Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2007, Article ID 74064, 14 pages
doi:10.1155/2007/74064

Research Article

A Bit Stream Scalable Speech/Audio Coder Combining Enhanced Regular Pulse Excitation and Parametric Coding

Felip Riera-Palou[1,2] and Albertus C. den Brinker[1]

[1] *Philips Research Laboratories, Digital Signal Processing Group, Prof. Holstlaan 4, 5656 AA Eindhoven, The Netherlands*
[2] *Department of Mathematics and Informatics, University of the Balearic Islands, Carretera de Valldemossa km 7.5, 07122 Palma de Mallorca, Spain*

Received 2 October 2006; Revised 16 March 2007; Accepted 29 June 2007

Recommended by Tan Lee

This paper introduces a new audio and speech broadband coding technique based on the combination of a pulse excitation coder and a standardized parametric coder, namely, MPEG-4 high-quality parametric coder. After presenting a series of enhancements to regular pulse excitation (RPE) to make it suitable for the modeling of broadband signals, it is shown how pulse and parametric codings complement each other and how they can be merged to yield a layered bit stream scalable coder able to operate at different points in the quality bit rate plane. The performance of the proposed coder is evaluated in a listening test. The major result is that the extra functionality of the bit stream scalability does not come at the price of a reduced performance since the coder is competitive with standardized coders (MP3, AAC, SSC).

1. INTRODUCTION

During the late eighties and early nineties, and with the explosive growth in the use of Internet, the need for efficient audio representations became more evident and numerous compression methods were proposed. Coders developed within MPEG-2, like MP3 or AAC [1], are popular techniques in use today. Both techniques (MP3, AAC) are examples of lossy coding algorithms where the decoded signal is not a perfect copy of the original material as some information is thrown away during the encoding. Information is discarded by exploiting the characteristics of the human hearing system so as to minimize the audible effects caused by the missing data. Despite these perceptual considerations, these coders aim essentially at a waveform match between the coded and the original signals.

More recently, an alternative audio coding paradigm has received substantial attention from the research community. This technique, generically called parametric coding, fits the input signal to a predetermined model simplifying in this way its representation [2, 3]. An example of this type of coder is the sinusoidal coder (SSC) recently introduced by Philips into MPEG-4 as Extension 2 (high-quality parametric coding) [4]. Using SSC, compression factors higher than 50 (24 Kbit/s for a stereo CD stream) have been realized while still maintaining a good quality in the reconstructed signal, although significantly lower than that of the original material.

It is assumed that the quality/bit rate tradeoffs for parametric and waveform coders follow the curves shown in Figure 1. This graph indicates that for low bit rates, parametric coders (SSC, speech vocoders) outperform waveform techniques (transform coders, code excited linear prediction (CELP) coders). On the other hand, it is also evident that parametric coders, due to the constraints inherent in the model, have difficulties in attaining the highest-quality levels, and thus require much higher bit rates than waveform coders when aiming at the high audio quality. In contrast, waveform coders attain excellent quality levels at relatively high bit rates but this abruptly drops when going to very low bit rates. The axes in Figure 1 have been deliberately left without specific numbers as it is not yet clear where the two approaches cross each other and which are the true slopes of the curves. Experimental indications for these trends can be found in [4–7]. The target objective of a coder able to perform well over a large range of bit rates also tackles the problem of a universal coder able to deal with general audio signals as well as speech. Traditionally, speech coding has relied on the speech

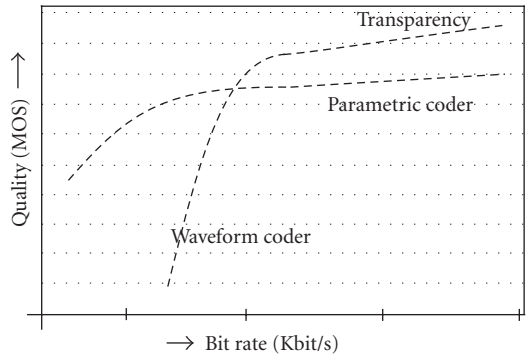

Figure 1: Schematic view of the quality/bit rate tradeoff for parametric and waveform coding paradigms.

production mechanism model [8] to derive very efficient encoding strategies which allow excellent quality to be attained at very low bit rates. Unfortunately, these methods are not suitable for general audio signals which, typically, do not adhere to any speech-dependent model. Hybrid approaches combining audio (transform coding) and speech methods (algebraic CELP) have been reported in [5, 9, 10]. These hybrid coders determine the optimal coding setting per frame on the basis of the input signal.

The main objective of this paper is to present the detailed design of an alternative hybrid coder combining parametric and waveform techniques. To this end, the SSC coder (parametric) is combined with an enhanced regular pulse excitation (ERPE) coder (waveform). The reason for choosing a pulse excitation method has a lot to do with the choice of SSC as the initial core coder: SSC is a sinusoidal coder supplemented with a noise coding module (details given in Section 3.1) implying that coding takes place, primarily, in the frequency domain and due to its tight bit-rate budget, significant parts of the original signal are substituted by properly filtered noise. In order to pursue higher-quality levels, it is reasonable to assume that the characteristics of the supplementary method should be complementary to those of SSC, hence, a time-domain coder method aiming at a waveform match seems the most logical choice. Pulse excitation coders like pulse code modulation (PCM), adaptive differential PCM (ADPCM), multipulse excitation (MPE), RPE or ACELP are thus alternatives. We did not consider PCM or ADPCM as these achieve a relatively poor quality bit-rate compromise. We started off by taking RPE and extended it to enhanced RPE (ERPE, see Section 2) to attain sufficient quality. ERPE can be seen as a combination of MPE and RPE. As such it is reminiscent of the ACELP structure which also uses sparse pulse-like excitations (see Section 2). Unlike previous hybrid approaches, the proposed coder does not rely on any decision regarding the type of input signal and its behavior is solely determined by the selected bit rate.

The proposed coding architecture is bit stream scalable allowing a coded file to be decoded at a variety of bit rates. Bit stream scalability is attractive in scenarios where it is desired to let the end user decide the operating bit rate (e.g., internet radio) or where, due to network issues, information layers have to be discarded at intermediate nodes. In order for the proposed coding architecture to be competitive with standard coders (tuned for a specific bit rate), the quality loss associated with the scalability (scalability loss) should be minimized.

As a last introductory remark, we note that most modern coders have to deal with stereo signals. In order to allow our hybrid coder to tackle stereo signals while keeping the total bit rate low, use has been made of one of the stereo coding tools in MPEG-4, namely, parametric stereo (PS).

This paper is structured as follows: Section 2 reviews the concept of analysis-by-synthesis (AbS) coding with particular emphasis on regular pulse excitation. A series of improvements which have recently proposed to enhance the RPE modeling capability when dealing with broadband signals are also discussed in this section. Section 3 briefly reviews the parametric elements of the hybrid coder which have been taken from the MPEG-4 standard, specifically, the SSC coder and the PS stereo coding tool. In Section 4, the structure of the hybrid SSC-ERPE coder/decoder is presented. Listening test results comparing the different SSC-ERPE layers with standardized coders are given in Section 5. Finally, Section 6 summarizes the main results of this work.

2. ENHANCED REGULAR PULSE EXCITATION (ERPE)

2.1. Analysis-by-synthesis (AbS) coding with RPE

Most narrowband (8 kHz sampling) speech coding techniques are based on three concepts which are strongly related: linear prediction, excitation modeling, and analysis-by-synthesis (AbS) with all three mechanisms operated in such a way as to minimize a prescribed perceptual distortion measure. Successive samples in a speech or audio signal usually exhibit a high degree of correlation indicating that the original time domain samples are, to a certain extent, redundant. It seems clear that this redundancy can be exploited in order to achieve a reduction in bit rate. Linear prediction [11, 12] aims at reducing the intersample redundancy by forming a prediction of a given sample as a linear combination of past samples. This is implemented by a linear prediction analysis filter acting upon the input signal and producing a decorrelated residual, which, at the decoder, can be used to drive a synthesis filter (the inverse of the analysis filter) to produce an approximation of the original signal. Recent years have witnessed a stream of new developments in linear prediction techniques based on the characteristics of human perception. Following the original work of Strube [13], prediction techniques [14, 15] based on the idea of frequency warping [16] were introduced and it has been demonstrated how warping can lead to significant improvements over conventional linear prediction (LP) when dealing with broadband signals. The improvement is due to the possibility of adjusting the predictor modeling capability to be more accurate over a specific frequency range at the cost of loosing some precision in the other frequency regions. Since many audio/speech signals have a low frequency character, where hearing is most sensitive, warping can be used to model more accurately this region at the cost of a certain

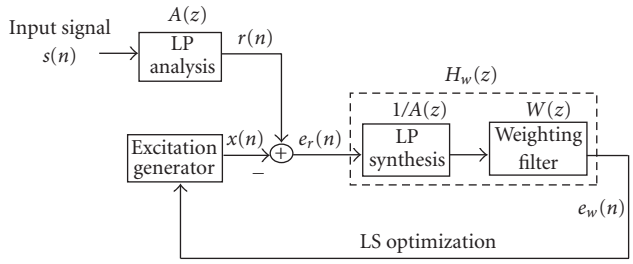

Figure 2: Analysis-by-synthesis encoder.

modeling loss in the high frequencies which are anyway, less perceptually relevant. Owing to its desirable features, a form of warped prediction, Laguerre-based linear prediction [14], has been incorporated (see Section 4).

In order to pursue high-compression ratios, the linear prediction residual needs to be processed so that the required bit rate for its transmission is as low as possible. The conventional method is to search for a signal from a restricted class (codebook) which when passed through the synthesis filter produces the best perceptual approximation of the original waveform. With this objective in mind, a decoder is incorporated in the encoder to evaluate distinct excitation candidates, giving rise to the concept of analysis-by-synthesis (AbS).

Figure 2 shows the typical AbS scheme usually found in linear prediction-based coders. The input signal $s(n)$ is passed through a linear prediction analysis filter with frame-dependent transfer function $A(z)$, whose output $r(n)$ is a spectrally flattened and temporally decorrelated residual signal. An excitation sequence $x(n)$ is generated such that the difference $e_r(n)$ with respect to the residual signal $r(n)$, after synthesis and perceptual weighting $e_w(n)$ is minimised, usually in a least-squares (LS) sense. The weighting filter $W(z)$ takes care of incorporating certain human hearing properties when minimising the distortion and it is usually derived from the computed LP coefficients [8, page 32].

The synthesis and perceptual filters, shown in Figure 2, have transfer functions given by $1/A(z)$ and $W(z)$, respectively, and the cascade of both filters is denoted as $H_w(z)$. The signal $e_w(n)$ represents the perceptually weighted error whose power is being minimized. Given that coding is usually carried out frame-by-frame, the optimization of the pulse sequence, $x(n)$, is performed using a finite data record with the influence of the excitation from one frame to the next one being propagated via the filter states. Assuming a frame length of N samples, the vector $\mathbf{e_w}(k)$, made of the N successive perceptual error samples over frame k, can be expressed as

$$\mathbf{e_w}(k) = \mathbf{e_0}(k) + \mathbf{H_w}(k)\mathbf{e_r}(k), \tag{1}$$

where $\mathbf{e_0}(k)$ is a vector corresponding to the response of the filter $H_w(z)$ due to its initial filter states and $\mathbf{e_r}(k)$ is an $N \times 1$ vector given by

$$\mathbf{e_r}(k) = \mathbf{r}(k) - \mathbf{x}(k), \tag{2}$$

where the vector $\mathbf{r}(k)$ corresponds to N samples of the residual signal ($r(n)$ in Figure 2) and $\mathbf{x}(k)$ denotes the corresponding computed quantized excitation vector. $\mathbf{H_w}(k)$ is an $N \times N$ matrix containing the samples of the frame-dependent impulse response $h_w(n)$ of the filter $H_w(z)$ with the following structure:

$$\mathbf{H_w}(k) = \begin{pmatrix} h_w(0) & 0 & \cdots & 0 & 0 \\ h_w(1) & h_w(0) & \ddots & \vdots & \vdots \\ \vdots & h_w(1) & \ddots & 0 & \vdots \\ \vdots & \vdots & & h_w(0) & 0 \\ h_w(N-1) & h_w(N-2) & \cdots & h_w(1) & h_w(0) \end{pmatrix}. \tag{3}$$

The different residual modeling methods such as regular pulse excitation (RPE) [17], multipulse excitation (MPE) [18] or algebraic code-excited linear prediction (ACELP) [19, 20] basically differ in the constraints imposed on the resulting excitation. Given the range of available residual modeling techniques, some justification is required as to why the hybrid coder proposed in this paper uses RPE. Both RPE and MPE are pure pulse coding techniques, that is, the excitation is solely formed by a number of nonzero pulses, which in the case of RPE are placed on a regular grid while in MPE are freely positioned on the excitation frame. It has been reported [8, page 82] that RPE and MPE, at equal bit rates, perform similarly. It is shown in Section 2.2 that our proposed enhanced RPE (ERPE) coder can somehow be seen as a combination of RPE and MPE, hence, exploiting the benefits of both methods. Another alternative for residual signal modeling would be ACELP, where the encoder looks for the best signal in sparse pulse-like codebooks which are structured in tracks to facilitate the excitation search [21]. On top of that, there is a long-term predictor (LTP) incorporated. For narrowband signals, this is an effective method as is clear from the quality provided by the AMR coder [22]. However, in the context of the proposed hybrid parametric-waveform coder, RPE was found to be a superior alternative to ACELP as proved by an informal listening test between an SSC-RPE coder and an SSC-ACELP coder in which the waveform method, RPE or ACELP, was used to model the lower quarter frequency range (0–11025 Hz). The reason is probably the relative ineffectiveness of the LTP for a residual signal created by SSC since presumably a large part of the long-term predictability is already removed by the sinusoidal coder.

When using conventional RPE, only J equidistant nonzero values are allowed per frame (decimation N/J), which are chosen to minimize $\|\mathbf{e_w}(k)\|^2$ and are given by [17]

$$\begin{aligned} \mathbf{x_p}(k) = {} & \left(\mathbf{M}(k)^t\mathbf{H_w}(k)^t\mathbf{H_w}(k)\mathbf{M}(k)\right)^{-1} \\ & \times \left(\mathbf{M}(k)^t\mathbf{H_w}(k)^t\right)\left(\mathbf{e_0}(k) + \mathbf{H_w}(k)\mathbf{r}(k)\right), \end{aligned} \tag{4}$$

where $\mathbf{M}(k)$ is an $N \times J$ location matrix signalling which positions in the excitation sequence are nonzero. The full N-sample quantized excitation, $\mathbf{x}(k)$, can be computed as

$$\mathbf{x}(k) = \mathbf{M}(k)Q[\mathbf{x_p}(k)], \tag{5}$$

where $Q[\cdot]$ denotes the quantization procedure. In the case of RPE, different location matrices corresponding to different grid positions are tested and the one with minimum error (i.e., lowest $\|\mathbf{e_w}(k)\|^2$) is selected. The quantized pulse amplitudes $Q[\mathbf{x_p}(k)]$ and the optimum grid position are the only parameters that need to be transmitted (alongside the LP coefficients). Given that the pulse amplitudes $\mathbf{x_p}(k)$ contribute the most to the total bit rate, it is important to consider how they are encoded. An important tradeoff in pulse excitation is that of number of pulses J per frame versus the number of quantization levels employed in their quantization Q. Currently, the determination of the optimum quantization/decimation point remains unsolved. Our experiments have suggested that given a target bit budget and in order to achieve very high reproduction quality for broadband audio at an attractive bit rate it is best to use densely packed excitations (i.e., low decimations) with very coarse quantization. Following extensive experimentation, it has been decided to fix each pulse to +1, −1 or 0. This is in line with previous works [23, 24] where CELP codebooks were proposed to be populated with ternary sequences. Additionally, a gain (denoted by g_{RPE}) is computed to scale the excitation to the optimum amplitude (in an LS sense). Therefore, it holds that $Q[\mathbf{x_p}(k)] = g_{\mathrm{RPE}}(k)\mathbf{x_q}(k)$ with $\mathbf{x_q}(k)$ being a ternary sequence. The gain $g_{\mathrm{RPE}}(k)$ should in turn be quantized but far more levels can be allowed since only one value per frame is required. From the described RPE and quantization procedure, it holds that the prediction residual is modelled by a suitably scaled ternary sequence where nonzero entries are only allowed to be on a regularly spaced grid.

2.2. Enhanced regular pulse excitation (ERPE)

The application of speech coding techniques (linear prediction, pulse excitation and AbS) to broadband audio coding has been addressed by several researchers in the past [25, 26] with the resulting coding schemes achieving good quality scores at around 100–128 Kbit/s. Our own experiments determined that pulse excitation on its own is not able to attain high quality levels at attractive bit rates for certain types of excerpts, most notably, clearly tonal fragments (e.g., pitchpipe from [27]) tended to be low quality. In an attempt to overcome the RPE limitations and make it more suitable for broadband signals, two extensions have been proposed in the literature: addition of extra pulses [28] and an improved optimisation of the pulse sequence computation [29]. The combination of both techniques results in the enhanced RPE (ERPE) coder which, for completeness, is reviewed next.

The LP analysis filter shown in Figure 2 targets the minimization of short-term correlation (typically less than 1 millisecond). However, it is well known that certain signals such as speech exhibit also a high degree of long-term correlation. The presence of long-term correlation, very common and perceptually important, for instance, in voiced speech segments, is often revealed as pulse train-like structures in the residual signal $r(n)$. These periodic pulses tend to create problems in the excitation modeling stage leading to a poor compromise excitation especially for a low number of quantization levels. In order to preserve this type of correlation, special processing actions are taken. In speech coding, the residual $r(n)$ is filtered using a long-term predictor (LTP) analysis filter [30] which takes care of the long-term periodicities. The LTP analysis filter is functionally identical to the LP analysis filter but since it aims at long-time correlations, it usually consists of a long tap-delay line with only a few consecutive nonzero coefficients. The position and values of the nonzero coefficients are updated every input signal frame and they will depend on the input signal periodicity. A synthesis LTP filter at the decoder, prior to the LP synthesis one, restores the long-term correlation in the LP residual.

The LTP gain for broadband signals is low due to the presence of high frequencies which obscure the signal periodicity typically present at low frequencies (e.g., due to speaker's pitch). Nevertheless, long-term correlation is present in the LP residual appearing as periodic pulse-like trains in $r(n)$. Frames containing pulse-like structures generate sets of RPE pulses with large dynamic range, resulting in poor compromise excitations when coarsely quantized. Long LTP filters (more than 3 coefficients) can attain higher gains but exacerbate the stability problems in the LTP synthesis introducing the need for complex stabilization procedures [31].

The solution we propose consists of skipping the LTP and instead provides the RPE excitation with additional degrees of freedom suitable to model effectively pulse-like trains (i.e., long-term correlations) in the LP residual $r(n)$. To this end, the RPE excitation for a frame is complemented with R additional independent pulses with free gains and positions resulting in an extended excitation of the form

$$\mathbf{x}_{\mathrm{ext}}(k) = \mathbf{M}(k)g_{\mathrm{RPE}}(k)\mathbf{x_q}(k) + \sum_{m=1}^{R} g_m(k)\mathbf{p}(d_m(k)), \qquad (6)$$

where the first term in (6) represents the (quantized) RPE component and the second term corresponds to the sum of R N-length vectors, $\mathbf{p}(d_m(k))$, with each vector consisting of zeros except at position $d_m(k)$ where it has unity value. Each of these vectors, labeled subsequently as impulse excitation vectors, correspond to an extra pulse. The scalars $g_m(k)$ denote the (quantized) gains associated with the extra pulses. These extra pulses are quantized independently. Limiting the number of extra pulses to just two (i.e., $R = 2$) makes the extra bit rate rather small (comparable to the LTP bit rate). Given the frame duration of 5 milliseconds, the two extra pulses per frame allow pulse trains with frequencies of up to 370 Hz (i.e., most of the human pitch range) to be modeled. Additionally, it has been experimentally observed that the extra pulses contribute very significantly to the proper modeling of transient phenomena with sharp onsets present in many audio excerpts (e.g., castanets from [27]). Note that (6) could be interpreted as a combination of RPE and multipulse excitation, an idea which was already considered in [32]. It should also be mentioned that an alternative solution specifically geared towards pitch modeling in broadband signals was proposed in the adaptive multirate wideband (AMR-WB) codec [22]. In AMR-WB, the search of the pitch lag is conducted in various sequential stages. The first stage provides a coarse pitch estimate based on a low-pass

Input: $\mathbf{r}$ (residual frame to be modeled), J, $\mathbf{e_0}$
For every offset j **do**
Construct location matrix $\mathbf{M}$ for j
Compute optimum RPE unquantised amplitudes $\Rightarrow \mathbf{x_p}$
Select positions of the R largest magnitude pulses $\Rightarrow d_1, \ldots, d_R$
Generate R impulse excitation vectors $\Rightarrow \mathbf{p}_m(d_m)$, $m = 1, \ldots, R$
Generate RPE excitation vector: $\mathbf{x_q}$
Compute optimum gains $\Rightarrow g_{\text{RPE}}, g_1, \ldots, g_R$
Compose total excitation $\Rightarrow \mathbf{x}_{\text{ext}} = g_{\text{RPE}}\mathbf{M}\mathbf{x_q} + g_1\mathbf{p}_1 + \cdots + g_R\mathbf{p}_R$
Compute norm of reconstruction error for current offset $j \Rightarrow \|\mathbf{e_w}\|^2$
end

Algorithm 1

filtered version of the input signal. Based on the initial estimate, subsequent stages provide a more refined estimation. Additionally, in ARM-WB the pitch prediction filter is made frequency-dependent to account for the nonregular harmonic structure of broadband signals. The advantage of the extra pulses method proposed here is its generality in the sense that it cannot only perform the pitch modeling, but also contribute to the better waveform match of other phenomena (such as transients) which might be present in general audio signals.

The computation of the optimum RPE excitation and additional pulses for each residual frame, $\mathbf{r}(k)$, is computationally very complex as all combinations of RPE sequences and extra pulse positions should be examined. To lower the computational burden and the associated bit rate, the extra pulses are restricted to lie on the selected RPE grid. This amounts to performing a conventional RPE search where for each RPE candidate, its two largest pulses are quantized separately while the rest of the pulses all have the same gain. This strategy is based on the idea that the (two) largest RPE pulses are the ones contributing the most to the error minimization. The resulting algorithm to compute an RPE excitation with R extra pulses is shown in Algorithm 1 [28], where for convenience, we have dropped the frame dependence.

The j index which attains the lowest error and the associated parameters $Q[\mathbf{x}_p]$, $g_{\text{RPE}}, g_1, \ldots, g_R$, $d_1, \ldots, d_R$, is all that needs to be transmitted.

The RPE sequence and extra pulses gains, $g_1, \ldots, g_K$, are computed using least-squares (LS) optimization on the error given by (1) for each processed frame. This algorithm, as shown by the results in Section 6, has proved effective in modeling long-term correlations and thus achieving better performance. Also, the encoding of the positions, d_m, of the extra pulses requires a lower bit rate when constrained to the RPE grid. Note that the complexity of this algorithm is only marginally higher than that of conventional RPE.

As it has already been mentioned, the RPE technique works on a segmental basis by calculating the optimal excitation for a given input signal segment. Nevertheless, the computation of the excitation for adjacent frames is not independent from the current one as they are related via the filter states (the term $\mathbf{e_0}(k)$ in (1) formally represents this dependence).

It has been experimentally observed that the resulting optimum excitation for frame k may induce a large error component due to the initial filter states in frame $k+1$, that is, large samples in $\mathbf{e_0}(k+1)$, which in turn may result in a considerable error $\mathbf{e_w}(k+1)$. This issue was already recognized in [33] but no satisfactory solution has been proposed. Additionally, the introduction of extra pulses with free gains in the excitation as previously explained can exacerbate this problem. A simple solution could be the avoidance of pulses, conventional or additional, positioned towards the end of the excitation, as these are the pulses that will mainly influence $\mathbf{e_w}(k+1)$. Such a restriction, however, is unnecessary for most frames and actually degrades the performance. A more elegant approach is to define a new error expression that takes into account the error induced in the next frames [29]. To this end, the new performance measure is given by (cf. (1))

$$\widetilde{\mathbf{e}}_{\mathbf{w}}(k) = \mathbf{V}\left(\widetilde{\mathbf{e}}_{\mathbf{0}}(k) + \widetilde{\mathbf{H}}_{\mathbf{w}}(k)\widetilde{\mathbf{e}}_{\mathbf{r}}(k)\right), \tag{7}$$

where, assuming a frame length of N samples, $\mathbf{V}$ is an $N+F$ diagonal weighting matrix, with $F \geq 1$, used to weigh the importance of the induced filter states for future frames. The variables $\widetilde{\mathbf{e}}_{\mathbf{0}}(k)$ and $\widetilde{\mathbf{H}}_{\mathbf{w}}(k)$ are equivalent to those in (1) but the filter responses in $\widetilde{\mathbf{H}}_{\mathbf{w}}(k)$ and those used to calculate $\widetilde{\mathbf{e}}_{\mathbf{0}}(k)$ are now extended to $N+F$ samples. The vector $\widetilde{\mathbf{e}}_{\mathbf{r}}(k)$ is identical to $\mathbf{e_r}(k)$ in (1) with F additional padding zeros. The additional length over which the error is measured, F, determines how far in the future the effects of the excitation being computed are accounted for. When $F < N$, the effect is limited only to the next frame. Setting $F > N$ will include the effects induced in more than one frame ahead.

Minimization of $\|\widetilde{\mathbf{e}}_{\mathbf{w}}(k)\|^2$ with respect to the pulse amplitudes results in an expression [29] very similar to (4):

$$\begin{aligned}\widetilde{\mathbf{x}}_{\mathbf{p}}(k) = {} & \left(\widetilde{\mathbf{M}}(k)^t\widetilde{\mathbf{H}}_{\mathbf{w}}(k)^t\mathbf{V}^t\mathbf{V}\widetilde{\mathbf{H}}_{\mathbf{w}}(k)\widetilde{\mathbf{M}}(k)\right)^{-1} \\ & \times\left(\widetilde{\mathbf{M}}(k)^t\widetilde{\mathbf{H}}_{\mathbf{w}}(k)^t\right)\mathbf{V}^t\mathbf{V}\left(\widetilde{\mathbf{e}}_{\mathbf{0}}(k) + \widetilde{\mathbf{H}}_{\mathbf{w}}(k)\widetilde{\mathbf{r}}(k)\right),\end{aligned} \tag{8}$$

where the variables $\widetilde{\mathbf{M}}(k)$ and $\widetilde{\mathbf{r}}(k)$ are given by

$$\begin{aligned}\widetilde{\mathbf{M}}(k) &= \left[\mathbf{M}(k)^t\ \mathbf{0}_{J\times F}\right]^t, \\ \widetilde{\mathbf{r}}(k) &= \left[\mathbf{r}(k)^t\ \mathbf{0}_{1\times F}\right]^t.\end{aligned} \tag{9}$$

The full excitation sequence $\widetilde{\mathbf{x}}(k)$ can be computed as

$$\widetilde{\mathbf{x}}(k) = \widetilde{\mathbf{M}}(k)Q[\widetilde{\mathbf{x}}_{\mathbf{p}}(k)]. \tag{10}$$

Notice that only the first N values of $\tilde{\mathbf{x}}(k)$ are nonzero. Despite the $N + F$ column length of the matrices and vectors in (8), only J pulse amplitudes are computed, making the new method only marginally more complex than the conventional approach. In particular, the matrix inversion in (8) has the same order as the one in (4). We note that an extension of the matrix $\mathbf{H}_{\mathbf{w}}(k)$ and thus, an extension of the error observation interval was already considered in [34] in the context of speech coding. The current algorithm differs from that in [34] since there, the essence was introducing symmetry in the H matrix in order to achieve a fast algorithm.

Experiments have confirmed [28, 29] that the extra pulses and improved optimization are effective in achieving a better modeling of the residual, thus removing certain artifacts that are otherwise present when using the conventional RPE method.

2.3. Scalable pulse layering

The ability to encode material at different bit rates is generally seen as an attractive feature as it allows to target specific requirements such as bandwidth or quality. This feature is generally called bit rate scalability. A different type of scalability, termed bit stream scalability (also known as embedded coding), allows the decoding of a coded signal at different bit rates. In a way, bit stream scalability can be seen as more flexible than bit rate scalability as it allows the material to be delivered at different rates without the need of additional reencoding. This is useful to let the end user decide the operating bit rate and to reduce the effects of information discarding at intermediate network nodes. Note also that if a coder is bit stream scalable, it will also be bit rate scalable while the converse is generally not true.

Some work on bit stream scalability for RPE coders was already presented in [35] in the context of narrowband speech coding. In [35], different RPE stages are concatenated with each stage modeling the residual error coming out of the previous stage increasing, in this way the quality of the reconstructed material. Obviously, each successive RPE layer brings along an increase in bit rate. The decoding bit rate can vary depending on the number of decoded layers. Experiments with wideband audio material were conducted using this approach to obtain a bit stream scalable audio coder. It was observed that each successive RPE stage did indeed lead to an increase in quality. High-quality encoding could be achieved by stacking a sufficient number of high-decimation layers. However, it was also found that directly encoding the input signal in a nonscalable manner using a low decimation factor in such a way that the bit rate was comparable to that of the layered solution often yielded significantly better quality. As an example, encoding the LP residual using decimation 2 led to better quality than using four RPE layers of decimation 8 using the method described in [35]. The drop in quality is referred to as "scalability loss."

The root of the scalability loss problem in the coding scheme from [35] can easily be understood in a two-layer RPE coder. When the first layer provides a good approximation of the incoming signal (LP residual), the resulting second residual will have lower energy and will be easier to code for a posterior RPE stage than the LP residual itself. In this case, the coder works as intended: the first RPE model provides a reasonably good and low bit rate representation of the LP residual whereas a subsequent RPE stage models "finer" details of the LP residual. In contrast, when the first RPE model is a poor approximation of the incoming signal, the subsequent RPE stage has the task of modeling the errors introduced by the first RPE layer which, as it has been observed experimentally, induces an input signal for the second layer which can be even larger (in energy) than the LP residual itself. As an example of this phenomenon, Figure 3 shows the squared input signal (i.e., energy samples) to a decimation-2 LP-RPE coder (solid line) and the energy of the difference between the original and the reconstructed signals. It can be seen how in this particular frame fragment, the resulting modeling error has even higher energy than the original input signal. In architectures where the the different layers contribute equally to the overall bit rate, little can be done to improve the scalability loss. However, if the second layer has a larger bit rate at its disposal, performance can be improved by letting the second layer act upon the original input signal directly rather that the residual from the first layer. In this situation, the coder would need to be able to determine how good the first RPE model is and then let the next RPE stage act either on a newly formed residual from first RPE model or directly on the LP residual. A similar idea, although in a rather different setup, has been accepted within MPEG-4 applied to the coder proposed in [9] in which the different coefficients of the transform coder following the CELP coder can be configured to act on either the CELP residual or the original signal (frequency selective switch).

We have refined the method in [35] in the following way. Instead of a binary decision to use either the original residual or the residual after the first (E)RPE stage, we introduce a system which is able to use intermediate options. This is done by

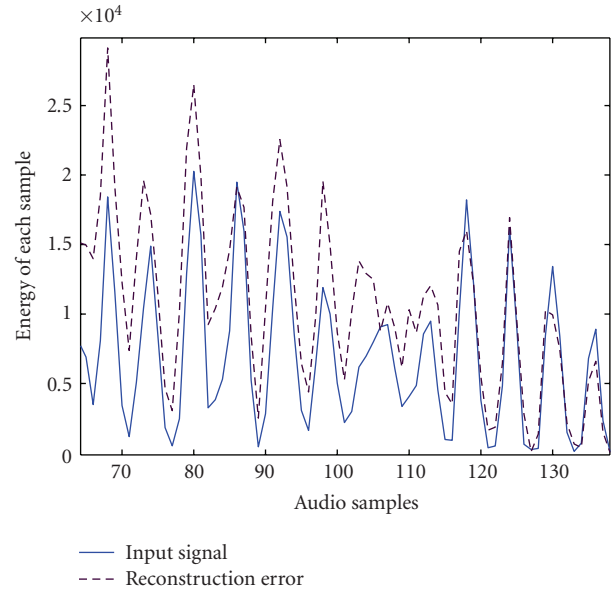

Figure 3: Illustration of the cause of the scalability loss.

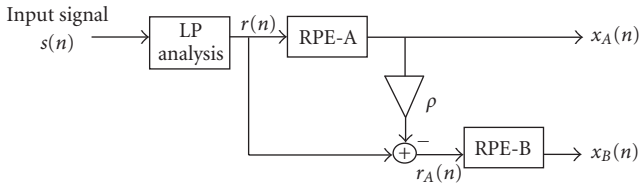

Figure 4: Scheme of the proposed bit stream scalable RPE coder.

including a weighting mechanism on the generated RPE layer model when forming the new residual. The proposed coder is shown in Figure 4. After decorrelating the input signal using linear prediction, a first RPE model (RPE-A) from which a first excitation signal, $x_A(n)$, can be constructed. This initial RPE model is computed using the procedure outlined by (1) through (5). Subtracting this first excitation signal, $x_a(n)$, from the LP residual $r(n)$, a second residual is formed which is subsequently weighed. The weight, denoted by ρ and called hereafter the mixing factor, is responsible for evaluating the quality of the first RPE model. This newly generated and weighted residual, $r_A(n)$, is then fed to a second RPE stage (RPE-B) giving rise to a second excitation signal $x_B(n)$. This architecture can be extended to an arbitrary number of layers.

The computation of the mixing factor is crucial to achieve the desired bit stream scalability. Since the purpose of the mixing factor is to indicate how well the first RPE stage models the LP residual, it seems natural to use some form of correlation between $r(n)$ and $x_A(n)$. To this end, and assuming framed versions of these signals (i.e., vectors), we compute the mixing factor that minimizes the square of the 2-norm of $r_A(n)$ over frame k:

$$\epsilon = ||\mathbf{r_A}(k)||^2 = ||\mathbf{r}(k) - \rho \mathbf{x_A}(k)||^2. \tag{11}$$

Minimizing ϵ with respect to ρ results in the mixing factor

$$\rho = \frac{\mathbf{r}(k)^t \mathbf{x_A}(k)}{||\mathbf{x_A}(k)||^2}. \tag{12}$$

We stress that the main reason to compute the mixing factor in the residual domain is to ensure that multiplication by ρ reduces the energy of the resulting residual. Figure 5 shows the frame evolution of the mixing factor ρ for two excerpts, German male and Orchestra, when combining two RPE stages with decimations 8 and 2. The computed mixing factor, ρ, represents the quality of the first RPE excitation generated (decimation 8). It can be appreciated that while for speech the first layer often provides a very good representation of the LP residual, in the case of Orchestra, the first excitation is of little value when computing the RPE-2 excitation.

The encoder generates a bit stream including the parameters to generate both excitations ($x_A(n)$ and $x_B(n)$) and it is then up to the decoder to decide how many layers are decoded (starting always from the first generated layers). The decoding process is illustrated in Figure 6. The decoded signal can be generated using only the first excitation layer $x_A(n)$

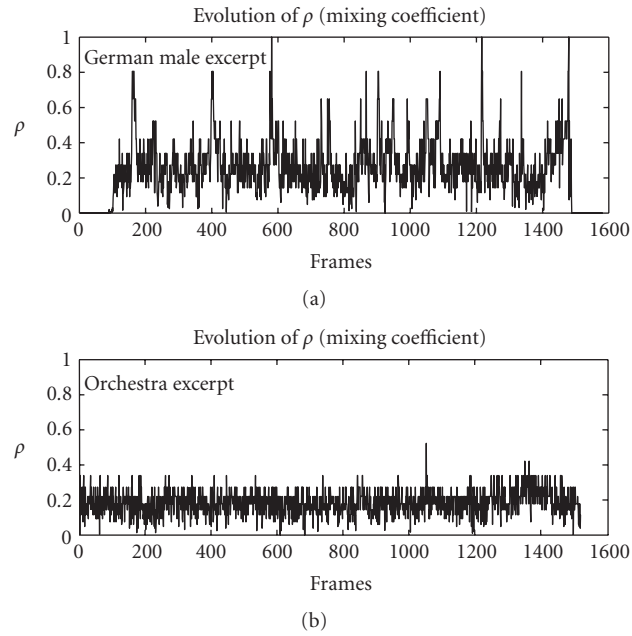

Figure 5: Evolution of the mixing factor for German male (a) and Orchestra (b).

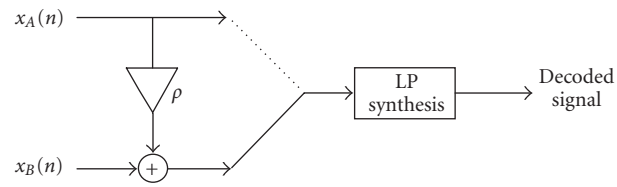

Figure 6: Scheme of the decoder including the mixing parameter.

or the combination of the two layers realizing in this way the concept of bit stream scalability. The mixing factor ρ is required at the decoder to weigh the different layer(s) and therefore needs to be transmitted. However, its high correlation from frame to frame makes this extra bit rate negligible.

Experiments have shown that the mixing factor is an efficient solution to combine two or more RPE (or more generally, excitations) layers in a way such that the first layers always contribute, to a greater or lesser extent, to enhance the quality provided by the subsequent layers. Despite the fact that the proposed method reduces the scalability loss significantly with respect to previous approaches, it should be pointed out that it cannot be completely eliminated.

Concluding this section, a list of the different ERPE parameters that need to be transmitted (per excitation frame) and their associated bit rate when using entropy coding are presented in Table 1. The results shown in this table refer to an ERPE model with decimation 8 without extra pulses and to an ERPE model with decimation 2 with two extra pulses. The values shown here represent averages computed from the different excerpts used in the listening test.

Table 1: Bit rate for the different parameters of ERPE models for a 60-sample frame (5.4 milliseconds of audio sampled at 11025 kHz).

Parameters	ERPE-8	ERPE-2 + 2 EP
RPE part gain (g_{RPE})	4 bits	4 bits
Offset ($\mathbf{M}$)	3 bits	1 bit
Pulse amplitudes ($\mathbf{x_q}$)	12 bits	48 bits
Extra pulses (g_1, g_2, d_1, d_2)	0 bits	20 bits
Mixing factor (ρ)	4.75 bits	4.75 bits

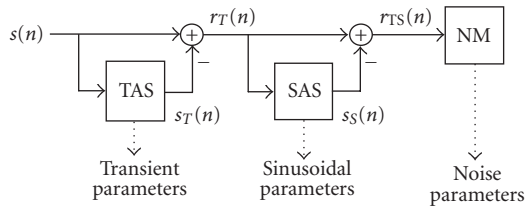

Figure 7: MPEG-4 SSC encoder.

3. MPEG-4 PARAMETRIC CODING TOOLS

3.1. MPEG-4 high-quality parametric coding (SSC)

The SSC coder fits a model which consists of transients, sinusoids, and noise to the input audio signal. The steps involved in this model fit are shown in Figure 7. First, transient elements (sudden and large changes in energy) are detected, modeled ($s_T(n)$), and subtracted from the input signal by means of transient analysis and synthesis procedure (TAS) generating in this way the residual $r_T(n)$. The perceptually most relevant sinusoids of this residual (tonal components) are also detected, modeled ($s_S(n)$), and subtracted to generate a second residual, $r_{\text{TS}}(n)$, using a sinusoidal analysis and synthesis (SAS) stage. This residual, $r_{\text{TS}}(n)$, is then input to a noise modeling procedure (NM) where its temporal and spectral envelope parameters are extracted.

The high efficiency of parametric coders like SSC stems from the fact that only the extracted parameters, and not the model itself, need to be transmitted. At the decoder, a local model receives the transmitted parameters and synthesizes an approximation of the original signal. In formal listening tests using critical material, SSC has achieved a mean opinion score (MOS) of *Fair* to *Good* at 24 Kbit/s stereo. Extensive information on the techniques used in the SSC coder and listening test results can be found in [4, 36].

3.2. MPEG-4 parametric stereo (PS)

The most straighforward method to encode the two signals corresponding to left and right channels of a stereo signal would be to code them separately by feeding each of them to a monoaudio coder. This rather simple approach is not at all efficient as it completely neglects the correlation between the two channels which is often present. There is a variety of stereo coding methods that exploits the interchannel correlation like mid/side coding [37] or intensity stereo [38]. Still, the resulting bit rates are too high for the coder under study, and moreover, these techniques are not easily adapted to the parametric coding based on SSC. For these reasons, it was decided to use a parametric stereo coding tool defined within MPEG-4, namely, MPEG-4 PS (parametric stereo) [39–41]. The PS encoder takes as input the left and right channels and produces as output a primary signal and a stereo parameter stream. The primary signal, formed by a suitable combination of the left and right channels, is a conventional mono audio signal at the same bit rate as any of the original input signals (e.g., 44 100 samples/s, 16 bit/sample). The parameter stream requires a much lower bit rate (typically between 1.5 and 16 Kbit/s) and contains the spatial information used by the PS decoder to reconstruct the stereo image from the (mono) primary signal. The primary (mono) signal can then be subsequently fed to a monocoder, which in our case is the SCC-RPE coder, to lower its transmission budget. At the decoding side, the monoaudio decoder produces an approximation of the primary signal and the PS decoder uses this reconstructed signal and the received spatial parameters to reproduce the left and right channels.

4. THE SSC-ERPE CODER

Having covered the basics of SSC, PS (Section 3), and ERPE (Section 2), we consider now how these techniques can be efficiently combined to realize a bit stream scalable coder. The main limiting factor in SSC when aiming at high quality is that the assumption of the residual signal r_{TS} (see Figure 7) being noise is often not valid. Frequently, r_{TS} still contains tonal components which have not been modeled by the sinusoidal stage (SAS in Figure 7) due to bit rate constraints. Notice that r_{TS} contains all signal details not modeled by transients or sinusoids, and consequently, if this residual was supplied to the decoder and the transient ($s_T(n)$) and sinusoidal components ($s_S(n)$) were added, the original audio signal would be perfectly reconstructed. The main idea of the hybrid bit stream scalable coder is to use the MPEG-4 SSC coder as depicted in Figure 7 as a base coding layer on top of which layers are added which progressively improve the description of r_{TS}. Before delving into the details of this approach, we mention that, in theory, it is possible to generate extra layers containing additional sinusoidal bit rate in order to better condition the residual r_{TS} for the noise modeler. This procedure, however, is expected to be an inefficient method (in terms of bit rate) to increase the final quality (see Figure 1 and [6, 7]).

A better approach is to supplement the noise modeller with some form of waveform coding, in our case ERPE, to describe the residual r_{TS} more accurately. Since obviously any extra refinement on r_{TS} costs bits, it is important to allocate the extra bits as efficiently as possible. To this end, and following a basic psychoacoustical principle that human hearing is most sensitive at low frequencies, r_{TS} is partitioned into 4 different frequency bands (0–5.5 kHz, 5.5–11 kHz, 11–16.5 kHz, and 16.5–22 kHz) to allow different degrees of modeling in each band. In this way, more bits can be used for the lower part of the spectrum. The band split

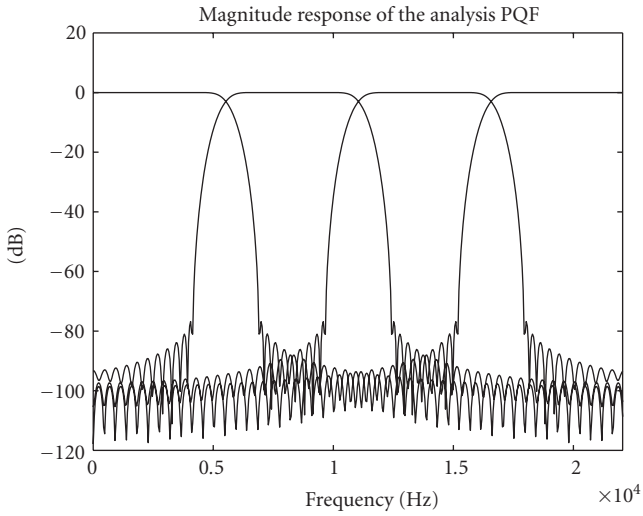

Figure 8: Magnitude frequency response of the analysis filterbank.

is performed using a 4-band quadrature mirror filterbank (QMF) specified by the set of impulse responses [42]:

$$h_k(n) = 2w(n)\cos\left(\frac{\pi}{4}(k-0.5)\left(n-\frac{L-1}{2}\right)+\theta_k\right), \quad (13)$$

where $k = 1\cdots 4$ is the band number, $\theta_k = (-1)^{k-1}(\pi/4)$, $w(n)$ is a linear phase FIR prototype low pass filter with normalized cutoff frequency $\pi/8$, and L is the number of taps of this FIR prototype. The magnitude frequency response of this QMF when setting $L = 96$ is depicted in Figure 8 for a sampling frequency of 44.1 kHz. The resulting filtered and downsampled residuals are given by

$$r_{\mathrm{TS},k}(n) = \left[r_{\mathrm{TS}}(n) * h_k(n)\right] \downarrow 4 \quad \text{with } k = 1\cdots 4. \quad (14)$$

The combined SSC-ERPE coder (first introduced in [43]) including the QMF filterbank is shown in Figure 9. By looking at this figure, a direct way to design the additional layers on top of SSC is to encode the different subband residuals $r_{\mathrm{TS},k}$ with $k = 1\cdots 4$ using a band-specific ERPE coder (ERPE-B$_k$ in the figure). This subband division and the ERPE scalability mechanism described in Section 2.3 allow the bit stream scalability to be easily implemented. Typically, the lower-frequency subbands are more critical when addressing quality issues, and therefore, the initial enhancement layers generated by the ERPE coders should mainly target bands 1 and 2 (0–5.5 kHz and 5.5–11 kHz, resp.). We emphasize that by ERPE we do not imply just a single modeling stage but most likely a succession of pulse excitation coding stages concatenated using the scalable pulse layering approach described in Section 2.3. We also note that the input signal to the QMF is not spectrally flat. Therefore, a linear prediction stage is included in each ERPE system. In the lowest frequency band, this is a Laguerre-based LP system [14] with $\lambda = 0.3$ and order 12 and for the second and third bands a conventional LP system with orders 9 and 3, respectively. This prediction stage has been shown to facilitate the subsequent pulse modeling phase while the required prediction parameters barely contribute to the total bit rate budget (e.g., for the lowest frequency band, where the highest prediction order is used, the bit budget for the prediction parameters added up to 1.5 Kbit/s).

Obviously, there are many configurations of ERPE stages with different decimation factors allowing many different quality-bit rate operating points to be attained. The one that was finally chosen for the SSC-ERPE coder to be assessed via a listening test is shown schematically in Figure 10. This particular configuration consists of 5 ERPE layers (layers 1–5) on top of the SSC coder (layer 0) tuned at a reduced mono-bit rate of 18 Kbit/s (rather than the 22 Kbit/s used in the MPEG-4 standard). In this figure, the annotated bit rates correspond to the total bit rates up to a given layer which allows the material to be decoded using that layer and all the lower ones. The number in brackets next to the ERPE labels indicates the decimation factor used for the basic RPE component. Notice that ERPE blocks with the same decimation factor can lead to slightly different bit rates; these are due to differences in the number of extra pulses employed (see Section 2.2) and/or the number of coefficients employed in the whitening stage (order of the Laguerre-based linear predictor) of a particular subband. It can be seen how the first two extension layers solely target the first subband as this has proved to be the most perceptually relevant for most classes of audio files. These two concatenated stages consist of an ERPE coder with decimation 8 and a subsequent ERPE coder with decimation 2, respectively. Upper layers are defined by additional ERPE stages in subbands 1, 2, and 3 as indicated in Figure 10. Notice that in the proposed architecture, the 4th subband is always modeled by noise as this has proved sufficient to attain excellent quality on a multiple stimuli hidden reference and anchor (MUSHRA) score.

Figure 11 depicts the frame structure of the layered audio representation generated by the SSC-ERPE coder. The decoder can then decide how many layers are used to reconstruct the audio signal by extracting the relevant parts of each frame.

The operation of the SSC-ERPE decoder shown in Figure 12 is as follows: using the SSC noise parameters contained in the base layer (layer 0, i.e., SSC), locally generated white Gaussian noise (WGN) is spectrally and temporally shaped (block NS). This noise signal is then fed to the analysis QMF described by (13) which decomposes it into 4 frequency bands. This decomposition is needed in order to reuse the full-band noise model already present in MPEG-4 SSC.

According to the layer being decoded, the switches at the input of a synthesis QMF will be configured to select for each band either the subband noise signal or the subband ERPE model generated using the corresponding excitation parameters and prediction coefficients. The QMF synthesis equation is given by

$$g_k(n) = 2w(n)\cos\left(\frac{\pi}{M}(k+0.5)\left(n-\frac{L-1}{2}\right)-\theta_k\right), \quad (15)$$

where M, k, θ_k, $w(n)$, and L have the same meaning as in the analysis equation. Within a given frequency subband, the

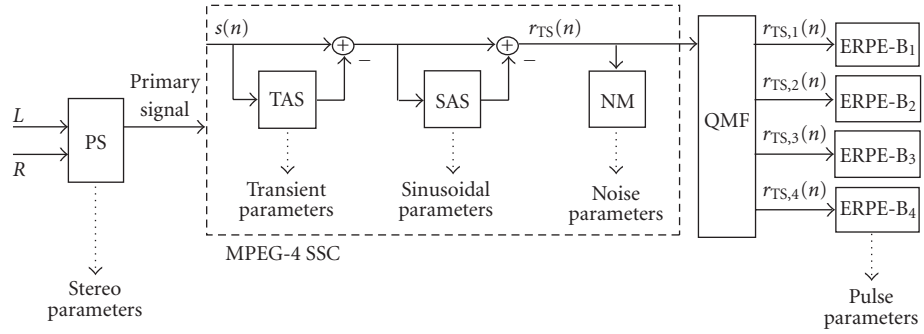

Figure 9: SSC-ERPE encoder.

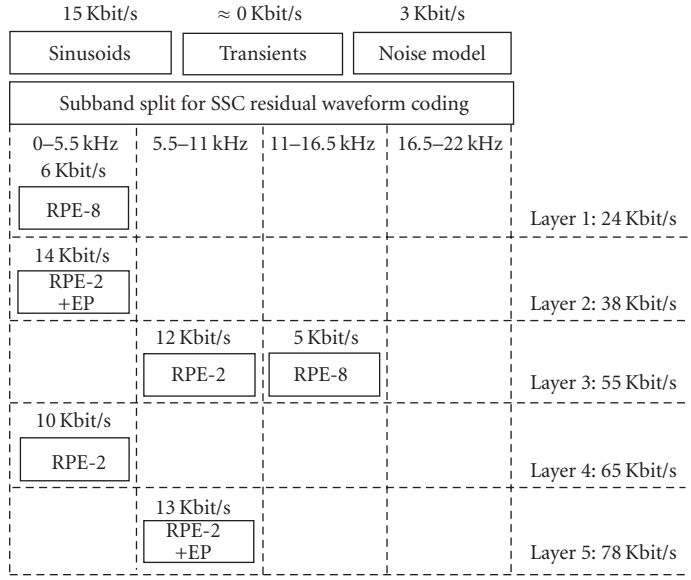

Figure 10: Bit rates (mono) distribution among the five SSC-ERPE extension layers.

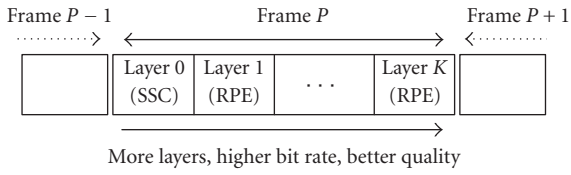

Figure 11: SSC-ERPE frame suitable for variable rate decoding.

mixing factor mechanism described in Section 2.3 is used to provide extra scalability. An approximation of $r_{TS}(n)$ comes out of the synthesis QMF to which generated transients (block TS) and sinusoids (block SS) are added to render the reconstructed audio signal $\hat{s}(n)$. Notice that when all switches are configured to select the subband noise, the scheme reduces to the conventional MPEG-4 SSC decoder. This fact guarantees that the SSC-ERPE decoder is backward compatible in the sense that it can still decode standard MPEG-4 SSC material. As a final step, the reconstructed monosignal is fed, jointly with the PS stereo parameter stream, to the PS decoder to obtain the decoded stereo signal.

5. LISTENING TEST RESULTS

In order to evaluate the performance of the SSC-ERPE coding architecture, a formal listening test was conducted. This test served the double purpose of establishing the quality achieved by the different SSC-ERPE layers and to compare them with current (standardized) coding techniques. To this end, five extensions layers were added on top of the 18 Kbit/s SSC leading to a bit stream that can be decoded at 24, 38, 55, 65, and 78 Kbit/s for a monosignal. To these figures, the PS parameter stream bit rate budget should be added. This has been set to 4 Kbit/s for the first two layers (resulting in stereo audio bit rates of 28 and 42 Kbit/s) and to 8 Kbit/s for the last three layers (stereo audio bit rates of 63, 73, and 86 Kbit/s). This choice of different PS bit rates is justified in view of the monoaudio quality achieved at each layer: the higher the quality, the more sensitive human hearing is to spatial artifacts and consequently, more bits are required for the spatial parameters when pursuing the highest quality levels (layers 3, 4, and 5). Overall, the bit rates of the different layers were chosen so that a gradual increase in the bit rate/quality plane was achieved. Figure 10 illustrates the bit rate contribution

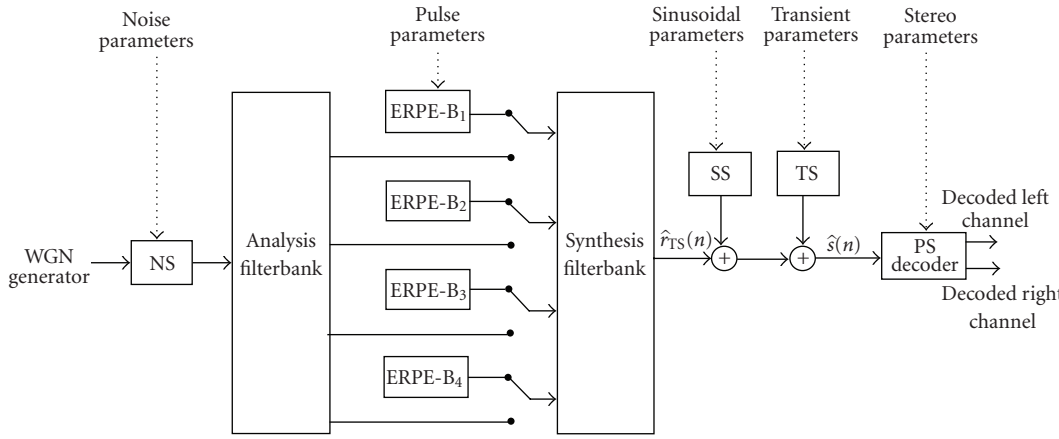

Figure 12: SSC-ERPE decoder.

Table 2

MUSHRA	MOS
0–20	Bad
20–40	Poor
40–60	Fair
60–80	Good
80–100	Excellent

for each coding element in every layer of the coder subject to the test. We note that, currently, the scalability of stereo parameters is not addressed within MPEG-4, therefore, the different PS parameter streams (4 Kbit/s and 8 Kbit/s) have been derived independently. Nevertheless, modifying MPEG-4 PS to produce bit stream scalable parameters is, in principle, feasible.

The listening test was performed with headphones in accordance to ITU-R Recommendation BS.1534 using the MUSHRA scale (0–100). In an MUSHRA test, for each excerpt to be tested, the subject is presented with a known original and blind versions from the coders to be tested. Also included among the coder versions are two anchors (consisting of the originals bandlimited to 7 kHz and 10 kHz, resp.) and a hidden reference. The subject has to rank each coder version on a quality scale from 0 to 100 in Table 2 where the relations between MUSRHA scale and the more common MOS scale applies.

The listening material consisted of 12 critical excerpts (German male, castanets, etc.) typically used within MPEG for codec evaluation. Given the critical nature of these excerpts, they do not reflect average audio material and therefore, the quality of the encoded material will be relatively low when compared to typical audio material. However, exactly these files should profit from an additional coding stage and, moreover, the contributions of the different layers to the quality are most clearly assessable. Nine subjects took part in the study out of which 5 were experienced (working also in audio coding) and 4 were nonexperienced. For each excerpt, the listener was presented with the original and the following versions: SSC-ERPE (layers 1–5), MP3 at 128 Kbit/s (stereo), AAC at 64 Kbit/s (stereo), MPEG-4 SSC at 24 Kbit/s (stereo), a hidden reference and two anchors. All test material was generated with coder implementations developed at Philips. Notice that we did not include high efficiency AAC (HE-AAC) in the listening test. The reason for not being in the test is that a good HE-AAC encoder was not yet available to us at the time of the listening test. Nevertheless, if the claim that HE-AAC at 64 Kbit/s offers the same quality as MP3 at 128 Kbit/s is taken for granted [44], conclusions can still be extracted about how it would compare with SSC-ERPE. Finally, notice that the three competing coders are, in theory, tuned for the best possible performance at the specified bit rates.

Figure 13 shows the results of the test averaged over all excerpts and all subjects for the different coders. This figure shows the average MUSHRA scores for each coder with their 95% confidence interval (CI) versus bit rate. The first important point to note is that the extension layers do indeed provide a graceful increase in the quality/bit rate plane. It is also apparent that the quality increase saturates around the 73 Kbit/s layer making the last extension layer unnecessary. The second important fact shown by Figure 13 is that SSC-ERPE is competitive when compared to standard coders like MP3 and AAC. More specifically, it can be seen that the 63 Kbit/s SSC-ERPE lies in between the quality levels offered by AAC at 64 Kbit/s and MP3 at 128 Kbit/s. The overlap in their CI's prevents us from saying that one is superior to the others. The highest SSC-ERPE layers (73 and 86 Kbit/s) clearly outperform AAC at 64 Kbit/s and achieve identical quality levels as MP3 at 128 Kbit/s. It is also worth mentioning that the CI's for the upper extension layers are very narrow, indicating robustness across excerpts and listeners.

In order to illustrate the coding quality attained for the different excerpts and also the improvement gained by using more layers, Figures 14 and 15 show the MUSHRA scores attained for each excerpt when using SSC-ERPE layer 1 (28 Kbit/s stereo) and SSC-ERPE layer 4 (73 Kbit/s stereo). It can be seen how at 28 Kbit/s, despite achieving a mean score of 60 (i.e., Fair-Good), quality is still fairly low for some of the material, more specifically, for speech and castanets. When decoding takes into account up to SSC-ERPE layer 4

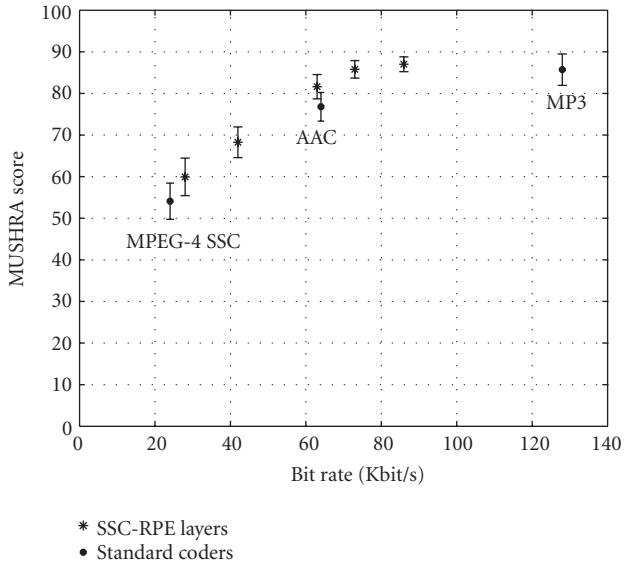

Figure 13: Overall listening test results. These results come from averaging across all excerpts and all subjects. The 7 and 10 kHz anchors (not shown in the graph for the sake of clarity) achieved mean MUSHRA scores of 61 and 74, respectively, whereas the hidden reference (not shown) attained a score of 97.

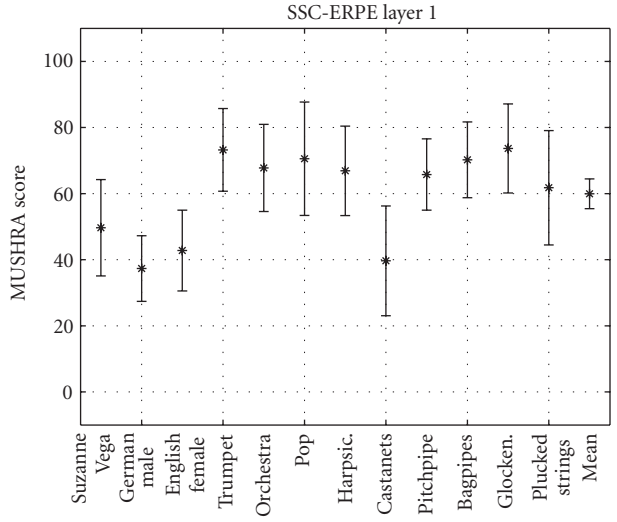

Figure 14: Listening test results per excerpt for SSC-ERPE Layer 1 (28 Kbit/s stereo).

almost all of the material is consistently ranked as excellent quality.

We conclude this section by noting that given the research status of the proposed coder, the optimisation of important implementation aspects such as delay and complexity have not been considered. Nevertheless, the processing requirements of the SSC-ERPE coder are dominated by those of the SSC and PS subsystems for which real-time implementations running on a conventional PC are available, therefore we consider a real-time SSC-ERPE encoder to be within reach of current technology.

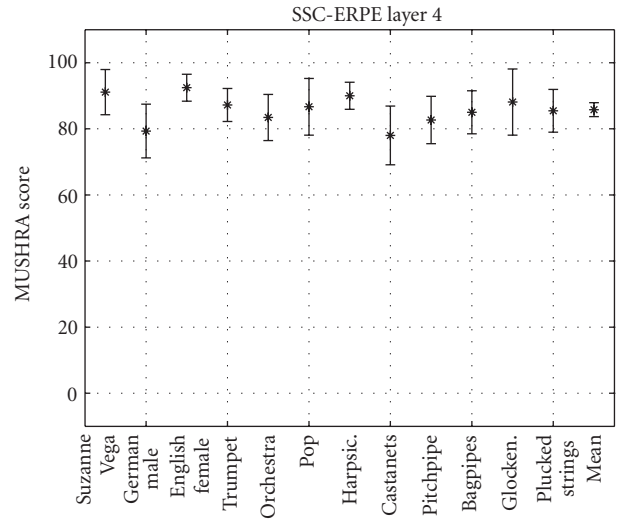

Figure 15: Listening test results per excerpt for SSC-ERPE Layer 4 (73 Kbit/s stereo).

6. CONCLUSIONS

This paper has introduced the SSC-ERPE coder, a technique that combines parametric (MPEG-4 SSC) and waveform (pulse excitation) coding. A parametric stereo coding tool has also been integrated to allow the efficient coding of stereo signals. It has been shown how this combination is suitable for the design of a scalable coder where the operating bit rate can be set at the decoder. The resulting coder is backward compatible with the standard MPEG-4 SSC. A formal listening test shows that SSC-ERPE offers a graceful increase in the quality/bit rate plane and is competitive with standard coders tuned at a particular bit rate.

ACKNOWLEDGMENT

The first author's work is currently supported by a Ramón y Cajal fellowship.

REFERENCES

[1] P. Noll, "MPEG digital audio coding," *IEEE Signal Processing Magazine*, vol. 14, no. 5, pp. 59–81, 1997.

[2] B. Edler, H. Purnhagen, and C. Ferekidis, "ASAC—Analysis/synthesis codec for very low bit rates," in *The 100th AES Convention*, p. 4179, Copenhagen, Denmark, April 1996.

[3] X. Serra, "Musical sound modeling with sinusoids plus noise," in *Musical Signal Processing*, C. Roads, S. Pope, A. Picialli, and G. D. Poli, Eds., pp. 91–122, Swets & Zeitlinger, Lisse, The Netherlands, 1997.

[4] Audio Subgroup, "Report on the verification test of MPEG-4 parametric coding for high-quality audio," ISO/IEC JTC1/SC29/WG11 N6675, 2004.

[5] N. H. van Schijndel and S. van de Par, "Rate-distortion optimized hybrid sound coding," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '05)*, pp. 235–238, New Paltz, NY, USA, October 2005.

[6] F. P. Myburg, "Design of a scalable parametric audio coder," Ph.D. dissertation, Universiteit Eindhoven, Eindhoven, The Netherlands, 2004.

[7] T. S. Verma and T. H. Y. Meng, "6Kbps to 85Kbps scalable audio coder," in *Proceedings of IEEE Interntional Conference on Acoustics, Speech, and Signal Processing (ICASSP '00)*, vol. 2, pp. 877–880, Istanbul, Turkey, June 2000.

[8] W. B. Kleijn and K. K. Paliwal, Eds., *Speech Coding and Synthesis*, Elsevier, Amsterdam, The Netherlands, 1995.

[9] S. A. Ramprashad, "The multimode transform predictive coding paradigm," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 2, pp. 117–129, 2003.

[10] R. Salami, R. Lefebvre, A. Lakaniemi, K. Kontola, S. Bruhn, and A. Taleb, "Extended AMR-WB for high-quality audio on mobile devices," *IEEE Communications Magazine*, vol. 44, no. 5, pp. 90–97, 2006.

[11] J. Makhoul, "Linear prediction: a tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[12] J. D. Markel and A. H. Gray Jr., *Linear Prediction of Speech*, Springer, Berlin, Germany, 1976.

[13] H. W. Strube, "Linear prediction on a warped frequency scale," *The Journal of the Acoustical Society of America*, vol. 68, no. 4, pp. 1071–1076, 1980.

[14] A. C. den Brinker, V. Voitishchuk, and S. J. L. van Eijndhoven, "IIR-based pure linear prediction," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 1, pp. 68–75, 2004.

[15] A. Härmä, M. Karjalainen, L. Savioja, V. Välimäki, U. K. Laine, and J. Huopaniemi, "Frequency-warped signal processing for audio applications," *Journal of the Audio Engineering Society*, vol. 48, no. 11, pp. 1011–1031, 2000.

[16] A. V. Oppenheim, D. H. Johnson, and K. Steiglitz, "Computation of spectra with unequal resolution using the fast Fourier transform," *Proceedings of the IEEE*, vol. 59, no. 2, pp. 299–301, 1971.

[17] P. Kroon, E. Deprettere, and R. Sluyter, "Regular-pulse excitation—a novel approach to effective and efficient multipulse coding of speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 5, pp. 1054–1063, 1986.

[18] B. S. Atal and J. R. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '82)*, pp. 614–617, Paris, France, May 1982.

[19] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): high-quality speech at very low bit rates," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '85)*, pp. 937–940, Tampa, Fla, USA, April 1985.

[20] J.-P. Adoul, P. Mabilleau, M. Delprat, and S. Morissette, "Fast CELP coding based on algebraic codes," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '87)*, pp. 1957–1960, Dallas, Tex, USA, April 1987.

[21] R. V. Cox and P. Kroon, "Low bit-rate speech coders for multimedia communication," *IEEE Communications Magazine*, vol. 34, no. 12, pp. 34–41, 1996.

[22] B. Bessette, R. Salami, R. Lefebvre, et al., "The adaptive multirate wideband speech codec (AMR-WB)," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 8, pp. 620–636, 2002.

[23] R. A. Salami, "Binary code excited linear prediction (BCELP): new approach to CELP coding of speech without codebooks," *Electronics Letters*, vol. 25, no. 6, pp. 401–403, 1989.

[24] C. S. Xydeas, M. A. Ireton, and D. K. Baghbadrani, "Theory and real time implementation of a CELP coder at 4.8 & 6.0 kbits/sec using ternary code excitation," in *Proceedings of IERE International Conference on Digital Processing of Signals in Communications*, pp. 167–174, Loughborough, UK, September 1990.

[25] S. Singhal, "High quality audio coding using multipulse LPC," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '90)*, vol. 2, pp. 1101–1104, Albuquerque, NM, USA, April 1990.

[26] X. Lin, R. A. Salami, and R. Steele, "High quality audio coding using analysis-by-synthesis technique," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '91)*, vol. 5, pp. 3617–3620, Toronto, Ont, Canada, May 1991.

[27] G. T. Waters, Ed., "Sound quality assessment material recordings for subjective tests," Users' handbook for the EBU—SQAM Compact Disk Tech. 3253-E, Technical Centre of the European Broadcasting Union, Brussels, Belgium, 1988.

[28] F. Riera-Palou, A. C. den Brinker, and A. J. Gerrits, "Modelling long-term correlations in broadband speech and audio pulse coders," *Electronics Letters*, vol. 41, no. 8, pp. 508–509, 2005.

[29] F. Riera-Palou, A. C. den Brinker, A. J. Gerrits, and R. J. Sluijter, "Improved optimisation of excitation sequences in speech and audio coders," *Electronics Letters*, vol. 40, no. 8, pp. 515–517, 2004.

[30] R. P. Ramachandran and P. Kabal, "Pitch prediction filters in speech coding," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 4, pp. 467–478, 1989.

[31] R. P. Ramachandran and P. Kabal, "Stability and performance analysis of pitch filters in speech coders," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 7, pp. 937–946, 1987.

[32] T. V. Sreenivas, "Modelling LPC-residue by components for good quality speech coding," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '88)*, pp. 171–174, New York, NY, USA, April 1988.

[33] P. Kroon, "Time-domain coding of (near) toll quality speech at rates below 16 kb/s," Ph.D. dissertation, Delft University of Technology, Delft, The Netherlands, 1985.

[34] W. B. Kleijn, D. J. Krasinski, and R. H. Ketchum, "Improved speech quality and efficient vector quantization in SELP," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '88)*, pp. 155–158, New York, NY, USA, April 1988.

[35] S. Zhang and G. Lockhart, "Embedded RPE based on multistage coding," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 4, pp. 367–371, 1997.

[36] E. G. P. Schuijers, A. W. J. Oomen, A. C. den Brinker, and A. J. Gerrits, "Advances in parametric coding for high-quality audio," in *Proceedings of IEEE Benelux Workshop on Model Based Processing and Coding of Audio*, pp. 73–79, Leuven, Belgium, November 2002.

[37] J. D. Johnson and A. J. Ferreira, "Sum-difference stereo transform coding," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '92)*, vol. 2, pp. 569–572, San Francisco, Calif, USA, March 1992.

[38] J. Herre, K. Brandenburg, and D. Lederer, "Intensity stereo coding," in *The 96th AES Convention*, p. 3799, Amsterdam, The Netherlands, March 1994.

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

[40] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers, "Parametric coding of stereo audio," *EURASIP Journal on Applied Signal Processing*, vol. 2005, no. 9, pp. 1305–1322, 2005.
[41] ISO/IEC, "Coding of audio-visual objects—part3: audio, AMENDMENT 2: parametric coding of high quality audio," ISO/IEC Int. Std. 14496-3:2001/Amd2:2004, July 2004.
[42] P. P. Vaidyanathan, *Multirate Systems And Filter Banks*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1993.
[43] F. Riera-Palou, A. C. den Brinker, and A. J. Gerrits, "A hybrid parametric-waveform approach to bit stream scalable audio coding," in *Proceedings of the 38th Asilomar Conference on Signals, Systems and Computers*, vol. 2, pp. 2250–2254, Pacific Grove, Calif, USA, November 2004.
[44] R. Amorim, "Results of 64kbps Public Listening Test," http://www.rjamorim.com/test/64test/results.html, 2003.

Felip Riera-Palou was born in Palma (Mallorca, Spain) in 1973. He received the B.S. degree in computer engineering from the University of the Balearic Islands (UIB), (Mallorca, Spain) in 1997, the M.S. and Ph.D. degrees in communication engineering from the University of Bradford (United Kingdom) in 1998 and 2002, respectively, and the M.S. degree in Statistics from the University of Sheffield (United Kingdom) in 2006. From May 2002 to March 2005, he was with Philips Research Laboratories, Eindhoven (The Netherlands) first as a Postdoctoral Fellow (Marie Curie program, European Union) and later as a member of technical staff. At Philips, he was involved in research programs related to broadband audio/speech compression and speech enhancement for mobile handsets. In April 2005, he became a Research Associate (Ramon y Cajal program, Spanish Ministry of Science and Education) in the Mobile Communications Group of the Department of Mathematics and Informatics at UIB where his work focuses on signal processing techniques for future wireless communication systems.

Albertus C. den Brinker was born in Heerlen, the Netherlands, in 1957. He received the M.S. degree in electrical engineering in 1983 from Eindhoven University of Technology (EUT). In 1989, he received the Ph.D. degree for his work on dynamic models of the human visual system. From 1987 to 1999, he worked in the Signal Processing Group at the Department of Electrical Engineering, Eindhoven University of Technology. His educational activities included electrical network theory and digital filter theory. His research activities concerned diverse topics from fields like approximation, identification, digital signal processing, and applications of orthogonal series expansions to (local) signal analysis and adaptive filtering. In 1999, he joined the Digital Signal Processing Group at Philips Research Laboratories, Eindhoven, where he is Head of the Cluster Signal Processing Techniques for Audio and Speech. One of the activities within the Cluster concerns standardization of audio coders, especially standardization within MPEG. Recently, major contributions were made to MPEG-4 Amendment 2 (high-quality parametric audio coding) and MPEG Surround. He publishes regularly in international scientific journals and proceedings of scientific conferences and is (co-)author of several patents.