Chapter 11

TRADE-OFFS IN SENSITIVITY, COMPONENT SPREAD AND COMPONENT TOLERANCE IN ACTIVE FILTER DESIGN

George Moschytz
Swiss Federal Institute of Technology

11.1. Introduction

The concept of sensitivity in analog circuit design has taken on increasing importance with the inclusion of analog and mixed-mode (i.e. analog and digital) circuits on an integrated circuit (IC) chip. Spectacular as the continued trend towards complex “systems on a chip” may be, the achievable accuracy of analog component values remains poor, and is not likely to improve significantly in the near future. Although it is true that a certain amount of tuning and trimming is possible by switching critical resistor and/or capacitor arrays to “ball-park” values, the achievable accuracy of the component values remains quite limited. Whereas in discrete-component design it is “only” a question of cost whether 1% or even 0.1% components are used to achieve critical specifications, on-chip such accuracy is attainable only by laborious and cost-intensive (laser, sand-blasting, etching, etc.) operations. An alternative is to live with the high component tolerances and to try to find circuits that are as insensitive to component values as possible. In active-RC filter design, for example, insensitive circuits do exist, but often a satisfactory degree of tolerance insensitivity is paid for by an increase in component spread. An increase in component spread, however, generally increases component tolerance, which in turn decreases functional accuracy and performance, thus closing a typical vicious cycle so often encountered in analog circuit design. A trade-off is clearly in order, if not inevitable. It is this trade-off loop that is the subject of this chapter. Because sensitivity and noise are somehow, and somewhat elusively related, we shall, toward the end of the chapter, briefly include this additional very important facet of analog design in our discussion.
11.2. Basics of Sensitivity Theory

The relative sensitivity of a function \( F(x) \) to variations of a variable \( x \) is defined as

\[
S^F_x = \frac{\frac{dF}{dx}}{\frac{d}{dx}x} = \frac{\frac{dF}{dx}}{F} = \frac{\frac{d\ln[F(x)]}{dx}}{\frac{d\ln x}{dx}}
\]  

(11.1)

This expression provides a value for the relative change of the function \( F(x) \) to a relative change of a parameter \( x \), on which \( F(x) \) depends. Although the absolute sensitivity of \( F(x) \) to \( x \), that is,

\[
F' = \frac{dF}{dx}
\]  

(11.2)

and the semi-relative sensitivity

\[
S^F_x = \frac{\frac{dF}{dx}}{x}
\]  

(11.3)

may often be of more relevance to a practical problem, it is the ease of using the sensitivity expression given by (11.1), and the simplicity with which the other two, that is, (11.2) and (11.3), can be derived from it, that is responsible for the importance of the relative sensitivity. This ease of usage is manifest by the table of sensitivity relations shown in Table 11.1. Most expressions in deterministic filter-theory can be broken down into simple expressions, for which the relationships listed in Table 11.1 apply.

We now consider the voltage or current transfer function of an \( n \)th-order transfer function, \( T(s) \):

\[
T(s) = \frac{N(s)}{D(s)} = \frac{b_m s^m + b_{m-1} s^{m-1} + \cdots + b_1 s + b_0}{a_n s^n + a_{n-1} s^{n-1} + \cdots + a_1 s + a_0}
\]  

(11.4)

where \( s = \sigma + j\omega \) is the complex frequency and \( N(s) \) and \( D(s) \) are polynomials in \( s \) with real coefficients \( a_i \) and \( b_j \). Expressing \( N(s) \) and \( D(s) \) in their factored form, we obtain the zeros \( z_i \) and poles \( p_j \) of the transfer function, that is,

\[
T(s) = K \frac{\prod_{i=1}^{m} (s - z_i)}{\prod_{j=1}^{n} (s - p_j)}
\]  

(11.5)

To obtain the frequency response of the filter in the steady state, we assume a sinusoidal input signal by letting \( s = j\omega \) in (11.5) and obtain:

\[
T(j\omega) = K \frac{\prod_{i=1}^{m} (j\omega - z_i)}{\prod_{j=1}^{n} (j\omega - p_j)} = |T(j\omega)| \cdot e^{j\phi(\omega)}
\]  

(11.6)
Taking the logarithm of \( T(j\omega) \), we obtain

\[
\ln T(j\omega) = \ln |T(j\omega)| + j \arg T(j\omega)
= \alpha(\omega) + j\phi(\omega) \tag{11.7}
\]

where \( \alpha(\omega) \) and \( \phi(\omega) \) are the gain and phase response of the filter in nepers and degrees, respectively.

Using the sensitivity relations in Table 11.1, we can readily express the sensitivity of \( T(s) \) with respect to some component \( x \) in terms of the poles and zeros, namely, with (11.1), (11.3) and (11.5)

\[
S_x^T(s) = S_x^K - \sum_{i=1}^{m} \frac{S_x^{z_i}}{s - z_i} + \sum_{j=1}^{n} \frac{S_x^{p_j}}{s - p_j} \tag{11.8}
\]

where the so-called root sensitivity of a root \( q_v \) (where \( q_v \) is a pole or zero) is given by the semi-relative sensitivity:

\[
S_x^{q_v} = \frac{d q_v}{d x} \tag{11.9}
\]

\( S_x^T(s) \) is often referred to as the transfer sensitivity which, in the form (11.8), is a partial fraction expansion in terms of the roots of \( T(s) \). As such, the root sensitivities in (11.8) are the residues of the transfer sensitivity, thus:

\[
S_x^{q_v} = \mp \text{Res} \left( S_x^T(s) \right)_{s=q_v} = \mp (s - q_v) S_x^T(s) \bigg|_{s=q_v} \tag{11.10}
\]
where the minus sign holds for a zero, the plus for a pole. Letting \( s = j\omega \) to obtain the sensitivity of the frequency response, we obtain from (11.1), (11.3), (11.6) and (11.7)

\[
S_x^{\varphi(\omega)} = \text{Re} \left[ S_x^T(j\omega) \right] = S_x^T(j\omega) \tag{11.11a}
\]

and

\[
S_x^{\varphi(\omega)} = \text{Im} \left[ S_x^T(j\omega) \right] = \varphi \cdot S_x^T(j\omega) \tag{11.11b}
\]

Thus, the amplitude and phase sensitivity function results directly from the real and imaginary part of the transfer sensitivity function \( S_x^T(s) \), for \( s = j\omega \).

Finally, for any pole \( p = \sigma_p + j\tilde{\omega}_p \) we have, for the radial pole frequency \( \omega_p \), and the pole \( Q, q_p \):

\[
\omega_p = \sqrt{\sigma_p^2 + \tilde{\omega}_p^2} \tag{11.12a}
\]

and

\[
q_p = \frac{\omega_p}{2\sigma_p} \tag{11.12b}
\]

Thus, with (11.9) and (11.10), it follows that

\[
\frac{\Delta p}{p} = \text{Re} \left[ \frac{\Delta p}{p} \right] + j \text{Im} \left[ \frac{\Delta p}{p} \right] = \frac{1}{p} \left\{ \text{Res} \: S_x^T(s) \right\}_{s=p} \tag{11.13a}
\]

and with the relations in Table 11.1, we obtain:

\[
\text{Re} \left[ \frac{\Delta p}{p} \right] = \frac{\Delta \omega_p}{\omega_p} \tag{11.13b}
\]

and

\[
\text{Im} \left[ \frac{\Delta p}{p} \right] = \frac{\Delta q_p/q_p}{\sqrt{4q_p^2 - 1}} \approx -\frac{1}{2q_p} \frac{\Delta q_p}{q_p} \tag{11.13c}
\]

Thus, also the variation of a pole, which can be expressed by the variation \( \omega_p \) and pole \( Q, q_p \), can be directly obtained from the transfer sensitivity, namely by computing its residue for \( s = p \). Conversely, the amplitude and phase sensitivity, and variations due to component tolerances, can be derived from the relative pole (and zero) variation \( \Delta p/p \) (and \( \Delta z/z \)). In general, the pole variation plays a more important role than the zero variation because it is also responsible for the filter (or system) stability, and because the pole variations will effect the filter passband in contrast to the zero variations which effect primarily the filter stopband.
11.3. The Component Sensitivity of Active Filters

We can express the relative variation of the transfer function $T(s)$ of an active RC filter in terms of pole and zero variations with the help of equations (11.1), (11.3) and (11.8). Thus,

$$\frac{\Delta T(s)}{T(s)} = \frac{\Delta K}{K} - \sum_{i=1}^{m} \left( \frac{z_i}{s-z_i} \right) \frac{\Delta z_i}{z_i} + \sum_{j=1}^{n} \left( \frac{p_j}{s-p_j} \right) \frac{\Delta p_j}{p_j} \quad (11.14)$$

Note that the quantities in parentheses depend on the poles and zeros of the initial transfer function $T(s)$. In general, these are given by the filter specifications and cannot be changed. This leaves the quantities $\Delta z_i/z_i$ and $\Delta p_j/p_j$ to be minimized in order to minimize the variation $\Delta T(s)/T(s)$. This quantity, in turn, contains both the amplitude variation $\Delta \alpha(\omega)$ and $\Delta \varphi(\omega)$ when $s$ is set equal to $j\omega$. This can really be seen from (11.1), (11.3) and (11.7), if we consider the variation of $T(j\omega)$ caused by the variation of a component $x$:

$$\frac{\Delta T(\omega)}{T(\omega)} = \frac{d \ln T(\omega)}{d \ln x} \frac{\Delta x}{x} = \left[ \frac{d \ln |T(\omega)|}{d \ln x} + j \frac{d \varphi(\omega)}{d \ln x} \right] \frac{\Delta x}{x}$$

$$= \left( \frac{d \alpha(\omega)}{d \ln x} + j \frac{d \varphi(\omega)}{d \ln x} \right) \frac{\Delta x}{x}$$

$$= \left( S_x^{\alpha(\omega)} + j S_x^{\varphi(\omega)} \right) \frac{\Delta x}{x} \quad (11.15)$$

It follows that the amplitude variation $\Delta \alpha(\omega)$ and the phase variation $\Delta \varphi(\omega)$ due to the variation of a component $x$ is given by

$$\Delta \alpha(\omega) = S_x^{\alpha(\omega)} \frac{\Delta x}{x} = \text{Re} \left\{ \frac{\Delta T(\omega)}{T(\omega)} \right\} = \frac{\Delta |T(\omega)|}{|T(\omega)|} \quad (11.16a)$$

and

$$\Delta \varphi(\omega) = S_x^{\varphi(\omega)} \frac{\Delta x}{x} = \text{Im} \left\{ \frac{\Delta T(\omega)}{T(\omega)} \right\} \quad (11.16b)$$

Relating this result to the expression given by (11.14), and considering the variation only in the vicinity of a dominant pole $p_{\text{dom}}$ (which is generally in the passband frequency range near the cut-off frequency) we obtain:

$$\frac{\Delta T(\omega)}{T(\omega)} = \alpha(\omega) + j \varphi(\omega) \approx \left( \frac{p_{\text{dom}}}{j\omega - p_{\text{dom}}} \right) \frac{\Delta p_{\text{dom}}}{p_{\text{dom}}} \quad (11.17)$$

As mentioned above, the quantity in parentheses is given by filter specifications, and the variation of the dominant pole is given by (12.13). It is this pole variation, and to a lesser degree that of the other non-dominant poles, that can
be minimized in order to minimize the effect on amplitude and phase caused by component variations. In what follows, we shall examine in somewhat more detail, how the pole variation (dominant or otherwise) is effected by component tolerances and variations. To do so, it follows from (12.13) that we must examine in more detail the variations $\Delta \omega_p/\omega_p$ and $\Delta q_p/q_p$ (the absolute value of $\omega_p$ and $q_p$ being given by the filter specifications).

From (11.1), it follows that

$$\Delta \omega_p/\omega_p = \sum_{i=1}^{r} S_{R_i}^{\omega_p} \Delta R_i/R_i + \sum_{j=1}^{c} S_{C_j}^{\omega_p} \Delta C_j/C_j + \sum_{k=1}^{g} S_{\beta_k}^{\omega_p} \Delta \beta_k/\beta_k$$

(11.18)

where we assume $r$ resistors, $c$ capacitors, and $g$ amplifiers with gain $\beta_k$ ($k = 1, \ldots, g$) in the filter network. In general, there will be only one or two amplifiers generating a complex-conjugate pole pair. Furthermore, most well-tried circuits are characterized by the fact that the pole frequency can be made to be independent of gain, thereby eliminating the third summation in (11.18). The first two summations will depend on the technology used to fabricate the active RC filter. If a technology is used that permits close tracking, with temperature, say, of the resistors and capacitors, so that within close limits $\Delta R_i/R_i \equiv \Delta R/R$ and $\Delta C_j/C_j \equiv \Delta C/C$, then it can readily be shown that (11.18) simplifies to

$$\Delta \omega_p/\omega_p = -\left(\frac{\Delta R}{R} + \frac{\Delta C}{C}\right)$$

(11.19)

This quantity relies for its minimization on the compensation of temperature coefficients (TCR and TCC), aging characteristics and other effects influencing the resistor and capacitor values.

As in (11.18), the relative $q_p$ variation is given by:

$$\Delta q_p/q_p = \sum_{i=1}^{r} S_{R_i}^{q_p} \Delta R_i/R_i + \sum_{j=1}^{c} S_{C_j}^{q_p} \Delta C_j/C_j + \sum_{k=1}^{g} S_{\beta_k}^{q_p} \Delta \beta_k/\beta_k$$

(11.20)

In contrast to the expression given by (11.18), in which the $\omega_p$ sensitivity to gain ($\beta_k$) variations generally plays an insignificant role (if it plays a role at all), the situation is exactly reversed here. The variation of the resistors and capacitors will effect $q_p$ very little, if at all – as in the case of tracking components. The sensitivity of $q_p$ to gain, however, cannot possibly be eliminated because, as we shall see below, it is through the gain that the complex-conjugate poles, necessary for any kind of filter selectivity, are obtained. This can be illustrated by considering a second-order, single-amplifier active RC filter, often referred to as a single-amplifier biquad (SAB).
Consider, for example, the transfer function of a second-order bandpass filter:

\[ T(s) = K \frac{\omega_p s}{s^2 + 2\sigma_p s + \omega_p^2} = K \frac{\omega_p s}{s^2 + (\omega_p/\alpha_p)s + \omega_p^2} \]  \hspace{1cm} (11.21)

Assume that the center frequency \( f_p \) is 500 Hz and the 3 dB bandwidth \( B \) is 50 Hz. It can then readily be shown that \( \omega_p = 2\pi \cdot 500 \text{ rad/sec} \) and that \( 2\sigma_p = 2\pi \cdot B = 2\pi \cdot 50 \text{ rad/sec} \). Introducing the pole \( \alpha_p \), we obtain \( \alpha_p = 10 \). Consider now, the case that our bandpass filter is realized by a second-order passive RC network. It can be shown that in this case, a \( \alpha_p \) equal to ten is impossible to achieve. In fact, the pole \( \alpha_p \) of any pole pair realized by a passive RC network is always less than 0.5. We indicate this by designating the pole \( \alpha_p \) of a pole pair realized by a passive RC network by \( \alpha_{p*} \). Thus, by definition

\[ \alpha_{p*} > 0.5 \]  \hspace{1cm} (11.22)

It follows from the example above that with a passive RC second-order bandpass filter, the attainable 3 dB bandwidth will always be larger than 1000 Hz. Thus, the filter selectivity, which in the bandpass case is precisely the ratio of center frequency to 3 dB bandwidth, is so extraordinarily poor as to be essentially useless.

How then does an active RC biquad, say with one amplifier of gain \( \beta \), achieve useful selectivity, that is, a pole \( \alpha_p \) which is larger than 0.5 (such as in our example, \( \alpha_p = 10 \))? The answer is by inserting the passive RC network into a negative or positive-feedback loop, depending on whether the gain \( \beta \) is inverting or non-inverting. (Obviously the topology of the RC network must take the polarity of the amplifier into account.)

If we have negative feedback, the passive RC pole \( \alpha_p \), \( \alpha_{p*} \) will be increased to the desired value \( \alpha_{p*} \) by an expression of the form:

\[ \alpha_p \approx \alpha_{p*} \cdot \beta^i \]  \hspace{1cm} (11.23)

where \( i \) is either 0.5 or unity, depending on the class of biquad used. In any case, the gain \( \beta \) required to obtain \( \alpha_{p*} \) is proportional to \( \alpha_{p*}/\alpha_p \). For our example, for which \( \alpha_p = 10 \), and a class of biquad for which \( i = 1 \), the minimum gain \( \beta \) required will be 10/0.5 = 20; depending on how small \( \alpha_{p*} \) is, the required gain \( \beta \) may also be much larger. Clearly, since amplifier gain cannot be made arbitrarily large, particularly if the pole frequency is increased and the dissipated power is to be limited and – typically, as small as possible – it is important to design the passive RC network such that \( \alpha_{p*} \) is as large as possible, that is, as close to 0.5 as possible.

If the gain of the active RC filter is obtained by a feedback amplifier, where \( \beta \) is the closed-loop gain and \( A \) the open-loop gain, it can readily be shown
that the relative variation of $q_p$ to variations of gain will be:

$$\frac{\Delta q_p}{q_p} = i \frac{\Delta \beta}{\beta}$$

$$\approx i \left( \frac{\beta}{1 + LG} \right) \frac{\Delta A}{A} = i \left( \frac{q_p/\hat{q}_p}{1 + LG} \right)^{1/i} \frac{\Delta A}{A}$$

(11.24)

where LG is the loop gain of the feedback amplifier and $i$ equals 0.5 or unity, depending on the biquad class used. Thus, also the variation of $q_p$ caused by amplifier variations will be inversely proportional to $q_p$, meaning that also with respect to the variations of gain (and, incidentally, of other components as well) $\hat{q}_p$ should be made as large as possible, that is, as close to 0.5 as the RC circuit, that is, the component spread, will permit. How this is achieved will be discussed in the next section. First, however, we shall examine the case for the RC network in a positive-feedback loop, that is, for $\beta$ realized by a non-inverting amplifier.

For positive feedback, it can be shown that the relationship between the desired $Q$, $q_p$, and the necessary gain $\beta$ to achieve it, is:

$$q_p = \frac{k_1}{k_2 - \beta k_3}$$

(11.25)

where $k_1$, $k_2$ and $k_3$ depend only on the passive RC network in the feedback loop. Note that the pole $Q$ of the passive RC network is given by:

$$\hat{q}_p = q_p (\beta = 0) = \frac{k_1}{k_2}$$

(11.26)

Calculating the variation of $q_p$ to variations of $\beta$, it follows from the sensitivity relations in Table 11.1 that:

$$\frac{\Delta q_p}{q_p} = \left( \frac{q_p}{\hat{q}_p} - 1 \right) \frac{\Delta \beta}{\beta}$$

(11.27)

Thus, in the positive-feedback case also, it follows that in order to minimize variations of $q_p$ due to variations of gain and other components, the pole $Q$ of the passive RC network, $\hat{q}_p$, should be as close to 0.5 as possible.

At this point it is of interest to compare the sensitivity of positive and negative feedbacks based filters. From (11.22) and (11.25), it follows that

\[\text{Negative feedback: } S_{\hat{q}_p}^{q_p} = i \quad i = 0.5, 1 \quad (11.28a)\]

\[\text{Positive feedback: } S_{\hat{q}_p}^{q_p} = \frac{q_p}{\hat{q}_p} - 1 \quad (11.28b)\]
At first glance, it would seem that negative-feedback based circuits have a much smaller sensitivity to gain – and other component – variations than those based on positive feedback. This conclusion is misleading however, if the gain \( \beta \) is obtained as the closed-loop gain of a feedback-based amplifier. To see this we briefly examine the sensitivity of closed-loop gain \( \beta \) to open-loop gain \( A \) in a typical operational amplifier. Whether used in the inverting or non-inverting mode, this sensitivity is essentially given by

\[
S^\beta_A \approx \left. \frac{1}{1 + LG} \frac{\Delta A}{A} \right|_{LG \gg 1} \approx \frac{1}{LG} \frac{\Delta A}{A}
\]  

(11.29)

Furthermore, the loop gain \( LG \) is readily shown to be approximately

\[
LG \approx \frac{A}{\beta}
\]

(11.30)

so that

\[
\frac{\Delta q_p}{q_p} = S^{q_p}_\beta \cdot S^\beta_A \Delta A / A
\]

(11.31)

which with (11.29) and (11.30) becomes

\[
\frac{\Delta q_p}{q_p} \approx \left[ \beta \cdot S^{q_p}_\beta \right] \frac{\Delta A}{A^2} = \Gamma^{q_p}_\beta \cdot \frac{\Delta A}{A^2}
\]

(11.32)

In other words, the variation of \( q_p \) depends on \( \Gamma^{q_p}_\beta \) which is the gain-sensitivity product (GSP) of \( q_p \) with respect to \( \beta \). This changes the conclusion from above quite drastically. If we consider the GSP, which we must, rather than the sensitivity, then instead of (11.28), we now have:

**Negative feedback:**  \( \Gamma^{q_p}_\beta = i \beta \approx i A \quad i = 0.5, 1 \) \hspace{1cm} (11.33a)

**Positive feedback:**  \( \Gamma^{q_p}_\beta \approx \beta \left( \frac{q_p}{q_p} - 1 \right) \)

(11.33b)

The reason we let \( \beta \approx A \) in (11.33a) is that negative-feedback circuits generally operate with the gain in, or close to, the open-loop mode. On the other hand, positive-feedback networks must restrict the closed-loop gain \( \beta \) to values close to unity, that is, from (11.25)

\[
\beta < \frac{\kappa_2}{\kappa_3}
\]

(11.34)

which is generally between unity and two. If we consider the open-loop gain of an op-amp to be in the order of 100 to 1000, it readily follows that the GSP for
positive feedback is certainly comparable to, if not considerably smaller than, that of negative feedback. This explains why the positive-feedback biquads are, in fact, more frequently used than their negative-feedback counterparts. (Incidentally, it should be pointed out already here, and will be discussed briefly in Section 11.7, that minimizing the GSP of an active RC biquad filter also reduces the output thermal noise.)

For higher order filters, it is often difficult to factor the polynomials \( N(s) \) and \( D(s) \) in equation (11.4) into complex-conjugate pairs, or, equivalently, into expressions involving the radian pole frequencies \( \omega_p j \) and pole \( Q_s, q_{pj} \), where \( j = 1, \ldots, n/2 \). (This assumes that \( n \) is even. If it is not, then there is still an additional negative real term.) In this case, \( T(s) \) is given in the form of (11.4) and its variation is expressed in terms of coefficient sensitivities thus:

\[
\frac{\Delta T(s)}{T(s)} = \sum_{i=0}^{m} S_{bi}^{T(s)} \frac{\Delta b_i}{b_i} + \sum_{j=0}^{n} S_{aj}^{T(s)} \frac{\Delta a_j}{a_j}
\]

\[
= \sum_{i=0}^{m} S_{bi}^{N(s)} \frac{\Delta b_i}{b_i} + \sum_{j=0}^{n} S_{aj}^{D(s)} \frac{\Delta a_j}{a_j} \tag{11.35}
\]

The sensitivity terms in (11.35), which represent the sensitivity of the transfer function to coefficient variations, and are themselves frequency-dependent functions, are given by the filter specifications and the resulting transfer function. It is therefore the sensitivity of the coefficients to circuit components that can be minimized, that is,

\[
\frac{\Delta b_i}{b_i} = \sum_{\mu=1}^{r} S_{R_{b\mu}^{j}} \frac{\Delta R_{b\mu}}{R_{b\mu}} + \sum_{v=1}^{c} S_{C_{bi}^{j}} \frac{\Delta C_{v}}{C_{v}} + \sum_{k=1}^{g} S_{\beta_{k}^{j}} \frac{\Delta \beta_{k}}{\beta_{k}} \tag{11.36a}
\]

and

\[
\frac{\Delta a_j}{a_j} = \sum_{\mu=1}^{r} S_{R_{a\mu}^{j}} \frac{\Delta R_{a\mu}}{R_{a\mu}} + \sum_{v=1}^{c} S_{C_{ai}^{j}} \frac{\Delta C_{v}}{C_{v}} + \sum_{k=1}^{g} S_{\beta_{k}^{j}} \frac{\Delta \beta_{k}}{\beta_{k}} \tag{11.36b}
\]

It can be shown that the minimization of these quantities entails a procedure identical to that of maximizing the RC pole \( Q_s \), which we have designated by \( \hat{q}_p \). The problem is that for higher order filters, breaking the corresponding polynomials \( N(s) \) and \( D(s) \) into these roots and root pairs becomes quite intractable, so that the \( \hat{q}_p \) quantities cannot be analytically obtained. Nevertheless, the procedure for maximizing \( \hat{q}_p \) outlined in the next section is valid also for higher order networks, even if the individual \( \hat{q}_p \) values cannot be identified.
11.4. Filter Selectivity, Pole $Q$ and Sensitivity

Consider the second-order passive RC bandpass filter shown in Figure 11.1. The transfer function is given by

$$\hat{T}(s) = K \frac{\omega_p s}{s^2 + (\omega_p/\hat{q}_p)s + \omega_p^2}$$

(11.37)

where

$$K = \sqrt{\frac{R_2C_2}{R_1C_1}}$$

(11.38a)

$$\omega_p = \frac{1}{\sqrt{R_1R_2C_1C_2}}$$

(11.38b)

$$\hat{q}_p = \frac{1}{x + 1/x + \varepsilon}$$

(11.38c)

$$x = \sqrt{\frac{R_1C_1}{R_2C_2}}$$

$$\varepsilon = \sqrt{\frac{R_1C_2}{R_2C_1}}$$

(11.38d)

To examine the bounds on $\hat{q}_p$, we consider its inverse

$$\frac{1}{\hat{q}_p} = x + \frac{1}{x} + \varepsilon$$

(11.39)

The well-known function $x + x^{-1}$ has a minimum of two for $x = 1$. Since (11.39) includes the term $\varepsilon$, this minimum can never be reached; its value will always be larger than two, or, conversely, the value of $\hat{q}_p$ will always be smaller than 0.5. Although this is a specific example, the result is representative for a basic theorem with regard to passive RC networks. This theorem states that the poles of an RC network are restricted to being single and on the negative-real

![Figure 11.1. Second-order passive RC bandpass filter.](image)
Chapter 11

axis in the complex frequency plane. Thus, for the two negative real poles \( P_1 \) and \( P_2 \) in Figure 11.2(a), we obtain a polynomial

\[
D(s) = (s + P_1)(s + P_2) = s^2 + (P_1 + P_2)s + P_1 P_2
\]

\[
= s^2 + \frac{\omega_p^2}{\hat{q}_p} s + \omega_p^2
\]  

(11.40)

where

\[
\omega_p^2 = P_1 P_2
\]  

(11.41a)

and

\[
\hat{q}_p = \frac{\sqrt{P_1 P_2}}{P_1 + P_2}
\]  

(11.41b)

so that

\[
\frac{1}{\hat{q}_p} = \sqrt{\frac{P_1}{P_2}} + \sqrt{\frac{P_2}{P_1}}
\]  

(11.41c)

Thus, \( \hat{q}_p \) would reach its unattainable maximum value of 0.5 only if it were possible to make \( P_1 = P_2 \) which, as we stated above, would mean that we have a double pole on the negative-real axis, and which, with a passive RC network, is not possible. A glance at our example above and of equations (11.38c and d) shows that such a double pole would occur when \( \varepsilon \), that is, the ratio \( R_1/R_2 \) or \( C_2/C_1 \), approaches zero. This, of course, demands a non-realizable spread of the resistors and/or capacitors. How large a spread of these components is acceptable in order to minimize \( \varepsilon \) and to approach a double pole (or, in other words, to permit \( \hat{q}_p \) to approach 0.5) is a question of the technology used. It represents one of the fundamental trade-offs of low-sensitivity active RC filter – and oscillator – design.

There is another way of illustrating the importance of trying to approach a double pole, (or the value of \( \hat{q}_p = 0.5 \)) with the passive RC network in a SAB
with closed-loop gain $\beta$. Let us assume that the desired bandpass filter biquad, has the transfer function

$$T(s) = \frac{\omega_p s}{s^2 + (\omega_p/q_p)s + \omega_p^2}$$  \hspace{1cm} (11.42)$$

with the pole–zero pattern in the complex frequency, or s-plane, as shown in Figure 11.2(b). Note that the only difference in the transfer function of the passive RC bandpass filter $\hat{T}(s)$ (see equation (11.37)) and that of the desired active RC bandpass filter $T(s)$ in equation (11.42) is in the pole $Q$, namely $\hat{q}_p$ and $q_p$, respectively. By definition, $\hat{q}_p < 0.5$ and, for any useful application, we can assume that $q_p > \hat{q}_p$. This difference will be apparent in the shape of the amplitude response as shown in Figure 11.3. For the passive RC case, the $3 \, \text{dB}$ bandwidth $B$ will be larger than $2f_p$, for the desired active RC filter the $3 \, \text{dB}$ bandwidth $B$ will, typically be substantially smaller than $f_p$, thereby providing a filter selectivity that is correspondingly higher. To consider, now, how we obtain the complex-conjugate pole pair of Figure 11.2(b) from the combination of a passive RC network whose poles are as in Figure 11.2(a), combined with an amplifier with (inverting or non-inverting) gain $\beta$, we consider a typical root locus of the resulting feedback network with respect to $\beta$. This is shown in Figure 11.4. Note that the effect of the gain $\beta$ is to create a pair of so-called closed-loop poles $p, p^*$ from the open-loop poles $P_1$ and $P_2$. It does so by moving the closed-loop poles along the root locus, which will differ according to such factors as the sign of the feedback amplifier (positive or negative), and the type of passive RC network in the feedback path. However, the common feature for all the possible biquad root loci will be that with increasing gain $\beta$, the closed-loop poles will first be shifted towards each other on the negative-real frequency axis, away from the negative-real open-loop poles, and towards a coalescence point (point $C$ in Figure 11.4). From there, with increasing gain $\beta$, a pair of complex-conjugate poles will be generated, with the required $q_p$ necessary to satisfy the transfer function in equation (11.42). Clearly, the further apart the two open-loop poles $P_1$ and $P_2$ are on the negative-real axis in the $s$-plane, the more gain $\beta$ is required to

![Figure 11.3. Amplitude response of second-order filter: (a) passive RC, (b) active RC (or LCR).](image-url)
reach the final values, namely the closed-loop poles \( p \) and \( p^* \) on the root locus. Designing open-loop poles far apart is wasteful of gain, since (almost) up to the coalescence point, the passive RC open-loop poles can be obtained by a passive RC network, that is, without any gain at all. Open-loop poles that are far apart on the negative-real axis are also detrimental to the stability of the filter, since the higher the required closed-loop gain, the smaller the stabilizing loop gain (see equation (11.29)), that is, the higher the sensitivity of the closed-loop gain to variations in the open-loop gain. In short, the highest possible \( q_p \), that is, open-loop poles as close together as possible on the negative-real axis, will minimize the closed-loop gain necessary to obtain the prescribed \( q_p^* \).

11.5. Maximizing the Selectivity of RC Networks

In the case of a second-order passive RC network, the maximum selectivity that can possibly be obtained is when the pole \( Q \) of the network, namely \( \hat{q}_p \), approaches 0.5. This is equivalent to obtaining a double pole on the negative-real axis in the \( s \)-plane which, as we have seen in the preceding section, can be achieved only in the limit by an infinite component spread; it is therefore impossible to actually realize in practice. We may now ask whether there is a simple way of obtaining a double pole \( (\hat{q}_p = 0.5) \) if we permit a simple active device to be included in the circuit? The answer is that by inserting a unity-gain amplifier in the RC ladder of Figure 11.1, as in Figure 11.5, we readily obtain a negative-real double pole.

We then obtain the transfer function:

\[
T(s) = K \frac{\omega_p s}{(s + \omega_p)^2} = K \frac{\omega_p s}{s^2 + 2\omega_p s + \omega_p^2}
\]  

\((11.43)\)
where, referring to Figure 11.5, \( \omega_p = (RC)^{-1} \) and the corresponding pole \( Q \), \( q_p = 0.5 \).

The task of the unity-gain buffer is to prevent the second RC ladder section from loading the first. Clearly, an \( n \)-th-order ladder network with \( n \) equal poles would, therefore, require \( n - 1 \) buffer amplifiers with one in between each RC L-section of the ladder network. Similar decoupling of the individual L-sections of an RC ladder network can be achieved by impedance scaling upwards (or “impedance tapering”, as we shall call it), the second L-section by a factor \( \rho \), as shown in Figure 11.6. From equations (11.38c and d) it follows that in this case \( x = 1 \) and \( \varepsilon = 1/\rho \), thus

\[
\hat{q}_p = \frac{1}{2 + 1/\rho}
\]

which for \( \rho \to \infty \) approaches 0.5.

The actual plot of \( \hat{q}_p \) versus \( \rho \) is shown in Figure 11.7, where we see that already for a value of \( \rho = 3 \), \( \hat{q}_p = 0.43 \), and for \( \rho = 10 \), \( \hat{q}_p = 0.48 \). For a higher order RC ladder network the third L-section would be impedance scaled by a factor \( \rho^2 \) and so on, with the \( n \)-th section being impedance scaled by a factor \( \rho^{n-1} \). Because each L-section is impedance scaled by a power of \( \rho \) higher than the previous section, its impedance level is tapered by an increasing power of \( \rho \) from left to right. The higher the impedance tapering factor \( \rho \), the closer the
negative-real poles of the RC ladder are clustered together on the negative-real axis of the $s$-plane, and the closer individual pole pairs are to having a $\hat{q}_p$ approaching 0.5. Thus, for the reasons given earlier, impedance tapering any RC network will minimize sensitivity to component variations by increasing the $\hat{q}_p$ of negative-real pole pairs. This is true not only for RC ladder networks, but for more general RC networks, such as bridged-T and twin-T networks, as well.

Consider for example, the twin-T network shown in Figure 11.8(a). As shown, the transfer function is:

\[
\hat{T}(s) = \frac{s^2 + \omega_N^2}{s^2 + (\omega_N/\hat{q}_N)s + \omega_N^2}
\]  

(11.45)

where the notch frequency $\omega_N$ and the corresponding pole $Q$, $\hat{q}_N$ are, respectively,

\[
\omega_N = \frac{1}{RC}
\]  

(11.46a)

\[
\hat{q}_N = 0.25
\]  

(11.46b)

Broken up into two symmetrical sections as in Figure 11.8(b), and impedance scaling the right section by a factor $\rho$ (Figure 11.8(c)), we obtain the so-called

---

1 Note that the twin-T network is actually a third-order network, but that for the values shown, the negative-real pole and zero are canceled out.
potentially symmetrical twin-T, where the notch frequency $\omega_N$ remains as in equation (11.46a) but $\hat{q}_N$ becomes:

\[
\text{Twin-T: } \hat{q}_N \mid_{\text{Potentially symmetrical}} = \frac{1}{2\left(1 + \rho\right)} \tag{11.47}
\]

which for $\rho \gg 1$ approaches 0.5. For $\rho = 3$ we obtain $\hat{q}_N = 0.375$, and for $\rho = 10$ we obtain $\hat{q}_N = 0.45$.

Similarly, the bridged-T shown in Figure 11.9(a) has the transfer function:

\[
\text{Bridged-T: } \hat{T}(s) = \frac{s^2 + (\omega_N/q_Z)s + \omega_N^2}{s^2 + (\omega_N/\hat{q}_N)s + \omega_N^2} \tag{11.48}
\]

where

\[
\omega_N = \frac{1}{RC} \tag{11.49a}
\]

\[
q_Z = 1 \tag{11.49b}
\]

\[
\hat{q}_N = \frac{1}{3} \tag{11.49c}
\]

Deriving the potentially symmetrical bridged-T as shown in Figures 11.9(b)−(d), we obtain the same values for $\omega_N$ and $q_Z$ as above, but for the pole $Q$, $\hat{q}_N$ we obtain:

\[
\text{Bridged-T: } \hat{q}_N \mid_{\text{Potentially symmetrical}} = \frac{\rho}{1 + 2\rho} \tag{11.50}
\]
For $\rho = 3$ we obtain $\hat{q}_N = 0.43$, and for $\rho = 10$, $\hat{q}_N = 0.48$.

With the examples above, namely of the RC ladder, twin-T and bridged-T networks, we have seen that by impedance scaling (which becomes impedance tapering in the case of the ladder, and potential symmetry in the case of the twin- and bridged-T networks), we have introduced a type of figure-of-merit, namely $\hat{q}$, which must approach its upper bound of 0.5 in order to minimize sensitivity and maximize selectivity. Approaching this upper bound always entails an increase in component spread by impedance scaling. This impedance scaling is performed in order to provide an impedance mismatch between individual circuit sections such as to minimize the loading of one section of a network on the preceding section. The degree of mismatch attainable through impedance tapering depends on the degree of component spread permissible for a given technology. The resulting trade-off can be considered only in the context of a specific application and technological realization, but with any kind of frequency-selective circuit, including various kinds of oscillators, it is bound to come up during the design process.

### 11.6. Some Design Examples

Consider the second-order lowpass filter shown in Figure 11.10(a). The transfer function is given by

$$T(s) = \frac{\beta \omega_p^2}{s^2 + (\omega_p/q_p)s + \omega_p^2}$$

(11.51)
Figure 11.10. Second-order lowpass filter with ideal op-amp and voltage gain $\beta$, (a) with tapering factors $r$ and $\rho$ for resistors and capacitors respectively, (b) impedance tapered by $\rho = r = 4$, (c) non-tapered: $\rho = r = 1$, (d) non-tapered with $C_{\text{tot}} = 625$ pF.

where

$$\omega_p = \frac{1}{\sqrt{R_1 R_2 C_1 C_2}}$$ (11.52a)

$$q_p = \frac{\kappa_1}{\kappa_2 - \beta \kappa_3}$$ (11.52b)

$$\kappa_1 = \sqrt{R_1 R_2 C_1 C_2}$$ (11.52c)

$$\kappa_2 = R_1 C_1 + R_2 C_2 + R_1 C_2$$ (11.52d)

$$\kappa_3 = R_1 C_1$$ (11.52e)
and
\[ q_p = \frac{\sqrt{R_1 R_2 C_1 C_2}}{R_1 C_1 + R_2 C_2 + R_1 C_2} \]  
(11.53)

Letting
\[ R_1 = R; \quad C_1 = C; \quad R_2 = \rho R; \quad C_2 = \frac{C}{\rho} \]  
(11.54)

and
\[ \omega_0 = \frac{1}{RC} \]  
(11.55)

we obtain
\[ \omega_p = \omega_0 = \frac{1}{RC} \]  
(11.56)

and
\[ \beta = 2 + \frac{1}{\rho} - \frac{1}{q_p} \]  
(11.57)

Assuming the following filter specifications
\[ \omega_p = \omega_0 = 2\pi \cdot 86 \text{ kHz} \]
\[ q_p = 5 \]
\[ C = 500 \text{ pF} \]  
(11.58)

we obtain from (11.56) \( R = 3.7 \text{ k}\Omega \). Assuming an impedance tapering factor \( \rho = 4 \), we obtain, from (11.57), \( \beta = 2.05 \). The resulting filter is shown Figure 11.10(b). For the equivalent circuit with \( \rho = 1 \), we obtain \( \beta = 2.8 \) and the circuit of Figure 11.10(c). Figure 11.11(a) shows the amplitude response of this filter with ideal nominal components. Figures 11.11(b)–(g) show Monte Carlo runs with various combinations of resistor, capacitor and gain tolerances for the circuit tapered with \( \rho = 4 \), and non-tapered (\( \rho = 1 \)), respectively. Noting the difference in the ordinate scale, it is clear that the tapered circuit is significantly less sensitive to component tolerances than the non-tapered circuit.

We now consider the third-order lowpass filter shown in Figure 11.12. The corresponding transfer function is given by:
\[ T(s) = \frac{\beta a_0}{s^3 + a_2 s^2 + a_1 s + a_0} \]  
(11.59a)
where

\[ a_0 = \frac{1}{R_1 R_2 R_3 C_1 C_2 C_3} \]  \hspace{1cm} (11.59b)

\[ a_1 = \frac{R_1 (C_1 + C_2 + C_3) + R_2 (C_2 + C_3) + R_3 C_3 - \beta C_2 (R_1 + R_2)}{R_1 R_2 R_3 C_1 C_2 C_3} \]  \hspace{1cm} (11.59c)

\[ a_2 = \frac{R_1 R_2 C_1 (C_2 + C_3) + R_1 R_3 C_3 (C_1 + C_2) + R_2 R_3 C_2 C_3 - \beta R_1 R_2 C_1 C_2}{R_1 R_2 R_3 C_1 C_2 C_3} \]  \hspace{1cm} (11.59d)
Although $T(s)$ can be expressed in terms of a complex-conjugate pole pair with pole frequency $\omega_p$ and $Q$, $q_p$, as well as a negative-real pole $\gamma$, that is,

$$T(s) = \frac{\beta \omega_p^2 \gamma}{(s + \gamma) \left( s^2 + (\omega_p/q_p)s + \omega_p^2 \right)}$$  \hspace{1cm} (11.60)

it is generally very difficult to find the relationship between these quantities and the components (i.e. resistors, capacitors and gain $\beta$). It is therefore more convenient to examine the variations of the polynomial coefficients $a_i$ in terms of the component tolerances, that is,

$$\frac{\Delta a_i}{a_i} = \sum_{\mu=1}^{3} S_{R_\mu} a_i \frac{\Delta R_\mu}{R_\mu} + \sum_{v=1}^{3} S_{C_v} a_i \frac{\Delta C_v}{C_v} + S_{\beta} a_i \frac{\Delta \beta}{\beta} \hspace{1cm} i = 0, 1, 2 \hspace{1cm} (11.61)$$

With the sensitivity relations given in Table 11.1, it can readily be shown that the variation of these coefficients can be minimized by tapering the third-order ladder network [1]. However, with networks of higher than second order, a tapering factor $\rho$ can generally not be arbitrarily selected, because the resulting component values may turn out to be non-realizable (e.g. negative or complex). It can be shown that tapering the impedance of only the capacitors, while selecting the two resistors $R_2$ and $R_3$ to be equal (or vice versa), is sufficient to desensitize the circuit effectively from the effect of component tolerances.

Consider, for example, a third-order Chebyshev lowpass filter with coefficients

$$a_0 = 0.749 \cdot 10^{17}$$
$$a_1 = 0.341 \cdot 10^{12}$$
$$a_2 = 0.59 \cdot 10^{6} \hspace{1cm} (11.62a)$$
or, with the equivalent dc gain and pole parameters:

\[ K = 2 \]
\[ \omega_p = 0.5037 \cdot 10^6 \]
\[ q_p = 1.7 \]
\[ \gamma = 0.295 \cdot 10^6 \]  

This corresponds to a filter with a maximum ripple of 0.5 dB in the passband up to 75 kHz, and a minimum attenuation of 38 dB in the stopband above 300 kHz. The amplitude response of this circuit is shown in Figure 11.13(a) and Monte Carlo runs for 5% component tolerances shown for a non-tapered circuit (Figure 11.13(b)), and a capacitively-tapered circuit in Figure 11.13(c) and (d). Again, the efficacy of tapering in order to reduce the sensitivity to component tolerances is quite apparent from these curves.

11.7. Sensitivity and Noise

It has long been suspected (but never undisputedly proved) that low-sensitivity active RC filters are also low in output thermal noise. A recent publication substantiates this assumption [2]. In what follows, we further demonstrate with two design examples that biquads, designed for minimum
sensitivity to component tolerances using the impedance tapering methods outlined above, are also superior, in terms of low output thermal noise, when compared with standard designs. In Figure 11.14, the output noise for the circuit of Figure 11.10(c) (non-tapered) and Figure 11.10(b) (tapered with tapering factor = 4) is shown. In Figure 11.15, the output noise for the circuit in Figure 11.12
for non-tapered (a), tapered with $\rho_c = 3$ (b), and tapered with $\rho_c = 5$ (c), is shown. The improvement in output thermal noise reduction is considerable, and comes free of charge, in that, as we have seen above, it requires simply the selection of appropriate component values to implement impedance tapering. It has been shown that the same phenomenon holds also for low-sensitivity (i.e. impedance tapered) higher order filters. As with biquads, they are also low in output thermal noise when desensitized to component variations by the use of impedance tapering. Just how much tapering is possible depends on the permissible component spread and comprises one of the principle trade-offs dealt with in this chapter.

11.8. Summary and Conclusions

We show in this chapter that the sensitivity of active RC filters can be minimized by appropriate impedance mismatching in the form of impedance tapering for RC ladder networks, and potential symmetry for bridged-T and parallel-ladder (e.g. twin-T) networks. This procedure, however, invokes a design trade-off in that it automatically increases the spread between the resistors and capacitors, which is generally constrained by the technology used to manufacture these components. An acceptable trade-off, therefore, depends on the technology used, and on the circuit characteristics, in that the sensitivity to component tolerances may be reduced, while the component quality and their tolerances may actually be increased, by the component spread. The trade-off is worth considering and dealing with, however, because, within the limits of an acceptable component spread, the reduction in sensitivity and output noise of the resulting active RC filters is considerable.

References


Chapter 12

CONTINUOUS-TIME FILTERS

Robert Fox
University of Florida

12.1. Introduction

As the rest of this book illustrates, analog circuit design is ripe with compromises and trade-offs, some arising from fundamental limitations and some from practical realities. (Actually, many parameters are best pushed to their maximum limits; those just aren’t as interesting.) In the following, we will examine several fundamental and practical trade-offs in the design of continuous-time filters. Some issues will only be mentioned; a few others will be illustrated in more detail. We will consider trade-offs in filter design, in circuit topology and in strategies for filter tuning.

When we think about specifications for analog integrated signal processing elements, power dissipation, power-supply voltage, noise, filter accuracy, dynamic range, distortion are the first to spring to mind. However, integrated circuit (IC) chip area, manufacturing cost and yield, testability, and the availability of suitable technologies should not be ignored. Another issue, rarely explicitly considered in such lists of important parameters, is the ease and related cost of design: a simple, under-performing design that is well understood may be superior in practice to one that is too complicated or that takes too long to design. Even esthetic or marketing issues can be important. However, we shall focus here on more practical engineering issues.

Filtering is one of the most important functions in analog signal processing. While many filtering tasks now use digital signal processing, continuous-time filters are still important. Continuous-time filters are commonly used as equalizers for magnetic disk drives and for anti-aliasing and reconstruction filters, etc. Continuous-time filters would be used in many more applications if difficult trade-offs in their design did not limit their usefulness.

12.2. Filter-Design Trade-Offs: Selectivity, Filter Order, Pole Q and Transient Response

To enhance the ability to pass a signal of one frequency while rejecting another (selectivity), we increase the order (number of transfer-function poles) of the filter. Increasing filter order adds complexity, chip area and power.
High-selectivity filtering also usually requires pole pairs to be complex conjugate, with some pole pairs having high quality factor ($Q$). The concept of $Q$ is used in many different contexts; its fundamental definition relates total stored energy to energy loss. The $Q$ of a complex pole pair is the ratio of the pole magnitude to twice the real part. Increasing pole $Q$ extends the duration of transient step and impulse responses. Also, it is easy to show [1] that noise power in active filters increases in proportion to $Q$.

12.3. Circuit Trade-Offs

12.3.1. Linearity vs Tuneability

One key property of a signal processor is its linearity. If a system is linear, when sinusoidal signals are applied as inputs, the output consists only of sinusoids of the same frequencies as the input. The effect of the circuit on the magnitude and phase of each sinusoidal component as a function of frequency is the transfer function. Distortion is a measure of the relative output power at frequencies that were not present in the input. Since in a linear system, the filter transfer coefficients should stay constant as the signal amplitude is varied, linearity can also be characterized by measuring how much these coefficients vary from constancy.

12.3.2. Passive Components

Continuous-time filters can be implemented in a variety of ways. Classic RC active filters use operational amplifiers, with high (nominally infinite) negative-feedback loop-gain, to obtain high linearity using constant-valued passive components. As we will see, integrated continuous-time filters require electronic tuning to correct for manufacturing process variability, temperature and power-supply voltage variations. The dependence of the filter coefficients on passive components thus conflicts with the need for tuneability.

Until a few years ago, integrated circuit technologies were usually designed to optimize the performance of digital circuitry. The shrinking proportion of analog functionality in mixed-signal ICs did not justify including special options useful only for analog. However, with the increasing use of foundry processes, there is pressure to include more options to make processes more general. As digital processes themselves have become more complicated (they often now have four or more metal layers), the relative cost to add process steps decreases. Thus new technologies often allow special analog options to provide linear passive components.

The most important passive components in IC technologies are resistors and capacitors. On-chip resistors can be made using various materials, sometimes using layers that are already available in the IC. The resistance is determined by
the conductivity and geometry of the layer, and so is not usually tuneable, but it can be quite signal independent. High-value resistors usually require large chip areas, and are thus costly. High-resistivity layers that would allow smaller area tend to match poorly, which degrades performance or yield.

Standard CMOS technologies provide limited options for linear capacitors. MOS capacitors using the thin gate oxide of the MOSFET have high capacitance per unit area, but the capacitance is voltage dependent and thus nonlinear. However, the voltage dependence is reduced in the accumulation region of operation, so such MOSCAPs operated in accumulation can be used in low-precision applications, although biasing can be awkward. On the other hand, in applications where tuneability is more important, voltage-dependence may be considered an advantage. In such cases, it is more common to use the well-defined voltage-dependent capacitance of a reverse-biased PN-junction to form a varactor. The metal-oxide metal (MOM) capacitance between wiring layers is quite linear, but to minimize cross talk between layers, these are designed to have low capacitance per unit area. Special processing may allow linear capacitor options for MOM or poly-poly capacitors with relatively high capacitance per unit area.

For processing frequencies in the GHz range, it is practical to include on-chip inductors, although they require a lot of chip area. The inductance is mostly determined by the geometry of the windings, so these are usually not tuneable.

The values of these elements will vary from lot to lot, from wafer to wafer, from die to die and from one location on a die to another. In general, the more “closely related” two nominally matched components are to each other, the closer they will match. Thus lot-to-lot variations in a high-value resistor might average 30% or more, whereas the mismatch of two resistors, carefully laid out close together on the same chip, could match within 0.1%. As we will see, some filter coefficients are dimensionless and others have dimensions of time or frequency. Products of resistance and capacitance or inductance and capacitance determine the time or frequency coefficients. It is because capacitance values do not predictably track resistance or inductance values that continuous-time filters usually require post-fabrication tuning.

### 12.3.3. Tuneable Resistance Using MOSFETs: The MOSFET-C Approach

The passive elements just considered are all two-terminal elements, which can never be at once both linear and tuneable. For an element to be both linear and tuneable, it must have a separate control input. For example, a MOSFET operated in triode region can implement a voltage-variable resistor. Consider the simple square-law model for the MOSFET, which predicts that a current

\[ I_D = \frac{1}{2} \mu C_{\text{ox}} \frac{W}{L} \cdot (V_{\text{gs}} - V_t)^2 \]
in active (saturated) operation \((V_{ds} > V_{Dsat})\) and

\[
I_D = \mu C_{ox} \frac{W}{L} \cdot (V_{gs} - V_t - V_{ds}/2) \cdot V_{ds}
\]

in the triode region \((V_{ds} < V_{Dsat})\). To the extent that these relationships are exact, there are a variety of ways that MOSFETs can be used to implement tuneable filter elements.

For example, Figure 12.1 shows a simple MOSFET-C integrator in which a pair of matched MOSFETs, operated in triode, are used to implement an electronically tuneable resistor [2]. The differential op-amp holds their sources at the same potential, and the input voltage signal is applied differentially between the drains. A control voltage \(V_C\) applied to the gates is used to vary the resistance of the source-drain channel. The MOSFET currents are quadratically related to the input signal voltage, with constant, linear and second-order terms. However, when the currents are subtracted, the constant terms and the nonlinear second-order terms cancel. This is a general property that arises from symmetry: differential operation ideally allows cancelation of all even-order nonlinearities. In practice, of course, because of mismatch, balance is never perfect, and real circuits always produce some odd-order nonlinearities.

The MOSFET-C approach can achieve good linearity at moderate frequencies. However, it requires differential op-amps with common-mode feedback (CMFB). Also, the amps must be able to drive resistive loads, which is difficult in CMOS technologies. Also, as in any active-RC filter, op-amp stability requirements limit use of the MOSFET-C approach for high-frequency applications.

### 12.4 The Transconductance-C (Gm-C) Approach

Gm-C filtering avoids many of these limitations. In this approach, transistors are used, often with little or no feedback, to implement electronically tuneable
voltage-controlled current sources, along with nominally linear capacitors to implement filter elements.

The ideal transconductor would implement a linear voltage-controlled current source with a predictably adjustable transconductance that is independent of the signal amplitude, with zero input-output phase shift and zero output admittance over all relevant frequencies. In practice, we must accept thermal and flicker noise, limited dynamic range, minimum power-supply voltage, finite output admittance and parasitic phase errors. A wide variety of circuits have been proposed to implement transconductors [3]. Rather than attempt detailed comparison, this survey will compare a few different approaches based on trade-offs in their fundamental operating principles.

12.4.1. Triode-Region Transconductors

The triode-region square-law model predicts that a triode-region transconductor in which a constant $V_{ds}$ controls the transconductance $G_m$, and $V_{gs}$ is the input, such as shown in Figure 12.2, would be linear even for single-ended operation. Differential operation should further enhance linearity. However, triode-region operation requires a low $G_m/I$ ratio, with correspondingly high $V_{gs}$. This leads to large vertical electric fields in the transistors, which leads to significant odd-order nonlinearity. These odd-order errors are not improved by differential operation. The op-amp loops in the figure could be replaced by simpler circuits with lower loop gain and higher speed (less parasitic phase shift). However, the reduced feedback raises the resistance seen by $M_1$ and $M_2$, which will degrade linearity.

![Diagram of triode-region transconductor](image)

*Figure 12.2. A triode-region transconductor.*
12.4.2. Saturation-Region Transconductors

A variety of transconductors have been proposed that use differential circuitry to exploit the saturation square-law equation. Since the square-law saturation-region current expression has only second-order nonlinearity, differentially operated saturation-region transconductors can, in this approximation, be perfectly linear. Among MOSFET-based topologies, this class of circuits has the highest transconductance for a given bias current. This leads to lower vertical electric fields and which reduces this source of odd-order nonlinearity. The simplest approach uses differentially operated common-source FETs or even CMOS logic inverters, with tuning provided by the common-mode voltage, which also sets the bias current. Since output resistance in saturation can be high, it may be possible to avoid cascoding the outputs, which would reduce high-frequency phase errors. This trade-off will be discussed later.

12.4.3. MOSFETs Used for Degeneration

The transconductors discussed so far have all used FETs with grounded sources. In such circuits, the common-mode input voltage must be controlled to set the operating currents (usually by CMFB in a previous stage). To allow a range of common-mode inputs, a source-coupled pair with tail-currents sources can be used. Source-coupled MOSFET differential pairs are highly nonlinear. Degeneration using fixed source resistors would improve linearity but reduce tuneability. A useful alternative is to degenerate the circuit using cross-coupled MOSFET’s, as shown in Figure 12.3. Simulations and accurate modeling are required to optimize the trade-offs between the sizes of the transistors [4].

![Figure 12.3. Transconductor using FET degeneration.](image-url)
12.4.4. BJT-Based Transconductors

Bipolar junction transistor (BJT) and BiCMOS technologies are more expensive than CMOS. In processes that include BJTs, they are the best choice for most analog applications, because of their near-optimum $G_m/I$ ratio. However, for transconductor applications, this can actually be a disadvantage. It means that nonlinearities in the current-voltage characteristic become large for very small voltage swings. The Ebers–Moll exponential model for the BJT contains both even and odd-order nonlinearities, so that differential operation cannot eliminate the odd-order distortion components. However, the large transconductance can be traded off for high linearity by including resistive degeneration, as in the MOSFET case, although this reduces tuneability.

To retain the linearity of these degenerated transconductors while providing some tuneability, the output currents can be passed to a translinear multiplier circuit used as a current mirror with electronically variable current gain. Unfortunately, the low input impedance of the multiplier in this approach lowers the input-stage voltage gain and degrades noise performance [5].

12.4.5. Offset Differential Pairs

Another way to linearize and degenerate a BJT-based transconductor uses a parallel connection of multiple differential pairs with intentionally introduced offsets to spread nonlinearities over a wider range of input voltages [6]. The offsets can be introduced by using multiple differential pairs with unequal emitter areas. In the limit of large number of such pairs, with the offsets chosen optimally, this gives the benefits of degeneration while providing wide-range tunability. Using even a few differential pairs significantly enhances linearity. This method does increase the required layout area and the increased capacitance limits frequency response.

12.5. Dynamic Range

The most fundamental trade-offs in Gm-C filter design involve dynamic range. The dynamic range of a signal processor is the ratio of the largest to the smallest signals that can be processed. The amplitude of the largest signals is limited by nonlinearity, which in practice is the amplitude for which some maximum tolerable level of distortion is reached. The lower end of the dynamic range is usually set by noise. Ideally the noise is dominated by unavoidable thermal noise. Flicker noise or extraneous noise injected from other parts of the system may further degrade the signal-to-noise ratio (SNR). While in most systems, dynamic range is equivalent to SNR, in some systems dynamic range may be limited by the ratio of the largest signal to the resulting distortion.
products rather than by the SNR. Also, as we will see, companding can allow dynamic range to exceed the SNR.

Analysis of the degenerated BJT transconductor (Figure 12.4) illustrates several fundamental trade-offs in continuous-time filter design [7]. The differential equivalent input voltage per unit bandwidth due to thermal noise is given by

$$\frac{v_{\text{ein}}^2}{\Delta f} = 2 \cdot 4kT \left[ \frac{1}{2g_m} + \frac{R}{2} \right]$$

$$= \frac{4kT}{g_m} \cdot [1 + g_m R]$$

$$= \frac{4kT}{G_m}$$ \hspace{1cm} (12.1)

where the transconductance of the circuit can be expressed as $G_m = g_m / (1 + g_m R) = g_m / N$, where $N = 1 + g_m R$ is the degeneration ratio. Now, $g_m = I_0/(kT/q)$ is the maximum transconductance achievable from a transistor. BJTs closely approach this limit. We can also write the degeneration factor as the ratio $N = g_m / G_m$.

If the transconductor drives a differential load capacitor of value $C$, the result is an integrator with unity-gain frequency $\omega_0 = G_m / C$. This integrator can be used in a first-order low-pass filter. Integrating over the effective noise bandwidth of a first-order low-pass with cut-off frequency $\omega_0$ gives the total squared input noise of $kT/C$. As with any class A circuit, the maximum current signal is limited by the bias current $I_0$. The maximum voltage swing is then $V_0 = G_m I_0 = N \cdot kT/C$. From this we can conclude that the maximum
signal-to-noise ratio is

$$\text{SNR}_{\text{max}} = \frac{V_0^2}{8kT} \cdot \frac{C}{N^2} = \frac{V_0^2}{8kT} \cdot \frac{TC}{8q^2} \quad (12.2)$$

This implies that dynamic range can be increased arbitrarily by increasing the degeneration factor. In practice, the maximum signal swing must be less than the total power supply voltage $V_{DD}$. When power and cut-off frequency are constrained, the benefits of degeneration are reduced, and the maximum SNR of such a filter is proportional to

$$\text{SNR}_{\text{max}} = \frac{P}{M \cdot kT \cdot f_0} \quad (12.3)$$

where $M = V_{DD}/V_0$ and $f_0$ is the cut-off frequency.

12.6. Differential Operation

Differential operation does not usually offer much of a trade-off, since its advantages so often overwhelm its disadvantages. Thus, differential operation is common in all aspects of analog signal processing, since in addition to improving linearity, it also enhances rejection of interfering signals transmitted through the substrate and through bias and power-supply lines; such signals are increasingly problematic in mixed-signal circuits where digital circuits, with their fast transitions. The main disadvantage of differential operation is the need for CMFB circuits. Their design can be more challenging than the differential circuits.

12.7. Log-Domain Filtering

A way to work around these noise and signal-range limitations, while providing high linearity in BJT-based filters is the log-domain approach [8,9]. Rather than individually linearize each transconductor, signals within a log-domain filter are allowed to be highly nonlinear, and pre- and post-distortion are used to cancel the nonlinearities of the overall filter.

Log-domain circuits can be arranged to operate in a class AB mode, in which they implement so-called instantaneous companding. Such circuits can process current-mode signals that are much larger than the bias currents. In class A mode, where signals are less than the bias currents, noise power is independent of signal amplitude, so the SNR is proportional to signal amplitude. In class AB operation, the noise increases with the instantaneous signal. As the signal amplitude increases, the average noise increases as well, and the SNR saturates. This noise modulation effect limits the usefulness of instantaneously companding filters in some applications: high-amplitude signals in the filter’s
stopband can intermodulate with the noise, producing unacceptable noise in the passband.

12.8. Transconductor Frequency-Response Trade-Offs

Two main effects limit transconductor frequency response. At low frequencies, the main effect arises due to finite output resistance \( R_{out} \), or equivalently, finite low-frequency gain \( g_m R_{out} \). Low dc gain typically lowers pole \( Q_s \). This is often a cause of limited low-frequency rejection in band-pass filters. The usual way to raise transconductor \( R_{out} \) is cascoding, in which the transconductor’s signal current is passed through a common-gate transistor. This multiplies the output resistance by the intrinsic gain of the added transistor. However, cascoding tends to degrade the high frequency-response of transconductors, since it adds at least one extra node to the circuit. The parasitic capacitance at the source of the added transistor adds high frequency phase shift that degrades filter response at high frequencies.

To allow filter operation at frequencies into the hundreds of MHz, the intermediate node added by cascoding should be avoided. In a transconductor without such intermediate nodes, phase errors at high frequencies arise only from second-order effects such as transmission-line effects due to distributed time delays along the gate. Without cascoding, however, low-frequency transconductor gain is limited to that of a single gain stage, typically less than 100, too low for any but very low-Q filter applications.

One solution is to use positive feedback using cross-coupled inverting transconductors, as shown in Figure 12.5. In [10], the individual transconductors were implemented using saturation-region transconductors identical to standard CMOS logic inverters. Filter frequencies were set by varying the transconductor \( V_{DD} \) inputs. The \( V_{DD} \) connections for the cross-coupled

![Figure 12.5. Cascode-free transconductor using positive feedback.](image-url)
transistors had separate $V_{DD}$ inputs, so that $G_{m4}$ could be set independently of $G_{m1}$ and $G_{m3}$. The differential-mode gain is $A_{DM} = G_{m1}/(G_{0} + G_{m3} - G_{m4})$. A $Q$-tuning circuit was used to adjust $G_{m4}$ to cancel $G_{0} + G_{m3}$. Mismatch limits the achievable cancelation.

High loop-gain CMFB is not needed, since the common-mode output resistance of the transconductor is low and the common-mode voltage gain $A_{CM} = G_{m1}/(G_{0} + G_{m3} + G_{m4})$ is approximately 1/2.

This approach is well suited for low-voltage operation. The required value of $V_{DD}$ is $2|V_{gs}|$ and the range for nominally linear signal swing is the sum of the NMOS and PMOS threshold voltages, $V_{TN} + |V_{tp}|$. However, the circuitry required to tune $V_{DD}$ means that the actual supply voltage must be much higher than $V_{DD}$. Another disadvantage of this circuit is the strong dependence of the operating current on $V_{DD}$, which impairs the power-supply rejection ratio (PSRR).

A circuit variation that addresses many of these limitations is shown in Figure 12.6, which uses a folded signal path with PMOS devices used only for biasing [11]. Frequency tuning is accomplished using current sources, eliminating the overhead for tuning $V_{DD}$. The minimum supply voltage for the circuit can be as low as $V_{DD_{min}} = V_{TN} + V_{DsatN} + |V_{DsatP}|$. The bias currents are no longer strongly affected by $V_{DD}$, which significantly improves PSRR. As drawn in the figure, there is no way to adjust $G_{m4}$ to maximize $Q$. A triode-connected FET could be included in $M_{4s}$ source to allow such adjustment, at some cost in linearity.

### Figure 12.6. Low-voltage cascode-free transconductor.

12.9. **Tuning Trade-Offs**

Any filter transfer function has several parameters that may require tuning. In a first-order filter, the pole frequency, the passband gain and possibly the stopband gain (or equivalently, a zero’s frequency) may require tuning. In higher order filters with complex poles, each biquadratic section potentially has several tuneable parameters: the cut-off frequencies and quality factors or...
$Q$ of the poles and those of any complex zeros, as well a gain factor. Pole or zero frequencies are always determined by ratios or products of two independent parameters, such as the $G_m/C$ ratio. These quantities generally do not track each other with temperature or process variations, and thus usually require tuning. Gain and quality factors, being dimensionless, are determined by ratios of like elements such as capacitances or of transconductances. With proper care to ensure adequate matching, such parameter ratios may not need to be tuned. However, the sensitivity of most filter parameter to element variations is proportional to $Q^2$. In highly selective filters, which usually include high-$Q$ poles, tuning of pole or zero $Q$ may be necessary. $Q$-tuning circuits often are very complex, and require large chip area. In any event, the effectiveness of any tuning approach is ultimately limited by mismatch. This is the most serious limitation on the use of continuous-time filtering for high-$Q$ applications.

Several tuning strategies may be considered, ranging from no tuning at all to tuning of the frequencies and $Q$s of multiple poles and zeros.

**No tuning.** If the specifications are loose enough and the process, temperature and voltage-induced variability is small enough, tuning may be unnecessary. This allows the use of passive conductors or highly degenerated transconductors, potentially offering wide dynamic range.

**Off-chip tuning.** In this approach, on-chip transconductances are controlled by current or voltage inputs. A feedback loop can be used to force transconductances to track an off-chip conductance whose temperature coefficient is chosen to compensate for the assumed temperature coefficient of on-chip capacitors. The extra pins and external components required are disadvantages of this approach.

**One-time post-fabrication tuning.** Coarse tuning can be achieved by selecting capacitors from an array using digital post-fabrication tuning or by blowing on-chip fuses. A related approach is to laser-trim an on-chip resistor used as a reference in a feedback arrangement as described in the previous section. Because each must be individually tuned, such strategies increase test and assembly costs. A compromise is to measure the required value for a few samples in each lot and apply the same correction to each part.

**Automatic tuning.** On-chip tuning strategies often use a phase-locked loop (PLL) in a master–slave arrangement to set cut-off frequencies. In a typical approach, a voltage-controlled oscillator (VCO) is formed using a loop of two identical Gm-C integrators. The VCO is phase-locked to an external frequency reference. The Barkhausen conditions for constant-amplitude oscillations of such a loop require 90-degree phase shift for each integrator, at the unity-gain
frequency. The voltage needed to meet these criteria within the oscillator (the master) is assumed to match that required for transconductors within the filter (the slaves), and can thus be used to tune the unity-gain frequency of the slaved transconductors. An oscillator based on a two-integrator loop generally includes an automatic gain control (AGC) circuitry to keep the amplitude constant. The output of this AGC circuitry can be used to adjust $Q$.

However, this master–slave arrangement has several fundamental limitations, and requires some difficult compromises. Mismatch between the master and the slave, inevitable in any analog system, leads to errors in tuning. To minimize these mismatches, the master and slave should be as close as possible to each other in all aspects, including their layouts and their operating conditions. This is clearly only approximately possible in a complex filter with multiple pole and zero frequencies. Furthermore, if the PLL operating frequency is close to the filter passband or stopband the PLL is likely to interfere with the filter.

12.10. Simulation Issues

IC design customarily depends heavily on computer simulation. In continuous-time filters, performance depends strongly on the details of transistor operation. The simple transistor models described previously are inadequate for designing continuous-time filters to meet linearity specifications. In particular, MOSFETs with short channels do not follow the square law, and in fact rarely follow any simple equation very accurately. Even in long-channel transistors, mobility reduction and other effects cause nonlinearities that simple models fail to predict. In this regard, BJT’s may be easier to design with, as their analytical models correspond fairly closely to their actual behavior.

References


