PART 3 Special Topics

9

Quantization effects

9.0 INTRODUCTION

In any digital filter bank implementation, the multipliers as well as internal signals have to be represented in quantized form. The effect of this quantization is that the filter output is different from the ideal one. Broadly speaking, we can classify the quantization effects into three categories, viz., coefficient sensitivity effects, roundoff noise and limit cycles. In this chapter we will analyze these effects quantitatively. In Appendix C we will deal with another important effect in filter bank systems, arising due to quantization of subband signals.

9.1 TYPES OF QUANTIZATION EFFECTS

Consider Fig. 9.1-1(a) which shows the implementation of a first order digital filter with transfer function $H(z) = 1/(1 - az^{-1})$. If the signal $y_i(n-1)$ and multiplier a are represented with a certain precision, then the product $ay_i(n-1)$ in general requires a higher precision. So the signal $y_i(n)$ requires higher precision than $y_i(n-1)$. Since $y_i(n)$ is circulated back during the next cycle, this process continues indefinitely, implying infinite bit accumulation.

In a practical system this is not feasible, and the result of a computation has to be quantized before recirculation. This is indicated schematically as shown in Fig. 9.1-1(b), where the box labeled Q represents a quantizer. The signal $y(n)$ which is recirculated is the quantized version of an intermediate signal $w(n)$, and we write $y(n) = Q[w(n)]$. In general there could be more than one quantizer in a system, but in order to avoid infinite bit accumulation it is *sufficient* to make sure that there are no *loops* without quantizers.

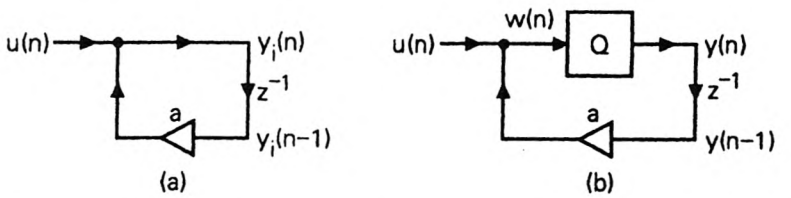

Figure 9.1-1 (a) A first order filter, and (b) implementation with a quantizer in the loop.

Effects of Multiplier (or Coefficient) Quantization

Quantization of the multiplier coefficients, for example, a in the above figure, results in a change of the transfer function from $H(z)$ to $H_q(z)$. Thus, the passband and stopband ripples after quantization can be significantly larger than the specified values. As an extreme case a stable filter may become unstable after coefficient quantization.

In a filter bank system, coefficient quantization can result in deeper consequences. For example, a QMF bank may lose the alias-free property, or perfect reconstruction (PR) property, because of multiplier quantization. It turns out, however, that in any perfect reconstruction system with paraunitary polyphase matrix, the paraunitary property (and hence the PR property) can be retained in spite of multiplier quantization. (In this sense the structure is 'robust' to quantization.) This will be justified only in Sec. 14.11 where we show how the paraunitary property of an $M \times M$ matrix can be retained in spite of coefficient quantization. A special case of this has already been noticed in Sec. 6.4.1 (two channel QMF lattice). In Sec. 5.3.5 we also studied a two channel IIR QMF bank which is free from aliasing as well as amplitude distortion in spite of coefficient quantization.

Effects of Signal Quantization

The effect of quantizing internal signals is more involved. Consider again Fig. 9.1-1(b). The quantity

$$q(n) = Q[w(n)] - w(n) \tag{9.1.1}$$

is called the *quantizer error,* and is a function of time n. We can model the structure as shown in Fig. 9.1-2. We say that $q(n)$ is the *noise source* associated with the quantizer. Notice that the output $y(n)$ in Fig. 9.1-2 is different from the ideal output $y_i(n)$ in Fig. 9.1-1(a). The difference $y(n) - y_i(n)$ is *not* equal to the quantizer error $q(n)$. This is because the effect of quantizer accumulates with time as explained below.

We can think of $q(n)$ as an input to the filter (just like $u(n)$ is). Its effect on the output is governed by the transfer function between $q(n)$ and

$y(n)$, called the *noise transfer function.* This is given by

$$G(z) = 1/(1 - az^{-1}), \tag{9.1.2}$$

which, in this example, happens to be the same as the filter transfer function $H(z)$. Let $e(n)$ denote the output of the system $G(z)$, in response to the input $q(n)$. Then the signal $y(n)$ can be written as $y(n) = y_i(n) + e(n)$, where $y_i(n)$ is the output of the ideal system of Fig. 9.1-1(a). So the noise source affects the filter output in a manner which depends on the noise transfer function.

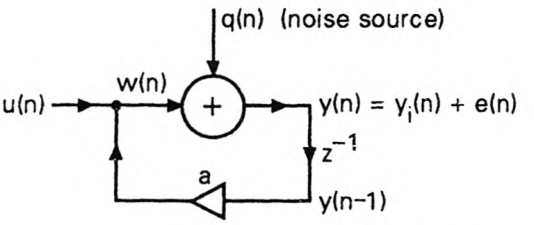

Figure 9.1-2 The roundoff noise model for the structure of Fig. 9.1-1(b).

In order to understand the effect of $q(n)$ on the filter output more quantitatively, it is common practice to model $q(n)$ as a random process (Appendix B), satisfying a set of simplifying assumptions. This allows us to estimate the variance of the noise which actually reaches the filter output, using simple and elegant techniques. The main point to note here is that the effect of signal quantization is to contribute a random component $e(n)$ to the filter output. Under some conditions, quantization also results in nonrandom components, called limit cycles. We will return to this in Sec. 9.6.

Subband Quantization

In subband coding applications, a third source of noise exists, due to quantization of the subband signals. This tends to dominate the total noise whenever it is present, but its effect is difficult to analyze. We will study this in Appendix C. In this chapter we will concentrate only on quantization noise due to filter implementation. Such a study is useful in many applications, for example, in voice privacy systems (Sec. 4.5.3) and transmultiplexers, where subband quantization effects do not dominate.

Chapter Outline

In Sec. 9.2 we present a brief summary of well known techniques for roundoff noise analysis. In Sec. 9.3–9.5 we use this to present a roundoff noise analysis for multirate filter banks. In Sec. 9.6 we consider limit cycles. We return to coefficient quantization effects in Sec. 9.7. It will be seen that many filter bank structures exhibit low passband sensitivity to coefficient quantization, particularly if the polyphase matrix $\mathbf{E}(z)$ is paraunitary.

Special prerequisites. We review the standard noise analysis techniques in Sec. 9.2. It is, however, assumed that the reader has some familiarity with fixed-point binary number representation, and random process representation of noise waveforms. Thus, we make free use of such terms as (a) b-bit fixed point arithmetic, (b) uncorrelated white noise source, (c) noise source with variance σ_q^2, and so on. There exist excellent treatements of this material [Oppenheim and Schafer, 1989], [Jackson, 1989], and [Rabiner and Gold, 1975]. Appendix B includes a brief review of random process, and we will freely use the definitions and properties in that appendix (e.g., uniform random variables, wide sense stationary random process, autocorrelation, power spectrum, white noise, and so on).

9.2 REVIEW OF STANDARD TECHNIQUES

9.2.1 Quantizers and Noise Models

All signals are represented as fixed point binary fractions, as shown in Fig. 9.2-1. This is a b-bit binary representation, with s representing the sign bit. We say that b is the *wordlength.* If $s = 0$ the number is nonnegative, whereas with $s = 1$ the number is nonpositive, and its decimal value depends on the convention chosen (e.g., two's complement convention, sign magnitude convention, etc.). All the numbers representable in this form are in the range $-1 \leq x < 1$ (with $x = -1$ permitted only in some conventions, e.g., two's complement). This is said to be the permissible *dynamic range.* The quantity

$$\Delta \triangleq 2^{-b}, \tag{9.2.1}$$

is the smallest positive number permitted, and is also the smallest possible increment. It is said to be the *quantization step* or stepsize.

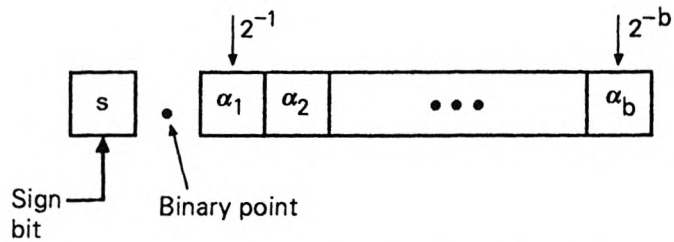

Figure 9.2-1 Format for the b-bit fixed-point binary fraction.

Quantizers

A quantizer is a device which takes an arbitrary real number and converts it into a b-bit fraction using some arithmetic rules. Thus, the quantizer input [e.g., $w(n)$ in Fig. 9.1-1(b)] need not be a b-bit fraction, but its output is. In this process, some error is introduced, and is denoted as $q(n)$

[see (9.1.1) and Fig. 9.1-2]. We say that $q(n)$ is the noise source due to the quantizer. If $w(n)$ does not belong to the permitted dynamic range, we say that a computational overflow has occured. The quantizer brings the number back to the permitted dynamic range by using certain rules (called *overflow handling rules*). So $y(n)$ still belongs to the dynamic range, but the error $q(n)$ is large.

Assume that there is no overflow, that is, $w(n)$ belongs to the dynamic range. In general $w(n)$ may still have more than b bits in its representation, that is, there could be some extra bits to the right of the bth bit in Fig. 9.2-1. When this is converted to a b bit number, the error $q(n)$ is 'small' and is of the order of the quantization step Δ. The exact details depend on the type of quantizer, that is, the rule used for quantization. Some rules are: (a) roundoff arithmetic where $y(n)$ is the quantized number closest to $w(n)$, (b) magnitude truncation, where the magnitude of quantized number $y(n)$ is no larger than that of the unquantized number $w(n)$, and (c) truncation arithmetic, where the extra bits to the right of the b bits are merely discarded.

The Noise Model Assumptions

Unless mentioned otherwise, we will assume roundoff arithmetic. In this case, we have

$$\frac{-\Delta}{2} \leq q(n) \leq \frac{\Delta}{2}. \tag{9.2.2}$$

We will assume that $q(n)$ is a random variable, uniformly distributed in the above range. Under this condition, it has zero mean and variance

$$\sigma_q^2 = \frac{2^{-2b}}{12}. \tag{9.2.3}$$

We make the further assumption that the sequence $q(n)$ is a white, wide sense stationary (WSS) random process. Summarizing, the quantizer noise source $q(n)$ is zero-mean white, with variance σ_q^2.

Multiple noise sources. In most practical structures, there are many quantizers. Fig. 9.2-2 shows the example of a cascade form structure (Sec. 2.1.3) with two quantizers. In such situations, each quantizer is replaced with a noise source, as shown by broken lines. We assume that each noise source satisfies the above model (i.e., white, etc.). To study the total effect of these at the filter output, we assume that any two noise sources are uncorrelated, and that each of them in turn is uncorrelated to the input $u(n)$. These assumptions will enable us to add the noise variances due to various sources, in order to obtain the total output noise variance.

Examples which violate these assumptions are not hard to generate (e.g., when the filter input is a sinusoid). However, in a large number of situations, the above assumptions have been verified to be reasonable [Barnes, et al., 1985]. In any case, noise analysis under these assumptions gives a very good qualitative idea of the nature of noise propagation. For example, one of the useful conclusions obtainable is that, in a direct form structure, the

output noise variace increases as the poles get closer to the unit circle [see discussions following (9.2.6) later].

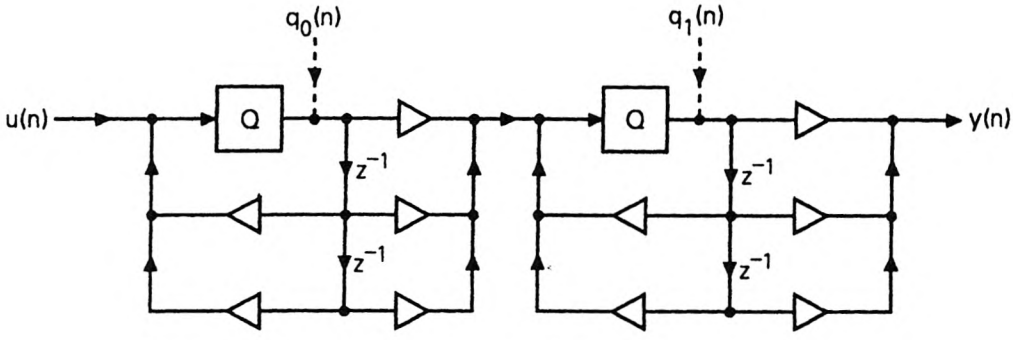

Figure 9.2-2 A cascade form structure, with two quantizers.

9.2.2 Noise Gain of a Filter

In Fig. 9.1-2 we associated a noise transfer function $G(z)$ with the noise source $q(n)$. This transfer function governs the extent to which $q(n)$ affects the output. More generally let there be many quantizers in the structure, each modeled by a noise source $q_k(n)$. Let $G_k(z)$ denote the transfer function from the noise source $q_k(n)$ to the filter output. We say that $G_k(z)$ is the noise transfer function for $q_k(n)$.

Let $e_k(n)$ denote the output of $G_k(z)$ in response to the input $q_k(n)$. Under the above noise model assumptions, $e_k(n)$ is a zero mean WSS random process with variance

$$\sigma_{e_k}^2 = \sigma_q^2 \underbrace{\sum_n |g_k(n)|^2}_{\text{noise gain}}, \tag{9.2.4}$$

where $g_k(n)$ is the impulse response of $G_k(z)$. The summation in (9.2.4) is the energy of $G_k(z)$. So the output noise variance is the quantizer noise variance amplified by the *energy* of the noise transfer function (which is therefore called the *noise gain*).

Each of the quantizer noise sources $q_k(n)$ contributes a noise component at the filter output. In view of the uncorrelated assumption the output noise $e(n)$ has total variance

$$\begin{aligned} \sigma_e^2 &= \sigma_q^2 \sum_{k=0}^{N-1} \sum_n |g_k(n)|^2 \\ &= \frac{2^{-2b}}{12} \sum_{k=0}^{N-1} \sum_n |g_k(n)|^2 \quad \text{(total output noise variance).} \end{aligned} \tag{9.2.5}$$

Returning to the example of Fig. 9.1-2, the impulse response $g(n)$ of the noise transfer function is $g(n) = a^n \mathcal{U}(n)$, so that the noise gain is

$$\sum_n |g(n)|^2 = \frac{1}{1 - |a|^2}. \tag{9.2.6}$$

This gain increases as the pole 'a' (which is inside the unit circle) gets closer and closer to the unit circle. As an example, if $a = 0.99$ then the noise gain ≈ 50. This demonstrates that the noise gain can be quite large indeed.

Effect of increasing the wordlength. If we increase the number of bits from b to $b+1$, this results in a four fold reduction in the noise variance (using (9.2.5)). On a dB scale, this is equivalent to $10 \log_{10} 4.0 = 6.02$ dB. So the output noise variance decreases by about 6 dB per every additional bit of internal precision.

9.2.3 Dynamic Range and Scaling

In a practical implementation we have to ensure that the signals do not overflow the dynamic range permitted by the number system. In order to study this issue, it is useful to find upper bounds on the magnitudes of various signals. In such an analysis, the presence of quantizers can be ignored, as they do not affect these bounds significantly.

Thus consider Fig. 9.1-1(b), and ignore the quantizer for this discussion. If the input is in the range $-1 \leq u(n) < 1$ (consistent with fixed point fractional representation), this does not imply that the signal $w(n)$ is in this range for all n. It can, however, be shown that $w(n)$ is bounded as

$$|w(n)| \leq \sum_n |f(n)| \tag{9.2.7}$$

where $f(n)$ is the impulse response from $u(n)$ to the node $w(n)$. In our example, this impulse response happens to be the same as $h(n)$, that is, $f(n) = a^n \mathcal{U}(n)$. So the right hand side of (9.2.7) reduces to $1/(1 - |a|)$. For example, if $a = 0.99$, this quantity equals 100. In other words, $|w(n)|$ can get as large as a hundred!

A simple way to ensure that $w(n)$ does not overflow (i.e., does not exceed the range $-1 \leq w(n) < 1$), is to insert a multiplier $1/L$ as shown in Fig. 9.2-3, with

$$L = \sum_n |f(n)|. \tag{9.2.8}$$

We then say that the structure has been scaled. The price we pay for this freedom from overflow is that the output *signal level* goes down. Since the roundoff noise level is unaffected by scaling, the *signal to noise ratio* is reduced. This is an example of interaction between roundoff noise and dynamic range in digital filter implementations.

It is in principle possible to reduce the roundoff noise level simply by inserting a scale factor at the output node, but this results in reduced signal level as well. If we try to restore the signal level by insertion of another multiplier at the filter input, this will affect the probability of internal overflow. So, 'finite word length' will have its effect one way or the other.

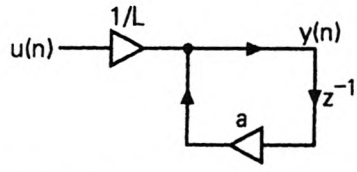

Figure 9.2-3 Scaling a first order digital filter.

Scaling a Structure

In practice, digital filter structures are more complicated than Fig. 9.1-1. There are several internal nodes, and one has to ensure that none of these suffers from computational overflow. Let $F_k(z)$ be the transfer function from the filter input to the kth internal node, and let $f_k(n)$ be its impulse response. We say that $F_k(z)$ is the scaling transfer function for the kth node. If

$$\sum_n |f_k(n)| < 1, \tag{9.2.9}$$

then the kth node [or the transfer function $F_k(z)$] is scaled to be free from overflow. If all nodes satisfy this property, then the entire structure is said to be scaled. Scaling can be accomplished by rearrangement of the internal structure, which may or may not involve explicit insertion of multipliers (as in Fig. 9.2-3).

Which nodes to scale? With certain types of arithmetic conventions (e.g., two's complement), it can be shown [Jackson, 1970] that only those nodes which are *inputs to multipliers* have to be scaled. For example, consider Fig. 9.2-4. Here every multiplier input is a delayed version of the signal $s(n)$. So it is sufficient to scale this node and, of course, the output node $y(n)$. Even if there is an overflow at any of the other nodes, it will not affect the final output $y(n)$. (This has to do with the fact that two's complement arithmetic has similarities to modulo arithmetic).

Types of Scaling

If each of the scaling transfer functions $F_k(z)$ satisfies (9.2.9), we say that the strcuture is *sum-scaled.* The structure is completely free from overflow, but the price paid is in terms of the signal to roundoff noise ratio at the filter output. There exist less stringent scaling rules (called $\mathcal{L}_p$ *scaling rules*) which are sufficient under some conditions. We will not go into these details (which can be found in [Jackson, 1970] and [Rabiner and Gold, 1975]) but

merely mention a scheme called the $\mathcal{L}_2$ *scaling policy.* In Problem 9.4 we cover some of the other scaling rules.

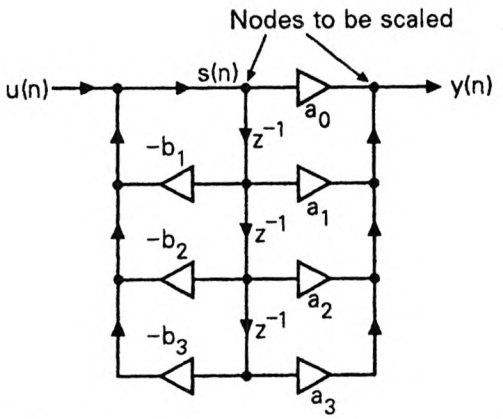

Figure 9.2-4 Pertaining to the choice of nodes to be scaled.

$\mathcal{L}_2$ scaling. In this scheme, instead of ensuring the condition (9.2.9), we ensure that

$$\sum_n |f_k(n)|^2 \leq 1, \tag{9.2.10}$$

(usually with equality), where $f_k(n)$ is the impulse response from the input to the kth node to be scaled. This is called $\mathcal{L}_2$ scaling because the above summation is the (square of) the $\mathcal{L}_2$ norm of $F_k(z)$. † If a node $x_k(n)$ is scaled (i.e., $F_k(z)$ is scaled) in the $\mathcal{L}_2$ sense, then it can be shown that $|x_k(n)| < 1$ as long as the filter input $u(n)$ has energy bounded by unity, that is,

$$\sum_n |u(n)|^2 < 1. \tag{9.2.11}$$

Use of $\mathcal{L}_2$ scaling. $\mathcal{L}_2$ scaling is less stringent than sum-scaling (which guarantees complete freedom from overflow), and therefore results in increased signal to roundoff noise ratio in absence of overflow. However, the chances of overflow are higher; note that the condition (9.2.11) is rather unrealistic. (For example if $u(n)$ is a sinusoid, its energy is infinite.) However, $\mathcal{L}_2$ scaling is still useful for several reasons.

First, in most practical cases, the sequence $f_k(n)$ is significant only over a finite duration. If the energy of $u(n)$ over such a duration is properly bounded, we can still control the possibility of overflow of $x_k(n)$.

Second, if we view the input as a wide sense stationary random process (which is sometimes a realistic assumption, at least over short segments of

† For integer p the $\mathcal{L}_p$ norm of $F(z)$ is defined as $[\int_0^{2\pi} |F(e^{j\omega})|^p \frac{d\omega}{2\pi}]^{1/p}$.

time), we can obtain some useful conclusions. In this case, the energy of $u(n)$ is not finite, but only the power spectrum of $u(n)$ is of relevance. It is possible to obtain a bound on the variance of $x_k(n)$ as follows:

$$\sigma_{x_k}^2 \leq \sum_n |f_k(n)|^2 \times \max_\omega \Big(S_{uu}(e^{j\omega}) \Big).$$

This variance can in turn be used to bound the *probability* of overflow at the node $x_k(n)$ [Jackson, 1970]. If all the scaling transfer functions satisfy (9.2.10) with equality, then we can reduce the probability of overflow at all the internal nodes to the *same value*, simply by inserting a common scale factor (as we did in Fig. 9.2-3), in front of the input. For the rest of the chapter, we consider only $\mathcal{L}_2$ scaling.

9.2.4 Some Useful Special Cases

FIR Direct Form Structures

Many of the filter banks we studied are FIR systems, for which noise analysis is fairly simple. Consider the FIR direct form structure shown in Fig. 9.2-5(a). The transfer function is $H(z) = \sum_{n=0}^{N} h(n) z^{-n}$.

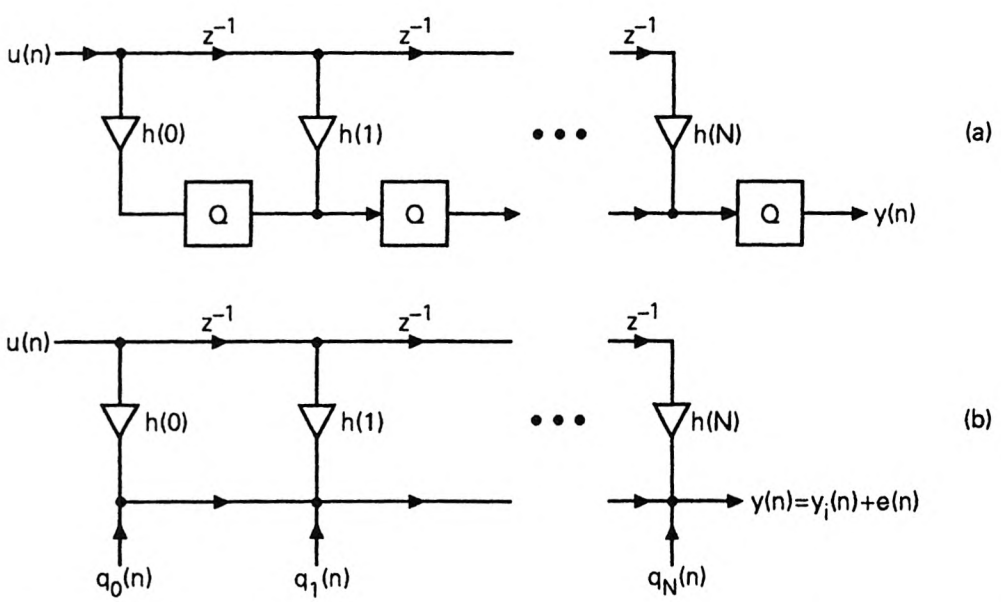

Figure 9.2-5 (a) The direct form FIR structure with quantizers, and (b) the noise model.

Here the output of every multiply/add operation is quantized, and the noise model is shown in Fig. 9.2-5(b). All noise sources $q_k(n)$ have the same noise transfer function, that is, $G_k(z) = 1$ for all k. Under the usual (white,

uncorrelated) assumptions, the output noise variance is thus $(N+1)\sigma_q^2$ where σ_q^2 is the quantizer noise variance (9.2.3). For the case of linear phase FIR filters where we require only half as many multiplications (e.g., see Fig. 2.4-3), the output noise variance is approximately half the above value.

Since $G_k(z) = 1$ and since $q_k(n)$ are white as well as uncorrelated, the output noise $e(n)$ is also white! This is true regardless of the transfer function $H(z)$ (which does not affect the noise transfer function). This is a somewhat unusual situation, which is not common with IIR filters. For example, in Fig. 9.1-2 the noise transfer function $G(z)$ is not constant, and the output noise is not white.

A second quantization scheme for the FIR case would be to carry higher internal precision and quantize only the output $y(n)$. This scheme has less output noise variance ($= \sigma_q^2$ only), at the expense of higher internal wordlength. The extra internal wordlength depends on the number of multipliers as well as the multiplier precisions, and complicates things in general. We do not consider it here.

Scaling. Since the inputs to multipliers are derived by delaying $u(n)$, these are already scaled. The only extra scaling necessary is to ensure that the output $y(n)$ does not overflow. This can be done by insertion of a scale factor $1/L$ as we did in Fig. 9.2-3. The value of L depends on the scaling policy chosen.

Allpass Cascade form

Consider Fig. 9.2-6(a) which represents a cascade of L first order filters $H_k(z)$, each implemented in direct form. Assuming that α_k are real, $H_k(z)$ are allpass so that the overall filter $H(z)$ is allpass (with real poles only). Such real-pole allpass functions find application in power symmetric IIR QMF banks, as seen in Sec. 5.3. In that section, a two channel IIR QMF bank was introduced with analysis filters

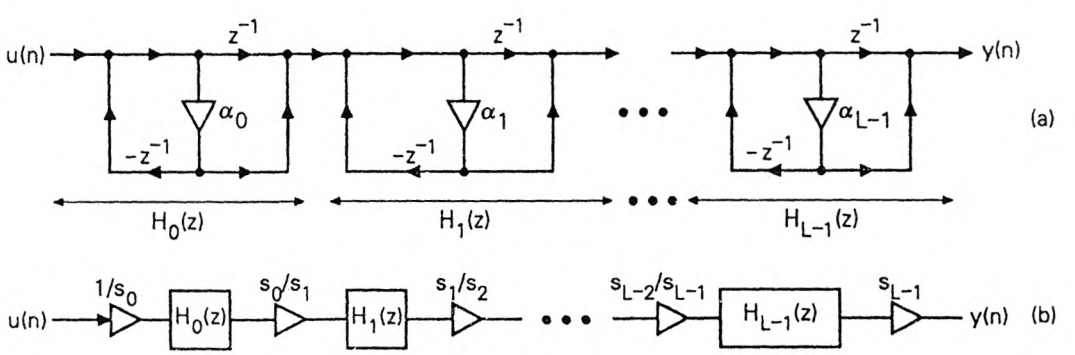

Figure 9.2-6 (a) A cascade of L first order allpass filters, and (b) insertion of scaling multipliers.

$$H_0(z) = \frac{a_0(z^2) + z^{-1}a_1(z^2)}{2}, \quad H_1(z) = H_0(-z).$$

The allpass functions $a_i(z)$ have only real poles, and can be implemented using the above cascade form.

Scaling. In this structure, the only nodes to be scaled are the inputs to the multipliers α_k. If we wish to scale these nodes in the $\mathcal{L}_2$ sense, then we define $s_k = 1/\sqrt{(1-\alpha_k^2)}$ and insert $1/s_k$ at the input of the kth section. We also insert s_k at the output of the section to ensure that $H_k(z)$ is unaffected. Simplifying, we obtain the scaled structure of Fig. 9.2-6(b).

Noise variance. The noise model of the scaled structure is shown in Fig. 9.2-7, assuming that a quantizer is inserted in each section (exactly as we did in Fig. 9.2-2). The noise transfer function for the noise source $q_k(n)$ is

$$G_k(z) = s_k \prod_{\ell=k}^{L-1} H_\ell(z) \tag{9.2.12}$$

which is allpass. Under the usual noise model assumptions, the output noise component $e(n)$ is therefore zero-mean and white, with total variance

$$\sigma_e^2 = \sigma_q^2 \sum_{k=0}^{L-1} s_k^2. \tag{9.2.13}$$

Complex case. These discussions can be generalized to the case where filter coeffcients and inputs are complex. In this case we have to define a complex quantizer (with b-bit real part and b-bit imaginary part). Under proper assumptions, many of the above results can be extended.

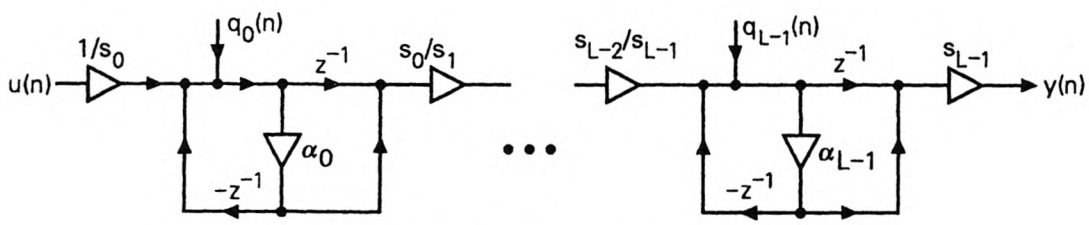

Figure 9.2-7 Noise model for the cascaded allpass structure.

9.3 NOISE TRANSMISSION IN MULTIRATE SYSTEMS

The study of noise generation and propagation in multirate systems is facilitated if we first note a number of useful properties exhibited by random processes in the presence of some familiar building blocks.

Decimators and Expanders

Let $x(n)$ be a wide sense stationary (WSS) random process with mean m and variance $\sigma^2 > 0$. The following properties are easily verified (Problem 9.5).

1. If $x(n)$ is input to an M-fold decimator, then the output $y(n) = x(Mn)$ is also WSS, with mean m and variance σ^2. In fact the autocorrelations of $y(n)$ and $x(n)$ are related as $R_{yy}(k) = R_{xx}(Mk)$ so that the power spectrum $S_{yy}(e^{j\omega})$ of $y(n)$ is related to the power spectrum $S_{xx}(e^{j\omega})$ of the input $x(n)$ in terms of the familiar aliasing relation (4.1.4) (i.e., simply replace $Y_D(e^{j\omega})$ and $X(e^{j\omega})$ in (4.1.4) with $S_{yy}(e^{j\omega})$ and $S_{xx}(e^{j\omega})$ respectively.)
2. If $x(n)$ is input to an M-fold expander ($M > 0$), then the output $y(n)$ is *not* WSS. For example, the random variable $y(0)$ [$= x(0)$] has variance σ^2, whereas $y(1) = 0$ (which is a 'random variable' with variance $= 0$). Since the variance is not constant with time, this rules out wide sense stationarity.

Expander/Delay-Chain Combination

Consider Fig. 9.3-1 where $x_k(n), 0 \leq k \leq M-1$ are WSS random processes. The signal $y(n)$ is an interlaced version of $x_k(n)$. (This is similar to the time-domain multiplexer in Fig. 4.5-4(a)). In general $y(n)$ is not WSS. For example, if $x_0(n)$ and $x_1(n)$ have unequal variances, then $y(n)$ has variance changing with time, and it cannot therefore be WSS.

A special example of interest arises when $x_k(n)$ are zero-mean, white-noise sources with variance σ^2 for all k. Assume further that $x_k(n)$ and $x_m(n)$ are uncorrelated for $k \neq m$. In this case, the output $y(n)$ is zero-mean and white (since $y(n_0)$ and $y(n_1)$ are uncorrelated for $n_0 \neq n_1$) with variance σ^2.

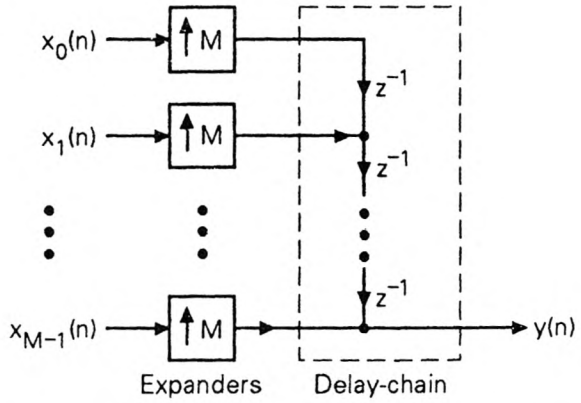

Figure 9.3-1 The time domain multiplexing circuit.

Paraunitary Systems

Some of the filter banks we have studied contain lossless (i.e., stable

paraunitary) building blocks. For example, consider Fig. 9.3-2 which is the polyphase implementation of a synthesis bank (Section 5.5). In many examples $\mathbf{R}(z)$ is paraunitary. We will derive a result which applicable in such situations.

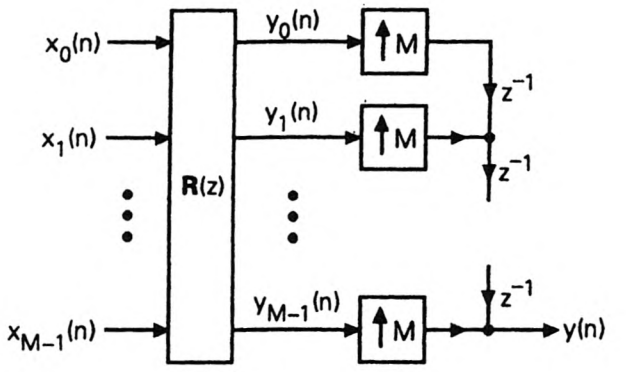

Figure 9.3-2 A synthesis bank in polyphase form.

Let $x_k(n), 0 \leq k \leq M-1$ be WSS random processes. These may, for example, represent the noise generated in the analysis bank. Suppose the following assumptions are true:

1. Each sequence $x_k(n)$ is white.
2. Any two of these sequences are uncorrelated, that is, $x_k(n_0)$ and $x_m(n_1)$ are uncorrelated unless $k = m$ and $n_0 = n_1$.
3. $x_k(n)$ have zero mean.
4. All the M sequences $x_k(n)$ have equal variance, that is, $\sigma_k^2 = \sigma^2$ for all k.

We then say that the vector

$$\mathbf{x}(n) \triangleq [\, x_0(n) \quad \ldots \quad x_{M-1}(n) \,]^T, \tag{9.3.1a}$$

which is a vector-random process (Section B.5, Appendix B), is WUZE(σ^2). This is an abbreviation for *white, uncorrelated, zero-mean,* and *equal* variance σ^2. A zero-mean WSS random (vector) process $\mathbf{x}(n)$ is WUZE(σ^2) if, and only if, $E[\mathbf{x}(m)\mathbf{x}^\dagger(k)] = \sigma^2\delta(m-k)\mathbf{I}$ (Problem 9.6).

Now suppose that $\mathbf{R}(z)$ is stable, and $\widetilde{\mathbf{R}}(z)\mathbf{R}(z) = d\mathbf{I}$ (i.e., lossless). Then the vector

$$\mathbf{y}(n) \triangleq [\, y_0(n) \quad \ldots \quad y_{M-1}(n) \,]^T \tag{9.3.1b}$$

has all the four properties of $\mathbf{x}(n)$. More precisely, $\mathbf{y}(n)$ is WUZE($d\sigma^2$). (This is a consequence of the theorem to be proved below.) In view of this, the output signal $y(n)$ [which is a time multiplexed version of $y_k(n)$] is a white, zero-mean process, with variance $d\sigma^2$. We now state and prove a more general result, which is useful in the study of filter banks.

♠**Theorem 9.3.1.** Let $\mathbf{T}(z)$ be a $p \times r$ transfer matrix and let $\mathbf{T}^T(z)$ be lossless, that is, stable with $\mathbf{T}(z)\widetilde{\mathbf{T}}(z) = d\mathbf{I}_p$. Let $\mathbf{x}(n)$ and $\mathbf{y}(n)$ denote the input and output (vector-)sequences. If $\mathbf{x}(n)$ is WUZE(σ^2), then the output is WUZE($d\sigma^2$). ◊

Proof. Let $\mathbf{S_{xx}}(e^{j\omega})$ and $\mathbf{S_{yy}}(e^{j\omega})$ denote the power spectral density matrices of the vector WSS processes $\mathbf{x}(n)$ and $\mathbf{y}(n)$. We then have

$$\mathbf{S_{yy}}(e^{j\omega}) = \mathbf{T}(e^{j\omega})\mathbf{S_{xx}}(e^{j\omega})\mathbf{T}^\dagger(e^{j\omega}) \tag{9.3.2}$$

In view of the WUZE property of $\mathbf{x}(n)$, its autocorrelation sequence is

$$\mathbf{R_{xx}}(k) = \sigma^2\delta(k)\mathbf{I}_r \tag{9.3.3}$$

so that $\mathbf{S_{xx}}(e^{j\omega}) = \sigma^2\mathbf{I}_r$ for all ω. Substituting in (9.3.2), we get $\mathbf{S_{yy}}(e^{j\omega}) = d\sigma^2\mathbf{I}_p$. This shows that $\mathbf{y}(n)$ is WUZE($d\sigma^2$) indeed. ▽▽▽

Here are some applications of this result in filter-banks. More can be found in the next two sections.

1. When $p = r$, losslessness of $\mathbf{T}^T(z)$ also implies that of $\mathbf{T}(z)$, and we can apply this to Fig. 9.3-2 with $\mathbf{R}(z) = \mathbf{T}(z)$. The special case where $\mathbf{T}(z)$ is a constant $M \times M$ unitary matrix also arises in filter bank theory (orthogonal transform coding, Appendix C).
2. Another useful example arises when $p = 1$ and $r = M$. In this case $\mathbf{T}(z)$ is an M channel *synthesis bank,* with power complementary property. If its input is WUZE(σ^2), then the output is zero-mean white with variance $d\sigma^2$.

9.4 NOISE IN FILTER BANKS

In a complete analysis/synthesis system (as in Fig. 5.4-1), roundoff noise is generated both by the analysis bank and the synthesis bank. In addition, the noise due to analysis bank propagates through the synthesis bank. We therefore have to consider not only the noise *generated* by individual filters, but also the way in which the synthesis bank transmits the noise entering its inputs. In transmultiplexers, where the analysis bank *follows* the synthesis bank (Fig. 5.9-1), the reverse situation prevails (Problem 9.7).

The effect of quantization of subband signals will be studied in Appendix C. In this section, we will concentrate only on quantization noise due to filter implementation. We will study the noise generated by some popular analysis banks introduced in Chap. 5 and 6. In the next section, the total noise due to analysis and synthesis filters will be considered.

Consider the QMF bank of Fig. 5.4-1, and let the analysis filters $H_k(z)$ be FIR with order N. Then each output has noise component which is white, with variance $(N+1)\sigma_q^2$ (Sec. 9.2.4). Since the decimated version of white noise is white, the noise $\epsilon_k(n)$ contaminating the decimated signal $v_k(n)$ is also white. The noise components $\epsilon_k(n)$ and $\epsilon_m(n)$ $(k \neq m)$ are in general

not uncorrelated (unless the filters $H_k(e^{j\omega})$ and $H_m(e^{j\omega})$ have completely non overlapping frequency responses, which is not the case in most filter banks; see Problem 9.12). Surprisingly however, in most filter banks, these noise components turn out to be uncorrelated for various other reasons (as we will elaborate).

Summary of Notations and Assumptions

a) σ_q^2 denotes the b-bit quantizer noise variance (9.2.3), and all noise components have zero-mean.
b) $H_k(z)$ denotes the kth analysis filter, and N denotes analysis filter order (which equals the synthesis filter order in all cases considered).
c) As in Fig. 5.4-1, $v_k(n)$ denotes the M-fold decimated version of the output of $H_k(z)$. Also, $\epsilon_k(n)$ is the noise component affecting $v_k(n)$ due to roundoff noise in the implementation of the analysis bank. In other words, $v_k(n) = v_{k,ideal}(n) + \epsilon_k(n)$.
d) All filter coefficients are assumed to be real for simplicity.

♠**Main points of this section.** We will justify the following conclusions pertaining to the noise generated by some of the popular analysis bank systems.

Case 1. *Two channel FIR system of Fig. 5.1-1(a), with* $H_1(z) = H_0(-z)$. This was considered in Sec. 5.2.2 (and listed as Method 1 in Table 6.7.2). The filter $H_0(z)$ is FIR with odd order N, and $h_0(n) = h_0(N-n)$. We assume that this is implemented in polyphase form [Fig. 5.2-2(b)], with $E_0(z)$ and $E_1(z)$ implemented in direct form. Then $\epsilon_0(n)$ and $\epsilon_1(n)$ can be assumed to be white and uncorrelated with each other, and have equal variance $(N+1)\sigma_q^2$.

Case 2. *Two-channel FIR perfect reconstruction (PR) system (direct-form).* (Sec. 5.3.6.) The filters are related by (5.3.28), and the order N is odd. Here $\epsilon_0(n)$ and $\epsilon_1(n)$ can be assumed to be white and uncorrelated with each other, with equal variance $(N+1)\sigma_q^2$.

Case 3. *Lattice implementation of the PR QMF bank.* (Sec. 6.4). We know that the above perfect reconstruction system can be implemented using the lattice structure of Fig. 6.4-1. The analyis bank is reproduced in Fig. 9.4-1 (with quantizers), by setting $\alpha = 1$ and $\eta = -1$ in Fig. 6.4-1. In this system $\epsilon_0(n)$ and $\epsilon_1(n)$ can again be assumed to be white and uncorrelated, with equal variance. But now the variance is $0.5(N+1)\sigma_q^2$.

Case 4. *M channel FIR PR system with paraunitary* $\mathbf{E}(z)$. (Sec. 6.5.) Let $\mathbf{E}(z)$ be implemented as a cascade of simpler paraunitary building blocks (e.g., as in Fig. 6.5-2). Assume that there is no quantization inside a building block, and that there are M b-bit quantizers at the output of each building block (Fig. 9.4-2). (This can be arranged by employing higher precision for all arithmetic inside the building block; the extra precision is finite, since there are no loops. The reader can modify the analysis for the case where there are more quantizers.). Then, $\epsilon_k(n)$ can again be assumed to be

white (and $\epsilon_k(n)$ uncorrelated with $\epsilon_\ell(m)$ for $k \neq \ell$), with equal variance $(N+1)\sigma_q^2/M$ for all k.

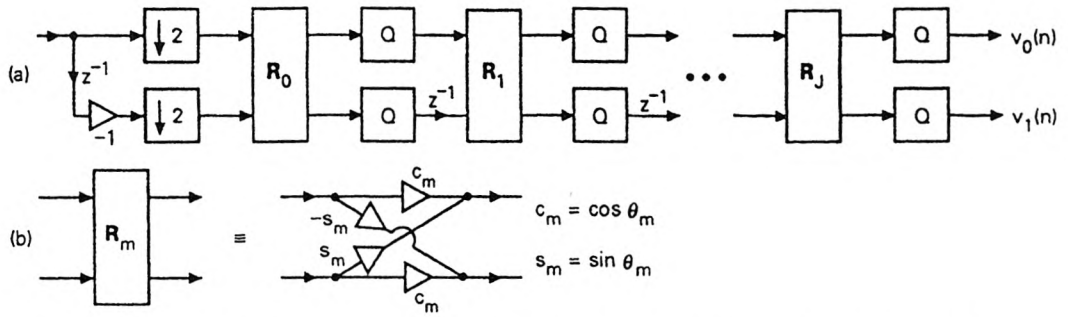

Figure 9.4-1 (a) Lattice structure for the analysis bank of the two channel PR QMF bnak. (b) Details of $\mathbf{R}_m$.

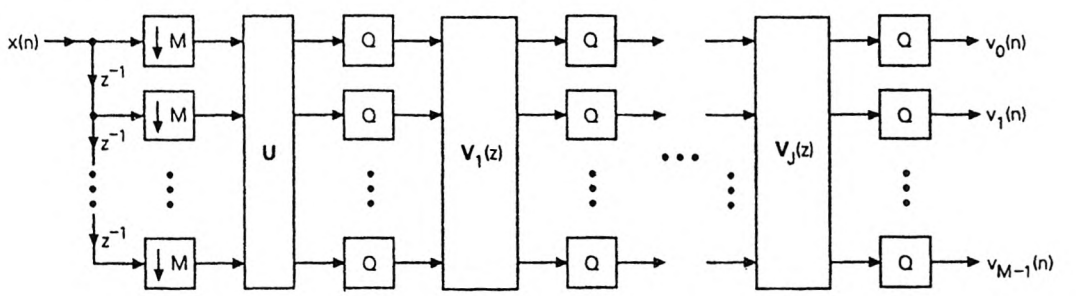

Figure 9.4-2 The anlaysis bank of a FIR PR QMF bank, with paraunitary polyphase matrix $\mathbf{E}(z)$. The paraunitary matrix is implemented as a cascade of simpler building blocks $\mathbf{U}$ and $\mathbf{V}_m(z)$.

Case 5. *Two channel IIR QMF bank with power symmetric analysis filters.* (Sec. 5.3). We know that this can be implemented as in Fig. 5.2-5 where $a_0(z)$ and $a_1(z)$ are unit-magnitude allpass. Each allpass filter can be implemented as in Fig. 9.2-7, with slight change of notations. Thus, let $s_{0,m}$ and $s_{1,m}$ be the scale factors used to scale the internal nodes. Also let k_i stand for the order of $a_i(z)$. In this case, $\epsilon_0(n)$ and $\epsilon_1(n)$ can be assumed to be white, and have equal variance given by

$$(\beta_0 + \beta_1)\sigma_q^2, \tag{9.4.1}$$

where

$$\beta_i = \sum_{m=0}^{k_i-1} s_{i,m}^2, \qquad i = 0, 1. \tag{9.4.2}$$

However $\epsilon_0(n)$ and $\epsilon_1(n)$ are not uncorrelated with each other.

Justifications

Case 1. Consider the polyphase implementation of Fig. 5.2-2(b). Since N is odd, the FIR filters $E_0(z)$ and $E_1(z)$ have $(N+1)/2$ coefficients each. So their outputs have noise components which are white, with equal variance $K\sigma_q^2$, with $K = 0.5(N+1)$. Now, the linear phase condition $h_0(n) = h_0(N-n)$ implies that the coefficients of $E_1(z)$ are time reversed versions of those of $E_0(z)$. Inspite of this, the noise components at the outputs of $E_0(z)$ and $E_1(z)$ can be assumed to be uncorrelated because, the samples entering $E_0(z)$ and $E_1(z)$ are even and odd numbered subsets of $x(n)$, respectively. The multipliers in $E_0(z)$ and $E_1(z)$ cannot be shared (as seen in Problem 5.3), and we cannot obtain a fifty-percent noise reduction (which is normally obtainable in linear-phase filters).

The 2×2 matrix $\mathbf{T}$ which follows $E_0(z)$ and $E_1(z)$ in Fig. 5.2-2(b) can be written as

$$\mathbf{T} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

This satisfies $\mathbf{T}\mathbf{T}^T = 2\mathbf{I}$, and we can invoke Theorem 9.3.1 to conclude that the noise components $\epsilon_0(n)$ and $\epsilon_1(n)$ are white and uncorrelated, with variance $2K\sigma_q^2 = (N+1)\sigma_q^2$.

Case 2. We know that the filter coefficients are related according to $h_1(n) = (-1)^n h_0(N-n)$, and since the same input $x(n)$ enters both filters, terms of the form $x(i)h_0(m)$ are shared by the filter outputs. So we cannot claim that the roundoff noise components at the outputs of the filters are uncorrelated. However, consider the decimated outputs

$$v_0(n_0) = \sum_{m_0=0}^{N} h_0(m_0)x(2n_0 - m_0), \quad v_1(n_1) = \sum_{m_1=0}^{N} h_1(m_1)x(2n_1 - m_1). \tag{9.4.3}$$

We will show that [for arbitrary $x(n)$] the same product $h_0(k)x(i)$ will not be shared by the two summations. Suppose this is not true. Then we must have

$$h_0(m_0) = \pm h_1(m_1), \quad \text{and} \quad x(2n_0 - m_0) = x(2n_1 - m_1), \tag{9.4.4}$$

for some n_0, n_1, m_0 and m_1. Since $h_1(n) = (-1)^n h_0(N-n)$, this implies $m_0 = N - m_1$ and $2n_0 - m_0 = 2n_1 - m_1$. This in turn means $2m_0 = 2(n_0 - n_1) + N$, which contradicts the fact that N is odd. Summarizing, we can assume that $v_0(n)$ and $v_1(n)$ are uncorrelated. Furthermore, we already know (Sec. 9.2.4) that $v_0(n)$ and $v_1(n)$ are white with variance $(N+1)\sigma_q^2$.

Case 4. Since Case 3 follows from Case 4, we now proceed directly to Case 4. The analysis bank is shown in Fig. 9.4-2, and there are $J+1$ paraunitary building blocks in cascade. The 0th building block is a constant unitary matrix $\mathbf{U}$, whereas the remaining ones are degree-one paraunitary systems. This cascade covers the situations in Fig. 9.4-1 as well, if we replace $\mathbf{U}$ and $\mathbf{V}_m(z)$ appropriately. At the output of the mth building block, we have the noise source vector $\mathbf{e}_m(n)$, generated by the M quantizers. According to our noise model assumptions, each component of this noise vector is white with variance σ_q^2, and any two components are uncorrelated. So $\mathbf{e}_m(n)$ is WUZE(σ_q^2). The transfer matrix from $\mathbf{e}_m(n)$ to the output terminal is a cascade of paraunitary systems and is therefore paraunitary. As a result, the noise vector $\mathbf{e}_{m,out}(n)$ which contaminates the output vector $\mathbf{v}(n) = [\,v_0(n) \quad \ldots \quad v_{M-1}(n)\,]^T$ is WUZE(σ_q^2). Since any two noise vectors $\mathbf{e}_m(n)$ and $\mathbf{e}_\ell(n)$ are uncorrelated, the total noise vector contaminating $\mathbf{v}(n)$ is WUZE($(J+1)\sigma_q^2$). Using the relation $N = MJ+M-1$ this can be written as WUZE($(N+1)\sigma_q^2/M$). Summarizing, the noise components $\epsilon_k(n)$ at the analysis bank output are white and uncorrelated, with variance $(N+1)\sigma_q^2/M$.

Case 5. The allpass filters $a_i(z)$ are products of first order real coefficient allpass filters, and can be implemented as in Fig. 9.2-7. From Sec. 9.2.4 we therefore conclude that the roundoff noise at the output of $a_i(z)$ is white with variance $\beta_i\sigma_q^2$, where β_i is as in (9.4.2). Using standard assumptions, it follows that the noise components at the two allpass filter outputs are uncorrelated. As a result, the noise components at the locations of $v_0(n)$ and $v_1(n)$ are white, with variance $(\beta_0+\beta_1)\sigma_q^2$.

9.5 FILTER BANK OUTPUT NOISE

In the QMF bank of Fig. 5.4-1, the roundoff noise components $\epsilon_k(n)$ generated by the analysis bank are propagated through the synthesis bank. This contributes a noise component $e_a(n)$ at the output node [i.e., node labeled $\hat{x}(n)$]. In addition to this, the roundoff operations in the implementation of the synthesis bank contribute a noise component $e_S(n)$. So the total noise component $e(n)$ affecting $\hat{x}(n)$ can be written as

$$e(n) = e_a(n) + e_S(n). \tag{9.5.1}$$

In other words, we can write $\hat{x}(n) = \hat{x}_i(n)+e(n)$ where $\hat{x}_i(n)$ is the filter bank output under ideal conditions (i.e., no quantizers). Under normal conditions, we can assume that $e_a(n)$ and $e_S(n)$ are uncorrelated, with zero mean. We will now study further properties of $e_a(n)$ and $e_S(n)$ for each of the five cases listed in Sec. 9.4. In order to compare various structures on a common ground, we will adopt some conventions:

9.5.1 Conventions and Assumptions

1. *Scaling.* We will insert scale factors at appropriate places to satisfy the following requirement: signals that are inputs to appropriate computa-

tional building blocks should be scaled in the $\mathcal{L}_2$ sense. This requirement means that if the filter-bank input $x(n)$ is white with unit variance, then the variance at the scaled node is also unity.

2. Whenever necessary, we will insert a scale factor in front of $\hat{x}(n)$ so that there is no discrepancy between $x(n)$ and $\hat{x}(n)$ (except for possible amplitude and/or phase distortions, etc.). For example, in a perfect reconstruction system we will have $\hat{x}(n) = cx(n - n_0)$, with $c = 1.0$.
3. We will neglect the noise generated by the above scale factors, as their contribution to total noise is small.

Figure 9.5-1 shows all the QMF banks of interest, with scale factors inserted according to these conventions. Quantizers, which are inserted as explained in earlier sections, are not shown just to keep the figures simple. We now make some observations and leave it to the reader to verify them.

1. **Fig. 9.5-1(a).** In this system, the inputs to the filters $E_0(z)$ and $E_1(z)$ in the analysis bank (in fact any nodes connected directly to $x(n)$) are automatically scaled (in the $\mathcal{L}_2$ sense). The same is approximately true of the filters $E_1(z)$ and $E_0(z)$ in the synthesis bank, under the assumption that the analysis filter $H_0(z)$ has energy ≈ 0.5. (This holds to the extent that $|H_0(e^{j\omega})| \approx 1$ in the passband and $|H_0(e^{j\omega})| \approx 0$ in the stopband). So we do not require any scale factor except the '2' inserted in front of $\hat{x}(n)$, to satisfy convention 2.
2. **Fig. 9.5-1(b).** Next consider Fig. 9.5-1(b). If we insert the three scale factors $\sqrt{2}$ as shown, then the inputs to $F_k(z)$ are scaled in the $\mathcal{L}_2$ sense, and furthermore convention 2 is satisfied.
3. **Fig. 9.5-1(c).** In Fig. 9.5-1(c), $\mathbf{E}(z)$ is implemented as in Fig. 6.5-2, where $\mathbf{U}$ is unitary and $\mathbf{V}_m(z)\widetilde{\mathbf{V}}_m(z) = \mathbf{I}$. We assume $\mathbf{U}$ is normalized such that $\mathbf{U}^\dagger\mathbf{U} = \mathbf{I}$. So $\mathbf{E}(z)\widetilde{\mathbf{E}}(z) = \mathbf{I}$. The same comments hold for the paraunitary system $\mathbf{R}(z)$. As a result we have $\hat{x}(n) = x(n - n_0)$, and no scale factors are necessary to satisfy convention 2. Also, the inputs to each of the paraunitary building blocks ($\mathbf{V}_m(z)$ and $\mathbf{U}$) are scaled in the $\mathcal{L}_2$ sense automatically.
4. **Fig. 9.5-1(d).** Finally, in Fig. 9.5-1(d), the allpass filters satisfy $|a_i(e^{j\omega})| = 1$. Insertion of $1/2$ as indicated ensures that the inputs to the allpass filters in the synthesis bank are scaled in the $\mathcal{L}_2$ sense. In addition, each allpass filter (implemented as in Fig. 9.2-7) has its own internal scale factors. For this figure, we have $\hat{X}(e^{j\omega}) = e^{j\phi(\omega)}X(e^{j\omega})$, consistent with the fact that aliasing and amplitude distortion have been eliminated.

9.5.2 Output Roundoff Noise e(n)

For each of the above cases, we can compute the output noise variance as follows. We will freely use the standard noise model assumptions, as well as the results in Sec. 9.2.4 (FIR roundoff noise, and allpass roundoff noise) and Sec. 9.3.

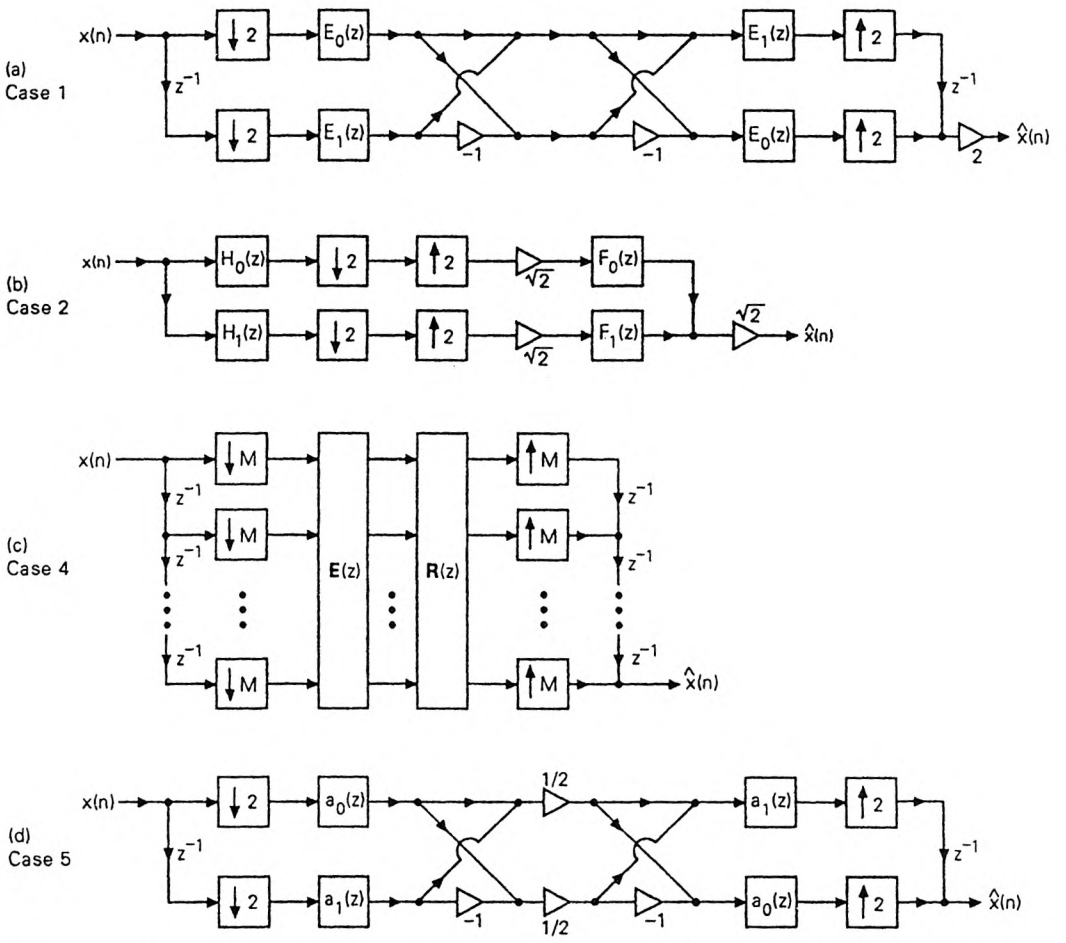

Figure 9.5-1 Examples of QMF banks, with scale factors inserted (a) case 1, (b) case 2, (c) case 4 and (d) case 5.

Case 1

Each polyphase component $E_k(z)$ in the synthesis bank, which is FIR with length $0.5(N+1)$, generates white noise $\delta_k(n)$ with variance $0.5(N+1)\sigma_q^2$. Under normal assumptions, $\delta_0(n)$ and $\delta_1(n)$ are uncorrelated. The

output noise $e_S(n)$, which is the interlaced version of $\delta_0(n)$ and $\delta_1(n)$ (scaled by two) is, therefore, white with variance $2(N+1)\sigma_q^2$.

The noises generated by $E_0(z)$ and $E_1(z)$ in the analysis bank are also white and uncorrelated, with variance $0.5(N+1)\sigma_q^2$. If $|H_0(e^{j\omega})| \approx 1$ in the passband, then $|E_k(e^{j\omega})|$ are 'approximately' constant (≈ 0.5). We can then verify that this contributes a noise component $e_a(n)$ ('almost' white) at the output of the filter bank, with variance $2(N+1)\sigma_q^2$. The total noise $e(n)$ is therefore essentially white, with variance $4(N+1)\sigma_q^2$.

Case 2

Each synthesis filter $F_k(z)$ is FIR with order N, and its input is the output of an expander. So only $0.5(N+1)$ multiplication are involved per computed output. Thus, the roundoff noise at the outputs of the two synthesis filters $F_k(z)$ are uncorrelated and white, with variance $0.5(N+1)\sigma_q^2$ each. So the output noise component $e_s(n)$ is white, with variance $2(N+1)\sigma_q^2$.

To study the effect due to analysis filter noise, consider Fig. 9.5-2, which is the synthesis bank in polyphase form. The noise entering the paraunitary system $\mathbf{R}(z)$ from the analysis bank is WUZE($2(N+1)\sigma_q^2$). By Theorem 9.3.1 the noise at the output of $\mathbf{R}(z)$ is WUZE($(N+1)\sigma_q^2$). (This is because $\widetilde{\mathbf{R}}(z)\mathbf{R}(z) = c\mathbf{I}$, with $c = 0.5$, which is consistent with $|F_k(e^{j\omega})| \leq 1$). So the noise component $e_a(n)$ is white with variance $2(N+1)\sigma_q^2$. Summarizing, the total output noise $e(n)$ is white with variance $4(N+1)\sigma_q^2$.

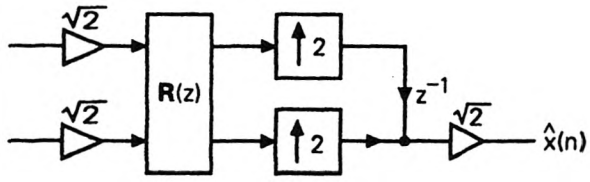

Figure 9.5-2 The synthesis bank of the PR QMF system, drawn in polyphase form.

Case 4

We will proceed to Case 4, since Case 3 is covered by this. The synthesis bank is implemented in a manner similar to Fig. 9.4-2 (i.e., as a cascade of paraunitary building blocks). Here $\mathbf{E}(z)\widetilde{\mathbf{E}}(z) = \mathbf{R}(z)\widetilde{\mathbf{R}}(z) = \mathbf{I}$. Proceeding as in Sec. 9.4 we conclude that the noise vector at the output of $\mathbf{R}(z)$, due to roundoff in synthesis bank, is WUZE($(N+1)\sigma_q^2/M$). So the noise component $e_S(n)$ which is the interlaced version of these, is white with variance $(N+1)\sigma_q^2/M$.

The noise entering $\mathbf{R}(z)$ from the analysis bank is also WUZE($(N+1)\sigma_q^2/M$). Using Theorem 9.3.1 as before, this noise vector propagates to the output of $\mathbf{R}(z)$ as WUZE($(N+1)\sigma_q^2/M$). The interlaced version $e_a(n)$ is

again white with variance $(N+1)\sigma_q^2/M$. So the total noise $e(n)$ is white, with variance $2(N+1)\sigma_q^2/M$. For the special case of the two channel lattice this reduces to $(N+1)\sigma_q^2$.

Case 5.

The allpass filters $a_i(z)$ in Fig. 9.5-1(d) are implemented in cascade form (Fig. 9.2-7), with scale factors $s_{i,m}$. The noise generated in the implementation of $a_i(z)$ contributes a white noise component at its output, with variance $\beta_i\sigma_q^2$, where $\beta_i = \sum_{m=0}^{k_i-1} s_{i,m}^2$. In the synthesis bank, the noises at the outputs of $a_i(z)$ get interlaced. Since $\beta_0 \neq \beta_1$, the interlaced version $e_S(n)$ is not stationary. However, we will see that the total noise $e(n)$ is stationary.

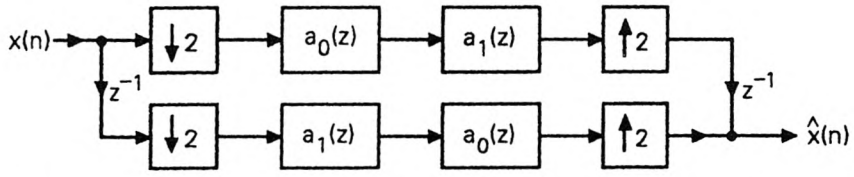

Figure 9.5-3 The power symmetric QMF bank, redrawn for the purpose of study.

Recall that the noise components $\epsilon_0(n)$ and $\epsilon_1(n)$ generated by the analysis bank are white but not uncorrelated. In order to study the properties of $e_a(n)$ it therefore turns out to be convenient to use the equivalent structure of Fig. 9.5-3. The noise from $a_0(z)$ in the analysis bank enters the filter $a_1(z)$ in the synthesis bank, and vice versa. So the noise at the output of $a_i(z)$ in the synthesis bank is a white noise component with variance $\beta_{1-i}\sigma_q^2$. If these are interlaced to obtain $e_a(n)$, the result is again non stationary. However, the total noise $e(n) = e_a(n) + e_S(n)$ has variance $(\beta_0+\beta_1)\sigma_q^2$, which is same for all n. Summarizing, $e(n)$ is white with variance $(\beta_0+\beta_1)\sigma_q^2$.

Summary. Table 9.5.1 summarizes the output noise variance for all the cases. It is interesting to note that the total noise $e(n)$ at the output of the QMF bank is white in each case. For $M = 2$ the variance for case 4 is ony $(N+1)\sigma_q^2$ which is four times smaller than for cases 1 and 2. This has to do with the choice of scale factors, and is not a very significant difference. It corresponds to about 6.02 dB improvement in noise (which is equivalent to one additional bit of internal word length).

9.6 LIMIT CYCLES

Signal quantization in a digital filter usually generates a random error at the filter output, as we have seen in the previous sections. Under some conditions, however, signal quantization causes periodic oscillations called *limit cycles*. The most well understood type of limit cycles are zero-input

limit cycles. As the name implies, these are self sustained oscillations which remain after the input $u(n)$ has been turned off.

TABLE 9.5.1 Summary of properties of output noise e(n) in various QMF banks.

Case considered	Case 1 2-channel FIR linear phase QMF (direct form polyphase)	Case 2 2-channel FIR PRQMF (direct form)	Case 4 M-channel FIR PRQMF with paraunitary $\mathbf{E}(z)$ (cascaded structure for $\mathbf{E}(z)$)	Case 5 2-channel IIR power symmetric QMF
Variance of output noise $e(n)$	$4(N+1)\sigma_q^2$	$4(N+1)\sigma_q^2$	$2(N+1)\sigma_q^2/M$	$(\beta_0+\beta_1)\sigma_q^2$ (see text)

In all cases $e(n)$ has zero mean. For cases 2,4 and 5 $e(n)$ is white.
For case 1, $e(n)$ is 'approximately' white (see text). N denotes filter order, and σ_q^2 is the basic quantizer noise variance.

Limit cycles arise because the quantizers, which are nonlinear elements, are inserted in feedback loops. Even though the linear system (i.e., structure without quantizer) is stable and therefore does not suffer from zero-input limit cycles, the system with quantizers can support such oscillations. Two types of limit cycles have been distinguished. The first is the granular or "roundoff" type, which is due to the "small" error introduced by the quantizer. The magnitude of this oscillation is proportional to the step size Δ, and can be reduced by adding more bits of precision. The second type, called overflow oscillations, can arise when the quantizer input exceeds the dynamic range. These are "large" oscillations, (with magnitude close to unity!) and cannot be reduced by adding more bits of precision. Examples of both types can be found in Oppenheim and Schafer [1989].

It is clear that FIR filter structures are free from limit cycles, since they have no feedback loops. Limit cycles arise only in IIR structures. In multirate filter bank systems, the only significant IIR filters we have seen are power symmetric filters (Sec. 5.3). These systems can be implemented in terms of allpass filters $a_0(z)$ and $a_1(z)$ as shown in Fig. 5.2-5. In these applications, $a_i(z)$ are real coefficient filters and furthermore can be factorized into first order allpass sections with *real coefficients* (Sec. 5.3.5). If these first order sections are free from limit cycles, then so is the complete structure. We are therefore interested in suppressing limit cycles in first order real coefficient sections. We will conclude this section by showing how.

First Order Sections

Consider Fig. 9.1-1(b) again, which shows a first order section with a quantizer Q. Under zero-input conditions the behavior of the closed loop

system is governed by the equations

$$y(n) = Q[w(n)], \tag{9.6.1}$$

and

$$w(n) = ay(n-1). \tag{9.6.2}$$

Assume $|a| < 1$ (i.e., the system without quantizer is stable). We then have

$$|w(n)| < |y(n-1)|, \tag{9.6.3}$$

unless $y(n-1) = 0$. Suppose now that the quantizer has the property

$$|Q[x]| \leq |x|, \tag{9.6.4}$$

for any number x. This is easily accomplished in practice. For example, if the quantizer is of the magnitude truncation type, this is satisfied for any x within the dynamic range. If x is outside the dynamic range (i.e., overflow situation), then (9.6.4) is still satisfied because $Q[x]$ is within the dynamic range.

Quantizers satisfying (9.6.4) are said to be passive. With such quantizers, we have

$$|y(n)| \leq |w(n)|. \tag{9.6.5}$$

Combining with (9.6.3) we see that $|y(n)| < |y(n-1)|$. But since $y(n)$ is a b-bit fraction with step size Δ, this implies

$$|y(n)| \leq |y(n-1)| - \Delta, \tag{9.6.6}$$

for any $y(n-1) \neq 0$. This means that as n increases, the magnitude of $y(n)$ keeps decreasing (at least by Δ each time) until it becomes zero (in a finite amount of time).

Summarizing, the first order section of Fig. 9.1-1(b) does not support zero-input limit cycles of either type as long as the quantizer is passive and $|a| < 1$.

9.7 COEFFICIENT QUANTIZATION

Detailed presentations of coefficient quantization effects in digital filters can be found in a number of references, for example, Oppenheim and Schafer [1975], and Rabiner and Gold [1975]. So our presentation is brief, and we will discuss only some special issues particularly relevant to multirate filter banks. When the multiplier coefficients in a filter structure are quantized, the transfer function changes, say from $H(z)$ to $H_q(z)$. This means that the magnitude as well as phase responses have changed. In some extreme cases, some of the poles, which are close to the unit circle, may move outside, resulting in unstable filters. The IIR direct-form structure (demonstrated in Fig. 2.1-5 for order $N = 2$) is known to be very sensitive to coefficient

quantization, particularly for large N. The effect is less severe for FIR direct form structures (Fig. 2.1-3), even though improved structures are available.

Magnitude response sensitivity. For linear phase FIR filter structures, the coefficient symmetry (hence linearity of phase) can be preserved in spite of quantization (e.g., see Fig. 2.4-3). So only the magnitude response $|H(e^{j\omega})|$ changes due to quantization. For filters which do not have linear phase (e.g., IIR), the phase response also changes, but this is usually not of concern since it is not linear anyway. It is therefore important to discuss only the sensitivity of the magnitude response $|H(e^{j\omega})|$.

Improved Structures

There exist structures for which the effects of quantization are less severe. In general cascade form structures (Sec. 2.1.3) are less sensitive to quantization. In these structures, quantization of a denominator coefficient affects only one complex-conjugate pole pair (or real pole, as the case may be). Similar comment holds for zeros. In Sec. 3.4.3 we introduced lattice structures which have some other advantages. One of these is that the transfer function remains stable as long as the quantized lattice coefficients k_m (Fig. 3.4-8) satisfy $|k_m| < 1$. From Sec. 3.6 we know that many IIR filters (including elliptic) can be implemented as a sum of two allpass filters (Fig. 3.6-2). Each allpass filter in turn can be implemented using the lattice. Such structures are therefore *stable even under quantization.* It is also known [Gray, Jr., 1980] that lattice structures are free from zero-input limit cycles. Other structures with improved finite-wordlength behavior are wave digital filters [Fettweis, 1971] and orthogonal filters [Deprettere and Dewilde, 1980].

In this section, we will show that a number of filter bank structures which we have presented in Chap. 5 and 6 exhibit low passband sensitivity. This means that the passband response of the quantized system is 'acceptably close' to the specified response. This is a consequence of a property called structural passivity, which we will elaborate. In two-channel PR QMF banks, since the analysis filters are power symmetric, this also implies that the stopband response is well-controlled under quantization, provided the structure retains power symmetry in spite of quantization.

9.7.1 Structural Passivity

Let m_i denote the multiplier coefficients in the structure, and assume $-1 < m_i < 1$. (This can always be arranged, since we can write m_i = integer plus fraction, and eliminate the integer part by using adders.) Suppose the structure is such that $|H(e^{j\omega})| \leq 1$ for all values of the coefficients in the range $-1 < m_i < 1$. (Assume further that the transfer function remains stable for $-1 < m_i < 1$.) We then say that the implementation is structurally *passive* (or *bounded*) [Vaidyanathan and Mitra, 1984]. This means, in particular, that the response is bounded by unity even if the multipliers are quantized, as long as the quantized value does not exceed

unity.

Consequence of Structural Passivity

Consider Fig. 9.7-1(a). This represents the ideal (unquantized) response of a digital elliptic filter. The magnitude attains a maximum of unity at the frequencies θ_k in the passband, that is, $|H(e^{j\theta_k})| = 1$. If we now quantize a coefficient m_i in the structure, the response $|H(e^{j\theta_k})|$ can only decrease, as demonstrated in Fig. 9.7-1(b). In other words, we have

$$\frac{\partial |H(e^{j\theta_k})|}{\partial m_i} = 0. \tag{9.7.1}$$

This means that the sensitivity of the magnitude response with respect to the coefficients, evaluated at the nominal coefficient values, is equal to zero. This is true with respect to every coefficient, and at every extremal frequency θ_k in the pass band. If there are several extrema in the passband, we can expect the sensitivity of $|H(e^{j\omega})|$ with respect to the coefficients m_i to be low for all frequencies in the passband. (This has also been verified by simulation, as we will demonstrate below.) This is the key behind the low passband sensitivity of many structures, for example, wave filters and orthogonal filters mentioned above.

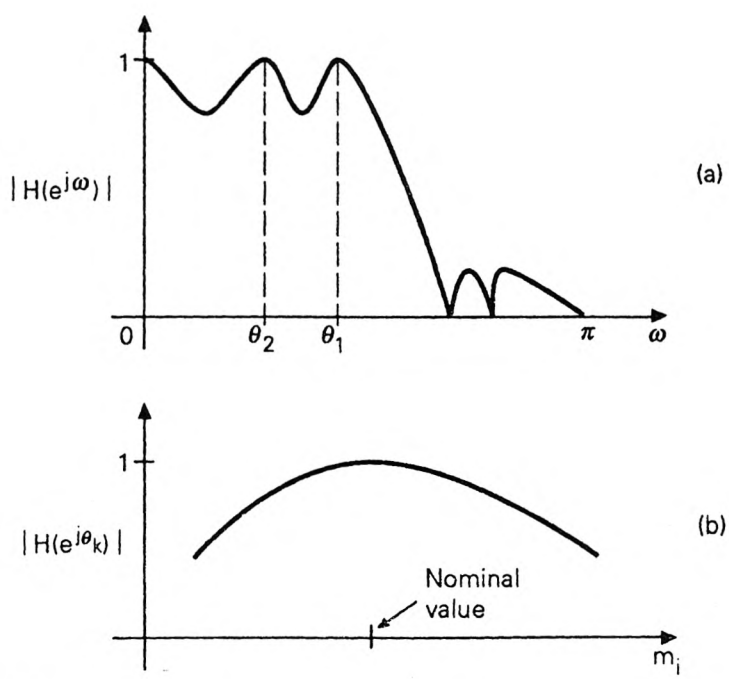

Figure 9.7-1 (a) Example of an elliptic filter response, and (b) demonstrating the effect of structural passivity.

9.7.2 Application in QMF Banks

Power Symmetric IIR QMF Bank

Consider the power symmetric IIR QMF bank discussed in Sec. 5.3. This system can be implemented as in Fig. 5.2-5 where $a_0(z)$ and $a_1(z)$ are unit magnitude allpass filters. In this system each allpass filter can be implemented as in Fig. 9.2-7, where the coefficients α_i satisfy $0 < \alpha_i < 1$. When these coefficients are quantized, the filters $a_0(z)$ and $a_1(z)$ still remain stable and satisfy $|a_i(e^{j\omega})| < 1$ (as long as $|\alpha_i| < 1$ continues to hold). Since

$$H_i(z) = \frac{a_0(z^2) \pm z^{-1} a_1(z^2)}{2}, \tag{9.7.2}$$

we have $|H_i(e^{j\omega})| \leq 1$, that is, the implementation is structurally passive.

To demonstrate the low sensitivity property, consider the case where $a_i(z)$ are first order filters so that $H_0(z)$ is as in (5.3.18). We now implement this system (i.e., Fig. 5.2-5), with the allpass filters $a_i(z)$ implemented in cascade form (i.e., as in Fig. 9.2-7). Fig. 9.7-2(a) shows $|H_0(e^{j\omega})|$ for the quantized as well as ideal systems. The quantization level is 6 bits per multiplier (i.e., $b = 6$ in Fig. 9.2-1). For comparison, Fig. 9.7-2(b) shows the response of a direct form structure, with multipliers quantized to the same level (for convenience all plots are normalized to have a peak value of unity). It is clear that the passband response of the quantized structurally-passive implementation is far superior to the direct-form structure. It is also worth noting that the direct form structure does not preserve the power symmetric property under quantization, unlike the structure of Fig. 5.2-5.

FIR Perfect-Reconstruction QMF Lattice

Consider now the two channel QMF lattice of Fig. 6.4-1. The analysis filters are determined by the $J + 1$ angles θ_m. The scale factor α does not affect the sensitivity of the response and can be assumed to be $1/\sqrt{2}$ for the purpose of discussion. We know from Sec. 6.4 that the analysis filters are power complementary and satisfy $|H_k(e^{j\omega})|^2 \leq 1$, regardless of the values of the angles θ_m. This structure therefore exhibits low passband sensitivity.

We now demonstrate this, using the more economic structure of Fig. 6.4-3(a), and quantizing the coefficients α_m. For this we consider a system designed using the technique described in Sec. 6.3.2, with filter order $N = 23$. Figure 9.7-3(a) shows $|H_0(e^{j\omega})|$, for the quantized as well as unquantized lattice structures. For the quantized lattice we use 8 bits per coefficient α_m. For comparison, Fig. 9.7-3(b) shows the response when the impulse response coefficients $h_0(n)$ are directly quantized (i.e., direct-form implementation). From the passband details it is clear that the lattice structure has much lower passband sensitivity compared to the direct form, demonstrating the effect of structural passivity. Once again, the direct form structure does not preserve the power symmetric property, unlike the lattice structure.

Further examples of structurally passive implementations can be found in Problem 14.29.

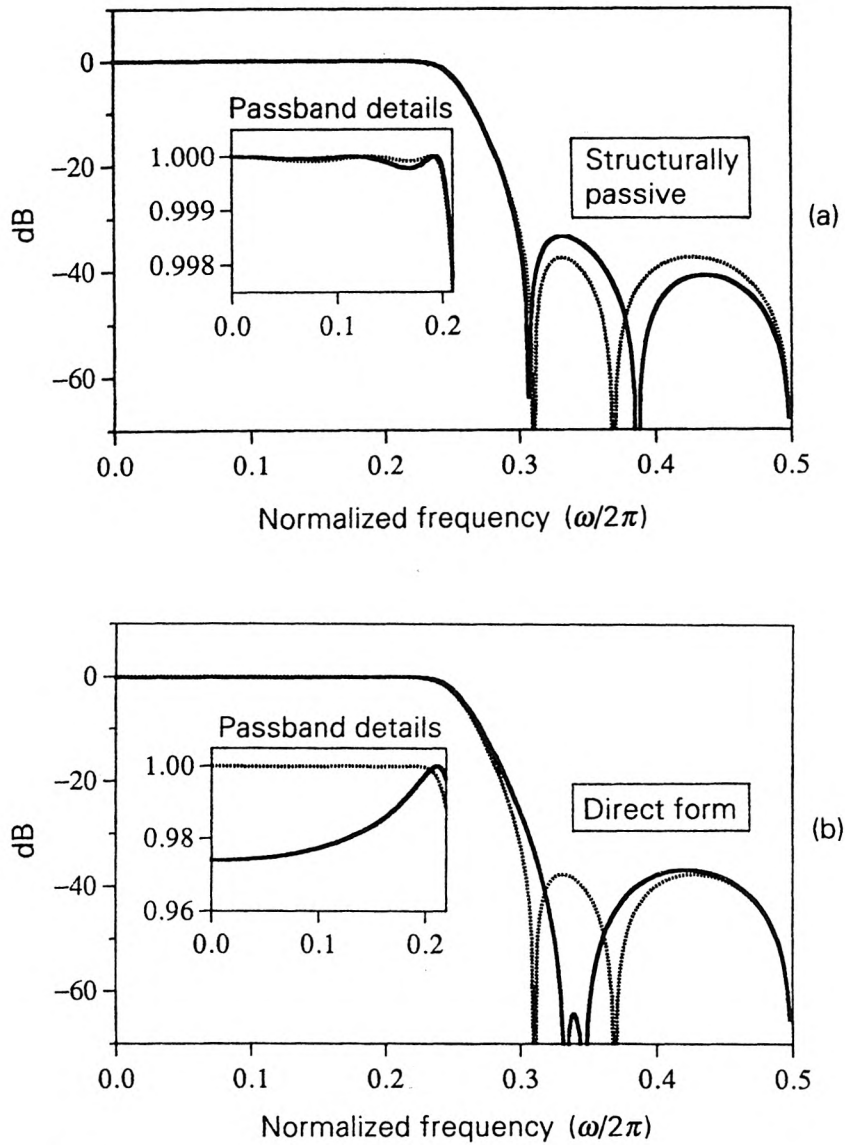

Figure 9.7-2 Magnitude response plots for quantized IIR power symmetric elliptic filters. (a) allpass-based structure (structurally passive) and (b) direct form structure. Broken lines indicate unquantized responses.

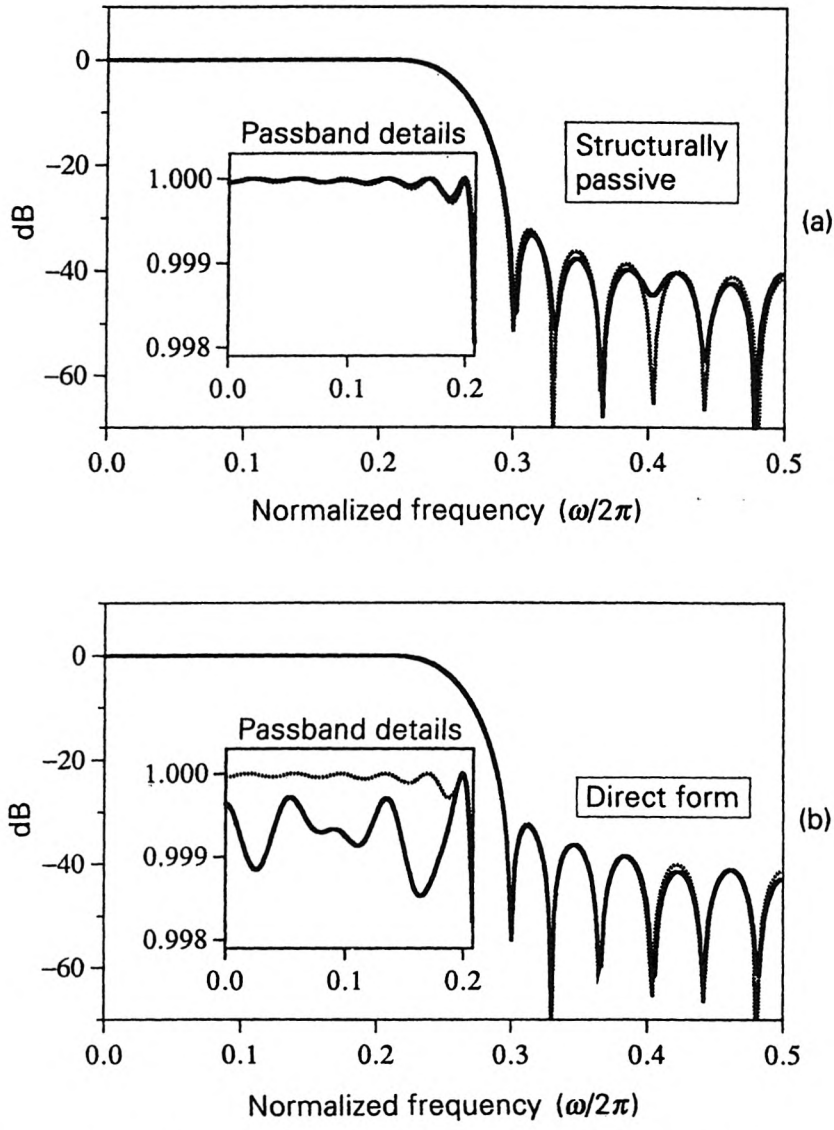

Figure 9.7-3 Magnitude responses for quantized FIR power symmetric filters. (a) QMF-lattice structure (structurally passive), and (b) direct form structure. Broken lines indicate unquantized responses.

PROBLEMS

Note. Familiarity with the material in Appendix B will be helpful while solving some of the following problems.

9.1. Consider the following lattice structure, where k is real with $k^2 < 1$, and $\hat{k} = \sqrt{1-k^2}$.

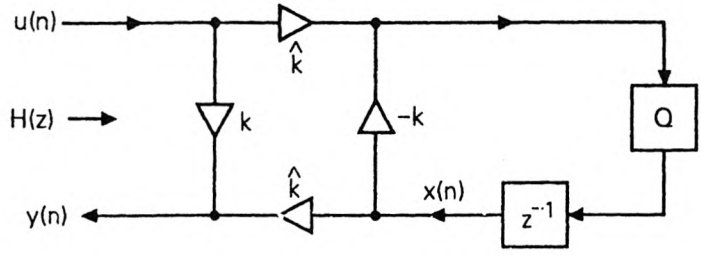

Figure P9-1

From Sec. 3.4 we know that the transfer function $H(z)$ is stable allpass. With a quantizer inserted in the feedback loop as shown, draw the noise model, and estimate the output noise variance under the usual assumptions.

9.2. Consider again the lattice structure in Fig. P9-1. There are four multipliers, two of which have input $u(n)$, and two of which have input $x(n)$. So in order to scale the structure, it is sufficient to scale the node $x(n)$. Show that this node is in fact already scaled in the $\mathcal{L}_2$ sense!

9.3. Consider Fig. 9.2-3, where a multiplier $1/L$ is inserted to scale the node which represents the output of the adder. For sum-scaling, we know we have to choose $L = 1/(1-|a|)$.

a) For $\mathcal{L}_2$ scaling, how would you choose L?

b) Now assume that the input $u(n)$ is a zero-mean white WSS random process with unit variance, and let $a = 0.99$. Estimate the variance of the output *signal* $y(n)$ for the two cases (i) sum-scaling and (ii) $\mathcal{L}_2$ scaling.

9.4. Let p be a positive integer. The $\mathcal{L}_p$ norm of a transfer function $F(z)$ is defined as

$$\|F\|_p = \left(\int_0^{2\pi} |F(e^{j\omega})|^p \frac{d\omega}{2\pi} \right)^{1/p}. \qquad (P9.4a)$$

Let $y(n)$ be the output of $F(z)$ in response to an input $u(n)$. Let $U(e^{j\omega})$ be the Fourier Transform of $u(n)$. It can then be shown that

$$|y(n)| \le \|F\|_p \|U\|_q, \qquad (P9.4b)$$

for any pair of positive integers p, q such that $p^{-1} + q^{-1} = 1$. Examples are $(p = 1, q = \infty)$, $(q = 1, p = \infty)$, and $(p = q = 2)$. If $\|F\|_p = 1$, we see that $|y(n)| \le \|U\|_q$ for all n. Under this condition we say that the node $y(n)$ is scaled in the $\mathcal{L}_p$ sense. If $\|U\|_q \le 1$ as well, then $|y(n)| \le 1$, that is, there is no overflow at node $y(n)$. (For simplicity, assume that $y(n) = \pm 1$ is not considered as overflow). Summarizing, $\mathcal{L}_p$ scaling prevents overflow if the input is such that $\|U\|_q \le 1$.

a) Show that $\|F\|_\infty$ is equal to the maximum value of $|F(e^{j\omega})|$.
b) In each of the following cases, what kind of $\mathcal{L}_p$ scaling will be appropriate (i.e., what p should be chosen) to avoid overflow? (i) $u(n)$ is a sequence with energy $\sum_n |u(n)|^2 = 1$, (ii) $u(n) = e^{j\omega_0 n}$ for some real ω_0, and (iii) $u(n)$ is such that $|U(e^{j\omega})| \leq 1$. (*Note.* In each of the above cases, sum-scaling (Sec. 9.2) could also have avoided overflow, but is more stringent than necessary.)

9.5. Let $x(n)$ be WSS with autocorrelation $R_{xx}(k)$, and let $y(n) = x(Mn)$. Show that the autocorrelation of $y(n)$ is given by $R_{yy}(k) = R_{xx}(Mk)$.

9.6. Let $\mathbf{x}(n)$ be a zero-mean WSS process. Show that it is WUZE(σ^2) if and only if $E[\mathbf{x}(n)\mathbf{x}^\dagger(n+i)] = \sigma^2\delta(i)\mathbf{I}$.

9.7. Consider the transmultiplexer structure of Fig. 5.9-1. Assume that the filters $H_k(z)$ and $F_k(z)$ are FIR with length $N+1 = MK$ for some K. Let each analysis filter $H_k(z)$ have energy $1/M$. Let $e_k(n)$ denote the roundoff noise component affecting the node labeled $\widehat{x}_k(n)$. Using the usual fixed-point b-bit roundoff noise model, estimate the variance of $e_k(n)$. With no further assumptions, can you say that $e_k(n)$ and $e_\ell(m)$ are uncorrelated for $k \neq \ell$?

9.8. Consider the two-stage decimation filter of Fig. 4.4-5(b). (This was the topic of Sec. 4.4.2 which should be reviewed at this time.) Here N_g and N_i are the orders of the FIR filters $G(z)$ and $I(z)$. Using the usual fixed-point roundoff noise model of Sec. 9.2, we wish to compute some noise variances, in terms of N_g, N_i, the quantizer noise variance σ_q^2, and the energy of $G(e^{j\omega})$. For simplicity, ignore the noise reduction obtainable by exploiting linear-phase symmetry.
a) Estimate the variance σ_1^2 of the roundoff noise at the final output, that is, output of M_2.
b) Instead of the above system suppose we use a single stage decimation filter for this problem, that is, Fig. 4.1-7 with $M = M_1M_2$. Let N denote the order of $H(z)$. Estimate the noise variance σ_2^2 at the output of M.
c) In Design example 4.4.2, we started with some specifications, and arrived at specific values for M_1, M_2, N_g, N_i and N. Using these values, find the improvement in noise variance due to multistage implementation, i.e., find σ_2^2/σ_1^2. You can make the assumption that the energy of $G(e^{j\omega})$ is 0.5, which is consistent with the specifications of this design.

9.9. Consider the fractional decimation circuit of Fig. 4.1-10(b). Assume $L = 2$ and $M = 3$, and let $H(z)$ be FIR with order $N = 59$. For this problem, ignore any simplicity offered by linear-phase symmetry. Assume the usual fixed-point roundoff noise model of Sec. 9.2.
a) Estimate the roundoff noise variance at the output node [labeled $y(n)$].
b) Now consider Fig. 4.3-8(d). This represents an efficient polyphase implementation of the above fractional decimation circuit. Estimate the roundoff noise variance at the output node.

9.10. Consider Fig. 8.5-2 which represents the cosine modulated analysis filter bank. Here each pair $G_k(z), G_{M+k}(z)$ is power complementary, and is implemented using the lattice Fig. 8.5-1. The cosine modulation matrix $\mathbf{T}$ is as in (8.4.9). (At this time you must review Section 8.4, in particular the meanings of $\mathbf{A}_0, \mathbf{A}_1, \mathbf{C}, \mathbf{S}$, and so on). Assume that (i) each lattice has quantizer inserted

similar to Fig. 9.4-1(a), (ii) there are no quantizers inside $\mathbf{T}$, and (iii) there are M quantizers for the M outputs of $\mathbf{T}$. Let $\epsilon_k(n)$ denote the total roundoff noise affecting the M outputs of the decimated analysis bank (i.e., outputs of $\mathbf{T}$). Using the standard fixed-point roundoff noise model of Sec. 9.2, estimate the variance of $\epsilon_k(n)$.

9.11. Let $\mathbf{x}(n)$ be a vector WSS random process with autocorrelation matrix $\mathbf{R}(k)$ and power spectral density matrix $\mathbf{S}(e^{j\omega})$. (These were defined in Appendix B. Note that if $\mathbf{x}(n)$ is $M \times 1$ then $\mathbf{R}(k)$ and $\mathbf{S}(e^{j\omega})$ are $M \times M$ matrices.) Show that $\mathbf{S}(e^{j\omega})$ is a positive semidefinite matrix for all ω. (*Hint.* Somehow try to relate this to a scalar WSS process $t(n)$, and use the fact that its power spectrum is nonnegative).

9.12. In this problem we consider the joint behavior of two WSS random process $x(n)$ and $y(n)$. Assume that they have zero mean.

a) Let the power spectra $S_{xx}(e^{j\omega})$ and $S_{yy}(e^{j\omega})$ be non overlapping, i.e., $S_{yy}(e^{j\omega})S_{xx}(e^{j\omega}) = 0$ for all ω. Show that this does not in general imply that the two random processes $x(n)$ and $y(n)$ are uncorrelated.

b) Suppose $x(n)$ and $y(n)$ are jointly WSS. Show then that the condition $S_{yy}(e^{j\omega})S_{xx}(e^{j\omega}) = 0$ does imply that the two processes are uncorrelated. (*Hint.* The result of Problem 9.11 might help!)

c) Suppose $x(n)$ and $y(n)$ are generated as follows,

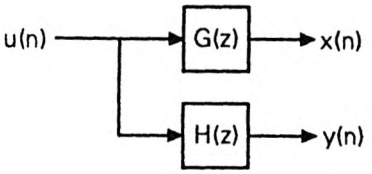

Figure P9-12

where $u(n)$ is a WSS process. Show that $x(n)$ and $y(n)$ are joinlty WSS. Hence show that if the filters $H(e^{j\omega})$ and $G(e^{j\omega})$ are nonoverlapping [that is, $H(e^{j\omega})G(e^{j\omega}) = 0$ for all ω], and $u(n)$ has zero-mean, then $x(n)$ and $y(n)$ are uncorrelated.

Note. If the zero-mean assumption is not true, the above statements should be modified by replacing "uncorrelated" with "orthogonal" everywhere.

9.13. Consider an M channel analysis bank $H_k(z), 0 \le k \le M-1$. Let the polyphase matrix $\mathbf{E}(z)$ be lossless with $\widetilde{\mathbf{E}}(z)\mathbf{E}(z) = \mathbf{I}$. In Sec. 6.2.2 we showed that this implies $\sum_n |h_k(n)|^2 = 1$ for each k, that is, each analysis filter has unit energy. Give a second proof of this using Theorem 9.3.1.