Author: J. Stemerdink	Date of issue: 16-10-2003 Kind of issue: public	Scope: Station processing studies Doc.nr.: LOFAR-ASTRON-MEM-109
	Status: Draft Revision nr.: 0.3	File:

Polyphase filter bank quantization error analysis

J. Stemerdink

Verified:			
Name	Signature	Date	Rev.nr.
A. Gunst			

Accepted:		
Team Manager	System Engineering Manager	Program Manager
M. van Veelen .... *date:*	C.M. de Vos .... *date:*	J. Reitsma .... *date:*

©ASTRON 2003

All rights are reserved. Reproduction in whole or in part is
prohibited without the written consent of the copyright owner.

Distribution list:

Group:	For Information:
ASTRON	WMC:
M. van Veelen	A. de Graaf
A. Gunst	A. Huijgen
C.M. de Vos	C. Huyts

Document revision:

Revision	Date	Section	Page(s)	Modification
0.1	14-8-2003	-	-	Creation
0.2	21-8-2003	-	-	Update

Abstract

In the LOFAR station a large number of digital data streams coming from the receivers are filtered and combined into a single digital signal stream. This processing takes place in the cascade of a filter-bank, a beam-former and a second filter-bank. In order to implement this station digital processing in a cost-efficient way the required accuracy of the signal and coefficient representation, and of their processing, has to be established.

This document describes a study of quantization effects in the polyphase filter bank which is currently seen as one of the possible candidates to be applied for the frequency analysis in the station digital processing system. By means of (Matlab) simulations of an example polyphase filter bank the different quantization types are analyzed.

Contents

1 Introduction

In the LOFAR station the signals coming from the receivers are filtered by a cascade of

1. the *first filter-bank*, which splits up the signal from each receiver (with 64 MHz bandwidth) into (256) sub-band signals with a bandwidth of 250 kHz per band;
2. the *beam-former* which in each sub-band combines the signals coming from the filter banks into a single sub-band signal, using a spatial filtering method to attenuate signals coming from unwanted directions;
3. the *second filter-bank*, which splits up the signals from the beam-former in each (250 kHz wide) sub-band into bands of 1 kHz wide.

In order to keep the computational complexity of this system low, the digital representation used to store and process the signals have to be limited. This should be done in such a way that the accuracy of the processing is not too much compromised. Therefore the effects of finite word lengths of coefficients and (intermediate) signals throughout the chain on the system performance (accuracy) should be asessed. Then the minimum word lengths (in terms of number of bits) needed to achieve the required system accuracy can be established at each stage of the chain.

A viable candidate for the filter-banks is the polyphase DFT filter-bank. Although not all of the filter requirements are met in the current design of this filter-bank, it is interesting as a case study to investigate the quantization effects in this particular type of filter-bank implementation. The insight obtained from this study will probably help to analyze other types of filter-banks when the applied type of filter-bank has to be changed in future. Some of the results from this study are likely to apply to this different filter-bank type.

This document describes a study of the effects of limited word lengths in the polyphase DFT filter bank. The processing in this type of filter bank consists mainly of FIR (Finite Impulse Response) filters and an FFT (Fast Fourier Transform; a well-known fast implementation of the DFT, the Discrete Fourier Transform). Therefore FIR and FFT quantization effects will be studied as well.

1.1 Quantization error measure: error-signal power

The quantization error of a quantized subsystem (hereafter to be named: test system) can be measured by comparing the output of that subsystem with that of its unquantized equivalent (hereafter named: reference system) to which the same input signal is administered. Some quantizers in the test system can be switched off while others are kept on, to study the effects of the individual quantizers or groups of quantizers on the output signal. A useful measure in this comparison is the quantization error signal power. If an input signal $x(t)$ is fed into a quantized sub-system S_q and also into a reference subsystem

S_r, the respective output signals $y_q(t) = S_q\{x(t)\}$ and $y_r(t) = S_r\{x(t)\}$ can be compared by subtracting the output signals into $e_q(t) = y_q(t) - y_r(t)$. The output quantization error signal power, measured from $t = 0$ to T_m can then be computed as the second moment of $e_q(t)$:

$$P_q = \frac{1}{T}\int_0^{T_m} |e_q(t)|^2 \mathrm{d}t \tag{1}$$

Note that $e_q(t)$ and other signals may be complex, so

$$|e_q(t)|^2 = e_q(t) \cdot e_q^*(t) \tag{2}$$

where $e_q^*(t)$ is the complex conjugate of $e_q(t)$.

In the first, preliminary simulations the *variance* (or second central moment) was taken as a measure, which resulted in strange effects: DC components became invisible. The variance can be defined as

$$\mathrm{var}\{e_q\} = \frac{1}{T}\int_0^{T_m} |e_q(t) - \mu_{eq}|^2 \mathrm{d}t \tag{3}$$

where μ_{eq} is the average value of $e_q(t)$ in the measurement interval:

$$\mu_{eq} = \frac{1}{T_m}\int_0^{T_m} e_q(t)\mathrm{d}t \tag{4}$$

Especially when a quantization error has a considerable non-zero average, which for example occurs when quantization takes place in 'floor'-mode, a considerable part of the quantization error becomes invisible in the variance measure. The power or second moment measure does not have this drawback, it includes the complete quantization error signal.

2 Quantization in FIR filters

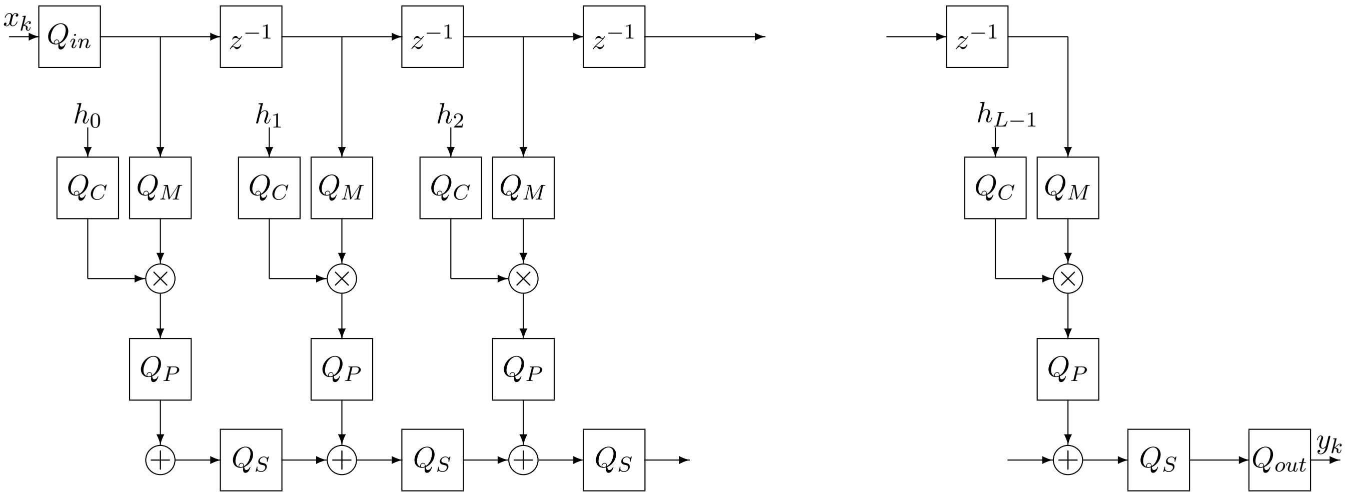


Figure 1: **Block diagram of a FIR filter with quantizers.**

Fig. 1 shows a block diagram of a quantized FIR filter. The input of the filter is quantized by an input quantizer Q_{in} before it enters the shift register. The tapped signals of this shift register are consequently quantized in the multiplicand quantizer Q_M before they are multiplied by the coefficients. Note that for a particular input signal value, this quantization with Q_M has the same output over and over again, after each shift in the shift register. It is therefore redundant, and can be performed once instead in the input quantizer Q_{in}. They are drawn here for generality reasons, and to keep in line with the Matlab quantized filter (qfilt) structure. The coefficients h_k are quantized with the coefficient quantizer Q_C, only once during the filter set-up. After multiplication the products are quantized with the product quantizer Q_P. During each filter cycle all the products are summed; each addition is followed by quantization in the sum quantizer Q_S.

In the diagram all quantizers with the same label are identical; for example, all quantizers marked Q_S are the same and quantize with the same number of bits.

A quantized FIR filter with response h can be defined in Matlab by

```
Hq = qfilt('fir',{h});
```

In this default (as to quantizer formats) definition, the quantizers Q_{in}, Q_C, Q_M, Q_{out} are set to 16 bits (1 sign bit + 15 fractional bits). The quantizers that affect the multiply-accumulate operation, Q_P ad Q_S, are set to 32 bits (1 sign bit, 1 integer bit, and 30 fractional bits). In this way no intermediate quantization takes place between the multiplication and the output quantizer Q_{out}, since the repeated multiply-accumulate operation takes place in 32 bits, which is enough to accomodate all the product bits. This multiply-accumulate operation can under certain (speed) conditions be implemented serially in a single multiplier and a single accumulator; then, the extra costs for doubling the word-length may be acceptable.

In the experiments described hereafter, the quantized FIR filters were defined with double word lengths in Q_P and Q_S. When the signal quantization of the filter is n_s and the coefficient quantization is n_c, the bit accuracies in the FIR filter will be:

$$\begin{aligned}
Q_{in} &: n_s \text{ bits} && (5)\\
Q_{out} &: n_s \text{ bits} && (6)\\
Q_M &: n_s \text{ bits} && (7)\\
Q_P &: 2 \times n_s \text{ bits} && (8)\\
Q_S &: 2 \times n_s \text{ bits} && (9)\\
Q_C &: n_c \text{ bits} && (10)
\end{aligned}$$

2.1 FIR signal quantization noise

If the filter is defined as described above, the quantization noise at the output (as a results of *signal* quantization) is well approximated by

$$\sigma_{qs}^2 \approx \frac{\Delta^2}{12} = \frac{1}{12} \cdot 2^{-2(n_s-1)} \tag{11}$$

where

$$\Delta = 2^{-n_{frac}} = 2^{-(n_s-1)} \tag{12}$$

is the quantization step-size of an n_s-bit quantizer with 1 sign bit and $n_s - 1$ fractional bits. The quantization noise is completely determined by the output quantizer Q_{out}. The product and sum quantizers Q_P and Q_S are dimensioned with such high accuracy that they don't affect the signal at all, so they can be ignored.

Apart from this quantization noise, there is also the effect of coefficient quantization. The quantization of the coefficients usually takes place during the filter design or the initialization of the filter; it affects the response of the filter (the impulse response and the frequency response) but does not affect the linearity of the filter. Coefficient quantization effects can be represented entirely by the quantized filter's frequency transfer function.

In the following fractional fixed-point representation will be used. Then, all signal numbers and coefficients should be of magnitude ≤ 1. This should apply to input signals, coefficients and output signals.

2.2 Quantization of FIR filter coefficients

Let us look at the FIR filter of Fig. 1, with real coefficients $\{h_k, k = 0...L-1\}$, with $|h_k| \leq 1$. Let the sequence x_k be the input sequence to this filter. If we ignore the quantizers for a moment, then the output sequence y_k will be given by

$$y_k = \sum_{i=0}^{L-1} h_i x_{k-i} \tag{13}$$

Let us consider an example of a low-pass FIR filter of order 127 (i.e. the filter has 128 taps and 128 coefficients). The impulse response of this sample FIR filter is shown in fig. 2. This filter will be used lateron as a so-called *prototype* filter to design the example polyphase-DFT filter bank.

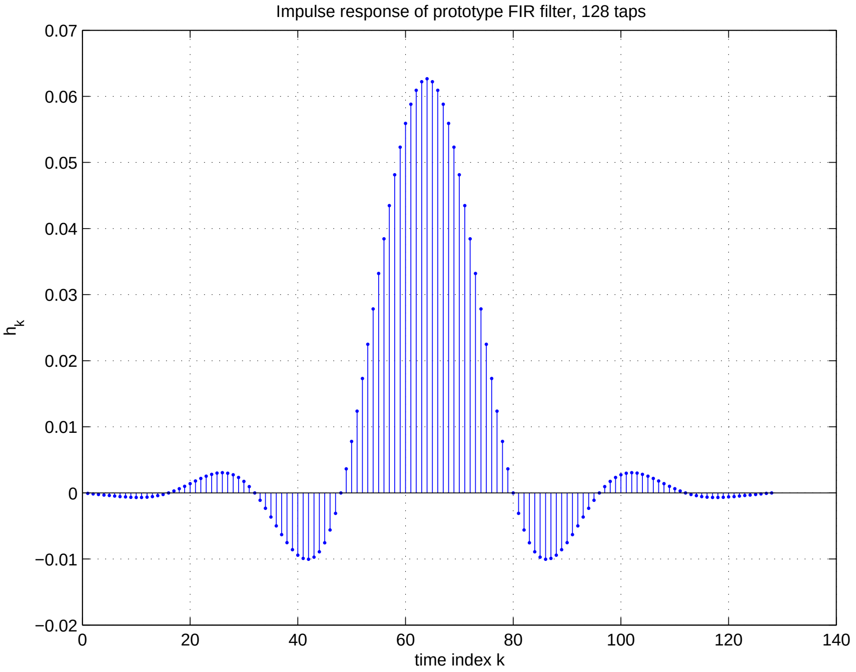


Figure 2: **Impulse response of a 128-tap low-pass FIR filter**

2.2.1 Preventing overflow

If the input sequence is normalized such that $|x_k| \leq 1$, then the output sequence y_k is guaranteed to be within the same range when h_k satisfy

$$\sum_{k=0}^{L-1} |h_k| \leq 1. \tag{14}$$

Hence, by scaling the coefficients in such a way that eq. (14) is satisfied, we can make the filter robust against overflow. A scale factor applied to the output signal can normalize the filter to its original response amplitude.

The filter in the example of fig. 2 does not satisfy eq. (14) . In this filter we find that

$$\sum_{k=0}^{L-1} |h_k| \approx 1.42 > 1. \tag{15}$$

We can scale down the coefficients h_k by a factor 2, so that (14) is satisfied. Then we can perform the quantization of the coefficients. In this example we shall use a 10-bit quantizer $Q_{10,9}$ with 9 fractional bits. The resulting impulse response is shown in fig. 3

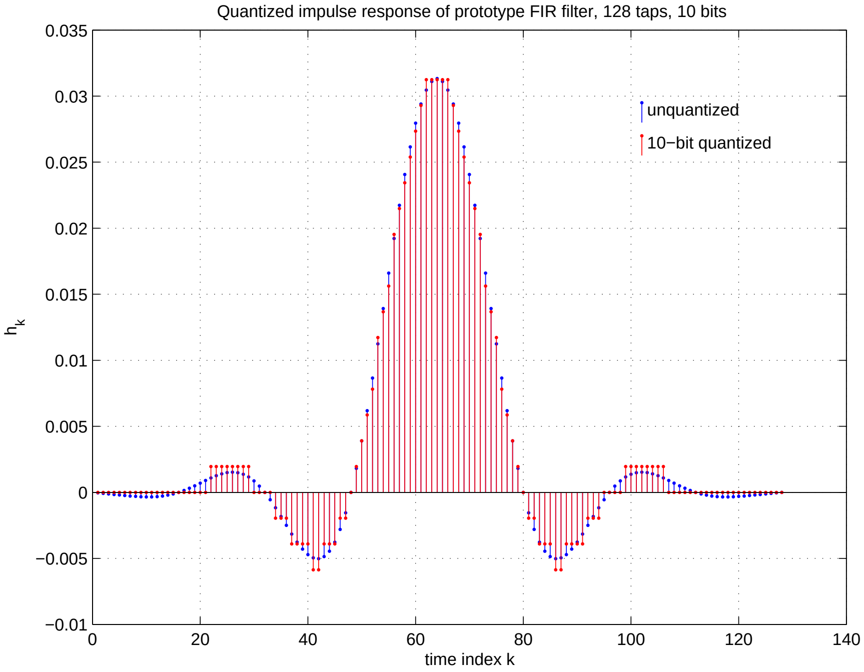


Figure 3: **Quantized impulse response of a 128-tap low-pass FIR filter (8 bits**

The corresponding spectral amplitude responses of the original (scaled) and quantized FIR filter are shown in fig. 4.

©ASTRON 2003

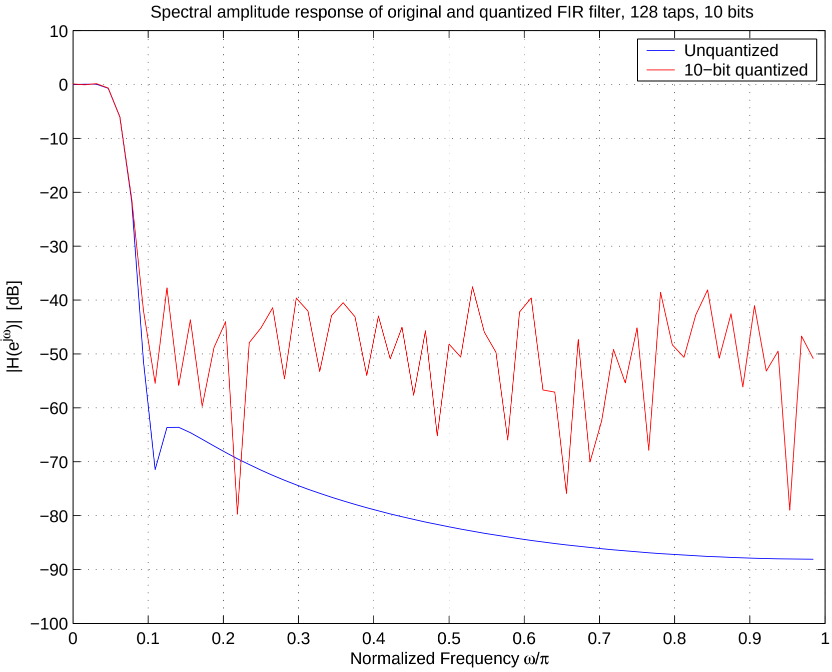


Figure 4: **Spectral amplitude response of 128-tap low-pass FIR filter**

In order to analyze the output error as a result of coefficient quantization we shall also consider the coefficient quantization error. Let $x(n)$ be the input of a FIR filter with coefficients h_k and also of a quantized FIR filter with coefficients $\hat{h}_k$. The output of the unquantized FIR filter is

$$y_n = \sum_{k=0}^{L-1} h_k \cdot x_{n-k} \tag{16}$$

Likewise, the output of the quantized filter is

$$\hat{y}_n = \sum_{k=0}^{L-1} \hat{h}_k \cdot x_{n-k} \tag{17}$$

The error in the output as a result of coefficient quantization is given by

$$\epsilon_n = \hat{y}_n - y_n = \sum_{k=0}^{L-1} \delta_k \cdot x_{n-k} \tag{18}$$

where $\delta_k = \hat{h}_k - h_k$ is the coefficient quantization error of coefficient h_k. Thus ϵ_n equals the output of an imaginary FIR filter with input x_n and coefficients (impulse response) δ_k. For our example FIR this imaginary impulse response is shown in fig. 5.

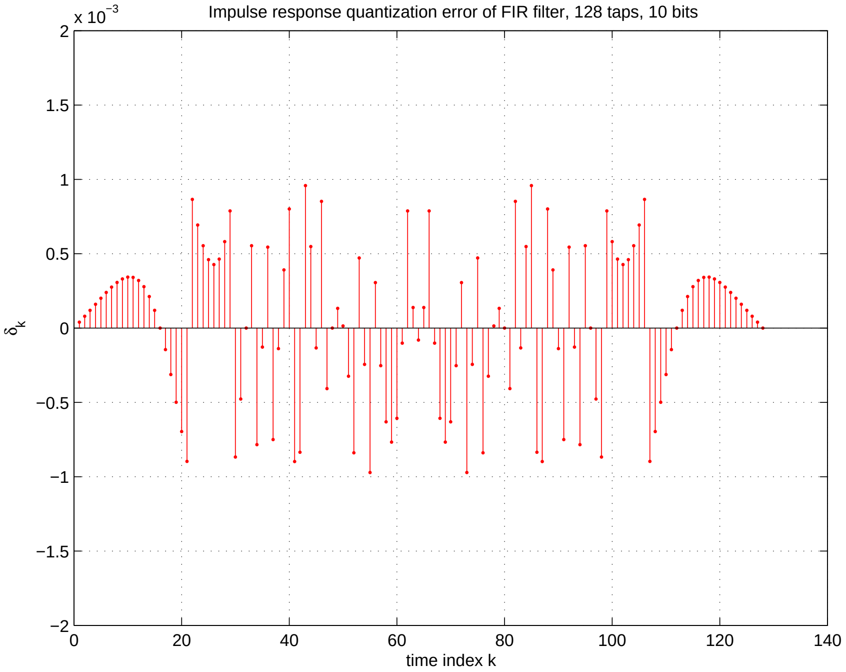


Figure 5: **Impulse response quantization error of a 128-tap low-pass FIR filter (10 bits**

Let x_n be a white noise input signal to both filters. Then the coefficient quantization error power can be computed as

$$\sigma_\epsilon^2 = \sigma_x^2 \cdot \sum_{\ell=0}^{L-1} |\delta_\ell|^2 \tag{19}$$

If the quantization is not too coarse, the coefficient quantization errors δ_k will be approximately uniformly distributed over an interval $\Delta_c = 2^{-(n_c-1)}$ where n_c is the number of coefficient quantization bits ($n_c - 1$ is the number of fractional bits). Then the coefficient quantization error power can be approximated as

$$\sigma_\epsilon^2 \approx \sigma_x^2 \cdot L \cdot \frac{\Delta_c^2}{12} = \sigma_x^2 \cdot L \cdot \frac{2^{-2(n_c-1)}}{12} \tag{20}$$

This approximation is fairly accurate for noise input signals, but not for narrow-band inputs like sine-waves. The exact formula for any kind of input signal would be

$$\sigma_\epsilon^2 = \frac{1}{\pi} \int_0^\pi \left|X(e^{j\omega})\right|^2 \cdot \left|\Delta(e^{j\omega})\right|^2 d\omega \tag{21}$$

where $X(e^{j\omega})$ and $\Delta(e^{j\omega})$ are the Fourier transform of x_n and δ_n, respectively. For a sine-wave input the outcome would be very dependent on the frequency. However, the white noise approximation is an average over all frequencies and therefore a reasonably good indicator for the coefficient quantization accuracy.

The spectral representation of the coefficient quantization error is shown in fig. 6.

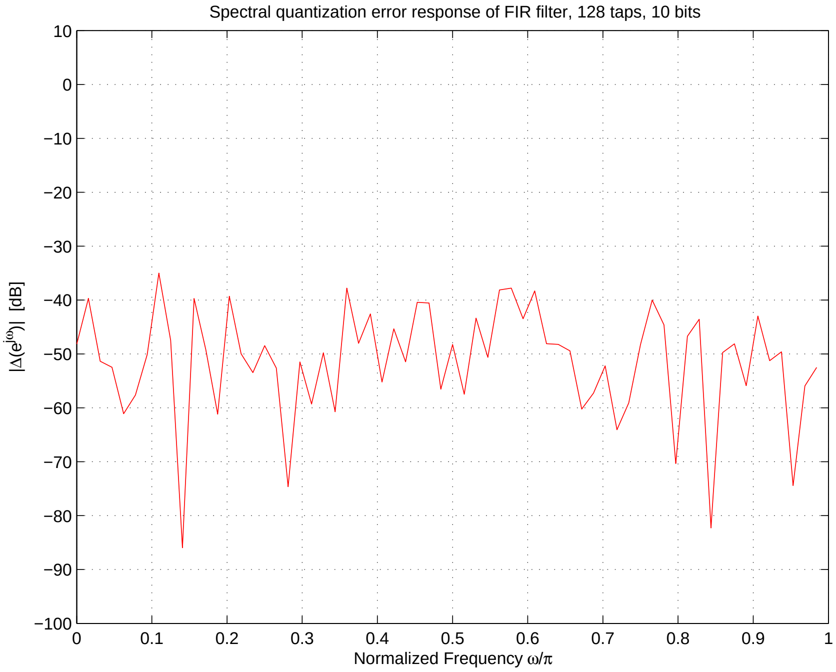


Figure 6: **Spectral response quantization error of 128-tap low-pass FIR filter**

2.2.2 Minimizing quantization errors

The accuracy of a quantized FIR filter can be optimized by scaling up the coefficients until just below the limit defined by equation (14); then the relative quantization errors (with respect to the output signal) will be minimum. In this way the dynamic range of the filter will be maximally exploited.

In the design of the polyphase filters a prototype low-pass FIR filter is designed first, and its impulse response is then transformed into M FIR filters (M=number of bands). Since the separate band-filters are much shorter than the prototype FIR, they can in general be scaled up further than the prototype FIR filter, before reaching this limit. Therefore, some improvement of the quantized polyphase FIR filter accuracy can be achieved by rescaling their coefficients after this transformation.

In the example of the filter bank described hereafter, the polyphase FIR coefficients could be scaled up by a factor 8. This increased the output levels of the filter by 18 dB, whereas the FIR quantization effect remained approximately the same.

3 Quantized Fast Fourier Transform (FFT)

The $M-$point Discrete Fourier Transform (DFT) of a vector $\mathbf{x} = (x_0, x_1, x_2, ...x_{M-1}$ is defined as

$$Y_k = \sum_{n=0}^{M-1} x_n \cdot e^{-\jmath \frac{2\pi}{M} kn} \quad , \quad k = 0, 1, ...M-1 \tag{22}$$

The inverse operation, the IDFT (Inverse Discrete Fourier Transform) is defined as

$$x_n = \frac{1}{M} \sum_{k=0}^{M-1} Y_k \cdot e^{\jmath \frac{2\pi}{M} kn} \quad , \quad n = 0, 1, ...M-1 \tag{23}$$

In the well-known FFT algorithm the DFT and IDFT operations are implemented in an extremely efficient way by factorizing the transform coefficients $e^{\pm \jmath \frac{2\pi}{M} kn}$ into terms called *twiddle factors* of the form

$$W_{2^m} = e^{-\jmath \frac{2\pi}{2^m}} = W_M^{2^m} \quad , \quad i = 0, 1, 2, \dots \log_2(M) \tag{24}$$

The coefficients can then be factorized into the terms

$$W_M,\ W_{M/2},\ W_{M/4},\ W_{M/8}, ...W_2,\ W_1 \tag{25}$$

By combining common multiplications by common twiddle factors the number of multiplications is reduced dramatically. The FFT is a network composed of so-called *butterfly* structures which contain one complex multiplication with a twiddle factor and two additions. Fig. 7 shows the signal flow diagram of a single butterfly.

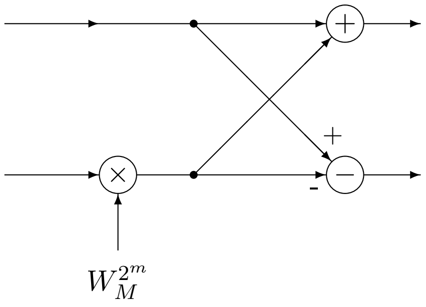


Figure 7: **FFT butterfly structure**

The Matlab structure `qfft` implements a quantized FFT, in which coefficients and signals are quantized in a similar manner as in the quantized filter `qfilt`.

The inputs of the FFT are quantized by an input quantizer Q_{in}, and the outputs by Q_{out}. Before the multiplication with a twiddle factor, signals are quantized by a multiplicand quantizer Q_M. The results of this multiplication are quantized by product quantizer Q_P. After the additions and subtractions the results are quantized by sum quantizer Q_S. The coefficients (twiddle factors) are quantized in a coefficient quantizer Q_C. A butterfly with quantizers is shown in fig. 8.

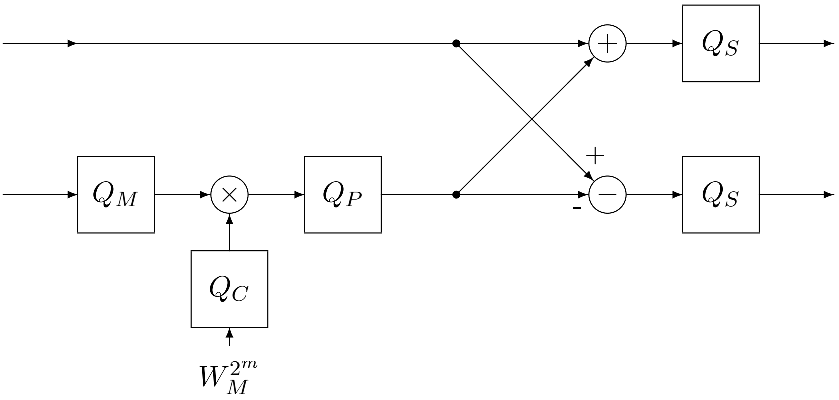


Figure 8: **Quantized FFT butterfly structure**

In the FFT in the example polyphase/DFT filter-bank described in the next section, the bit-length parameters were set according to (10), just like in the FIR filters.

4 Quantized polyphase/DFT filter bank

A block diagram of the polyphase/FFT filter bank is shown in fig. 9. The tapped delay line on the left, in combination with the M-fold decimation units, splits the incoming signal sample stream $x(n)$ into M parallel sample streams $u_0(n), u_1(n), ...u_{M-1}(n)$ which each have a sample rate which is a factor $1/M$ lower than the input signal $x(n)$.

$$u_\ell(n) = x(nM + M - 1 - \ell) \tag{26}$$

These M signal streams are each fed into a polyphase FIR filter of length N. The filters $E_\ell(z)$ are the polyphase band-filters derived from the prototype low-pass FIR filter. The z-transform $E_\ell(z)$ is defined as

$$E_\ell(z) = \sum_{n=0}^{N-1} e_\ell(n) z^{-n} \tag{27}$$

where the coefficients $e_\ell(n)$ are derived from the prototype low-pass filter $H(z)$

$$H(z) = \sum_{n=0}^{L-1} h(n) z^{-n} \tag{28}$$

©ASTRON 2003

The coefficients of the polyphase filters $e_\ell(n)$ are simply a rearrangement of the coefficients $h(n)$, $n = 0 ... L-1$:

$$e_\ell(n) \overset{\triangle}{=} h(nM+\ell), \quad 0 \leq \ell \leq M-1, \quad 0 \leq n \leq N-1 \tag{29}$$

In this schematic we can see that there are basically two parts where rounding errors can occur and re-quantization will become an issue: the polyphase FIR filters, and the DFT (implemented as an FFT here). Therefore, the quantization effects in these two parts will be analyzed.

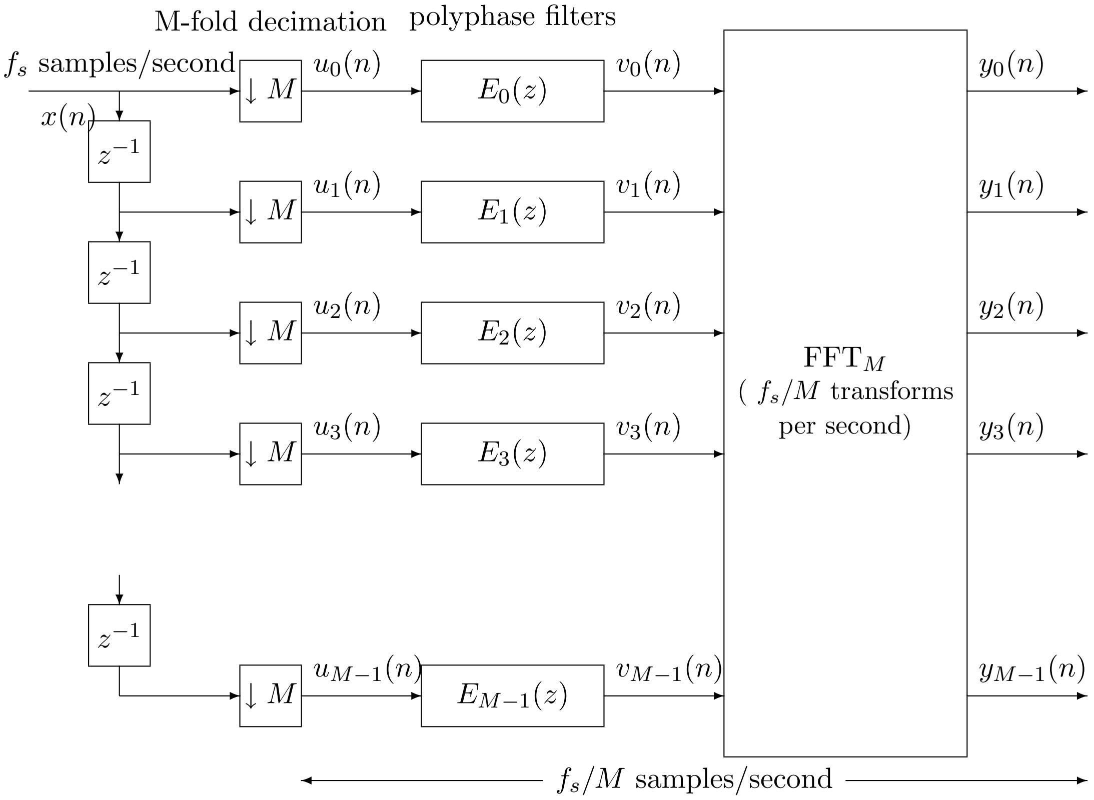


Figure 9: **Block diagram of a polyphase/DFT filter bank.**

4.1 Quantization of polyphase FIR coefficients

The problem of defining the quantized FIR coefficients is mainly a scaling problem. Equation (14) can be applied to the polyphase FIR filters $E_\ell(z)$ by defining a common scaling factor a_h by which all the coefficients $e_\ell(n)$ are scaled before quantization. For this purpose (14) can be restated as

$$\max_{\ell=0}^{M-1} \alpha_h \sum_{k=0}^{N-1} |e_\ell| \leq 1. \tag{30}$$

After selecting an appropriate value for α_h which satisfies this condition, the quantized coefficients $\hat{e}_\ell(n)$ can be computed as

$$\hat{e}_\ell(n) = Q_C\left(\alpha_h e_\ell(n)\right) \tag{31}$$

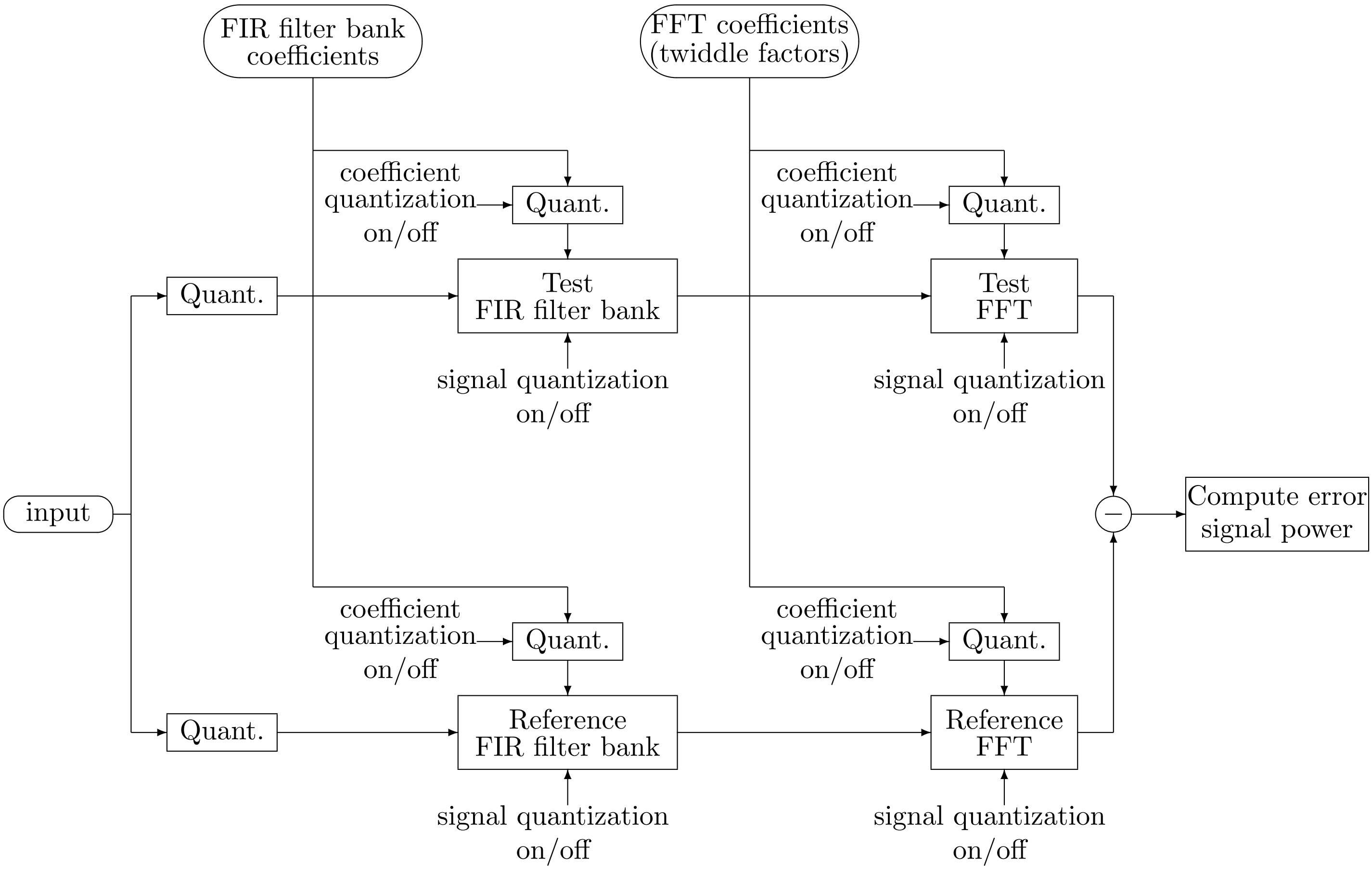
This factor is in general larger than the factor allowed to scale the prototype FIR. In the example prototype FIR described in 2.2, the coefficients h_k had to be scaled *down* by a factor $\alpha_{h,\max} = 1/4.222 = 0.7032$ (we chose the value 0.5 since that can be realized by a bit-shift); the upper-bound for the scaling factor given by (30) is $\alpha_{h,\max} = 9.75$. In the quantized polyphase simulations hereafter the value $\alpha_h = 8$ was chosen, which can be realized by a mere bit-shift. In comparison with the prototype the scale-factor can be chosen a factor 13.8 larger for the polyphase FIR filter bank than for the prototype FIR.

5 Quantization error analysis of the polyphase/DFT filter bank

5.1 The polyphase/DFT filter bank simulation model

In Matlab a simulation model was set up consisting of two polyphase/DFT filter bank systems, a ”test-system” and a ”reference-system”. In these two systems the quantization of the coefficients in the polyphase FIR filters, the coefficients (twiddle factors) in the FFTs, and the quantization of the (intermediate) signals in the polyphase FIR filters and in the FFTs can be switched on and off separately. The quantized filters and FFTs were implemented as qfilt and qfft objects from the Matlab Filter Design Toolbox; also the quantizer object from this toolbox was used.

©ASTRON 2003



FIR filter bank coefficients
FFT coefficients (twiddle factors)
coefficient quantization on/off
Quant.
coefficient quantization on/off
Quant.
Quant.
Test FIR filter bank
Test FFT
signal quantization on/off
signal quantization on/off
input
Compute error signal power
−
coefficient quantization on/off
Quant.
coefficient quantization on/off
Quant.
Quant.
Reference FIR filter bank
Reference FFT
signal quantization on/off
signal quantization on/off

Figure 10: Block diagram of polyphase/DFT filter-bank simulation set-up

By performing simulations with combinations of certain quantizations switched on and off, and comparing the outputs of the test system with those of the reference system, the effects of the various quantization steps can be evaluated. This is illustrated in the block diagram of figure 10.

5.2 Sine-wave sweep simulations

Using this simulation set-up, a number of simulations were performed, using a sine-wave sweep across the full frequency range (200 frequency points from 0 to half the sample frequency). In order to limit the simulation time, a relatively simple filter bank with 16 sub-bands ($M = 16$) and polyphase FIR filters with 8 taps ($N = 8$) was taken as an example. In order to make the quantization effects clearly visible and discernable from the stop-bad ripple, a 10-bits quantization was taken as a basis.

The 10-bit quantized sine wave was applied as an input signal to both the reference system and the test system simultaneously. A 16-band filter bank was used, where each polyphase FIR filters has 8 taps. Since the input signal is real, the output spectrum has an even symmetry. This symmetry can be described as

$$y_{M-k}(n) = y_k^*(n) \ , \quad k = 1, 2, ...\frac{M}{2} \tag{32}$$

Therefore only the results for positive frequencies were plotted. For each measurement frequency, the output signals of the reference system were subtracted from those of the test system, resulting in an "error" signal, representing the quantization error. The variance of this error signal was computed, resulting in a quantization error-signal power for each output, for each frequency. These data were then collected in a plot, along with the output signal power of each frequency band.

The plotted signal data represent output signal power spectrums since the plotted signal levels are the output powers of sine waves, although shifted in frequency. The frequency of the sine wave at the output of the band is given by the frequency with respect to the centre of that band. However, it is important to understand that the plotted error data are not representations of output error signal spectrums, since the frequency parameter on the x-axis is the frequency of the input sine wave. For each frequency the plotted error values are the error powers when the input is a sine wave with that particular frequency. The error signal, with the power that can be read from the graph at a certain frequency, has a complete frequency spectrum of its own, and can be concentrated at completely different frequencies. So the position of the plotted error data on the frequency axis bears no relation to the frequency content of the error itself.

All quantizers were set to 10 bits, except the product and sum quantizers, that are used in multiply-accumulate constructions: they were set to 20 bits. The 20-bit results are quantized to 10 bits in the output quantizer. The coefficient quantizers were set to round-mode 'round' (= rounding to the nearest integer) whereas all the signal quantizers were set to round-mode 'floor' (= rounding towards $-\infty$). The overflow-mode of all quantizers was set to 'saturate' (= clipping, as opposed to 'wrap', were overflowing values are 'wrapped around' to negative values).

5.2.1 The accumulated quantization errors at the filter bank output

Figure 11 shows the signal powers and quantization error powers at the output of the filter bank, as a result of all the quantization steps. In the reference system all the quantizations were switched off, and in the test system all quantizations were switched on. The plotted error signal thus represents the total quantization error as they appear in the output signals.

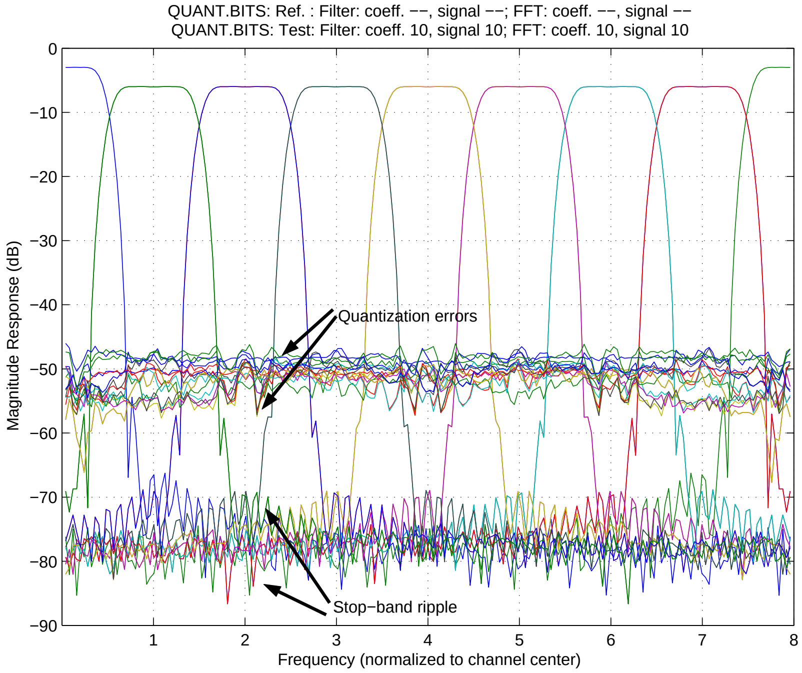


Figure 11: **The effect of coefficient and signal quantizations on the output signal.**
Output signal and output quantization error power spectra, obtained by a sine-wave sweep. In the reference system, filter coefficients (and FFT twiddle factors) and signals are unquantized. In the test system, all coefficients and signals are quantized (10 bits). Coefficients and intermediate signals are quantized in 'round'-mode.

Only 8 bands appear in the plot (actually 7 + 2 half bands), whereas $M = 16$; this is a result of the fact that this is a complex DFT filter bank. In fact the 16 bands are divided over the frequency range $[-\pi, +\pi]$, but since the input is real, the output is symmetric around frequency 0.

The input signal was a sine-wave with amplitude 1 which is equivalent with a signal power of $\frac{1}{2}$ or -3 dB. The maximum output power is therefore -3 dB. It can be seem that bands 0 and 8 indeed have this -3 dB output power level. For the other bands the output power is split up in equal halves among the symmetric bands k and $M - k$; therefore these bands have an output of another 3 dB lower, amounting to a -6 dB output power. Bands 0 and $M/2 = 8$ have a real output and therefore don't have a mirror band; in fact, they are their own mirror bands and therefore they have twice the output power that the other bands have.

In the following an attempt will be made to isolate the various contributions of the separate quantization parts to this total quantization error at the output, in order to determine which quantizations are critical. Using these results, we can then tune the word lengths in different parts of the filter bank, by increasing the word length of the most critical parts.

5.2.2 Polyphase FIR filter coefficient quantization errors at filter bank output

In figure 12 only the FIR filter coefficient quantizations in the test system were switched on, and all other quantizations in the test system and the reference system were switched off. In this way the FIR filter coefficient quantization effects are shown separately. In fact, they are not really quantization errors, in the sense that they cannot be seen as quantization noise, but rather deviations in the response of the filter bank. The "quantized" test filter bank is a completely linear system just like the "unquantized" reference filter bank; only its response shows slight deviations due to the quantizations of the coefficients.

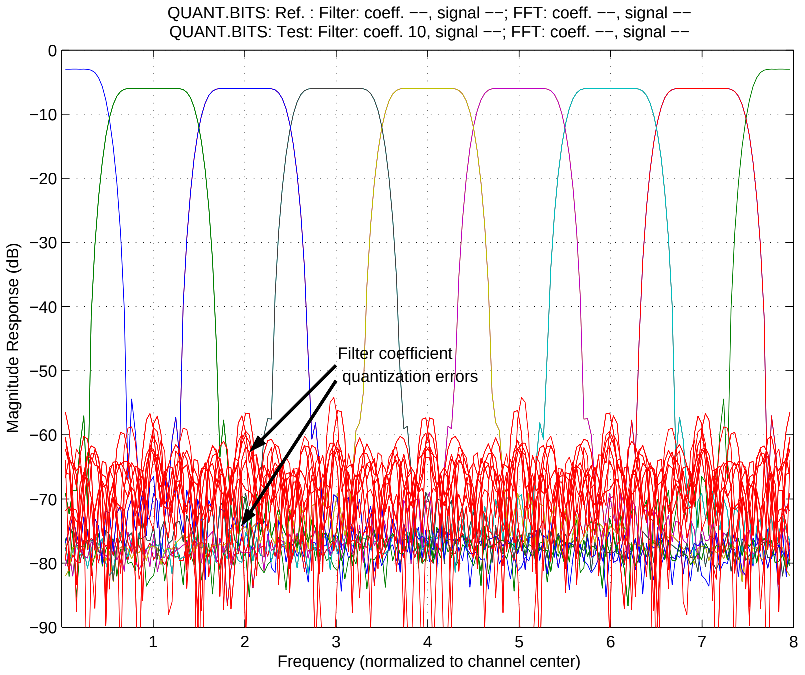


Figure 12: **The effect of filter coefficient quantization.**
Output signal and output quantization error power spectra, obtained by a sine-wave sweep. In the reference system, filter coefficients, FFT twiddle factors and signals are unquantized. In the test system, filter coefficients are quantized, filter signals are unquantized; in the FFT coefficients (twiddle factors) and signals are unquantized.

The FIR coefficient quantization error, at the output of the polyphase FIR filters, can be approximated by (20).

$$\sigma_\epsilon^2 \approx \sigma_x^2 \cdot N \cdot \frac{2^{-2(n_c-1)}}{12} \tag{33}$$

Taking into account that

- the M-point FFT increases the error power by a factor M;
- the FIR responses were scaled up by a factor α_h and therefore the FFT output is scaled down by a factor $1/\alpha_h$ to compensate for this,

we can rewrite this to an approximation of the quantization error at the output of the filter-bank:

$$\sigma_{\epsilon'}^2 \approx \sigma_x^2 \cdot \frac{M \cdot N \cdot 2^{-2(n_c-1)}}{12 \cdot \alpha_h^2} \tag{34}$$

If we fill in

- the values $\sigma_x^2 = \frac{1}{2}$ (sine-wave with amplitude 1),
- $M = 16$ (number of sub-bands),
- $N = 8$ (number of taps per polyphase FIR filter),
- $n_c = 10$ (number of filter coefficient quantization bits),
- $\alpha_h = 8$ (factor by which FIR coefficients were scaled before quantization, and by which the output is divided to compensate for this scaling)

we find the value

$$\sigma_{\epsilon'}^2 \approx 3.178 \times 10^{-7} \equiv -65.0 \text{ dB} \tag{35}$$

If we compute the average over all values of the output error from fig. 12, we find the value $2.84 \times 10^{-7} \equiv -65.4$ dB which matches remarkably well with the above approximation.

5.2.3 Polyphase FIR filter signal quantization errors at filter bank output

In figure 13 only the signals in the FIR filters are quantized, whereas no quantization takes place in the reference system. In this way, only the quantization errors as a result of the signal quantizations in the FIR filters occur in the output.

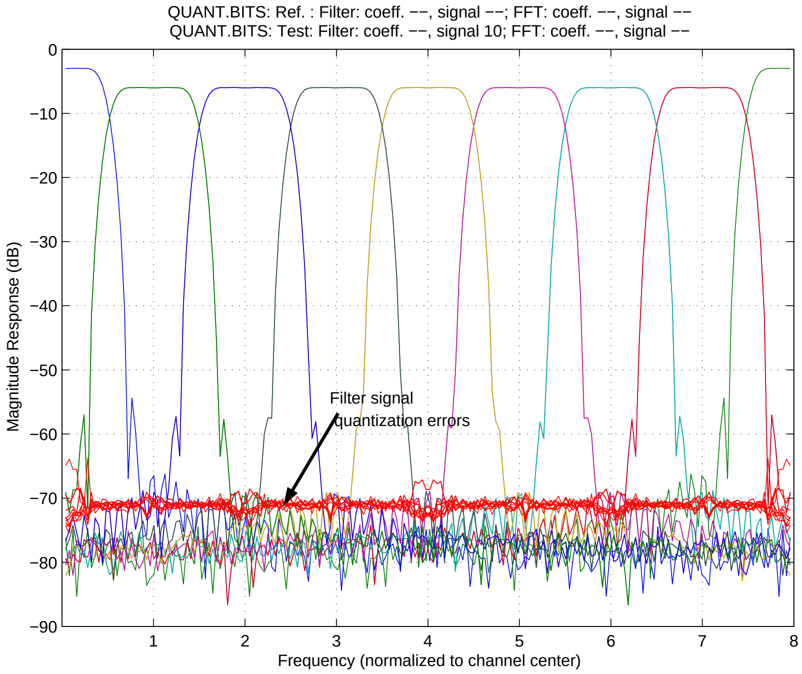


Figure 13: **The effect of filter signal quantization.**
Output signal and output quantization error power spectra, obtained by a sine-wave sweep. In the reference system, no quantization takes place.In the test system, only signals in the filters are quantized.

An alternative approach was used in figure 14. In the test system the filter coefficients and the filter signals were quantized, and the FFT coefficients and signals are unquantized. In the reference system only the FIR filter coefficients are quantized, so that they are identical to their test system counterparts. Therefore the only quantization error showing up in the plot is that of the signal quantization in the FIR filters.

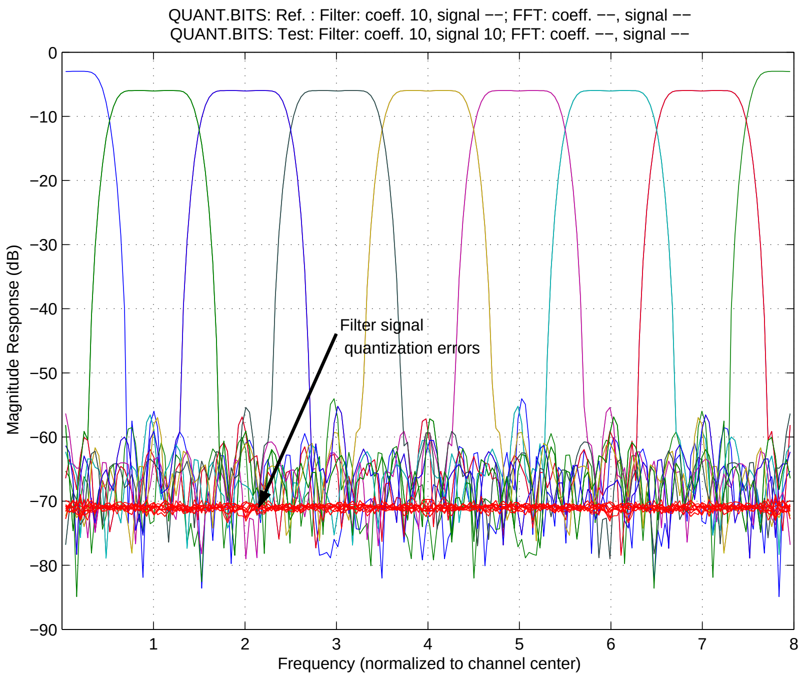


Figure 14: **The effect of filter signal quantization, alternative approach.**
Output signal and output quantization error power spectra, obtained by a sine-wave sweep. In the reference system, filter coefficients are quantized, and signals are not quantized; FFT twiddle factors and signals are unquantized. In the test system, filter coefficients and signals are quantized, in the FFT twiddle factors and signals are unquantized.

The polyphase FIR filter signal quantization results in quantization noise at the output of the filters, which is then passed on by the FFT and the scaling. The FIR filter output quantization noise can be accurately approximated by equation (11). In these simulations, where the output of the polyphase FIRs passes through the FFT and scaling before reaching the output, the quantization noise power as a result of the polyphase FIR signal quantization is seen at the output scaled with a constant factor of $\frac{M}{\alpha_h^2}$, where α_h is the scale factor used before the quantization of the polyphase FIR coefficients. This can be explained as follows. The forward FFT increases the (quantization) noise *power* by a factor M. After the FFT, all output signals (amplitude) are scaled by a factor $1/\alpha_h$ to compensate for the up-scaling by α_h of the impulse response of the polyphase FIR filter coefficients. Hence, the polyphase FIR signal quantization noise appears at the filter-bank's output with a power given as

$$\sigma_{qpf}^2 \approx \frac{M}{\alpha_h^2} \cdot \frac{1}{12} \cdot 2^{-2(n_s-1)} \tag{36}$$

In this particular case $M = 16$ and $\alpha = 8$, and $n_s = 10$, so the approximation becomes:

$$\sigma_{qpf}^2 \approx \frac{1}{4} \cdot \frac{1}{12} \cdot 2^{-2(n_s-1)} = \frac{1}{4} \cdot \frac{1}{12} \cdot 2^{-18} \approx 7.9473 \times 10^{-8} \equiv -71.0 \text{ dB} \tag{37}$$

This is exactly the quantization noise level that we observe in the plot.

5.2.4 FFT coefficient quantization errors at filter bank output

In figure 15 the effect of the FFT coefficients (twiddle factors) quantization is shown by switching off all the reference system quantizations and switching on only the FFT twiddle factor quantizations in the test system.

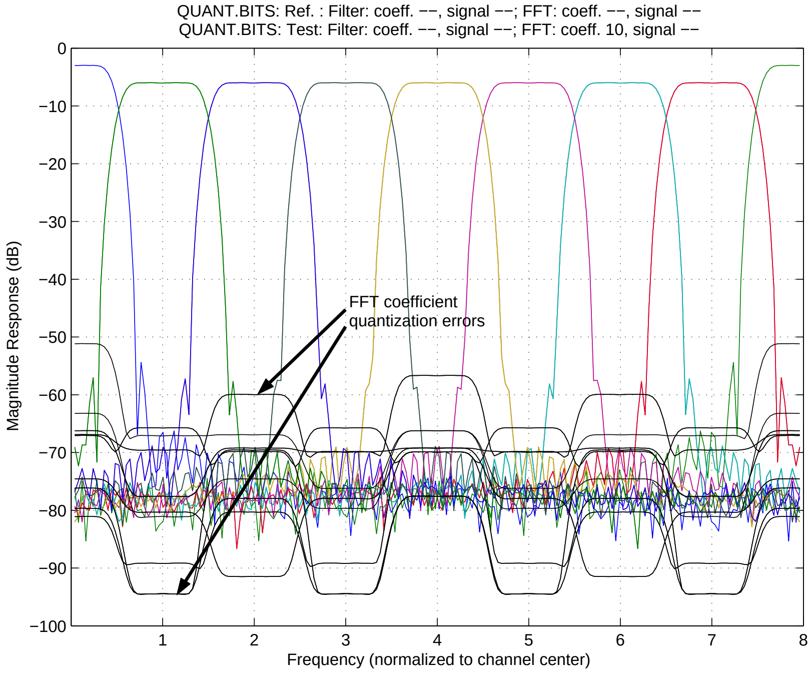


Figure 15: **The effect of FFT twiddle factor quantization.**
Output signal and output quantization error power spectra, obtained by a sine-wave sweep. In the reference system, all quantizations are switched off. In the test system, only the FFT twiddle factors are quantized.

In an alternative approach, the FIR filter coefficients and signals, as well as the FFT coefficients are quantized in the test system. In the reference system only the FIR filter coefficients and signals are quantized. The quantization error that shows up in the plot is now that of the FFT coefficient quantization. This is shown in fig. 16. The difference with fig. 15 lies in the fact that the input signals to both the reference and test system FFT now come from a quantized set of polyphase filters, which is a more realistic situation.
However, the quantization error results differ by no more than 0.5 dB.

From these plots it can be seen that the FFT coefficient quantization error depends strongly on the sub-band the input signal frequency is in; it also depends strongly on the output sub-band. The variations amount up to 45 dB.

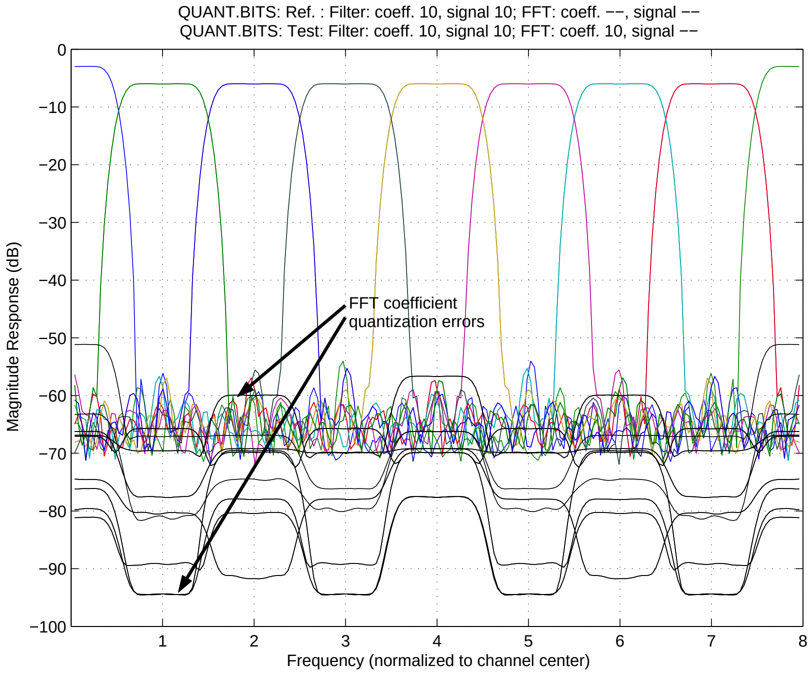


Figure 16: **The effect of FFT twiddle factor quantization.**
Output signal and output quantization error power spectra, obtained by a sine-wave sweep. In the reference system, filter coefficients and signals, are quantized; in the FFT, twiddle factors and signals are unquantized. In the test system, filter coefficients and signals are quantized, in the FFT twiddle factors are quantized and signals are unquantized.

5.2.5 FFT signal quantization errors at filter bank output

The error component resulting from the intermediate signal quantizations in the FFT is shown in figure 17. Here only the signal quantization in the FFT is switched on in the test system, and all quantizations are switched off in the reference system.

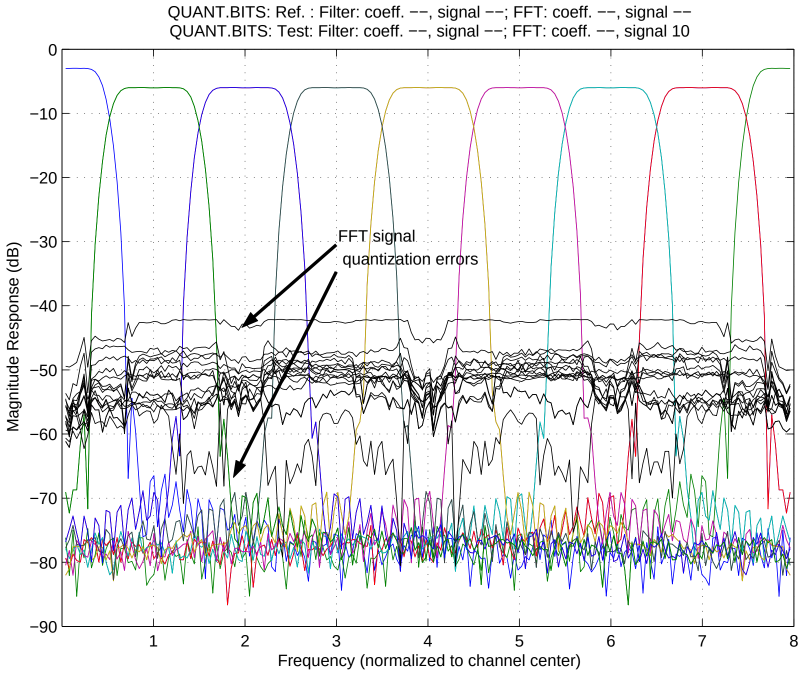


Figure 17: **The effect of FFT signal quantization.**
Output signal and output quantization error power spectra, obtained by a sine-wave sweep. In the reference system all quantizations are switched off. In the test system, only the FFT signals are quantized

It can be seen from this plot that the FFT signal quantization noise depends rather strongly on the output sub-band. The dependence on the input frequency is smaller than in the FFT coefficient quantization case. Further it can be seen that the FFT signal quantization error is an order of magnitude larger than the FFT coefficient quantization error.

The alternative approach is shown in figure 18. Here all quantizations were switched on in the test system. In the reference system all quantizations except the FFT signal quantizers were switched on. The error signal appearing here is the signal quantization error in the FFT. As can be seen, this approach results in a smaller dependence on the output sub-band.

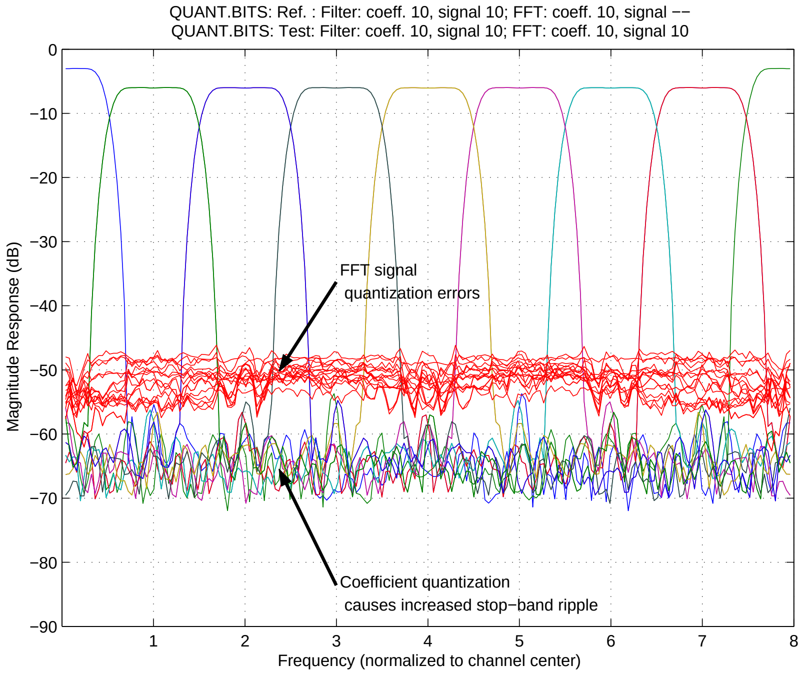


Figure 18: **The effect of FFT signal quantization.**
Output signal and output quantization error power spectra, obtained by a sine-wave sweep. In the reference system, filter coefficients, FFT twiddle factors and signals are unquantized. In the test system, filter coefficients and signals are quantized, in the FFT signals are unquantized

5.2.6 A first evaluation

A quick glance at the plots reveals that the signal quantization in the FFTs has the largest impact on the output signals. Compared to these quantization errors, the contributions of the other quantizations seem almost negligible.

Figure 19 gives a plot of the distributions of the various quantization errors. Here too it seems obvious that improving the overall accuracy will be most successful when concentrating on the FFT signal quantizations.

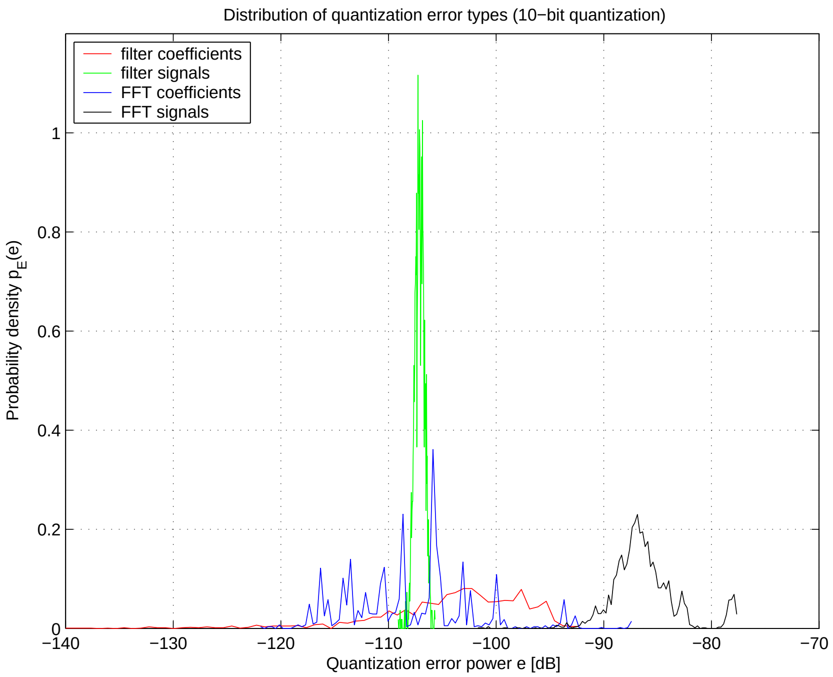


Figure 19: **Distribution of the 4 types of quantization errors, over all input frequencies and all output frequency bands.**

This is shown in fig. 20, where the total output quantization error is shown with the accuracy of the quantizers in the FFT raised to 14 bits (both coefficient and signal quantizations).
Comparison with fig. 11 shows an improvement of roughly 12 dB.

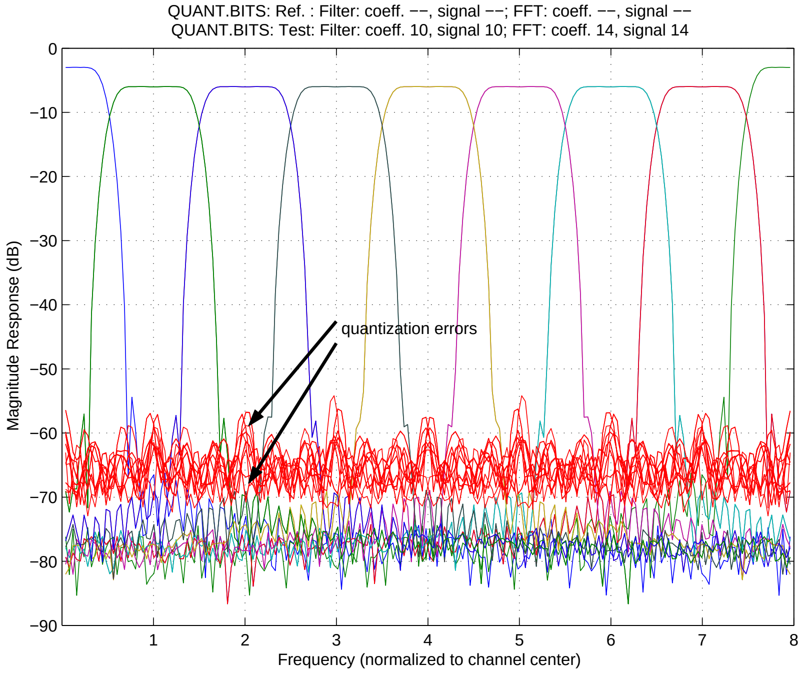


Figure 20: **The total effect of coefficient and signal quantizations, with 14 bits FFT quantization** Output signal and output quantization error power spectra, obtained by a sine-wave sweep. In the reference system all quantizations are switched off. In the test system, all quantizations are switched on. Quantizations in the filters have 10-bit accuracy, and those in the FFT have 14 bits accuracy.

5.3 Balancing the polyphase FIR vs. FFT quantizations

Considering the quantization of a chain of sub-systems, one might want to balance the number of bits used for the processing in each sub-system. If one sub-system in the chain spoils the total accuracy of the chain, it is often recommendable to increase the bit accuracy in that sub-system. In the underlying example, using the same number of bits in the whole filter bank resulted in a dominating effect of the signal re-quantization in the FFT sub-system.

By the same argumentation, one could consider to balance the number of coefficient bits and the number of signal re-quantization bits, in a filter or in the FFT. However, rougher quantization of coefficients result in a systematic error (an error in the linear response, actually), which will not be averaged-out in the correlation process. On the other hand, rougher quantization of the signals will result in higher quantization noise, which, hopefully, will be eliminated in the correlation process and will therefore be less detrimental. Therefore it is advisable to analyse the systematic quantization errors (response deviations) caused by coefficient quantization separately from the quantization noise caused by re-quantization of intermediate result signals.

In the following experiment, however, we treat the quantization errors in a sub-system (polyphase FIRs, FFT) as a whole. A simulation experiment was carried out where the FIR filter quantization accuracy and the FFT quantization accuracy were varied as separate parameters. The accurarcy of the coefficient quantization equals that of the signal quantization, in both sub-systems. For a certain number n_{pf} of FIR quantization bits, and a number n_{fft} of FFT quantization bits, a simulation of the test filter bank (quantized in n_{pf} and n_{fft} bits, respectively) was run against the unquantized reference filter bank. The input was a single-frequency sine-wave, which was applied to the test and reference filter banks simultaneously. The total output quantization error was measured for various values of n_{pf} and n_{fft}.

The results are shown in figure 21. Each curve in this figure represents the result for a certain value of n_{pf} (number of FIR bits) and for a range of values of n_{fft} (number of FFT bits).

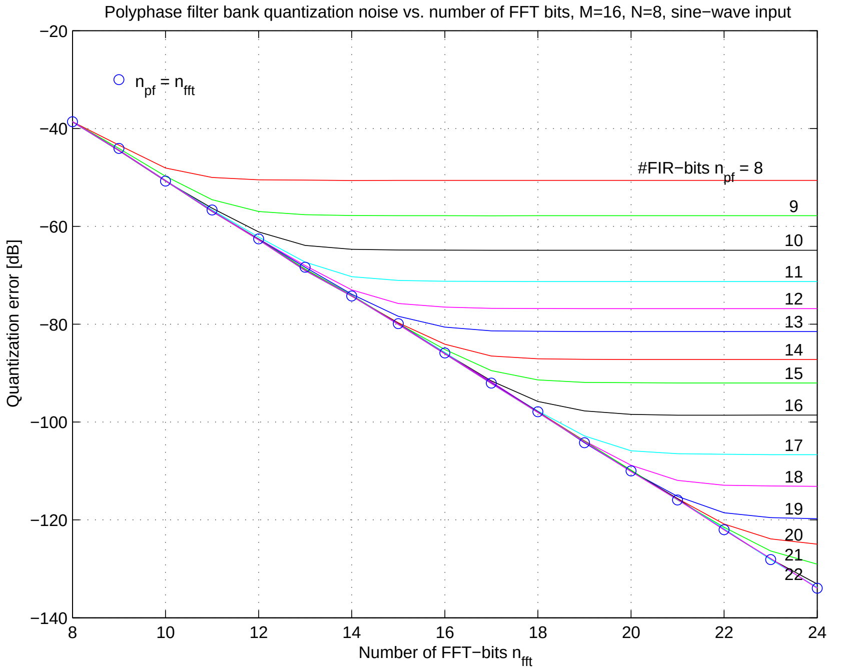


Figure 21: **Balancing the FFT quantizations against the polyphase FIR filter quantizations**

It can be seen from this plot that the number of FFT bits should be about 4 higher than the number of polyphase FIR bits in order to achieve the asymtotic minimum quantization noise. More exact computations have shown that 4 extra bits suffice to approximate the asymptotic optimum within 0.3 dB; to achieve the asymptotic value within 0.1 dB requires 5 extra bits.

One should keep in mind that these results only apply for one frequency, and that the quantitative results are probably different for other frequencies, or other kinds of input signals like noise.

5.4 A 256-band polyphase filter-bank

The above experiment was repeated with a polyphase filter-bank whith 256 sub-bands (with $M = 256$ and $N = 16$). The computational complexity of this filter-bank makes simulation times of a full-range sine-wave sweep almost unacceptably long. Therefore this experiment with a single frequency sine-wave input signal was carried out with varying number of FIR bits n_{pf} and number of FFT-bits n_{fft}. The results were plotted in figure 22.

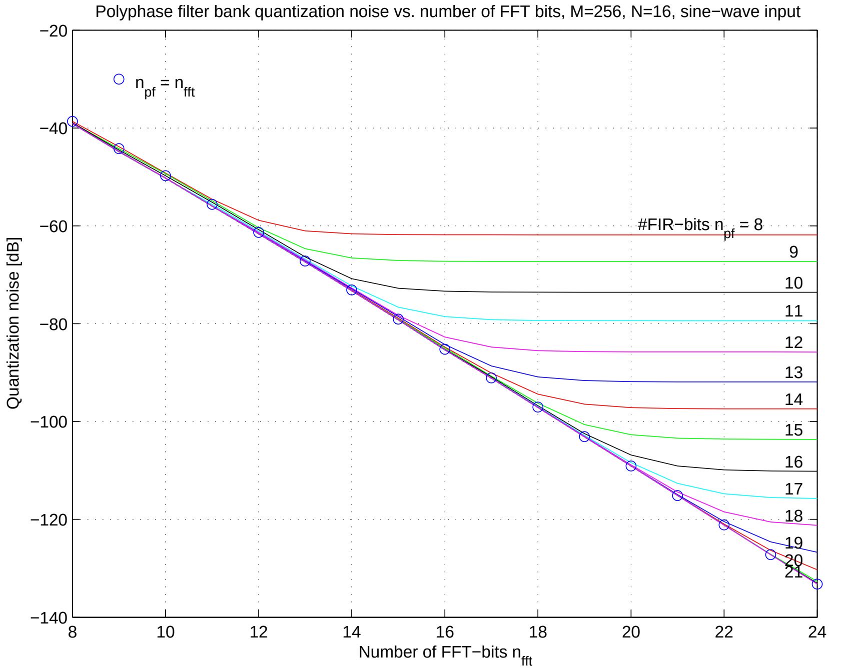


Figure 22: **Balancing the FFT quantizations against the polyphase FIR filter quantizations**

Comparing the quantization error levels at high n_{fft} values (the horizontal part of the curves) with those in figure 21, it can be seen that the quantization error for the same number of FIR-bits n_{pf} is about 12 dB lower for the 256-band filter-bank than for the 16-band filter-bank (when the number of FFT-bits is sufficiently large). In terms of numbers of bits: the required number of FIR-bits is 2 less in the 256-band filter-bank than in the 16-band filter-bank.

In this case the number of bits required for the FFT to attain the asymptotic minimum is about 6 to 7 bits. More accurate computations show that 6 extra bits are needed to approximate the asymptotic optimum within 0.3 dB; to achieve the asymptotic value within 0.1 dB requires 7 extra bits.

6 Conclusions

A study was made of the quantization effects in polyphase/DFT filter banks. The separate effects were analyzed of:

- coefficient quantization in the polyphase FIR filters;
- signal quantization in the polyphase FIR filters;
- coefficient (twiddle factor) quantization in the FFT;
- signal quantization in the FFT.

The quantization error resulting from coefficient quantization in the polyphase FIR filters can be approximated as given in eq. (34), repeated here:

$$\sigma_{\epsilon'}^2 \approx \sigma_x^2 \cdot \frac{M \cdot N \cdot 2^{-2(n_c-1)}}{12 \cdot \alpha_h^2} \tag{38}$$

where

- σ_x^2 is the input signal power,
- M is the number of sub-bands,
- N is the number of taps per polyphase FIR filter,
- n_c is the number of filter coefficient quantization bits,
- α_h is the factor by which FIR coefficients were scaled before quantization, and by which the output is divided to compensate for this scaling.

It was found that polyphase FIR signal quantization noise power at the filter-bank's output can be approximated by eq. (36) :

$$\sigma_{qpf}^2 \approx \frac{M}{\alpha_h^2} \cdot \frac{1}{12} \cdot 2^{-2(n_s-1)} \tag{39}$$

Simulations with equal word-lengths for signals and coefficients throughout the filter-bank structure, have shown that the most critical part is the signal quantization in the FFT. It is therefore recommended that the signal word-length in the FFT should be some bits higher than in the rest of the word-lengths in the filter-bank.