the QMF synthesis filter-bank, which is the final processing step in the decoder. When extending a mono aacPlus decoder to stereo by integrating PS decoding, another QMF synthesis channel has to be added, as well as additional filtering of a predetermined number of low-frequency QMF bands to achieve the hybrid filter-bank structure described above. Since the core decoder operates in mono mode, large memory buffers (that would be required for the second channel in normal stereo decoding) are available to be re-used by the PS system. This results in practically no extra RAM needed for an aacPlus decoder supporting PS (enhanced aacPlus). The same holds true for the computational complexity, where PS decoding gives about the same MIPS figures as for traditional stereo decoding. On the encoder side, even a significant decrease of complexity is achieved compared to standards stereo encoding, which might be of interesting for e.g. hand-held terminal applications. These arguments let the reverberation method as decorrelator, compared to just delays, be easily integrated while not exceeding current complexity constrains that are applicable for traditional stereo coding conditions.

The decorrelation method used in enhanced aacPlus is a combination of a short IIR reverberator and single delay lines. Table 2 shows the delay line configuration that are used, excluding the fractional delays. As can be seen in the table, the reverberation decay length is converging towards 20 ms as approximating the upper frequency limit of the reverb. This is a typical behavior of the chosen all-pass filter structure as the filter coefficient vector, g in Equation 9, approximates the zero vector. Consequently, a continuous transition between the frequency bands utilizing all-pass filters and delays is achieved.

The IIR reverberator used for frequencies under 8625 Hz according to configuration Table 2, is designed according to Sections 3.4 and 3.5. Since fractional delay is used as described in Equations 7 and 8, the delay lines

$$z^{-m(n)} \tag{13}$$

are replaced by

$$e^{i\alpha_k(n)} z^{-m(n)} \tag{14}$$

frequency region	decorrelation method	reverberation time RT60 (or delay)
$<$ 8625 Hz	IIR reverb	
at 100 Hz		∼80 ms
at 8625 Hz		∼20 ms
8625–13125 Hz	delay	20 ms
$>$ 13125 Hz	delay	2 ms

Table 2: *Stereo decorrelator configuration of enhanced aacPlus. Figures are based on 44.1 kHz sampling rate.*

where $\alpha_k(n)$ for $n = 0,1,2$ are the phase rotation values for the corresponding all-pass link n and QMF band k. Additionally, an overall fractional delay is introduced by multiplying with phase rotation factor ψ_k. This results in the complete all-pass transfer function shown in Equation 15.

$$H_k(z) = z^{-2}\psi_k \prod_{n=0}^{2} \frac{g_k(n) + e^{i\alpha_k(n)} z^{-m(n)}}{1 + g_k(n) e^{i\alpha_k(n)} z^{-m(n)}} \qquad (15)$$

Here $H_k(z)$ represents the reverberation transfer function for QMF band k. $H_k(z)$ is based on three all-pass links as defined in Equation 9, but is in Equation 15 completed with fractional delays and a two samples overall delay.

5 CONCLUSIONS

It has been found that the traditional design of reverberation tools is only poorly fulfilling the requirements of synthetic ambience generation in a parametric stereo coding system. Based on a discussion of all-pass decorrelator techniques, an efficient IIR design derived from basic artificial reverberation principles has been proposed. Despite compromises in favor of complexity minimizations, the suggested method successfully improves the reconstruction of ambience as a stereo decorrelator compared to earlier solutions.

Subjective listening test have confirmed the usefulness of the PS system [18], [19], [9], [8], utilizing the synthetic ambience generation presented in this paper, in combination with aacPlus [11] in terms of quality versus bit rate gain. The gain of using an artificial reverberation structure as decorrelation method is also an important contribution to that advantage in quality. Furthermore, an enhanced aacPlus decoder which includes the PS tool has approximately the same computational complexity as a normal stereo aacPlus decoder.

6 REFERENCES

[1] J. Herre and H. Purnhagen, "General audio coding," in *The MPEG-4 Book*, F. Pereira and T. Ebrahimi, Eds., chapter 11. Prentice Hall, Englewood Cliffs, NJ, US, 2002.

[2] ISO/IEC, "Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s – Part 3: Audio (MPEG-1 Audio)," ISO/IEC Int. Std. 11172-3:1992, 1992.

[3] ISO/IEC, "Generic coding of moving pictures and associated audio information – Part 7: Advanced Audio Coding (AAC)," ISO/IEC Int. Std. 13818-7:1997, 1997.

[4] M. R. Schroeder, "An artificial stereophonic effect obtained from using a single signal," in *Proc. 9th AES Convention*, Oct. 1957, Preprint 14.

[5] J. A. Moorer, "About this reverberation business," *Computer Music Journal*, vol. 3, no. 2, pp. 13–28, 1979.

[6] T. Ziegler, A. Ehret, P. Ekstrand, and M. Lutzky, "Enhancing mp3 with SBR: Features and capabilities of the new mp3PRO algorithm," in *Proc. 112th AES Convention*, Munich, Germany, May 2002, Preprint 5560.

[7] ETSI ES 201 980, "Text of ETSI ES 201 980 V1.2.2 (Digital Radio Mondiale (DRM) system specification)," ETSI ES 201 980 V1.2.2.

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

[9] ISO/IEC JTC1/SC29/WG11, "Text of ISO/IEC 14496-3:2001/FDAM2 (parametric coding for high quality audio)," ISO/IEC JTC1/SC29/WG11 N6130, Dec. 2003.

[10] 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, Mar. 2003, Preprint 5852.

[11] 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.

[12] 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.

[13] P. Ekstrand, "Bandwidth extension of audio signals by spectral band replication," in *Proc. 1st IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA-2002)*, Leuven, Belgium, Nov. 2002, pp. 73 – 79.

[14] ISO/IEC JTC1/SC29/WG11, "Text of ISO/IEC 14496-3:2001/FPDAM2 (parametric coding for high quality audio)," ISO/IEC JTC1/SC29/WG11 N5713, July 2003.

[15] J.-B. Rault, "MPEG4-Ext2: CE on complexity reduction in parametric stereo decoding," ISO/IEC JTC1/SC29/WG11 MPEG2003/M9742, July 2003.

[16] 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, Nov. 2003.

[17] M. R. Schroeder and B. F. Logan, "'colorless' artificial reverberation," *J. Audio Eng. Soc.*, vol. 9, no. 3, pp. 192–197, July 1961.

[18] H. Purnhagen, J. Engdegård, W. Oomen, and E. Schuijers, "Combining low complexity parametric stereo with high efficiency AAC," ISO/IEC JTC1/SC29/WG11 MPEG2003/M10385, Dec. 2003.

[19] W. Oomen, E. Schuijers, H. Purnhagen, and J. Engdegård, "MPEG4-Ext2: CE on low complexity parametric stereo," ISO/IEC JTC1/SC29/WG11 MPEG2003/M10366, Dec. 2003.