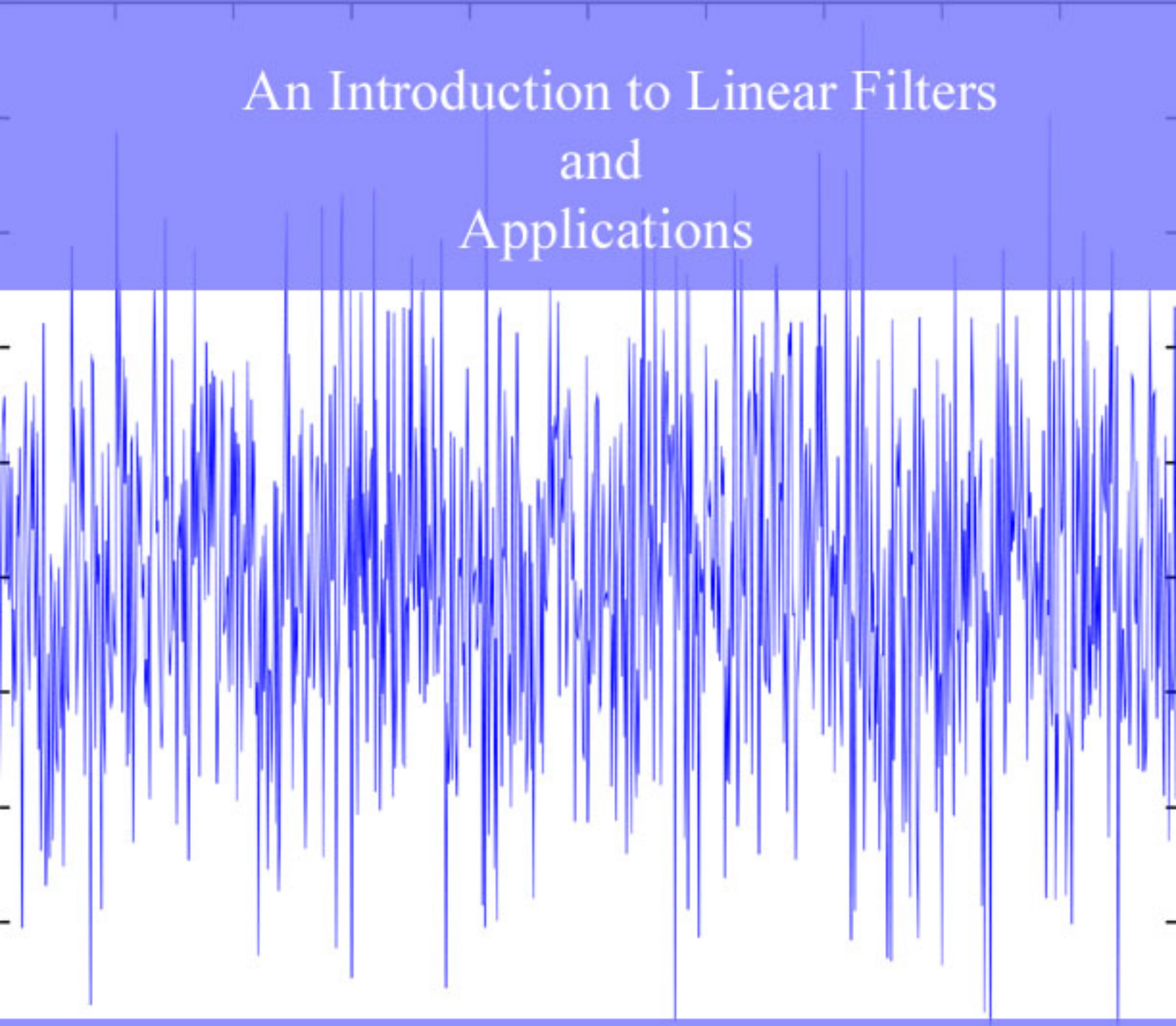
# An Introduction to Linear Filters and Applications



Parthenia Braxton

First Edition, 2012

ISBN 978-81-323-3024-0



*Published by:*
**Research World**
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

# Table of Contents

Chapter 1 - Linear Filter

Chapter 2 - Active Filter and Antimetric (Electrical Networks)

Chapter 3 - Passive Analogue Filter Development

Chapter 4 - Composite Image Filter

Chapter 5 - Bessel Filter and Bartlett's Bisection Theorem

Chapter 6 - Elliptic Filter

Chapter 7 - Fast Kalman Filter and Filter Bank

Chapter 8 - Gaussian Filter and Digital Biquad Filter

Chapter 9 - Prototype Filter

Chapter 10 - Network Synthesis Filters

Chapter 11 - Propagation Constant

Chapter 12 - Image Impedance

Chapter 1

# Linear Filter

**Linear filters** in the time domain process time-varying input signals to produce output signals, subject to the constraint of linearity. This results from systems composed solely of components (or digital algorithms) classified as having a linear response.

Most filters implemented in analog electronics, in digital signal processing, or in mechanical systems are classified as causal, time invariant, and linear. However the general concept of linear filtering is broader, also used in statistics, data analysis, and mechanical engineering among other fields and technologies. This includes noncausal filters and filters in more than one dimension such as would be used in image processing; those filters are subject to different constraints leading to different design methods, which are discussed elsewhere.

A linear time-invariant (LTI) filter can be uniquely specified by its impulse response $h$, and the output of any filter is mathematically expressed as the convolution of the input with that impulse response. The frequency response, given by the filter's transfer function $H(\omega)$, is an alternative characterization of the filter. The frequency response may be tailored to, for instance, eliminate unwanted frequency components from an input signal, or to limit an amplifier to signals within a particular band of frequencies. There are a number of particularly desirable or useful filter transfer functions.

Among the time-domain filters we here consider, there are two general classes of filter transfer functions that can approximate a desired frequency response. Very different mathematical treatments apply to the design of filters termed infinite impulse response (IIR) filters, characteristic of mechanical and analog electronics systems, and finite impulse response (FIR) filters, which can be implemented by discrete time systems such as computers (then termed *digital signal processing*).

## *Impulse response and transfer function*

The impulse response *h* of a linear time-invariant causal filter specifies the output that the filter would produce if it were to receive an input consisting of a single impulse at time 0. An "impulse" in a continuous time filter means a Dirac delta function; in a discrete time filter the Kronecker delta function would apply. The impulse response completely characterizes the response of any such filter, inasmuch as any possible input signal can be expressed as a (possibly infinite) combination of weighted delta functions. Multiplying the impulse response shifted in time according to the arrival of each of these delta functions by the amplitude of each delta function, and summing these responses together (according to the superposition principle, applicable to all linear systems) yields the output waveform.

Mathematically this is described as the convolution of a time-varying input signal *x(t)* with the filter's impulse response *h*, defined as:

$$y(t) = \int_0^T x(t-\tau)\, h(\tau)\, d\tau$$

$$y_k = \sum_{i=0}^{N} x_{k-i}\, h_i$$

The first form is the continuous-time form which describes mechanical and analog electronic systems, for instance. The second equation is a discrete-time version used, for example, by digital filters implemented in software, so-called *digital signal processing*. The impluse response *h* completely characterizes any linear time-invariant (or shift-invariant in the discrete-time case) filter. The input *x* is said to be "convolved" with the impulse response *h* having a (possibly infinite) duration of time *T* (or of *N* sampling periods).

The filter response can also be completely characterized in the frequency domain by its transfer function $H(\omega)$, which is the Fourier transform of the impulse response *h*. Typical filter design goals are to realize a particular frequency response, that is, the magnitude of the transfer function $|H(\omega)|$ ; the importance of the phase of the transfer function varies according to the application, inasmuch as the shape of a waveform can be distorted to a greater or lesser extent in the process of achieving a desired (amplitude) response in the frequency domain.

Filter design consists of finding a possible transfer function that can be implemented within certain practical constraints dictated by the technology or desired complexity of the system, followed by a practical design that realizes that transfer function using the chosen technology. The complexity of a filter may be specified according to the order of the filter, which is specified differently depending on whether we are dealing with an IIR or FIR filter. We will now look at these two cases.

## Infinite impulse response filters

Consider a physical system that acts as a linear filter, such as a system of springs and masses, or an analog electronic circuit that includes capacitors and/or inductors (along with other linear components such as resistors and amplifiers). When such a system is subject to an impulse (or any signal of finite duration) it will respond with an output waveform which lasts past the duration of the input, eventually decaying exponentially in one or another manner, but never completely settling to zero (mathematically speaking). Such a system is said to have an infinite impulse response (IIR). The convolution integral (or summation) above extends over all time: T (or N) must be set to infinity.

For instance, consider a damped harmonic oscillator such as a pendulum, or a resonant L-C tank circuit. If the pendulum has been at rest and we were to strike it with a hammer (the "impulse"), setting it in motion, it would swing back and forth ("resonate"), say, with an amplitude of 10cm. But after 10 minutes, say, it would still be swinging but the amplitude would have decreased to 5cm, half of its original amplitude. After another 10 minutes its amplitude would be only 2.5cm, then 1.25cm, etc. However it would never come to a complete rest, and we therefore call that response to the impulse (striking it with a hammer) "infinite" in duration.

The complexity of such a system is specified by its order N. N is often a constraint on the design of a transfer function since it specifies the number of reactive components in an analog circuit; in a digital IIR filter the number of computations required is proportional to N.

## Finite impulse response filters

A filter implemented in a computer program (or a so-called digital signal processor) is a discrete-time system; a different (but parallel) set of mathematical concepts defines the behavior of such systems. Although a digital filter can be an IIR filter if the algorithm implementing it includes feedback, it is also possible to easily implement a filter whose impulse truly goes to zero after N time steps; this is called a finite impulse response (FIR) filter.

For instance, suppose we have a filter which, when presented with an impulse in a time series:

0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.....

will output a series which responds to that impulse at time 0 until time 4, and has no further response, such as:

0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0.....

Although the impulse response has lasted 4 time steps after the input, starting at time 5 it has truly gone to zero. The extent of the impulse response is *finite*, and this would be

classified as a $4^{th}$ order FIR filter. The convolution integral (or summation) above need only extend to the full duration of the impulse response T, or the order N in a discrete time filter.

### Implementation issues

Classical analog filters are IIR filters, and classical filter theory centers on the determination of transfer functions given by low order rational functions, which can be synthesized using the same small number of reactive components. Using digital computers, on the other hand, both FIR and IIR filters are straightforward to implement in software.

A digital IIR filter can generally approximate a desired filter response using less computing power than a FIR filter, however this advantage is more often unneeded given the increasing power of digital processors. The ease of designing and characterizing FIR filters makes them preferable to the filter designer (programmer) when ample computing power is available. Another advantage of FIR filters is that their impulse response can be made symmetric, which implies a response in the frequency domain which has zero phase at all frequencies (not considering a finite delay), which is absolutely impossible with any IIR filter.

## *Frequency response*

The frequency response or transfer function $| H(\omega) |$ of a filter can be obtained if the impulse response is known, or directly through analysis using Laplace transforms, or in discrete-time systems the Z-transform. The frequency response also includes the phase as a function of frequency, however in many cases the phase response is of little or no interest. FIR filters can be made to have zero phase, but with IIR filters that is generally impossible With most IIR transfer functions there are related transfer functions having a frequency response with the same magnitude but a different phase; in most case the so-called minimum phase transfer function is preferred.

Filters in the time domain are most often requested to follow a specified frequency response. Then a mathematical procedure is used to find a filter transfer function which can be realized (within some constraints) and which approximates the desired response to within some criterion. Common filter response specifications are described as follows:

- A low-pass filter passes low frequencies while blocking higher frequencies.
- A high-pass filter passes high frequencies.
- A band-pass filter passes a band (range) of frequencies.
- A band-stop filter passes high and low frequencies outside of a specified band.
- A notch filter has a null response at a particular frequency. This function may be combined with one of the above responses.
- An all-pass filter passes all frequencies equally well, but alters the phase relationship among them.

- An equalization filter is not designed to fully pass or block any frequency, but instead to gradually vary the amplitude response as a function of frequency: filters used as pre-emphasis filters, equalizers, or tone controls are good examples.

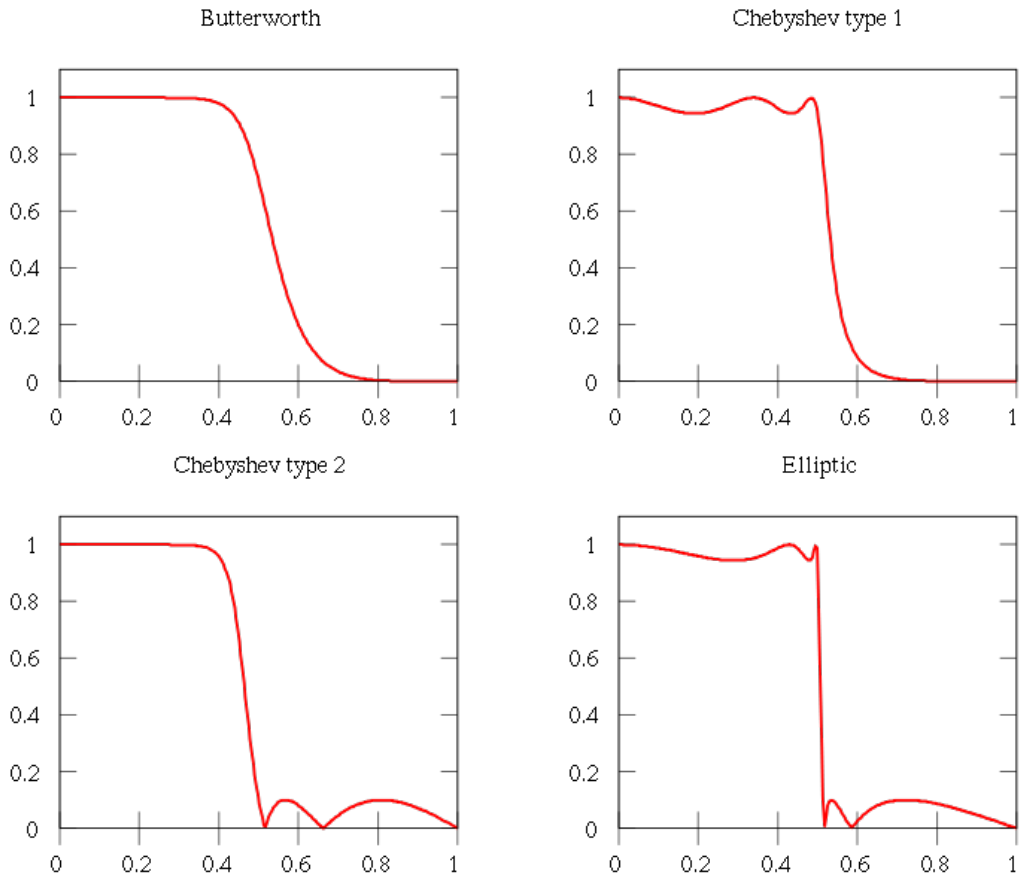
## FIR transfer functions

Meeting a frequency response requirement with an FIR filter uses relatively straight-forward procedures. In the most basic form, the desired frequency response itself can be sampled with a resolution of $\Delta f$ and fourier transformed to the time domain. This will obtain the filter coefficients $h_i$ which will implement a zero phase FIR filter which matches the frequency response at the sampled frequencies used. In order to better match a desired response, $\Delta f$ must be reduced. However the duration of the filter's impulse response, and the number of terms which must be summed for each output value (according to the above discrete time convolution) is given by $N = 1/(\Delta f\, T)$ where $T$ is the sampling period of the discrete time system (N-1 is also termed the *order* of an FIR filter). Thus the complexity of a digital filter and the computing time involved, grows inversely with $\Delta f$, placing a higher cost on filter functions which better approximate the desired behavior. For the same reason, filter functions whose critical response is at lower frequencies (compared to the sampling frequency *1/T*) require a higher order, more computationally intensive FIR filter. An IIR filter can thus be much more efficient in such cases.

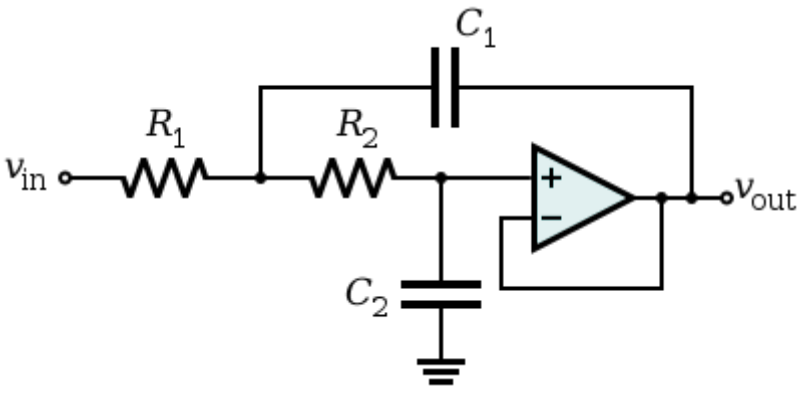
## IIR transfer functions

Since classical analog filters are IIR filters, there has been a long history of studying the range of possible transfer functions implementing various of the above desired filter responses in continuous time systems. Using transforms it is possible to convert these continuous time frequency responses to ones that are implemented in discrete time, for use in digital IIR filters. The complexity of any such filter is given by the *order* N, which describes the order of the rational function describing the frequency response. The order N is of particular importance in analog filters, because an $N^{th}$ order electronic filter requires N reactive elements (capactors and/or inductors) to implement. If a filter is implemented using, for instance, biquad stages using op-amps, N/2 stages will be needed. In a digital implementation, the number of computations performed per sample is proportional to N. Thus the mathematical problem is to obtain the best approximation (in some sense) to the desired response using a smaller N, as we shall now illustrate.

Below are the frequency responses of several standard filter functions which approximate a desired response, optimized according to some criterion. These are all fifth-order low-pass filters, designed for a cutoff frequency of .5 in normalized units. Frequency responses are shown for the Butterworth, Chebyshev, inverse Chebyshev, and elliptic filters.



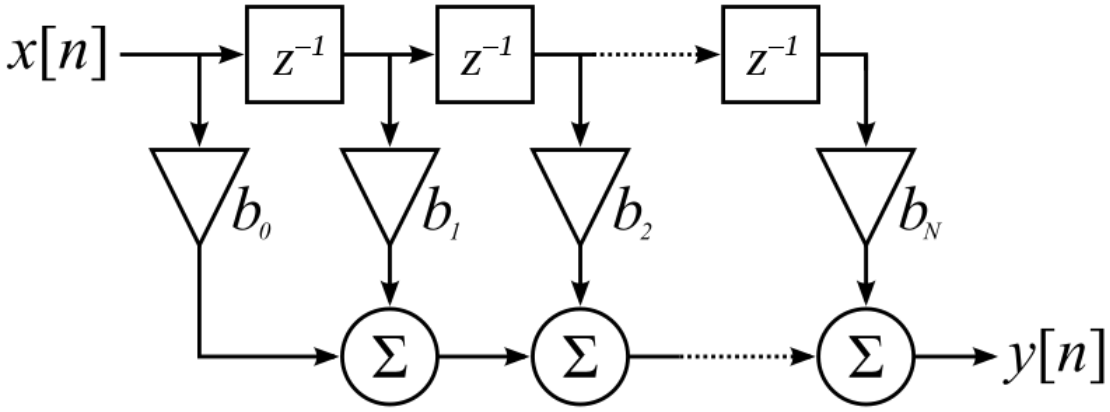

As is clear from the image, the elliptic filter is sharper than the others, but at the expense of ripples in both its passband and stopband. The Butterworth filter has the poorest transition but has a more even response, avoiding ripples in either the passband or stopband. A Bessel filter (not shown) has an even poorer transition in the frequency domain, but maintains the best phase fidelity of a waveform. Different applications will emphasize different design requirements, leading to different choices among these (and other) optimizations, or requiring a filter of a higher order.




Low-pass filter implemented with a Sallen–Key topology

## *Example implementations*

A popular circuit implementing a second order active R-C filter is the Sallen-Key design, whose schematic diagram is shown here. This topology can be adapted to produce low-pass, band-pass, and high pass filters.




A discrete-time FIR filter of order *N*. The top part is an *N*-sample delay line; each delay step is denoted $z^{-1}$.

An $N^{th}$ order FIR filter can be implemented in a discrete time system using a computer program or specialized hardware in which the input signal is subject to N delay stages. The output of the filter is formed as the weighted sum of those delayed signals, as is depicted in the accompanying signal flow diagram. The response of the filter depends on the weighting coefficients denoted $b_0$, $b_1$, .... $b_N$. For instance, if all of the coefficients were equal to unity, a so-called boxcar function, then it would implement a low-pass filter with a low frequency gain of N+1 and a frequency response given by the sinc function. Superior shapes for the frequency response can be obtained using coefficients derived from a more sophisticated design procedure.

## *Mathematics of filter design*

LTI system theory describes linear *time-invariant* (LTI) filters of all types. LTI filters can be completely described by their frequency response and phase response, the specification of which uniquely defines their impulse response, and *vice versa*. From a mathematical viewpoint, continuous-time IIR LTI filters may be described in terms of linear differential equations, and their impulse responses considered as Green's functions of the equation. Continuous-time LTI filters may also be described in terms of the Laplace transform of their impulse response, which allows all of the characteristics of the filter to be analyzed by considering the pattern of poles and zeros of their Laplace transform in the complex plane. Similarly, discrete-time LTI filters may be analyzed via the Z-transform of their impulse response.

Before the advent of computer filter synthesis tools, graphical tools such as Bode plots and Nyquist plots were extensively used as design tools. Even today, they are invaluable tools to understanding filter behavior. Reference books had extensive plots of frequency response, phase response, group delay, and impulse response for various types of filters, of various orders. They also contained tables of values showing how to implement such filters as RLC ladders - very useful when amplifying elements were expensive compared to passive components. Such a ladder can also be designed to have minimal sensitivity to component variation a property hard to evaluate without computer tools.

Many different analog filter designs have been developed, each trying to optimise some feature of the system response. For practical filters, a custom design is sometimes desirable, that can offer the best tradeoff between different design criteria, which may include component count and cost, as well as filter response characteristics.

These descriptions refer to the *mathematical* properties of the filter (that is, the frequency and phase response). These can be *implemented* as analog circuits (for instance, using a Sallen Key filter topology, a type of active filter), or as algorithms in digital signal processing systems.

Digital filters are much more flexible to synthesize and use than analog filters, where the constraints of the design permits their use. Notably, there is no need to consider component tolerances, and very high Q levels may be obtained.

FIR digital filters may be implemented by the direct convolution of the desired impulse response with the input signal. They can easily be designed to give a matched filter for any arbitrary pulse shape.

IIR digital filters are often more difficult to design, due to problems including dynamic range issues, quantization noise and instability. Typically digital IIR filters are designed as a series of digital biquad filters.

All low-pass second-order continuous-time filters have a transfer function given by

$$H(s) = \frac{K\omega_0^2}{s^2 + \frac{\omega_0}{Q}s + \omega_0^2}.$$

All band-pass second-order continuous-time have a transfer function given by

$$H(s) = \frac{K\frac{\omega_0}{Q}s}{s^2 + \frac{\omega_0}{Q}s + \omega_0^2}.$$

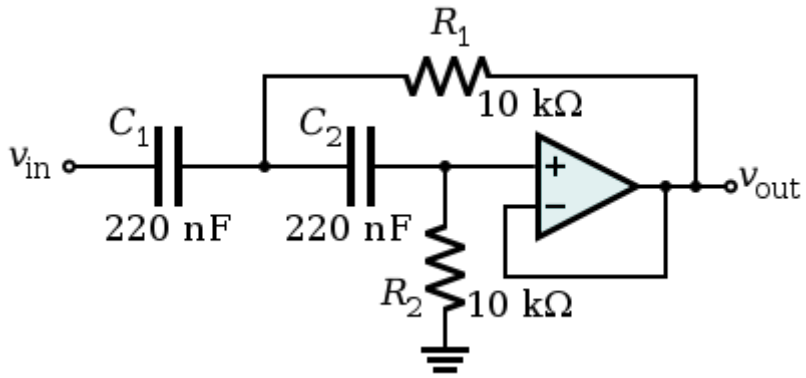
where

- $K$ is the gain (low-pass DC gain, or band-pass mid-band gain) ($K$ is 1 for passive filters)
- $Q$ is the Q factor
- $\omega_0$ is the center frequency
- $s = \sigma + j\omega$ is the complex frequency

**Chapter 2**

# Active Filter and Antimetric (Electrical Networks)

## Active Filter




An example of high-pass active filter of the Sallen Key topology. The operational amplifier, U1, is used as a buffer amplifier.

An **active filter** is a type of analog electronic filter, distinguished by the use of one or more active components i.e. voltage amplifiers or buffer amplifiers. Typically this will be a vacuum tube, or solid-state (transistor or operational amplifier).

Active filters have three main advantages over passive filters:

- Inductors can be avoided. Without them, passive filters cannot obtain a high Q (low damping), but inductors are often large and expensive (at low frequencies), may have significant internal resistance, and may pick up surrounding electromagnetic signals.
- The shape of the response, the Q (Quality factor), and the tuned frequency can often be set easily by varying resistors, in some filters one parameter can be adjusted without affecting the others. Variable inductances for low frequency filters are not practical.

- The amplifier powering the filter can be used to buffer the filter from the electronic components it drives or is fed from, variations in which could otherwise significantly affect the shape of the frequency response.

Active filter circuit configurations (electronic filter topology) include:

- Sallen and Key, and VCVS filters (low dependency on accuracy of the components)
- State variable and biquadratic filters
- Twin T filter (fully passive)
- Dual Amplifier Bandpass (DABP)
- Wien notch
- Multiple Feedback Filter
- Fliege (lowest component count for 2 opamp but with good controllability over frequency and type)
- Akerberg Mossberg (one of the topologies that offer complete and independent control over gain, frequency, and type)

All the varieties of passive filters can also be found in active filters. Some of them are:

- High-pass filters – attenuation of frequencies below their cut-off points.
- Low-pass filters – attenuation of frequencies above their cut-off points.
- Band-pass filters – attenuation of frequencies both above and below those they allow to pass.
- Notch filters – attenuation of certain frequencies while allowing all others to pass.

  Combinations are possible, such as notch and high-pass (for example, in a rumble filter where most of the offending rumble comes from a particular frequency), e.g.Elliptic filters.

## *Design of active filters*

To design filters, the specifications that need to be established include:

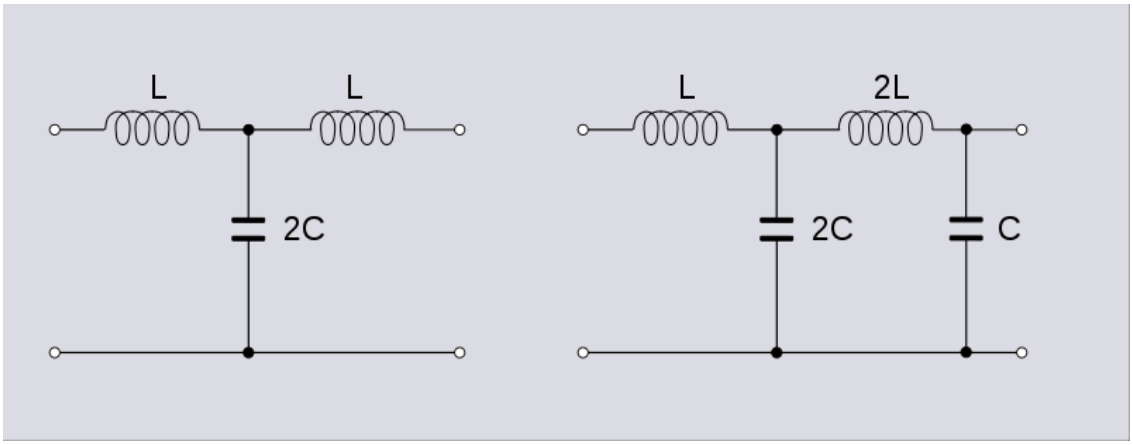
- The range of desired frequencies (the passband) together with the shape of the frequency response. This indicates the variety of filter and the center or corner frequencies.
- Input and output impedance requirements. These limit the circuit topologies available; for example, most, but not all active filter topologies provide a buffered (low impedance) output. However, remember that the internal output impedance of operational amplifiers, if used, may rise markedly at high frequencies and reduce the attenuation from that expected. Be aware that some high-pass filter topologies present the input with almost a short circuit to high frequencies.
- The degree to which unwanted signals should be rejected.
    - In the case of narrow-band bandpass filters, the Q determines the -3dB bandwidth but also the degree of rejection of frequencies far removed

from the center frequency; if these two requirements are in conflict then a staggered-tuning bandpass filter may be needed.
    - For notch filters, the degree to which unwanted signals at the notch frequency must be rejected determines the accuracy of the components, but not the Q, which is governed by desired steepness of the notch, i.e. the bandwidth around the notch before attenuation becomes small.
    - For high-pass and low-pass (as well as band-pass filters far from the center frequency), the required rejection may determine the slope of attenuation needed, and thus the "order" of the filter. A second-order all-pole filter gives an ultimate slope of about 12 dB per octave (40dB/decade), but the slope close to the corner frequency is much less, sometimes necessitating a notch be added to the filter.
- The allowable "ripple" (variation from a flat response, in decibels) within the passband of high-pass and low-pass filters, along with the shape of the frequency response curve near the corner frequency, determine the damping factor (reciprocal of Q). This also affects the phase response, and the time response to a square-wave input. Several important response shapes (damping factors) have well-known names:
    - Chebyshev filter – slight peaking/ripple in the passband before the corner; Q>0.7071 for 2nd-order filters
    - Butterworth filter – flattest amplitude response; Q=0.7071 for 2nd-order filters
    - Linkwitz–Riley filter – desirable properties for audio crossover applications; Q = 0.5 (critically damped)
    - Paynter or transitional Thompson-Butterworth or "compromise" filter – faster fall-off than Bessel; Q=0.639 for 2nd-order filters
    - Bessel filter – best time-delay, best overshoot response; Q=0.577 for 2nd-order filters
    - Elliptic filter or Cauer filters – add a notch (or "zero") just outside the passband, to give a much greater slope in this region than the combination of order and damping factor *without* the notch.

# Antimetric

An **antimetric** electrical network is one that exhibits anti-symmetrical electrical properties. The term is often encountered in filter theory, but it applies to general electrical network analysis. Antimetric is the diametrical opposite of symmetric; it does not merely mean "asymmtetric" (i.e., "lacking symmetry").

## *Definition*




Examples of symmetry and antimetry: both networks are low-pass filters but one is symmetric (left) and the other is antimetric (right). For a symmetric ladder the 1st element is equal to the *n*th, the 2nd equal to the (*n*-1)th and so on. For an antimetric ladder, the 1st element is the dual of the *n*th and so on.

References to symmetry and antimetry of a network usually refer to the input impedances of a two-port network when correctly terminated. A symmetric network will have the two equal impedances, $Z_{i1}$ and $Z_{i2}$. For an antimetric network, the two impedances must be the dual of each other with respect to some nominal impedance $R_0$. That is,

$$\frac{Z_{i1}}{R_0} = \frac{R_0}{Z_{i2}}$$

which is well-defined because $R_0 \neq 0$ and $Z_{i2} \neq 0$. Hence,

$$Z_{i1} Z_{i2} = {R_0}^2.$$

It is necessary for antimetry that the terminating impedances are also the dual of each other, but in many practical cases the two terminating impedances are resistors and are both equal to the nominal impedance $R_0$. Hence, they are both symmetric and antimetric at the same time.

Other network parameters may also be referred to as antimetric. For instance, for a two-port network described by scattering parameters (*S*-parameters),

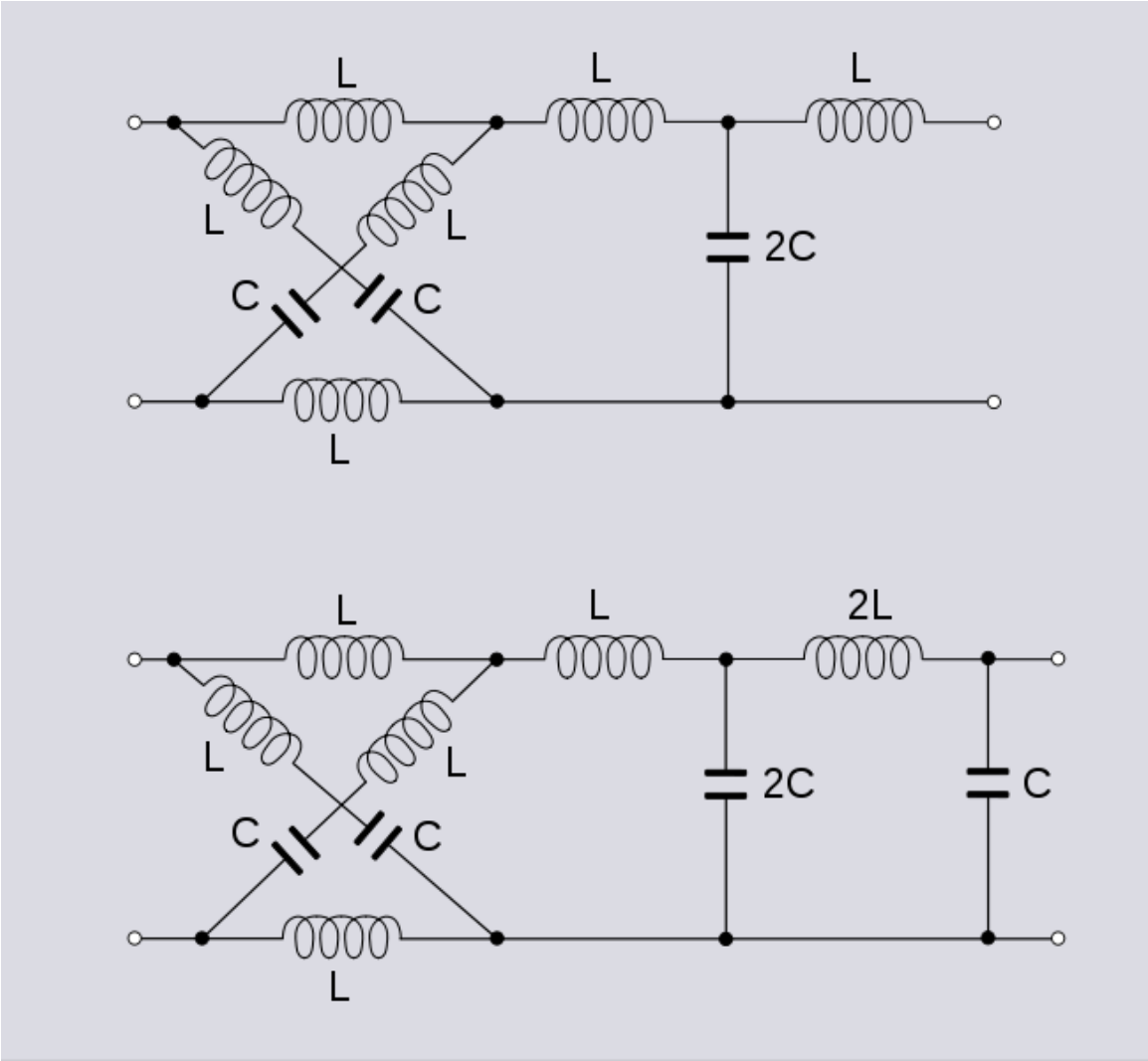
$$S_{11} = S_{22}$$

if the network is symmetric, and

$$S_{11} = -S_{22}$$

if the network is antimetric.

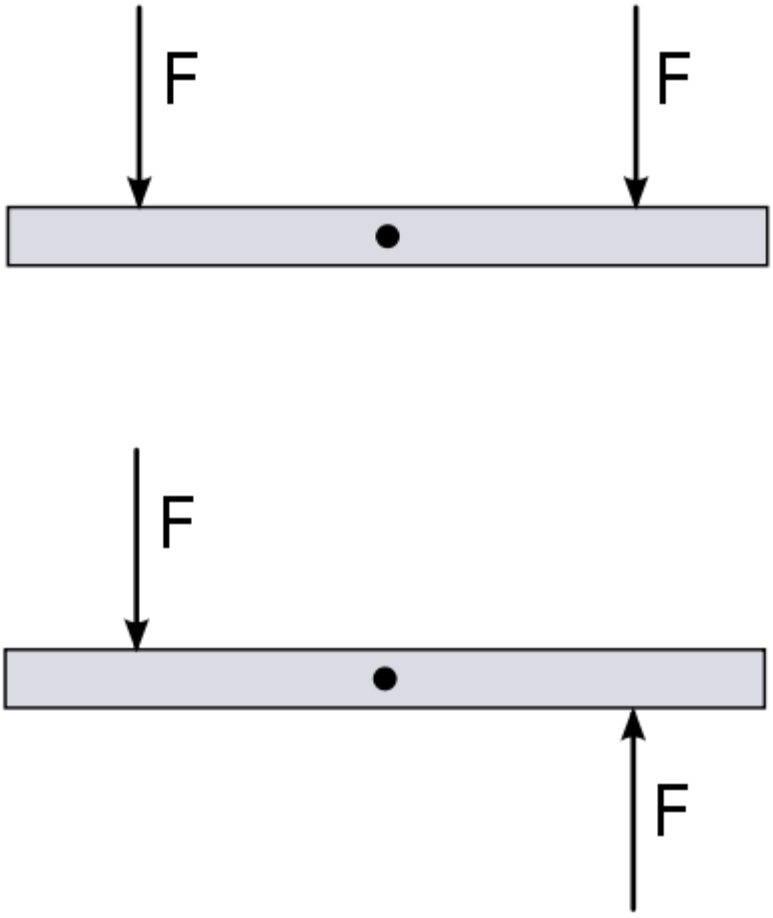
### *Physical and electrical antimetry*




Examples of symmetric (top) and antimetric (bottom) networks which do not exhibit topological symmetry nor antimetry.

Symmetric and antimetric networks are often also topologically symmetric and antimetric, respectively. That is, the physical arrangement of their components and values are symmetric or antimetric as in the ladder example above. However, it is not a necessary condition for electrical antimetry. For example, if the example networks from the preceding section have an additional T-section added to the left-hand side, then the networks remain topologically symmetric and antimetric. However, the network resulting from the application of Bartlett's bisection theorem applied to the first T-section in each network are neither physically symmetric nor antimetric but retain their electrical symmetric (in the first case) and antimetric (in the second case) properties.

### *Mechanics*




Examples of symmetric (top) and antimetric (bottom) forces acting on a pivoted beam.

Antimetry appears in mechanics as a property of forces, motions, and oscillations. Symmetric forces produce translational motion and normal stress, and antimetric forces produce rotational motion and shear stress. Any asymmetric pair of forces can be expressed as a linear combination of a symmetric and an antimetric pair.

**Chapter 3**

# Passive Analogue Filter Development

**Analogue filters** are a basic building block of signal processing much used in electronics. Amongst their many applications are the separation of an audio signal before application to bass, mid-range and tweeter loudspeakers; the combining and later separation of multiple telephone conversations onto a single channel; the selection of a chosen radio station in a radio receiver and rejection of others.

Passive linear electronic analogue filters are those filters which can be described with linear differential equations (linear); they are composed of capacitors, inductors and, sometimes, resistors (passive) and are designed to operate on continuously varying (analogue) signals. There are many linear filters which are not analogue in implementation (digital filter), and there are many electronic filters which may not have a passive topology. Analogue filters are most often used in wave filtering applications, that is, where it is required to pass particular frequency components and to reject others from analogue (continuous-time) signals.

Analogue filters have played an important part in the development of electronics. Especially in the field of telecommunications, filters have been of crucial importance in a number of technological breakthroughs and have been the source of enormous profits for telecommunications companies. It should come as no surprise, therefore, that the early development of filters was intimately connected with transmission lines. Transmission line theory gave rise to filter theory, which initially took a very similar form, and the main application of filters was for use on telecommunication transmission lines. However, the arrival of network synthesis techniques greatly enhanced the degree of control of the designer.

Today, it is often preferred to carry out filtering in the digital domain where complex algorithms are much easier to implement, but analogue filters do still find applications, especially for low-order simple filtering tasks and are often still the norm at higher frequencies where digital technology is still impractical, or at least, less cost effective.

Wherever possible, and especially at low frequencies, analogue filters are now implemented in a filter topology which is active in order to avoid the wound components required by passive topology.

It is possible to design linear analogue mechanical filters using mechanical components which filter mechanical vibrations or acoustic waves. While there are few applications for such devices in mechanics per se, they can be used in electronics with the addition of transducers to convert to and from the electrical domain. Indeed some of the earliest ideas for filters were acoustic resonators because the electronics technology was poorly understood at the time. In principle, the design of such filters can be achieved entirely in terms of the electronic counterparts of mechanical quantities, with kinetic energy, potential energy and heat energy corresponding to the energy in inductors, capacitors and resistors respectively.

## *Historical overview*

There are three main stages in the history of **passive analogue filter development**:

1. **Simple filters**. The frequency dependence of electrical response was known for capacitors and inductors from very early on. The resonance phenomenon was also familiar from an early date and it was possible to produce simple, single-branch filters with these components. Although attempts were made in the 1880s to apply them to telegraphy, these designs proved inadequate for successful frequency division multiplexing. Network analysis was not yet powerful enough to provide the theory for more complex filters and progress was further hampered by a general failure to understand the frequency domain nature of signals.
2. **Image filters**. Image filter theory grew out of transmission line theory and the design proceeded in a similar manner to transmission line analysis. For the first time filters could be produced that had precisely controllable passbands and other parameters. These developments took place in the 1920s and filters produced to these designs were still in widespread use in the 1980s, only declining as the use of analogue telecommunications has declined. Their immediate application was the economically important development of frequency division multiplexing for use on intercity and international telephony lines.
3. **Network synthesis filters**. The mathematical bases of network synthesis were laid in the 1930s and 1940s. After the end of World War II network synthesis became the primary tool of filter design. Network synthesis put filter design on a firm mathematical foundation, freeing it from the mathematically sloppy techniques of image design and severing the connection with physical lines. The essence of network synthesis is that it produces a design that will (at least if implemented with ideal components) accurately reproduce the response originally specified in black box terms.

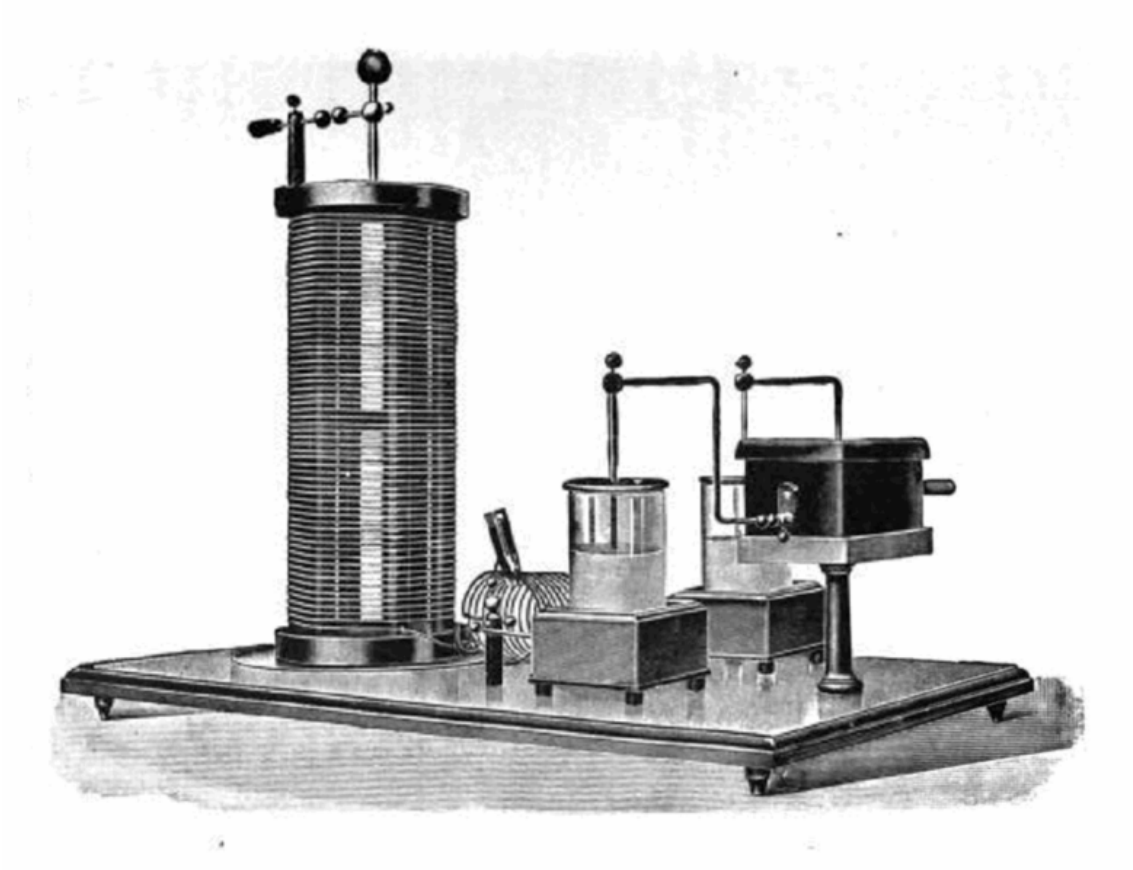
Throughout the letters R,L and C are used with their usual meanings to represent resistance, inductance and capacitance, respectively. In particular they are used in combinations, such as LC, to mean, for instance, a network consisting only of inductors

and capacitors. Z is used for electrical impedance, any 2-terminal combination of RLC elements and in some sections D is used for the rarely seen quantity elastance, which is the inverse of capacitance.

## *Resonance*

Early filters utilised the phenomenon of resonance to filter signals. Although electrical resonance had been investigated by researchers from a very early stage, it was at first not widely understood by electrical engineers. Consequently, the much more familiar concept of acoustic resonance (which in turn, can be explained in terms of the even more familiar mechanical resonance) found its way into filter design ahead of electrical resonance. Resonance can be used to achieve a filtering effect because the resonant device will respond to frequencies at, or near, to the resonant frequency but will not respond to frequencies far from resonance. Hence frequencies far from resonance are filtered out from the output of the device.

### Electrical resonance



A 1915 example of an early type of resonant circuit known as an Oudin coil which uses Leyden jars for the capacitance.

Resonance was noticed early on in experiments with the Leyden jar, invented in 1746. The Leyden jar stores electricity due to its capacitance, and is, in fact, an early form of capacitor. When a Leyden jar is discharged by allowing a spark to jump between the electrodes, the discharge is oscillatory. This was not suspected until 1826, when Felix Savary in France, and later (1842) Joseph Henry in the US noted that a steel needle placed close to the discharge does not always magnetise in the same direction. They both independently drew the conclusion that there was a transient oscillation dying with time.

Hermann von Helmholtz in 1847 published his important work on conservation of energy in part of which he used those principles to explain why the oscillation dies away, that it is the resistance of the circuit which dissipates the energy of the oscillation on each successive cycle. Helmholtz also noted that there was evidence of oscillation from the electrolysis experiments of William Hyde Wollaston. Wollaston was attempting to decompose water by electric shock but found that both hydrogen and oxygen were present at both electrodes. In normal electrolysis they would separate, one to each electrode.

Helmholtz explained why the oscillation decayed but he had not explained why it occurred in the first place. This was left to Sir William Thomson (Lord Kelvin) who, in 1853, postulated that there was inductance present in the circuit as well as the capacitance of the jar and the resistance of the load. This established the physical basis for the phenomenon - the energy supplied by the jar was partly dissipated in the load but also partly stored in the magnetic field of the inductor.

So far, the investigation had been on the natural frequency of transient oscillation of a resonant circuit resulting from a sudden stimulus. More important from the point of view of filter theory is the behaviour of a resonant circuit when driven by an external AC signal: there is a sudden peak in the circuits response when the driving signal frequency is at the resonant frequency of the circuit. James Clerk Maxwell heard of the phenomenon from Sir William Grove in 1868 in connection with experiments on dynamos, and was also aware of the earlier work of Henry Wilde in 1866. Maxwell explained resonance mathematically, with a set of differential equations, in much the same terms that an RLC circuit is described today.

Heinrich Hertz (1887) experimentally demonstrated the resonance phenomena by building two resonant circuits, one of which was driven by a generator and the other was tunable and only coupled to the first electromagnetically (i.e., no circuit connection). Hertz showed that the response of the second circuit was at a maximum when it was in tune with the first. The diagrams produced by Hertz in this paper were the first published plots of an electrical resonant response.

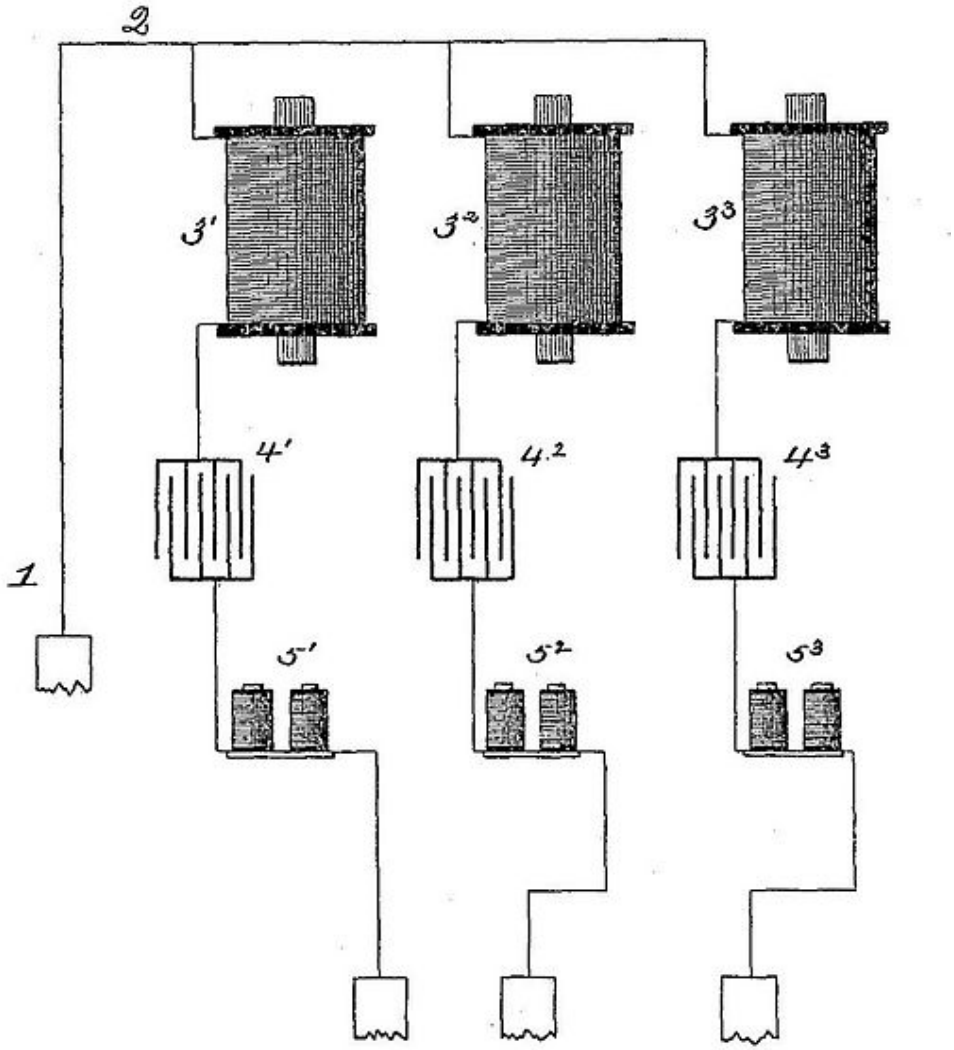
## Acoustic resonance

As mentioned earlier, it was acoustic resonance that inspired filtering applications, the first of these being a telegraph system known as the "harmonic telegraph". Versions are due to Elisha Gray, Alexander Graham Bell (1870s), Ernest Mercadier and others. Its

purpose was to simultaneously transmit a number of telegraph messages over the same line and represents an early form of frequency division multiplexing (FDM). FDM requires the sending end to be transmitting at different frequencies for each individual communication channel. This demands individual tuned resonators, as well as filters to separate out the signals at the receiving end. The harmonic telegraph achieved this with electromagnetically driven tuned reeds at the transmitting end which would vibrate similar reeds at the receiving end. Only the reed with the same resonant frequency as the transmitter would vibrate to any appreciable extent at the receiving end.

Incidentally, the harmonic telegraph directly suggested to Bell the idea of the telephone. The reeds can be viewed as transducers converting sound to and from an electrical signal. It is no great leap from this view of the harmonic telegraph to the idea that speech can be converted to and from an electrical signal.

## Early multiplexing



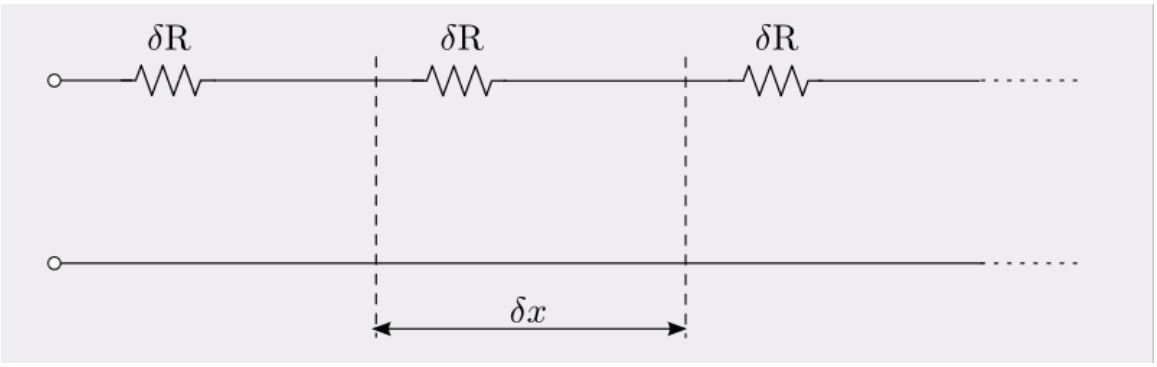

Hutin and Leblanc's multiple telegraph filter of 1891 showing the use of resonant circuits in filtering.

By the 1890s electrical resonance was much more widely understood and had become a normal part of the engineer's toolkit. In 1891 Hutin and Leblanc patented an FDM scheme for telephone circuits using resonant circuit filters. Rival patents were filed in 1892 by Michael Pupin and John Stone Stone with similar ideas, priority eventually being awarded to Pupin. However, no scheme using just simple resonant circuit filters can successfully multiplex (i.e. combine) the wider bandwidth of telephone channels (as opposed to telegraph) without either an unacceptable restriction of speech bandwidth or a channel spacing so wide as to make the benefits of multiplexing uneconomic.

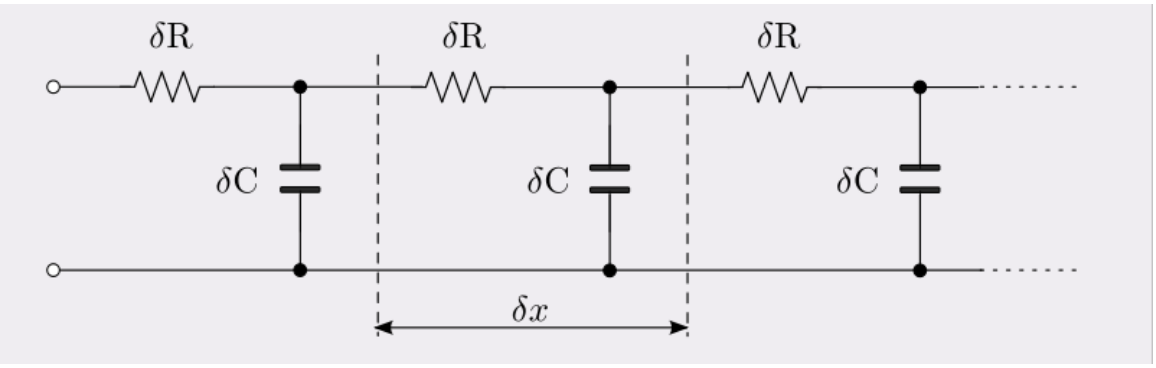
The basic technical reason for this difficulty is that the frequency response of a simple filter approaches a fall of 6 dB/octave far from the point of resonance. This means that if telephone channels are squeezed in side-by-side into the frequency spectrum, there will be crosstalk from adjacent channels in any given channel. What is required is a much more sophisticated filter that has a flat frequency response in the required passband like a low-Q resonant circuit, but that rapidly falls in response (much faster than 6 dB/octave) at the transition from passband to stopband like a high-Q resonant circuit. Obviously, these are contradictory requirements to be met with a single resonant circuit. The solution to these needs was founded in the theory of transmission lines and consequently the necessary filters did not become available until this theory was fully developed. At this early stage the idea of signal bandwidth, and hence the need for filters to match to it, was not fully understood; indeed, it was as late as 1920 before the concept of bandwidth was fully established. For early radio, the concepts of Q-factor, selectivity and tuning sufficed. This was all to change with the developing theory of transmission lines on which image filters are based, as explained in the next section.

At the turn of the century as telephone lines became available, it became popular to add telegraph on to telephone lines with an earth return phantom circuit. An LC filter was required to prevent telegraph clicks being heard on the telephone line. From the 1920s onwards, telephone lines, or balanced lines dedicated to the purpose, were used for FDM telegraph at audio frequencies. The first of these systems in the UK was a Siemens and Halske installation between London and Manchester. GEC and AT&T also had FDM systems. Separate pairs were used for the send and receive signals. The Siemens and GEC systems had six channels of telegraph in each direction, the AT&T system had twelve. All of these systems used electronic oscillators to generate a different carrier for each telegraph signal and required a bank of band-pass filters to separate out the multiplexed signal at the receiving end.

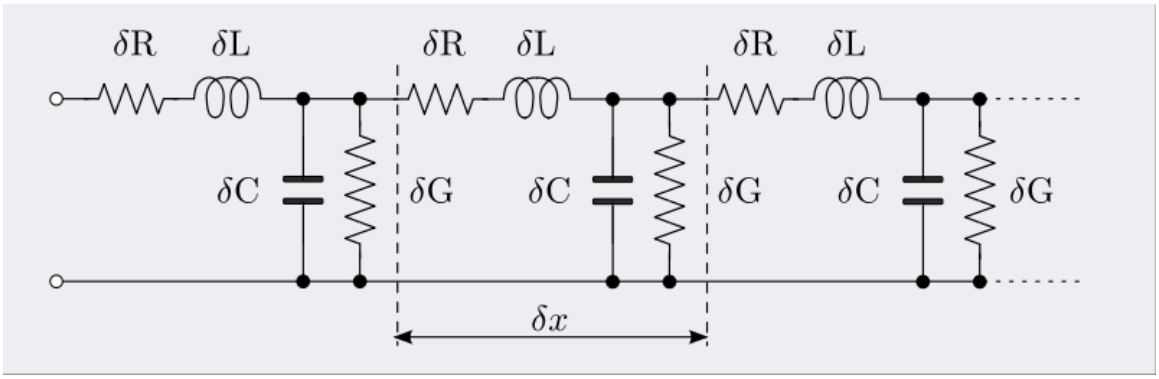
## *Transmission line theory*




Ohm's model of the transmission line was simply resistance.




Lord Kelvin's model of the transmission line accounted for capacitance and the dispersion it caused. The diagram represents Kelvin's model translated into modern terms using infinitesimal elements, but this was not the actual approach used by Kelvin.

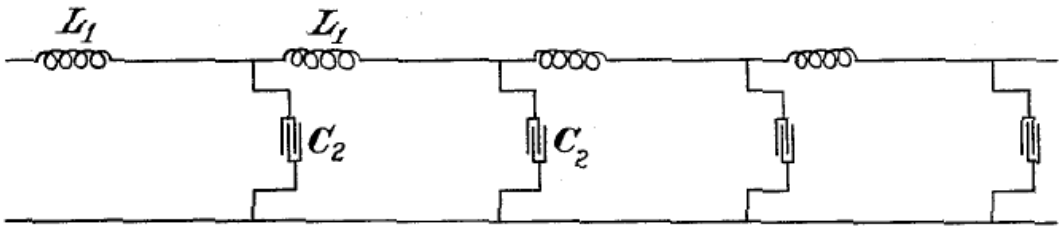



Heaviside's model of the transmission line. L, R, C and G in all three diagrams are the primary line constants. The infinitesimals δL, δR, δC and δG are to be understood as L$\delta x$, R$\delta x$, C$\delta x$ and G$\delta x$ respectively.

The earliest model of the transmission line was probably described by Georg Ohm (1827) who established that resistance in a wire is proportional to its length. The Ohm model thus included only resistance. Latimer Clark noted that signals were delayed and elongated along a cable, an undesirable form of distortion now called dispersion but then called retardation, and Michael Faraday (1853) established that this was due to the capacitance present in the transmission line. Lord Kelvin (1854) found the correct mathematical description needed in his work on early transatlantic cables; he arrived at an equation identical to the conduction of a heat pulse along a metal bar. This model incorporates only resistance and capacitance, but that is all that was needed in undersea cables dominated by capacitance effects. Kelvin's model predicts a limit on the telegraph signalling speed of a cable but Kelvin still did not use the concept of bandwidth, the limit was entirely explained in terms of the dispersion of the telegraph symbols. The mathematical model of the transmission line reached its fullest development with Oliver Heaviside. Heaviside (1881) introduced series inductance and shunt conductance into the model making four distributed elements in all. This model is now known as the telegrapher's equation and the distributed elements are called the primary line constants.

From the work of Heaviside (1887) it had become clear that the performance of telegraph lines, and most especially telephone lines, could be improved by the addition of inductance to the line. George Campbell at AT&T implemented this idea (1899) by inserting loading coils at intervals along the line. Campbell found that as well as the desired improvements to the line's characteristics in the passband there was also a definite frequency beyond which signals could not be passed without great attenuation. This was a result of the loading coils and the line capacitance forming a low-pass filter, an effect that is only apparent on lines incorporating lumped components such as the loading coils. This naturally led Campbell (1910) to produce a filter with ladder topology, a glance at the circuit diagram of this filter is enough to see its relationship to a loaded transmission line. The cut-off phenomenon is an undesirable side-effect as far as loaded lines are concerned but for telephone FDM filters it is precisely what is required. For this application, Campbell produced band-pass filters to the same ladder topology by replacing the inductors and capacitors with resonators and anti-resonators respectively. Both the loaded line and FDM were of great benefit economically to AT&T and this led to fast development of filtering from this point onwards.

## *Image filters*




Campbell's sketch of the low-pass version of his filter from his 1915 patent showing the now ubiquitous ladder topology with capacitors for the ladder rungs and inductors for the stiles. Filters of more modern design also often adopt the same ladder topology as used by Campbell. It should be understood that although superficially similar, they are really quite different. The ladder construction is essential to the Campbell filter and all the sections have identical element values. Modern designs can be realised in any number of topologies, choosing the ladder topology is merely a matter of convenience. Their response is quite different (better) than Campbell's and the element values, in general, will all be different.

The filters designed by Campbell were named wave filters because of their property of passing some waves and strongly rejecting others. The method by which they were designed was called the image parameter method and filters designed to this method are called image filters. The image method essentially consists of developing the transmission constants of an infinite chain of identical filter sections and then terminating the desired finite number of filter sections in the image impedance. This exactly corresponds to the way the properties of a finite length of transmission line are derived

from the theoretical properties of an infinite line, the image impedance corresponding to the characteristic impedance of the line.

From 1920 John Carson, also working for AT&T, began to develop a new way of looking at signals using the operational calculus of Heaviside which in essence is working in the frequency domain. This gave the AT&T engineers a new insight into the way their filters were working and led Otto Zobel to invent many improved forms. Carson and Zobel steadily demolished many of the old ideas. For instance the old telegraph engineers thought of the signal as being a single frequency and this idea persisted into the age of radio with some still believing that frequency modulation (FM) transmission could be achieved with a smaller bandwidth than the baseband signal right up until the publication of Carson's 1922 paper. Another advance concerned the nature of noise, Carson and Zobel (1923) treated noise as a random process with a continuous bandwidth, an idea that was well ahead of its time, and thus limited the amount of noise that it was possible to remove by filtering to that part of the noise spectrum which fell outside the passband. This too, was not generally accepted at first, notably being opposed by Edwin Armstrong (who ironically, actually succeeded in reducing noise with wide-band FM) and was only finally settled with the work of Harry Nyquist whose thermal noise power formula is well known today.

Several improvements were made to image filters and their theory of operation by Otto Zobel. Zobel coined the term constant k filter (or k-type filter) to distinguish Campbell's filter from later types, notably Zobel's m-derived filter (or m-type filter). The particular problems Zobel was trying to address with these new forms were impedance matching into the end terminations and improved steepness of roll-off. These were achieved at the cost of an increase in filter circuit complexity.

A more systematic method of producing image filters was introduced by Hendrik Bode (1930), and further developed by several other investigators including Piloty (1937-1939) and Wilhelm Cauer (1934-1937). Rather than enumerate the behaviour (transfer function, attenuation function, delay function and so on) of a specific circuit, instead a requirement for the image impedance itself was developed. The image impedance can be expressed in terms of the open-circuit and short-circuit impedances of the filter as $Z_i=\sqrt{Z_oZ_s}$. Since the image impedance must be real in the passbands and imaginary in the stopbands according to image theory, there is a requirement that the poles and zeroes of $Z_o$ and $Z_s$ cancel in the passband and correspond in the stopband. The behaviour of the filter can be entirely defined in terms of the positions in the complex plane of these pairs of poles and zeroes. Any circuit which has the requisite poles and zeroes will also have the requisite response. Cauer pursued two related questions arising from this technique: what specification of poles and zeroes are realisable as passive filters; and what realisations are equivalent to each other. The results of this work led Cauer to develop a new approach, now called network synthesis.

This "poles and zeroes" view of filter design was particularly useful where a bank of filters, each operating at different frequencies, are all connected across the same transmission line. The earlier approach was unable to deal properly with this situation,

but the poles and zeroes approach could embrace it by specifying a constant impedance for the combined filter. This problem was originally related to FDM telephony but frequently now arises in loudspeaker crossover filters.

## *Network synthesis filters*

The essence of network synthesis is to start with a required filter response and produce a network that delivers that response, or approximates to it within a specified boundary. This is the inverse of network analysis which starts with a given network and by applying the various electric circuit theorems predicts the response of the network. The term was first used with this meaning in the doctoral thesis of Yuk-Wing Lee (1930) and apparently arose out of a conversation with Vannevar Bush. The advantage of network synthesis over previous methods is that it provides a solution which precisely meets the design specification. This is not the case with image filters, a degree of experience is required in their design since the image filter only meets the design specification in the unrealistic case of being terminated in its own image impedance, to produce which would require the exact circuit being sought. Network synthesis on the other hand, takes care of the termination impedances simply by incorporating them into the network being designed.

The development of network analysis needed to take place before network synthesis was possible. The theorems of Gustav Kirchhoff and others and the ideas of Charles Steinmetz (phasors) and Arthur Kennelly (complex impedance) laid the groundwork. The concept of a port also played a part in the development of the theory, and proved to be a more useful idea than network terminals. The first milestone on the way to network synthesis was an important paper by Ronald Foster (1924), *A Reactance Theorem*, in which Foster introduces the idea of a driving point impedance, that is, the impedance that is connected to the generator. The expression for this impedance determines the response of the filter and vice versa, and a realisation of the filter can be obtained by expansion of this expression. It is not possible to realise any arbitrary impedance expression as a network. Foster's reactance theorem stipulates necessary and sufficient conditions for realisability: that the reactance must be algebraically increasing with frequency and the poles and zeroes must alternate.

Wilhelm Cauer expanded on the work of Foster (1926) and was the first to talk of realisation of a one-port impedance with a prescribed frequency function. Foster's work considered only reactances (i.e., only LC-kind circuits). Cauer generalised this to any 2-element kind one-port network, finding there was an isomorphism between them. He also found ladder realisations of the network using Thomas Stieltjes' continued fraction expansion. This work was the basis on which network synthesis was built, although Cauer's work was not at first used much by engineers, partly because of the intervention of World War II, partly for reasons explained in the next section and partly because Cauer presented his results using topologies that required mutually coupled inductors and ideal transformers. Although on this last point, it has to be said that transformer coupled double tuned amplifiers are a common enough way of widening bandwidth without sacrificing selectivity.

## *Image method versus synthesis*

Image filters continued to be used by designers long after the superior network synthesis techniques were available. Part of the reason for this may have been simply inertia, but it was largely due to the greater computation required for network synthesis filters, often needing a mathematical iterative process. Image filters, in their simplest form, consist of a chain of repeated, identical sections. The design can be improved simply by adding more sections and the computation required to produce the initial section is on the level of "back of an envelope" designing. In the case of network synthesis filters, on the other hand, the filter is designed as a whole, single entity and to add more sections (i.e., increase the order) the designer would have no option but to go back to the beginning and start over. The advantages of synthesised designs are real, but they are not overwhelming compared to what a skilled image designer could achieve, and in many cases it was more cost effective to dispense with time-consuming calculations. This is simply not an issue with the modern availability of computing power, but in the 1950s it was non-existent, in the 1960s and 1970s available only at cost, and not finally becoming widely available to all designers until the 1980s with the advent of the desktop personal computer. Image filters continued to be designed up to that point and many remained in service into the 21st century.

The computational difficulty of the network synthesis method was addressed by tabulating the component values of a prototype filter and then scaling the frequency and impedance and transforming the bandform to those actually required. This kind of approach, or similar, was already in use with image filters, for instance by Zobel, but the concept of a "reference filter" is due to Sidney Darlington. Darlington (1939), was also the first to tabulate values for network synthesis prototype filters, nevertheless it had to wait until the 1950s before the Cauer-Darlington elliptic filter first came into use.

Once computational power was readily available, it became possible to easily design filters to minimise any arbitrary parameter, for example time delay or tolerance to component variation. The difficulties of the image method were firmly put in the past, and even the need for prototypes became largely superfluous. Furthermore, the advent of active filters eased the computation difficulty because sections could be isolated and iterative processes were not then generally necessary.

## *Realisability and equivalence*

Realisability (that is, which functions are realisable as real impedance networks) and equivalence (which networks equivalently have the same function) are two important questions in network synthesis. Following an analogy with Lagrangian mechanics, Cauer formed the matrix equation,

$$[\mathbf{A}] = s^2[\mathbf{L}] + s[\mathbf{R}] + [\mathbf{D}] = s[\mathbf{Z}]$$

where [**Z**],[**R**],[**L**] and [**D**] are the *n*x*n* matrices of, respectively, impedance, resistance, inductance and elastance of an *n*-mesh network and *s* is the complex frequency operator

$s=\sigma+i\omega$. Here [**R**],[**L**] and [**D**] have associated energies corresponding to the kinetic, potential and dissipative heat energies, respectively, in a mechanical system and the already known results from mechanics could be applied here. Cauer determined the driving point impedance by the method of Lagrange multipliers;

$$Z_{\mathrm{p}}(s) = \frac{\det[\mathbf{A}]}{s\, a_{11}}$$

where $a_{11}$ is the complement of the element $A_{11}$ to which the one-port is to be connected. From stability theory Cauer found that [**R**], [**L**] and [**D**] must all be positive-definite matrices for $Z_{\mathrm{p}}(s)$ to be realisable if ideal transformers are not excluded. Realisability is only otherwise restricted by practical limitations on topology. This work is also partly due to Otto Brune (1931), who worked with Cauer in the US prior to Cauer returning to Germany. A well known condition for realisability of a one-port rational impedance due to Cauer (1929) is that it must be a function of $s$ that is analytic in the right halfplane ($\sigma>0$), have a positive real part in the right halfplane and take on real values on the real axis. This follows from the Poisson integral representation of these functions. Brune coined the term positive-real for this class of function and proved that it was a necessary and sufficient condition (Cauer had only proved it to be necessary) and they extended the work to LC multiports. A theorem due to Sidney Darlington states that any positive-real function $Z(s)$ can be realised as a lossless two-port terminated in a positive resistor R. No resistors within the network are necessary to realise the specified response.

As for equivalence, Cauer found that the group of real affine transformations,

$$[\mathbf{T}]^T[\mathbf{A}][\mathbf{T}]$$

where,

$$[\mathbf{T}] = \begin{bmatrix} 1 & 0\cdots 0 \\ T_{21} & T_{22}\cdots T_{2n} \\ \cdot & \cdots \\ T_{n1} & T_{n2}\cdots T_{nn} \end{bmatrix}$$

is invariant in $Z_{\mathrm{p}}(s)$, that is, all the transformed networks are equivalents of the original.

## Approximation

The approximation problem in network synthesis is to find functions which will produce realisable networks approximating to a prescribed function of frequency within limits arbitrarily set. The approximation problem is an important issue since the ideal function of frequency required will commonly be unachievable with rational networks. For instance, the ideal prescribed function is often taken to be the unachievable lossless transmission in the passband, infinite attenuation in the stopband and a vertical transition between the two. However, the ideal function can be approximated with a rational function, becoming ever closer to the ideal the higher the order of the polynomial. The

first to address this problem was Stephen Butterworth (1930) using his Butterworth polynomials. Independently, Cauer (1931) used Chebyshev polynomials, initially applied to image filters, and not to the now well-known ladder realisation of this filter.

## Butterworth filter

Butterworth filters are an important class of filters due to Stephen Butterworth (1930) which are now recognised as being a special case of Cauer's elliptic filters. Butterworth discovered this filter independently of Cauer's work and implemented it in his version with each section isolated from the next with a valve amplifier which made calculation of component values easy since the filter sections could not interact with each other and each section represented one term in the Butterworth polynomials. This gives Butterworth the credit for being both the first to deviate from image parameter theory and the first to design active filters. It was later shown that Butterworth filters could be implemented in ladder topology without the need for amplifiers, possibly the first to do so was William Bennett (1932) in a patent which presents formulae for component values identical to the modern ones. Bennett, at this stage though, is still discussing the design as an artificial transmission line and so is adopting an image parameter approach despite having produced what would now be considered a network synthesis design. He also does not appear to be aware of the work of Butterworth or the connection between them.

## Insertion-loss method

The insertion-loss method of designing filters is, in essence, to prescribe a desired function of frequency for the filter as an attenuation of the signal when the filter is inserted between the terminations relative to the level that would have been received were the terminations connected to each other via an ideal transformer perfectly matching them. Versions of this theory are due to Sidney Darlington, Wilhelm Cauer and others all working more or less independently and is often taken as synonymous with network synthesis. Butterworth's filter implementation is, in those terms, an insertion-loss filter, but it is a relatively trivial one mathematically since the active amplifiers used by Butterworth ensured that each stage individually worked into a resistive load. Butterworth's filter becomes a non-trivial example when it is implemented entirely with passive components. An even earlier filter which influenced the insertion-loss method was Norton's dual-band filter where the input of two filters are connected in parallel and designed so that the combined input presents a constant resistance. Norton's design method, together with Cauer's canonical LC networks and Darlington's theorem that only LC components were required in the body of the filter resulted in the insertion-loss method. However, ladder topology proved to be more practical than Cauer's canonical forms.

Darlington's insertion-loss method is a generalisation of the procedure used by Norton. In Norton's filter it can be shown that each filter is equivalent to a separate filter unterminated at the common end. Darlington's method applies to the more straightforward and general case of a 2-port LC network terminated at both ends. The procedure consists of the following steps:

1. determine the poles of the prescribed insertion-loss function,
2. from that find the complex transmission function,
3. from that find the complex reflection coefficients at the terminating resistors,
4. find the driving point impedance from the short-circuit and open-circuit impedances,
5. expand the driving point impedance into an LC (usually ladder) network.

Darlington additionally used a transformation found by Hendrik Bode that predicted the response of a filter using non-ideal components but all with the same *Q*. Darlington used this transformation in reverse to produce filters with a prescribed insertion-loss with non-ideal components. Such filters have the ideal insertion-loss response plus a flat attenuation across all frequencies.

## Elliptic filters

Elliptic filters are filters produced by the insertion-loss method which use elliptic rational functions in their transfer function as an approximation to the ideal filter response and the result is called a Chebyshev approximation. This is the same Chebyshev approximation technique used by Cauer on image filters but follows the Darlington insertion-loss design method and uses slightly different elliptic functions. Cauer had some contact with Darlington and Bell Labs before WWII (for a time he worked in the US) but during the war they worked independently, in some cases making the same discoveries. Cauer had disclosed the Chebyshev approximation to Bell Labs but had not left them with the proof. Sergei Schelkunoff provided this and a generalisation to all equal ripple problems. Elliptic filters are a general class of filter which incorporate several other important classes as special cases: Cauer filter (equal ripple in passband and stopband), Chebyshev filter (ripple only in passband), reverse Chebyshev filter (ripple only in stopband) and Butterworth filter (no ripple in either band).

Generally, for insertion-loss filters where the transmission zeroes and infinite losses are all on the real axis of the complex frequency plane (which they usually are for minimum component count), the insertion-loss function can be written as;

$$\frac{1}{1 + JF^2}$$

where *F* is either an even (resulting in an antimetric filter) or an odd (resulting in an symmetric filter) function of frequency. Zeroes of *F* correspond to zero loss and the poles of *F* correspond to transmission zeroes. *J* sets the passband ripple height and the stopband loss and these two design requirements can be interchanged. The zeroes and poles of *F* and *J* can be set arbitrarily. The nature of *F* determines the class of the filter;

- if *F* is a Chebyshev approximation the result is a Chebyshev filter,
- if *F* is a maximally flat approximation the result is a passband maximally flat filter,
- if 1/*F* is a Chebyshev approximation the result is a reverse Chebyshev filter,

- if $1/F$ is a maximally flat approximation the result is a stopband maximally flat filter,

A Chebyshev response simultaneously in the passband and stopband is possible, such as Cauer's equal ripple elliptic filter.

Darlington relates that he found in the New York City library Carl Jacobi's original paper on elliptic functions, published in Latin in 1829. In this paper Darlington was surprised to find foldout tables of the exact elliptic function transformations needed for Chebyshev approximations of both Cauer's image parameter, and Darlington's insertion-loss filters.

### Other methods

Darlington considers the topology of coupled tuned circuits to involve a separate approximation technique to the insertion-loss method, but also producing nominally flat passbands and high attenuation stopbands. The most common topology for these is shunt anti-resonators coupled by series capacitors, less commonly, by inductors, or in the case of a two-section filter, by mutual inductance. These are most useful where the design requirement is not too stringent, that is, moderate bandwidth, roll-off and passband ripple.

## *Other notable developments and applications*

### Mechanical filters

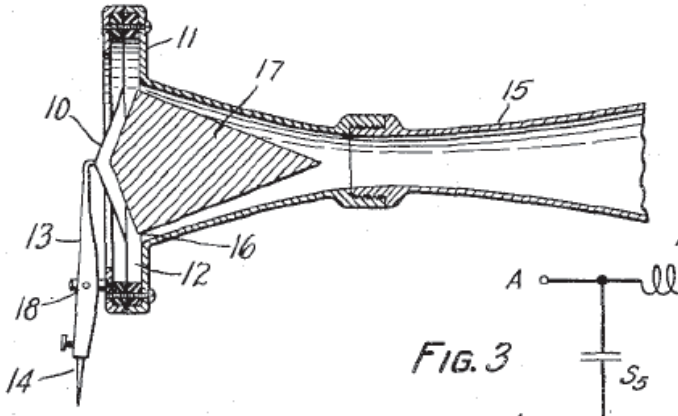



FIG. 2

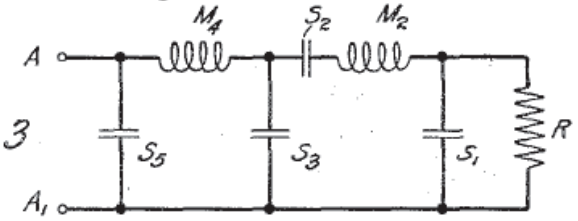



FIG. 3

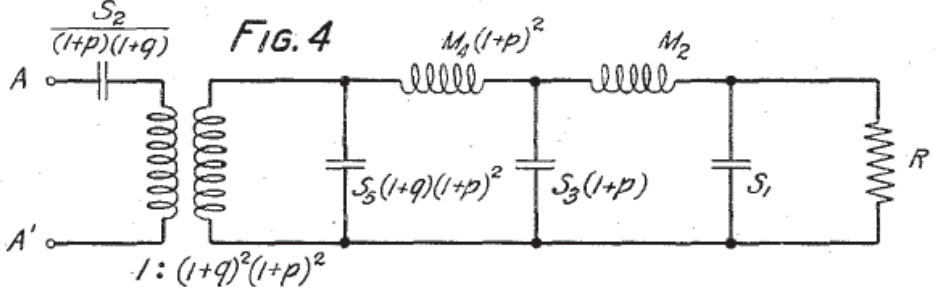



FIG. 4

Norton's mechanical filter together with its electrical equivalent circuit. Two equivalents are shown, "Fig.3" directly corresponds to the physical relationship of the mechanical components; "Fig.4" is an equivalent transformed circuit arrived at by repeated application of a well known transform, the purpose being to remove the series resonant circuit from the body of the filter leaving a simple *LC* ladder network.

Edward Norton, around 1930, designed a mechanical filter for use on phonograph recorders and players. Norton designed the filter in the electrical domain and then used the correspondence of mechanical quantities to electrical quantities to realise the filter using mechanical components. Mass corresponds to inductance, stiffness to elastance and damping to resistance. The filter was designed to have a maximally flat frequency response.

In modern designs it is common to use quartz crystal filters, especially for narrowband filtering applications. The signal exists as a mechanical acoustic wave while it is in the crystal and is converted by transducers between the electrical and mechanical domains at the terminals of the crystal.

## Transversal filters

Transversal filters are not usually associated with passive implementations but the concept can be found in a Wiener and Lee patent from 1935 which describes a filter consisting of a cascade of all-pass sections. The outputs of the various sections are summed in the proportions needed to result in the required frequency function. This works by the principle that certain frequencies will be in, or close to antiphase, at different sections and will tend to cancel when added. These are the frequencies rejected by the filter and can produce filters with very sharp cut-offs. This approach did not find any immediate applications, and is not common in passive filters. However, the principle finds many applications as an active delay line implementation for wide band discrete-time filter applications such as television, radar and high-speed data transmission.

## Matched filter

The purpose of matched filters is to maximise the signal-to-noise ratio (S/N) at the expense of pulse shape. Pulse shape, unlike many other applications, is unimportant in radar while S/N is the primary limitation on performance. The filters were introduced during WWII (described 1943) by Dwight North and are often eponymously referred to as "North filters".

## Filters for control systems

Control systems have a need for smoothing filters in their feedback loops with criteria to maximise the speed of movement of a mechanical system to the prescribed mark and at the same time minimise overshoot and noise induced motions. A key problem here is the extraction of Gaussian signals from a noisy background. An early paper on this was

published during WWII by Norbert Wiener with the specific application to anti-aircraft fire control analogue computers. Rudy Kalman (Kalman filter) later reformulated this in terms of state-space smoothing and prediction where it is known as the linear-quadratic-Gaussian control problem. Kalman started an interest in state-space solutions, but according to Darlington this approach can also be found in the work of Heaviside and earlier.

## *Modern practice*

LC passive filters gradually became less popular as active amplifying elements, particularly operational amplifiers, became cheaply available. The reason for the change is that wound components (the usual method of manufacture for inductors) are far from ideal, the wire adding resistance as well as inductance to the component. Inductors are also relatively expensive and are not "off-the-shelf" components. On the other hand, the function of LC ladder sections, LC resonators and RL sections can be replaced by RC components in an amplifier feedback loop (active filters). These components will usually be much more cost effective, and smaller as well. Cheap digital technology, in its turn, has largely supplanted analogue implementations of filters. However, there is still an occasional place for them in the simpler applications such as coupling where sophisticated functions of frequency are not needed.

**Chapter 4**

# Composite Image Filter

A **composite image filter** is an electronic filter consisting of multiple image filter sections of two or more different types.

The image method of filter design determines the properties of filter sections by calculating the properties they have in an infinite chain of such sections. In this, the analysis parallels transmission line theory on which it is based. Filters designed by this method are called *image parameter filters*, or just *image filters*. An important parameter of image filters is their image impedance, the impedance of an infinite chain of identical sections.

The basic sections are arranged into a ladder network of several sections, the number of sections required is mostly determined by the amount of stopband rejection required. In its simplest form, the filter can consist entirely of identical sections. However, it is more usual to use a composite filter of two or three different types of section to improve different parameters best addressed by a particular type. The most frequent parameters considered are stopband rejection, steepness of the filter skirt (transition band) and impedance matching to the filter terminations.

Image filters are linear filters and are invariably also passive in implementation.

## *History*

The image method of designing filters originated at AT&T, who were interested in developing filtering that could be used with the multiplexing of many telephone channels on to a single cable. The researchers involved in this work and their contributions are briefly listed below;

- John Carson provided the mathematical underpinning to the theory. He invented single sideband modulation for the purpose of multiplexing telephone channels. It was the need to recover these signals that gave rise to the need for advanced filtering techniques. He also pioneered the use of operational calculus (what has now become Laplace transforms in its more formal mathematical guise) to analyse these signals.
- George Campbell worked on filtering from 1910 onwards and invented the constant k filter. This can be seen as a continuation of his work on loading coils on transmission lines, a concept invented by Oliver Heaviside. Heaviside, incidentally, also invented the operational calculus used by Carson.
- Otto Zobel provided a theoretical basis (and the name) for Campbell's filters. In 1920 he invented the m-derived filter. Zobel also published composite designs incorporating both constant k and m-derived sections.
- R S Hoyt also contributed.

## *The image method*

The image analysis starts with a calculation of the input and output impedances (the image impedances) and the transfer function of a section in an infinite chain of identical sections. This can be shown to be equivalent to the performance of a section terminated in its image impedances. The image method, therefore, relies on each filter section being terminated with the correct image impedance. This is easy enough to do with the internal sections of a multiple section filter, because it is only necessary to ensure that the sections facing the one in question have identical image impedances. However, the end sections are a problem. They will usually be terminated with fixed resistances that the filter cannot match perfectly except at one specific frequency. This mismatch leads to multiple reflections at the filter terminations and at the junctions between sections. These reflections result in the filter response deviating quite sharply from the theoretical, especially near the cut-off frequency.

The requirement for better matching to the end impedances is one of the main motivations for using composite filters. A section designed to give good matching is used at the ends but something else (for instance stopband rejection or passband to stopband transition) is designed for the body of the filter.

## *Filter section types*

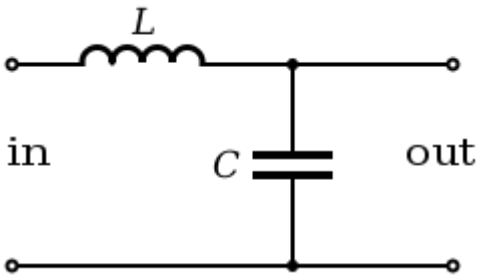
Each filter section type has particular advantages and disadvantages and each has the capability to improve particular filter parameters. The sections described below are the prototype filters for low-pass sections. These prototypes may be scaled and transformed to the desired frequency bandform (low-pass, high-pass, band-pass or band-stop).

The smallest unit of an image filter is an L half-section. Because the L section is not symmetrical, it has different image impedances ($Z_\mathrm{i}$) on each side. These are denoted $Z_\mathrm{iT}$and $Z_\mathrm{i\Pi}$. The T and the Π in the suffix refer to the shape of the filter section that would be formed if two half sections were to be connected back-to-back. T and Π are the

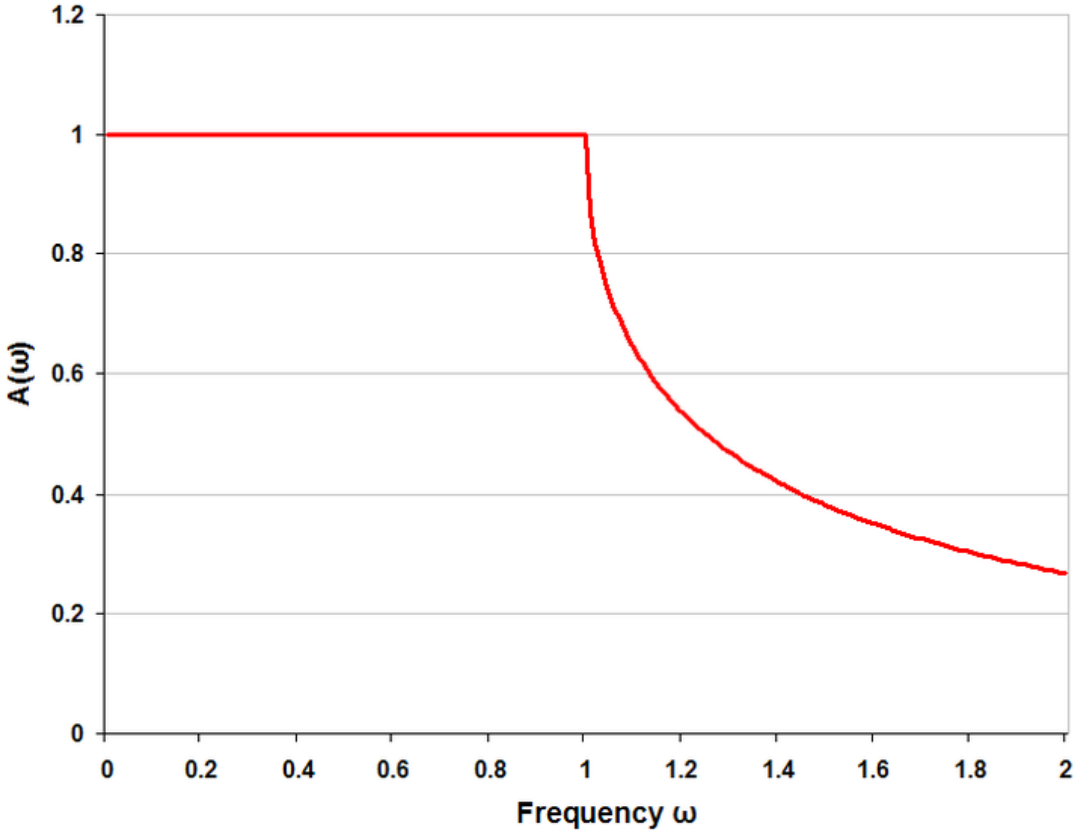
smallest symmetrical sections that can be constructed, as shown in diagrams in the topology chart (below). Where the section in question has an image impedance different from the general case a further suffix is added identifying the section type, for instance $Z_{\mathrm{iT}m}$.

## Constant k section

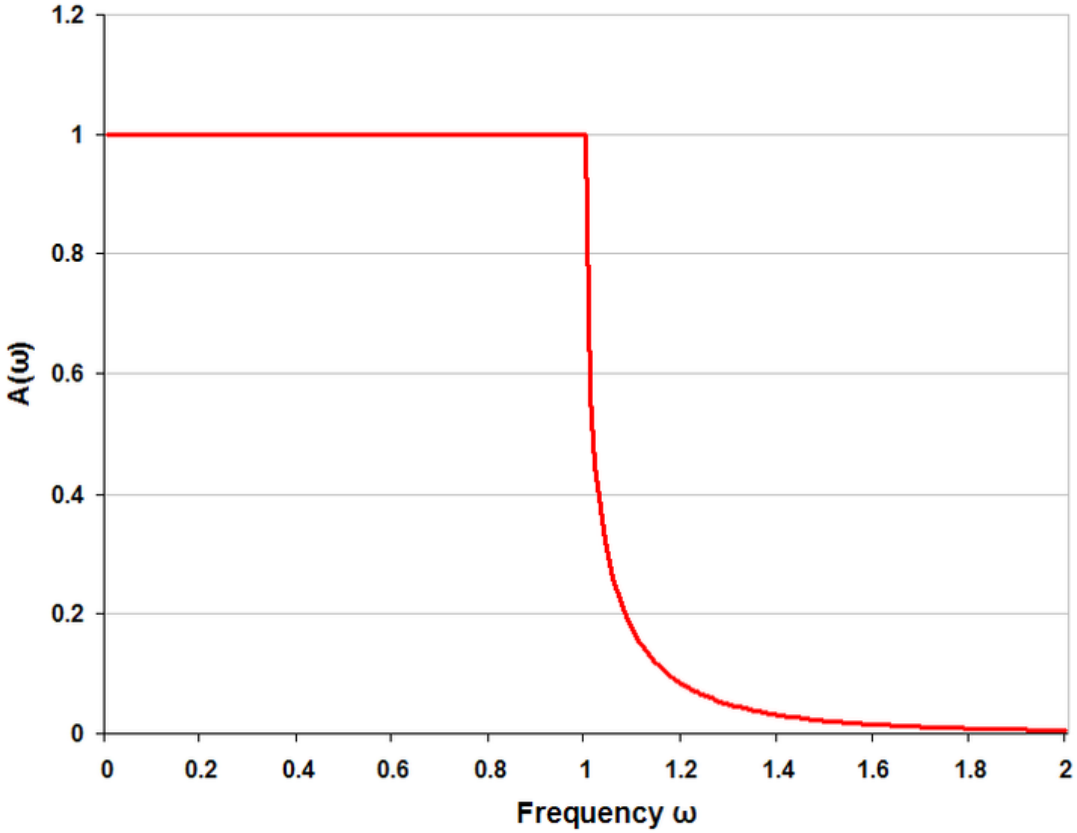
The **constant k** or **k-type** filter section is the basic image filter section. It is also the simplest circuit topology. The k-type has moderately fast transition from the passband to the stopband and moderately good stopband rejection.




k-type low-pass filter half section




k-type low-pass response, single half-section




k-type low-pass response with four (half) sections

## m-derived section

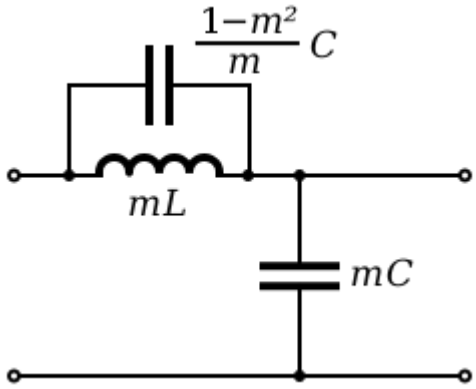
The **m-derived** or **m-type** filter section is a development of the k-type section. The most prominent feature of the m-type is a pole of attenuation just past the cut-off frequency inside the stopband. The parameter $m$ ($0<m<1$) adjusts the position of this pole of attenuation. Smaller values of $m$ put the pole closer to the cut-off frequency. Larger values of $m$ put it further away. In the limit, as $m$ approaches unity, the pole approaches $\omega$ of infinity and the section approaches a k-type section.

The m-type has a particularly fast cut-off, going from fully pass at the cut-off frequency to fully stop at the pole frequency. The cut-off can be made faster by moving the pole nearer to the cut-off frequency. This filter has the fastest cut-off of any filter design; note that the fast transition is achieved with just a single section, there is no need for multiple sections. The drawback with m-type sections is that they have poor stopband rejection past the pole of attenuation.

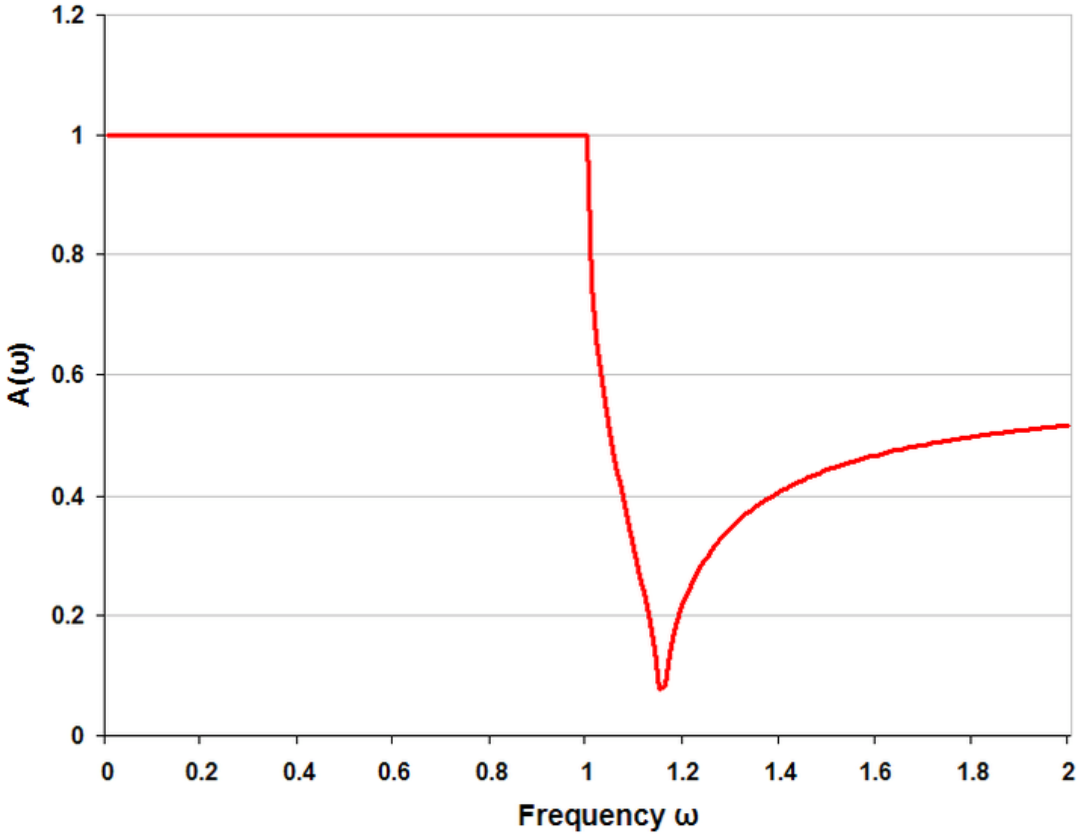
There is a particularly useful property of m-type filters with $m$=0.6. These have maximally flat image impedance $Z_{im}$ in the passband. They are therefore good for

matching in to the filter terminations, in the passband at least, the stopband is another story.

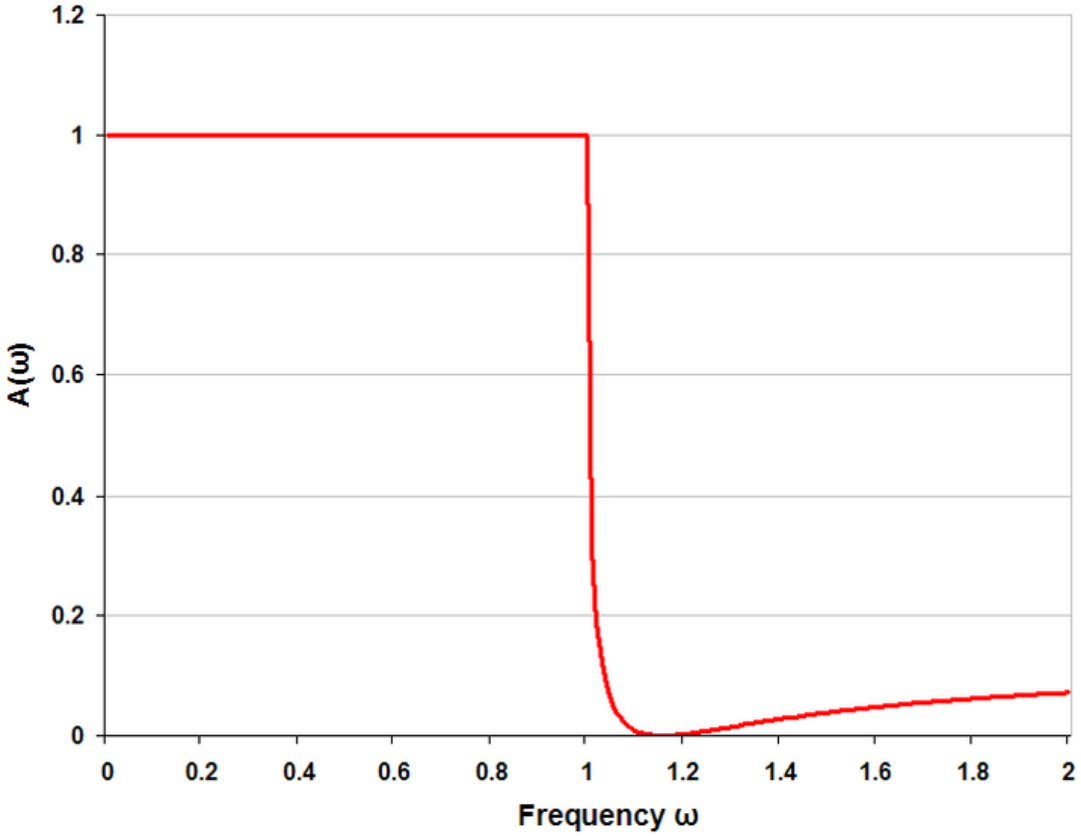
There are two variants of the m-type section, *series* and *shunt*. They have identical transfer functions but their image impedances are different. The shunt half-section has an image impedance which matches $Z_{\mathrm{i}\Pi}$on one side but has a different impedance, $Z_{\mathrm{i}\mathrm{T}m}$on the other. The series half-section matches $Z_{\mathrm{i}\mathrm{T}}$on one side and has $Z_{\mathrm{i}\Pi m}$on the other.



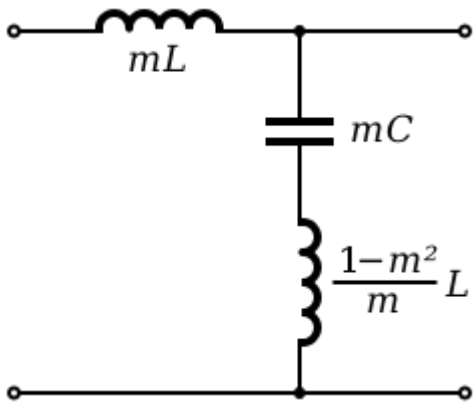

m-type low-pass filter shunt half section



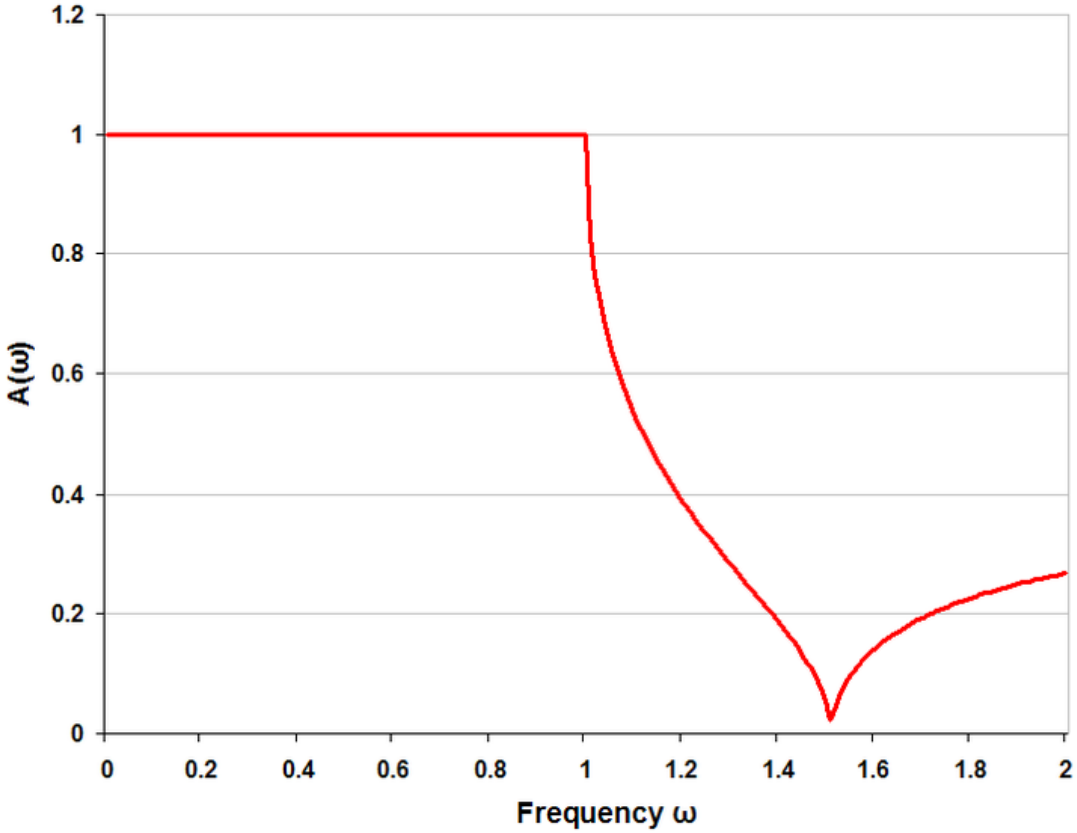

m-type low-pass response single half-section *m*=0.5



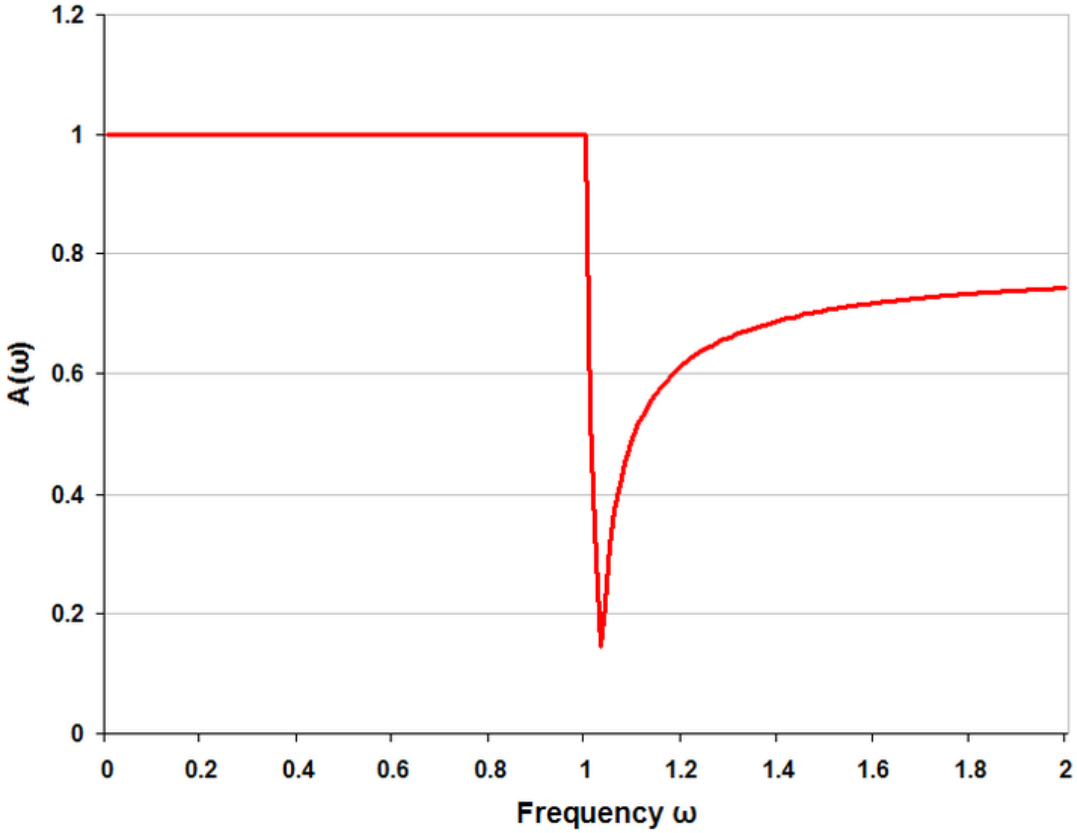

m-type low-pass response with four (half) sections $m$=0.5




m-type low-pass filter series half section




m-type low-pass response single half-section *m*=0.75




m-type low-pass response single half-section $m$=0.25

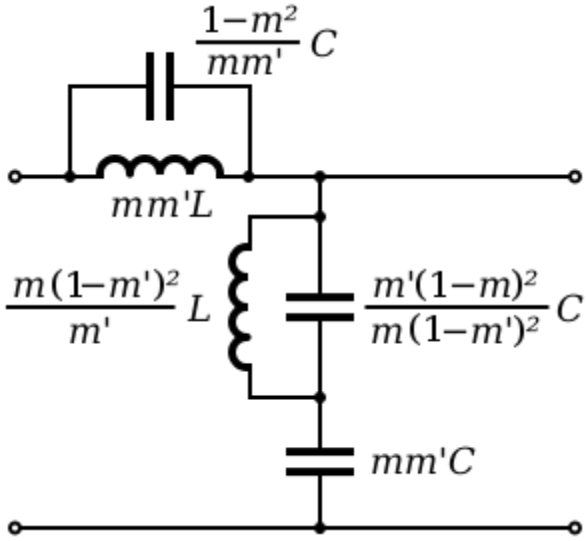
## mm'-type section

The **mm'-type** section has two independent parameters ($m$ and $m'$) that the designer can adjust. It is arrived at by double application of the $m$-derivation process. Its chief advantage is that it rather better at matching in to resistive end terminations than the k-type or m-type. The image impedance of a half-section is $Z_{\mathrm{i}m}$on one side and a different impedance, $Z_{\mathrm{i}mm'}$on the other. Like the m-type, this section can be constructed as a series or shunt section and the image impedances will come in T and Π variants. Either a series construction is applied to a shunt m-type or a shunt construction is applied to a series m-type. The advantages of the $mm'$-type filter are achieved at the expense of greater circuit complexity so it would normally only be used where it is needed for impedance matching purposes and not in the body of the filter.

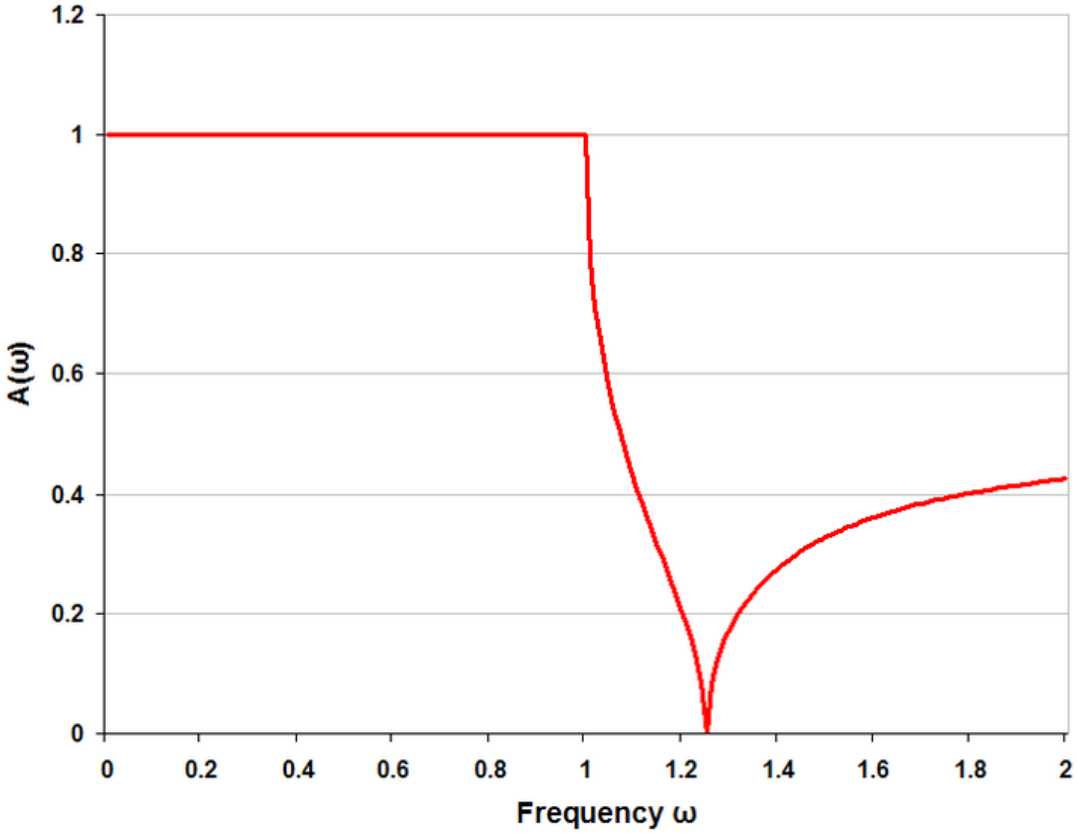
The transfer function of an mm'-type is the same as an m-type with $m$ set to the product $mm'$. To choose values of $m$ and $m'$ for best impedance match requires the designer to choose two frequencies at which the match is to be exact, at other frequencies there will be some deviation. There is thus some leeway in the choice but Zobel suggests the values $m$=0.7230 and $m'$=0.4134 which give a deviation of the impedance of less than 2% over

the useful part of the band. Since *mm*'=0.3, this section will also have a much faster cut-off than an m-type of *m*=0.6 which is an alternative for impedance matching.

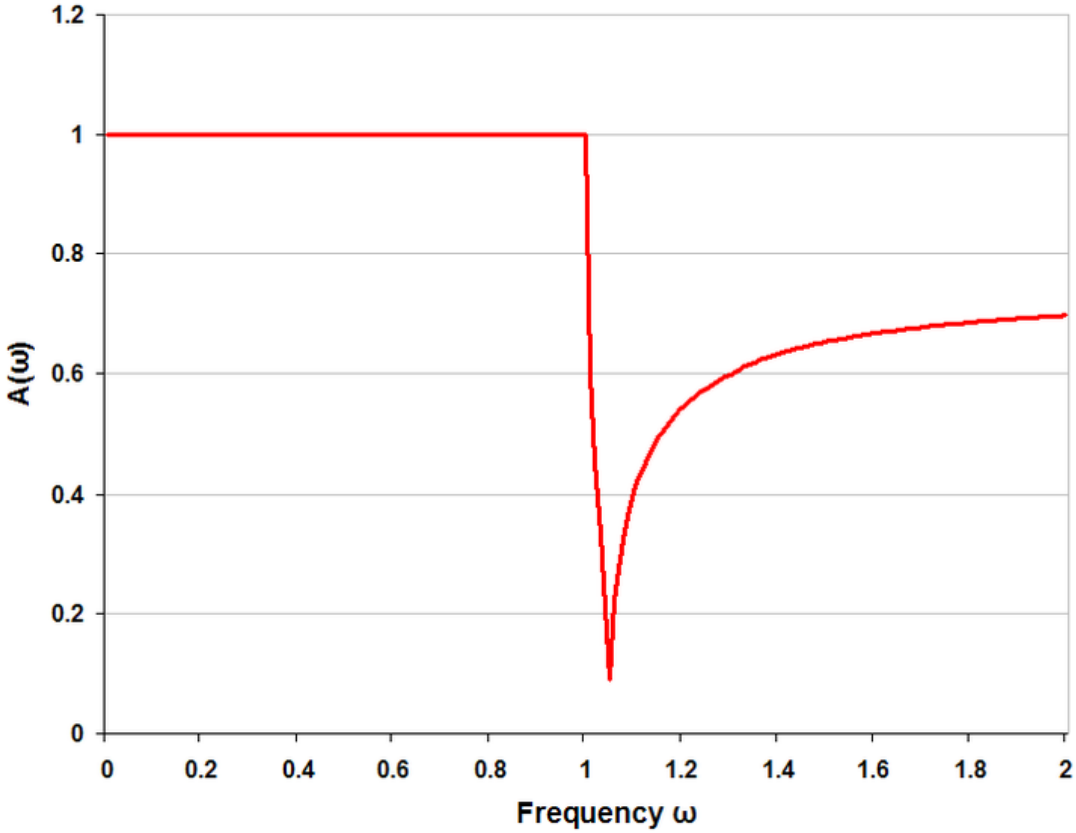
It is possible to continue the m-derivation process repeatedly and produce mm'm"-types and so on. However, the improvements obtained diminish at each iteration and are not usually worth the increase in complexity.




mm'-type low-pass filter series half section

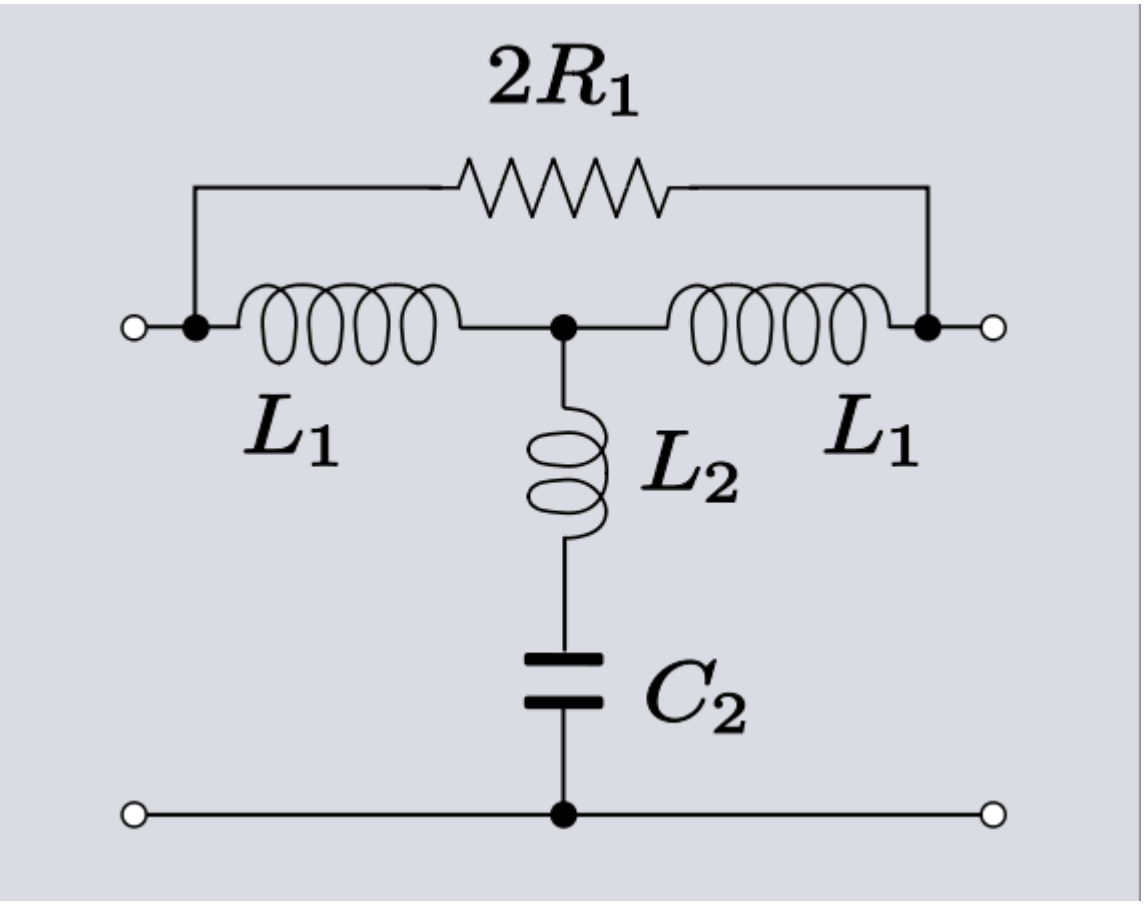



m-type low-pass response single half-section *m*=0.6




mm'-type low-pass response single half-section *mm*'=0.3

## Bode's filter



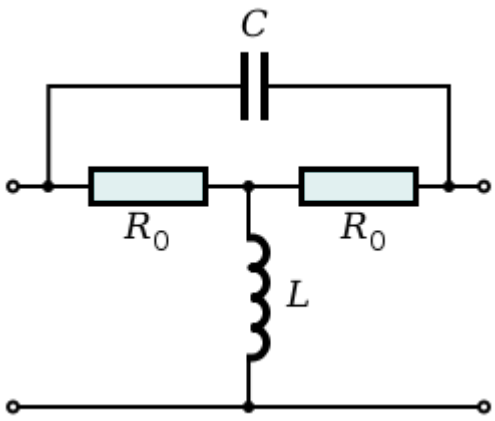

One incarnation of Bode's filter as a low-pass filter.

Another variation on the m-type filter was described by Hendrik Bode. This filter uses as a prototype a mid-series m-derived filter and transforms this into a bridged-T topology with the addition of a bridging resistor. This section has the advantage of being able to place the pole of attenuation much closer to the cut-off frequency than the Zobel filter, which starts to fail to work properly with very small values of *m* because of inductor resistance.

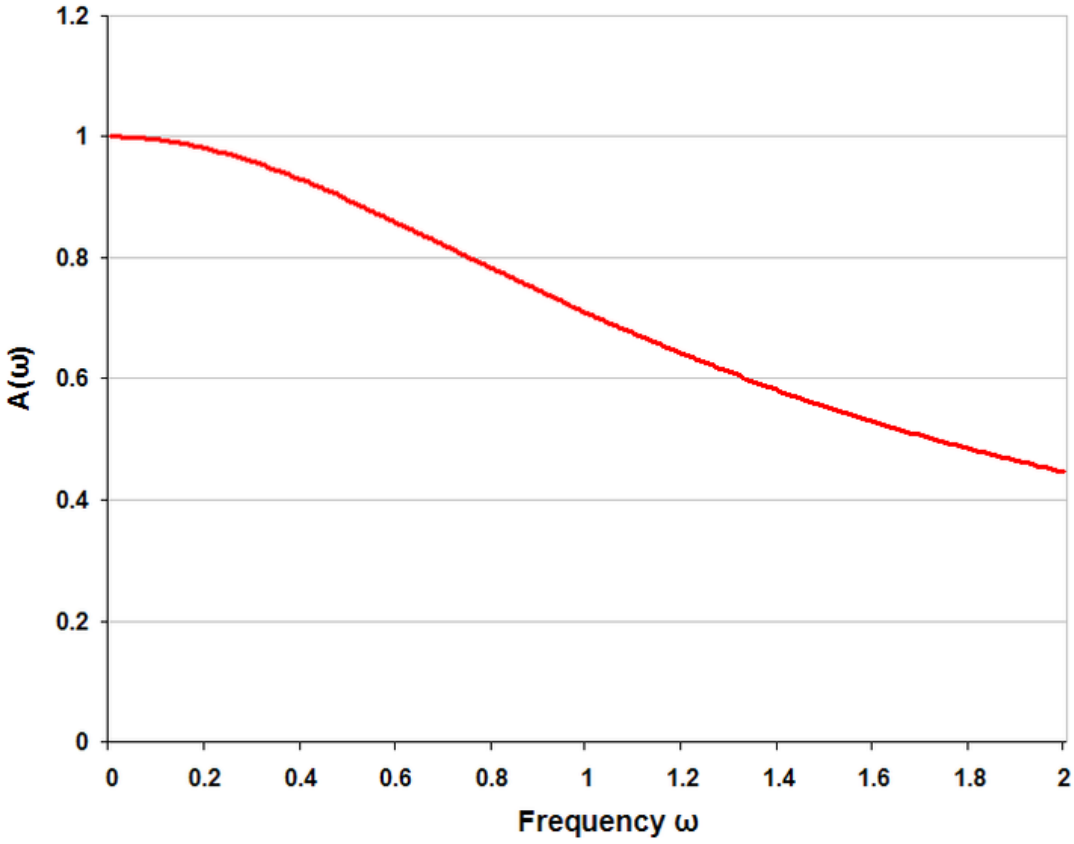
## Zobel network

The distinguishing feature of **Zobel network** filters is that they have a constant resistance image impedance and for this reason are also known as **constant resistance networks**. Clearly, the Zobel network filter does not have a problem matching to its terminations and this is its main advantage. However, other filter types have steeper transfer functions and sharper cut-offs. In filtering applications, the main role of Zobel networks is as equalisation filters. Zobel networks are in a different group from other image filters. The constant resistance means that when used in combination with other image filter sections

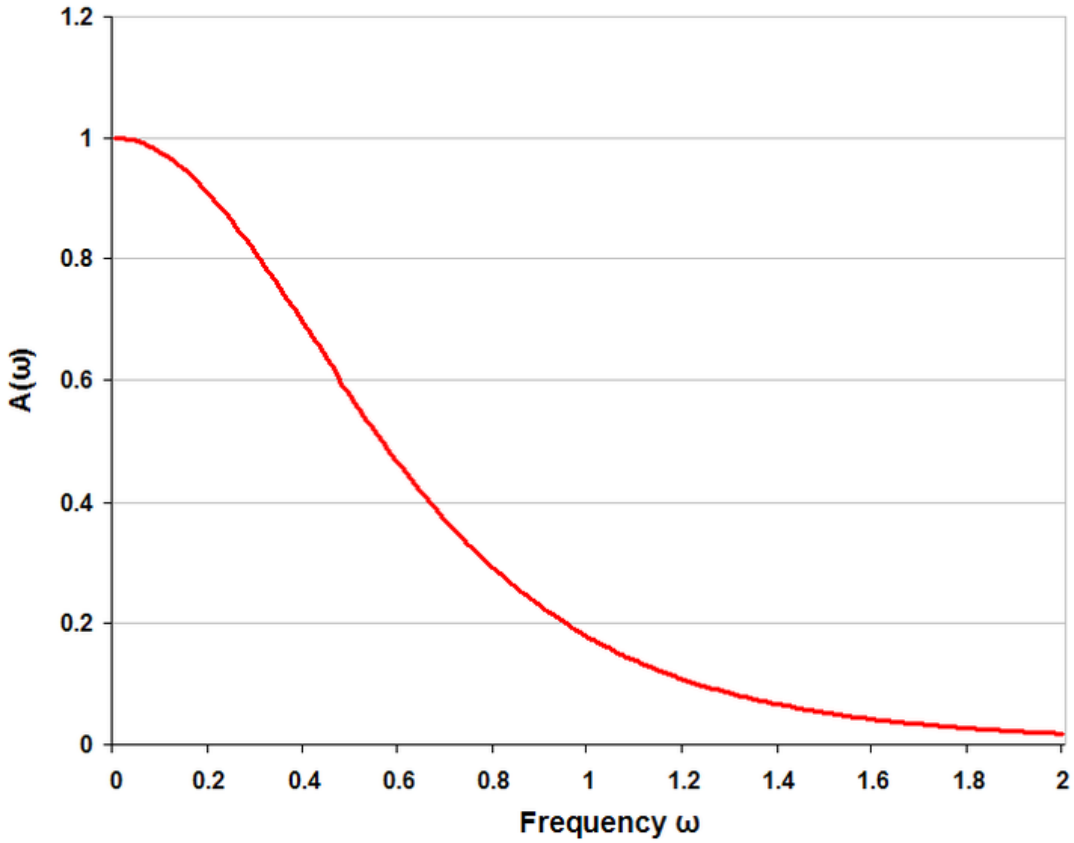
the same problem of matching arises as with end terminations. Zobel networks also suffer the disadvantage of using far more components than other equivalent image sections.




Zobel network bridge T high-pass filter section



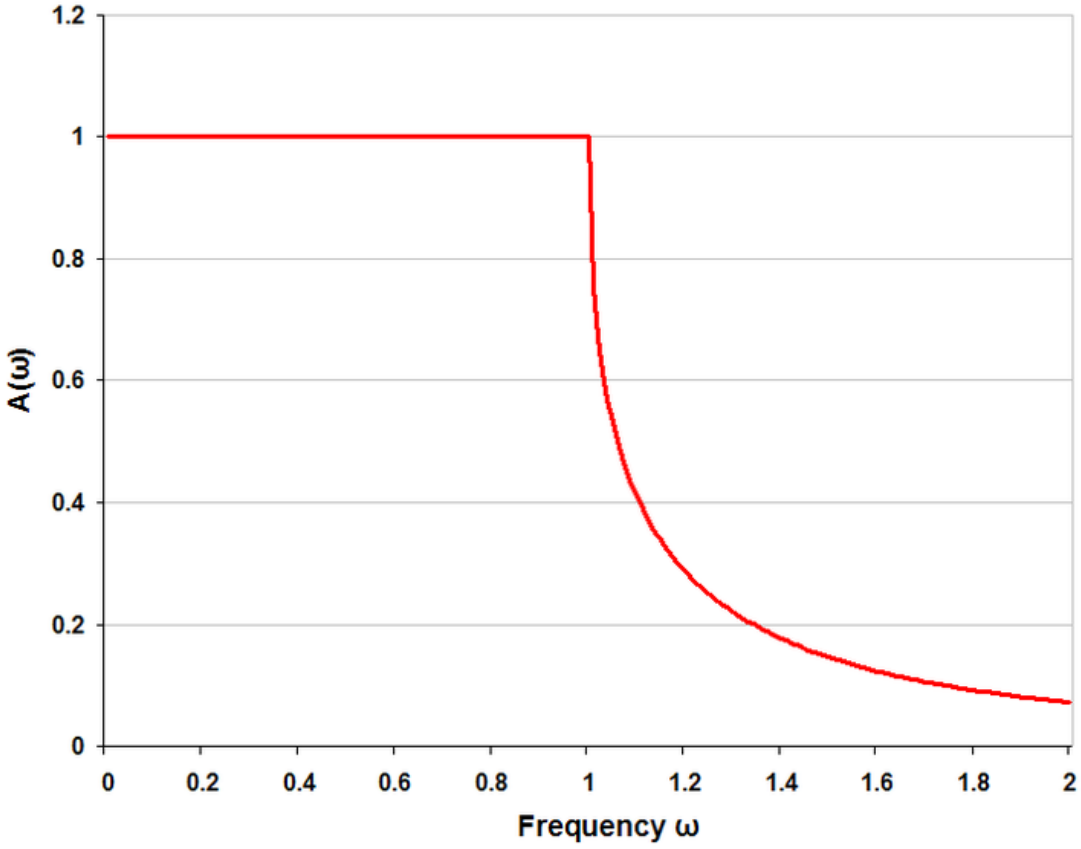

Zobel network low-pass response single section



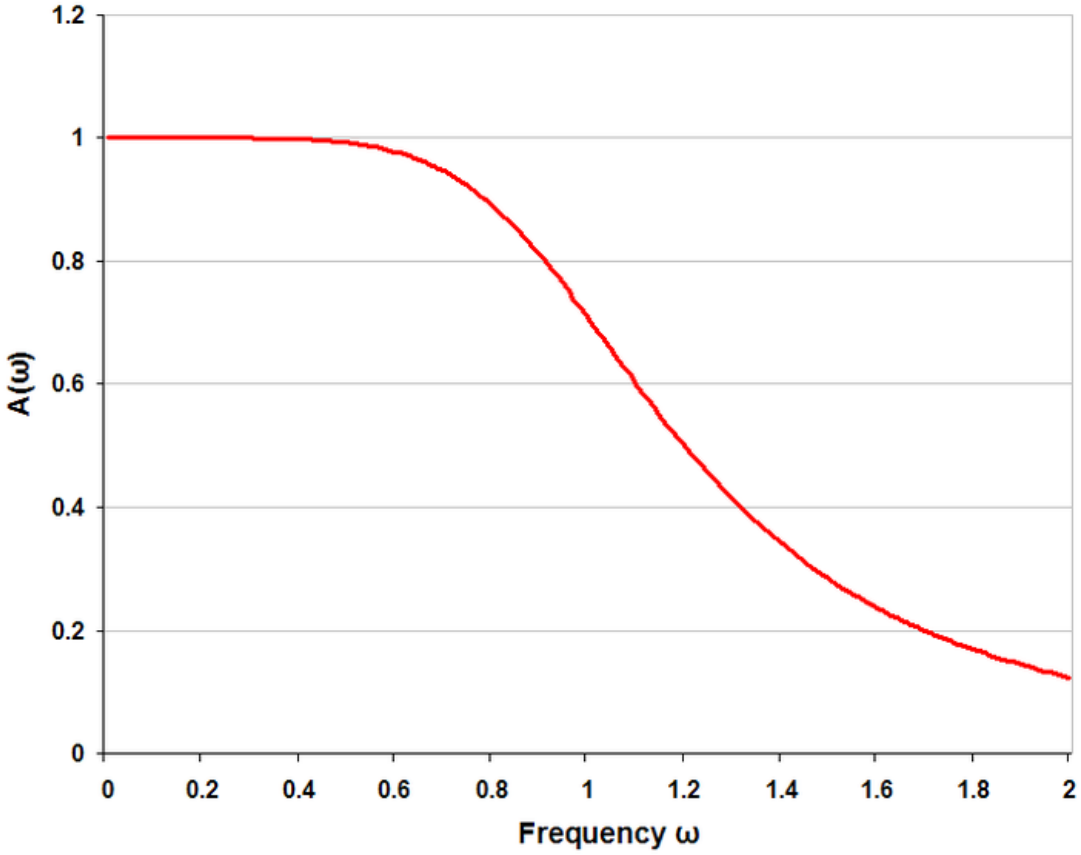

Zobel network low-pass response five sections

## Effect of end terminations

A consequence of the image method of filter design is that the effect of the end terminations has to be calculated separately if its effects on response are to be taken into account. The most severe deviation of the response from that predicted occurs in the passband close to cut-off. The reason for this is twofold. Further into the passband the impedance match progressively improves, thus limiting the error. On the other hand, waves in the stopband are reflected from the end termination due to mismatch but are attenuated twice by the filter stopband rejection as they pass through it. So while stopband impedance mismatch may be severe, it has only limited effect on the filter response.




Theoretical k-type low-pass T-filter (two half-sections) response when correctly terminated in image impedance




Practical k-type low-pass T-filter (two half-sections) response when terminated with fixed resistors

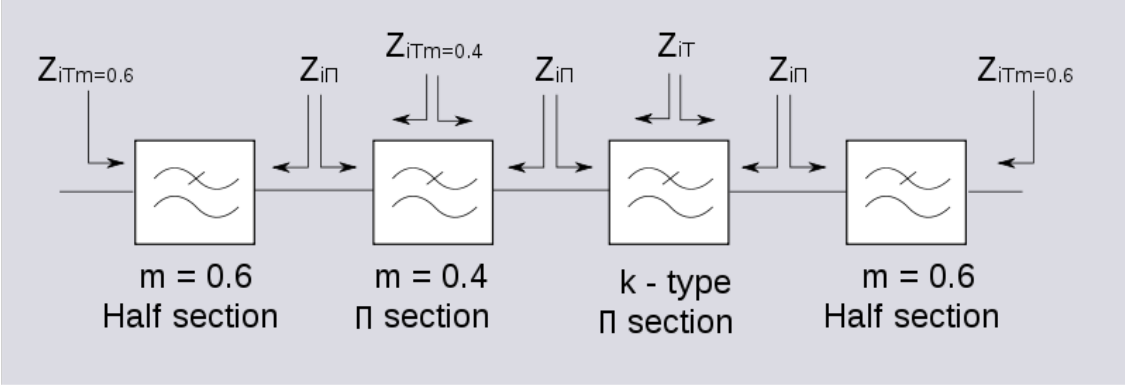
## *Cascading sections*

Several L half-sections may be cascaded to form a composite filter. The most important rule when constructing a composite image filter is that the image impedances must always face an identical impedance; like must always face like. T sections must always face T sections, Π sections must always face Π sections, k-type must always face k-type (or the side of an m-type which has the k-type impedance) and m-type must always face m-type. Furthermore, m-type impedances of different values of m cannot face each other. Nor can sections of any type which have different values of cut-off frequency.

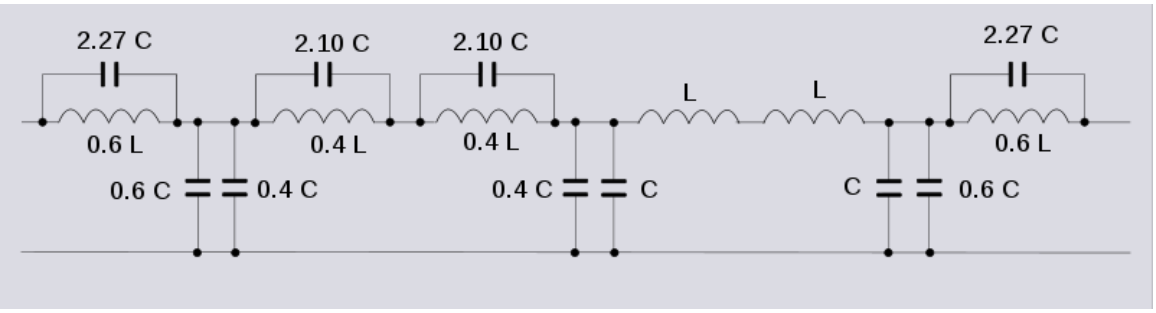
Sections at the beginning and end of the filter are often chosen for their impedance match in to the terminations rather than the shape of their frequency response. For this purpose, m-type sections of $m$ = 0.6 are the most common choice. An alternative is mm'-type sections of $m$=0.7230 and $m'$=0.4134 although this type of section is rarely used. While it has several advantages noted below, it has the disadvantages of being more complex and

also, if constant k sections are required in the body of the filter, it is then necessary to include m-type sections to interface the mm'-type to the k-types.

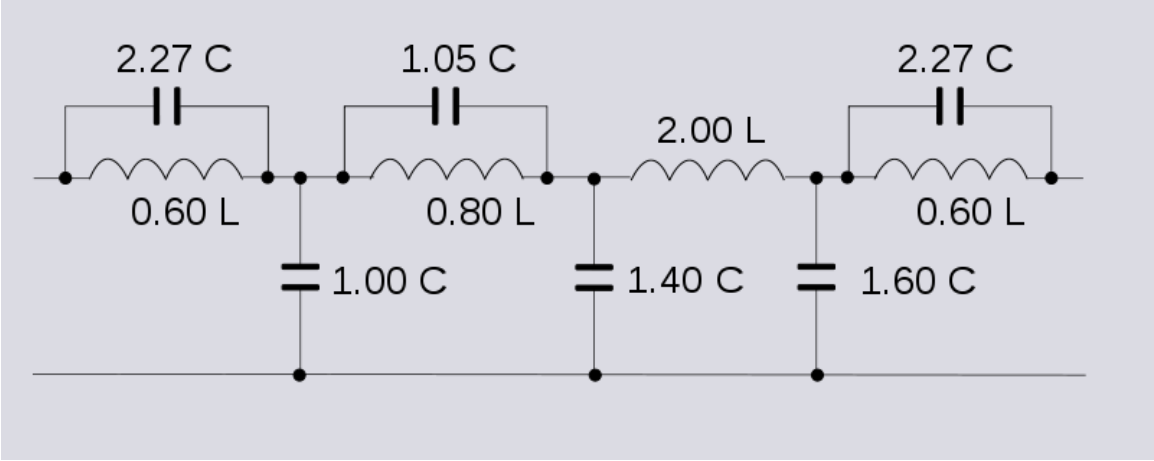
The inner sections of the filter are most commonly chosen to be constant k since these produce the greatest stopband attenuation. However, one or two m-type sections might also be included to improve the rate of fall from pass to stopband. A low value of m is chosen for m-types used for this purpose. The lower the value of *m*, the faster the transition, while at the same time, the stopband attenuation becomes less, increasing the need to use extra k-type sections as well. An advantage of using mm'-types for impedance matching is that these type of end sections will have a fast transition anyway (much more so than *m*=0.6 m-type) because *mm*'=0.3 for impedance matching. So the need for sections in the body of the filter to do this may be dispensed with.




Typical example of a composite image filter in block diagram form. The image impedances and how they match are shown.




The above filter realised as a ladder low-pass filter. Component values are given in terms of L and C, the component values of a constant k half-section.




The same filter minimised by combining components in series or parallel where appropriate.

Another reason for using m-types in the body of the filter is to place an additional pole of attenuation in the stopband. The frequency of the pole directly depends on the value of *m*. The smaller the value of *m*, the closer the pole is to the cut-off frequency. Conversely, a large value of *m* places the pole further away from cut-off until in the limit when *m*=1 the pole is at infinity and the response is the same as the k-type section. If a value of *m* is chosen for this pole which is different from the pole of the end sections it will have the effect of broadening the band of good stopband rejection near to the cut-off frequency. In this way the m-type sections serve to give good stopband rejection near to cut-off and the k-type sections give good stopband rejection far from cut-off. Alternatively, m-type sections can be used in the body of the filter with different values of m if the value found in the end sections is unsuitable. Here again, the mm'-type would have some advantages if used for impedance matching. The mm'-type used for impedance matching places the pole at *m*=0.3. However, the other half of the impedance matching section needs to be an m-type of *m*=0.723. This automatically gives a good spread of stopband rejection and as with the steepness of transition issue, use of mm'-type sections may remove the need for additional m-type sections in the body.

Constant resistance sections may also be required, if the filter is being used on a transmission line, to improve the flatness of the passband response. This is necessary because the transmission line response is not usually anywhere near perfectly flat. These sections would normally be placed closest to the line since they present a predictable impedance to the line and also tend to mask the indeterminate impedance of the line from the rest of the filter. There is no issue with matching constant resistance sections to each other even when the sections are operating on totally different frequency bands. All sections can be made to have precisely the same image impedance of a fixed resistance.

Chapter 5

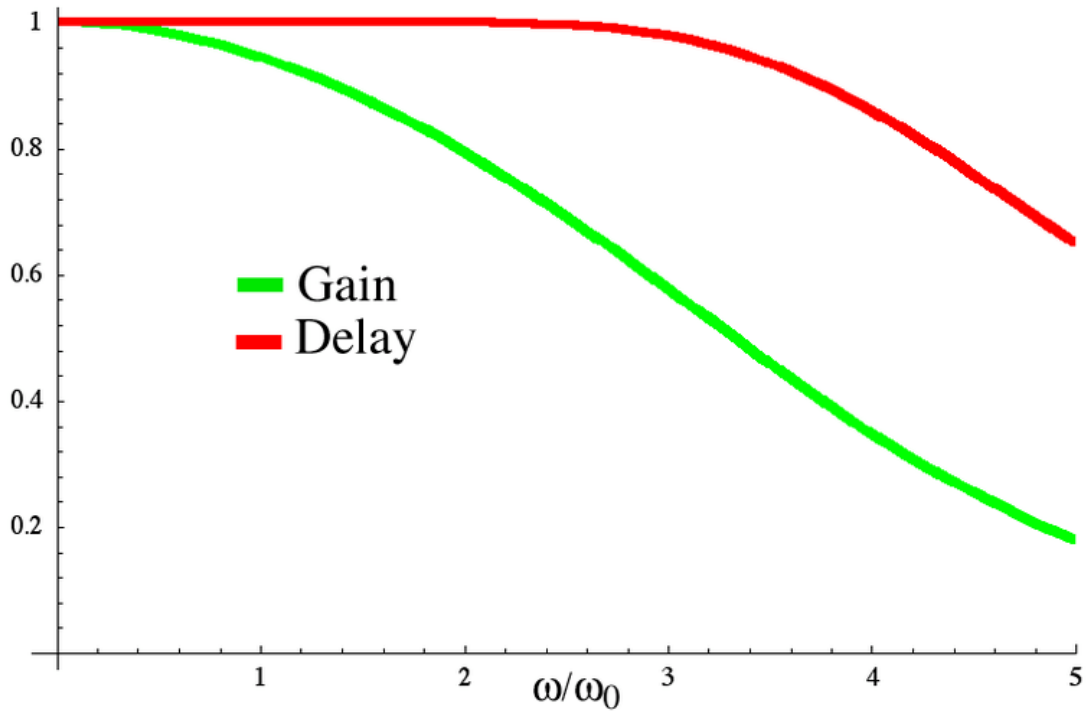
# Bessel Filter and Bartlett's Bisection Theorem

## Bessel Filter

In electronics and signal processing, a **Bessel filter** is a type of linear filter with a maximally flat group delay (maximally linear phase response). Bessel filters are often used in audio crossover systems. Analog Bessel filters are characterized by almost constant group delay across the entire passband, thus preserving the wave shape of filtered signals in the passband.

The filter's name is a reference to Friedrich Bessel, a German mathematician (1784–1846), who developed the mathematical theory on which the filter is based. The filters are also called Bessel–Thomson filters in recognition of W. E. Thomson, who worked out how to apply Bessel functions to filter design.

### *The transfer function*



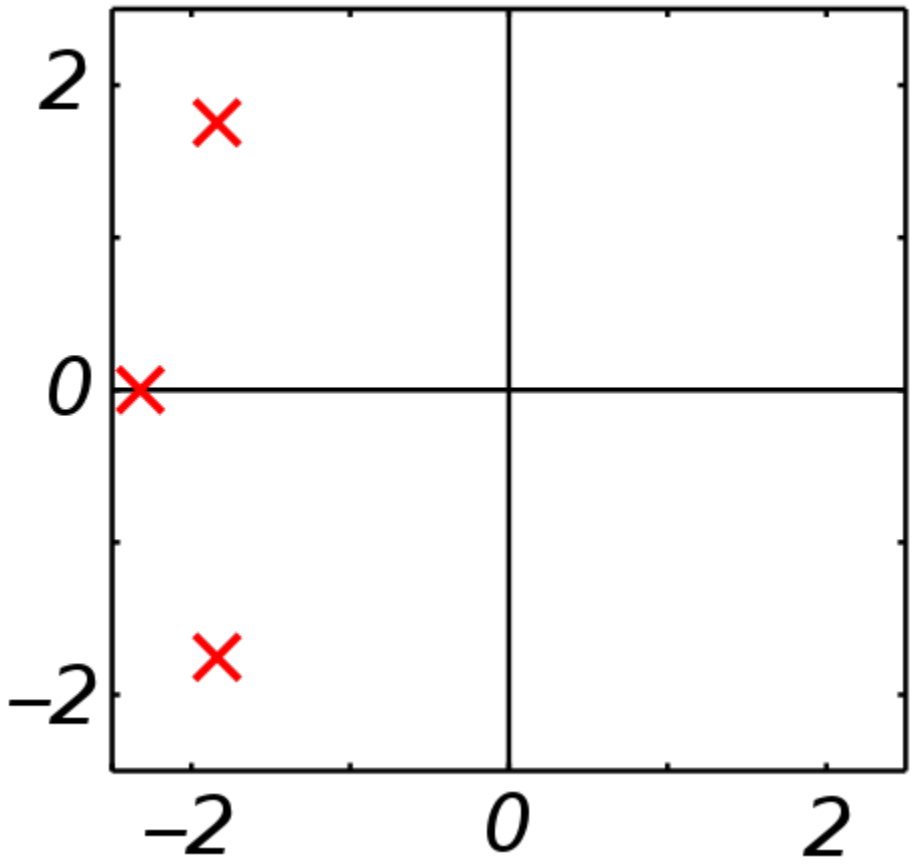

A plot of the gain and group delay for a fourth-order low pass Bessel filter. Note that the transition from the pass band to the stop band is much slower than for other filters, but the group delay is practically constant in the passband. The Bessel filter maximizes the flatness of the group delay curve at zero frequency.

A Bessel low-pass filter is characterized by its transfer function:

$$H(s) = \frac{\theta_n(0)}{\theta_n(s/\omega_0)}$$

where $\theta_n(s)$ is a reverse Bessel polynomial from which the filter gets its name and $\omega_0$ is a frequency chosen to give the desired cut-off frequency. The filter has a low-frequency group delay of $1 / \omega_0$.

### *Bessel polynomials*




The roots of the third-order Bessel polynomial are the poles of filter transfer function in the *s* plane, here plotted as crosses.

The transfer function of the Bessel filter is a rational function whose denominator is a reverse Bessel polynomial, such as the following:

$$\begin{aligned} n &= 1; \quad s + 1 \\ n &= 2; \quad s^2 + 3s + 3 \\ n &= 3; \quad s^3 + 6s^2 + 15s + 15 \end{aligned}$$

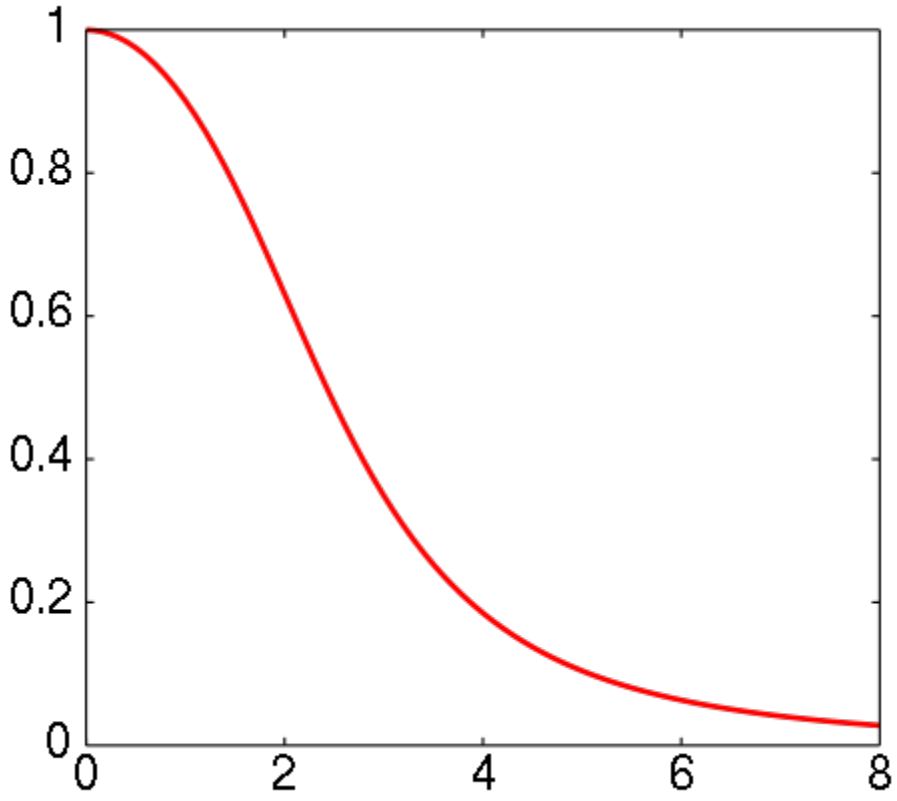
The reverse Bessel polynomials are given by:

$$\theta_n(s) = \sum_{k=0}^{n} a_k s^k,$$

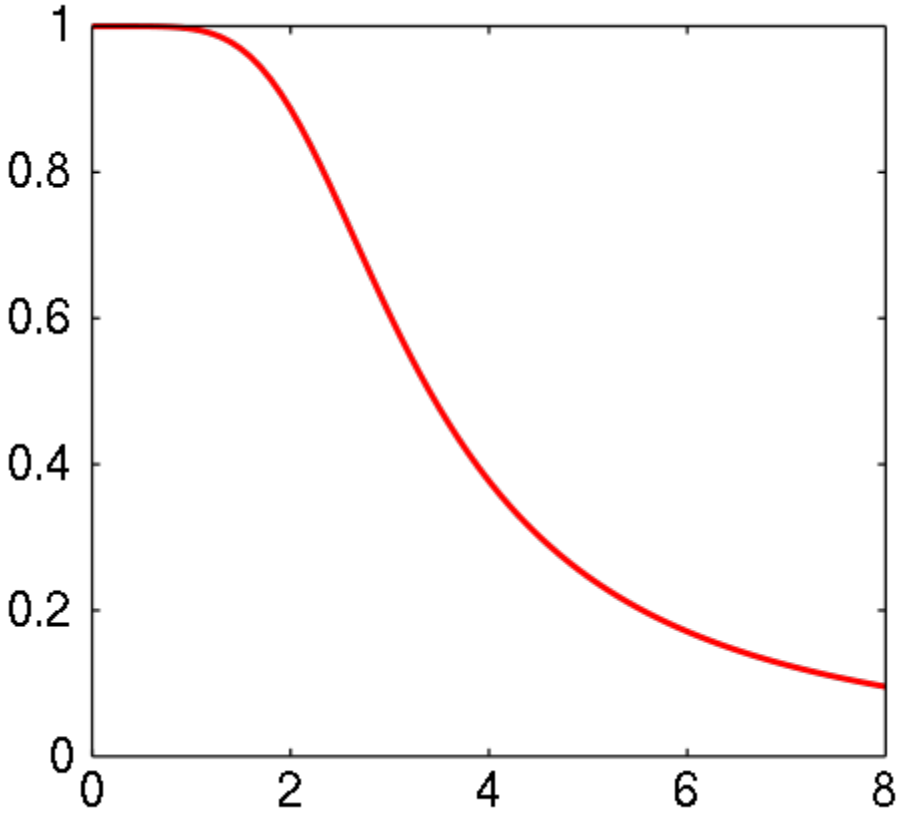
where

$$a_k = \frac{(2n-k)!}{2^{n-k}k!(n-k)!} \quad k = 0, 1, \ldots, n.$$

### *Example*




Gain plot of the third-order Bessel filter, versus normalized frequency




Group delay plot of the third-order Bessel filter, illustrating flat unit delay in the passband

The transfer function for a third-order (three-pole) Bessel low-pass filter, normalized to have unit group delay, is

$$H(s) = \frac{15}{s^3 + 6s^2 + 15s + 15}.$$

The roots of the denominator polynomial, the filter's poles, include a real pole at $s = -2.3222$, and a complex-conjugate pair of poles at $s = -1.8389 \pm j1.7544$, plotted above. The numerator 15 is chosen to give a gain of 1 at DC (at $s = 0$).

The gain is then

$$G(\omega) = |H(j\omega)| = \frac{15}{\sqrt{\omega^6 + 6\omega^4 + 45\omega^2 + 225}}.$$

The phase is

$$\phi(\omega) = -\arg(H(j\omega)) = -\arctan\left(\frac{15\omega - \omega^3}{15 - 6\omega^2}\right).$$

The group delay is

$$D(\omega) = -\frac{d\phi}{d\omega} = \frac{6\omega^4 + 45\omega^2 + 225}{\omega^6 + 6\omega^4 + 45\omega^2 + 225}.$$

The Taylor series expansion of the group delay is

$$D(\omega) = 1 - \frac{\omega^6}{225} + \frac{\omega^8}{1125} + \cdots.$$

Note that the two terms in $\omega^2$ and $\omega^4$ are zero, resulting in a very flat group delay at $\omega = 0$. This is the greatest number of terms that can be set to zero, since there are a total of four coefficients in the third order Bessel polynomial, requiring four equations in order to be defined. One equation specifies that the gain be unity at $\omega = 0$ and a second specifies that the gain be zero at $\omega = \infty$, leaving two equations to specify two terms in the series expansion to be zero. This is a general property of the group delay for a Bessel filter of order $n$: the first $n - 1$ terms in the series expansion of the group delay will be zero, thus maximizing the flatness of the group delay at $\omega = 0$.

# Bartlett's Bisection Theorem

**Bartlett's Bisection Theorem** is an electrical theorem in network analysis due to Albert Charles Bartlett. The theorem shows that any symmetrical two-port network can be transformed into a lattice network. The theorem often appears in filter theory where the lattice network is sometimes known as a filter X-section following the common filter theory practice of naming sections after alphabetic letters to which they bear a resemblance.

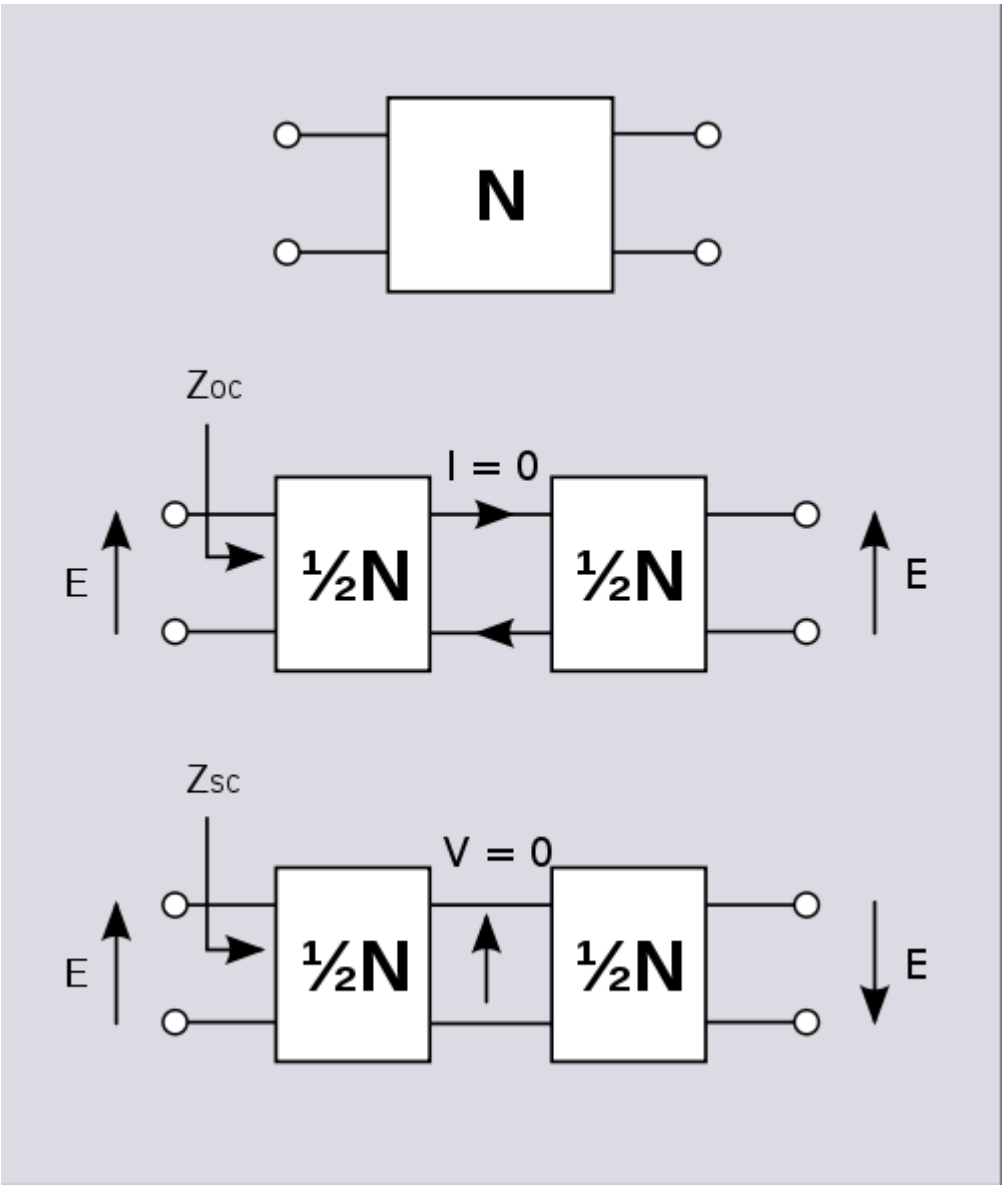
The theorem as originally stated by Bartlett required the two halves of the network to be topologically symmetrical. The theorem was later extended by Wilhelm Cauer to apply to all networks which were electrically symmetrical. That is, the physical implementation of the network is not of any relevance. It is only required that its response in both halves are symmetrical.

## *Applications*

Lattice topology filters are not very common. The reason for this is that they require more components (especially inductors) than other designs. Ladder topology is much more

popular. However, they do have the property of being intrinsically balanced and a balanced version of another topology, such as T-sections, may actually end up using more inductors. One application is for all-pass phase correction filters on balanced telecommunication lines. The theorem also makes an appearance in the design of crystal filters at RF frequencies. Here ladder topologies have some undesirable properties, but a common design strategy is to start from a ladder implemetation because of its simplicity. Bartlett's theorem is then used to transform the design to an intermmediate stage as a step towards the final implementation (using a transformer to produce an unbalanced version of the lattice topology).

## *Definition and proof*

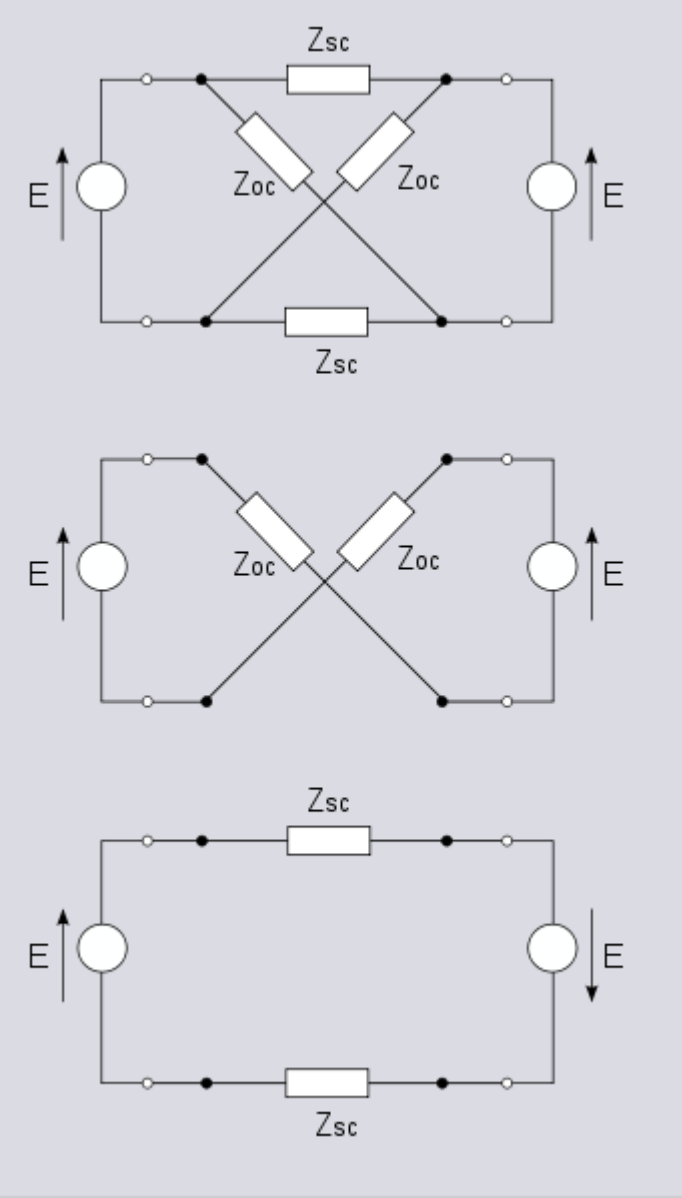


## Definition

Start with a two-port network, N, with a plane of symmetry between the two ports. Next cut N through its plane of symmetry to form two new identical two-ports, ½N. Connect two identical voltage generators to the two ports of N. It is clear from the symmetry that no current is going to flow through any branch passing through the plane of symmetry. The impedance measured into a port of N under these circumstances will be the same as the impedance measured if all the branches passing through the plane of symmetry were open circuit. It is therefore the same impedance as the open circuit impedance of ½N. Let us call that impedance $Z_{oc}$.

Now consider the network N with two identical voltage generators connected to the ports but with opposite polarity. Just as superposition of currents through the branches at the plane of symmetry must be zero in the previous case, by analogy and applying the principle of duality, superposition of voltages between nodes at the plane of symmetry must likewise be zero in this case. The input impedance is thus the same as the short circuit impedance of ½N. Let us call that impedance $Z_{sc}$.

Bartlett's bisection theorem states that the network N is equivalent to a lattice network with series branches of $Z_{sc}$ and cross branches of $Z_{oc}$.




## Proof

Consider the lattice network shown with identical generators, E, connected to each port. It is clear from symmetry and superposition that no current is flowing in the series branches $Z_{sc}$. Those branches can thus be removed and left open circuit without any effect on the rest of the circuit. This leaves a circuit loop with a voltage of 2E and an impedance of $2Z_{oc}$ giving a current in the loop of;

$$I = \frac{2E}{2Z_{oc}}$$

and an input impedance of;

$$\frac{E}{I} = Z_{oc}$$

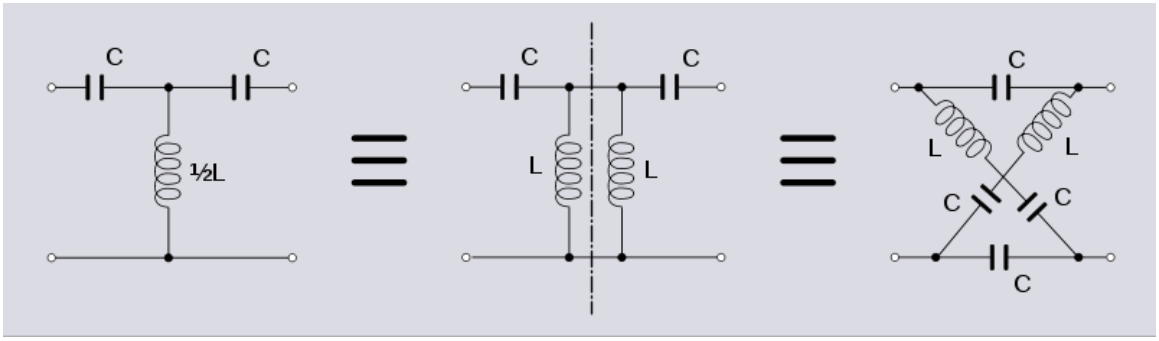
as it is required to be for equivalence to the original two-port.

Similarly, reversing one of the generators results, by an identical argument, in a loop with an impedance of $2Z_{sc}$ and an input impedance of;

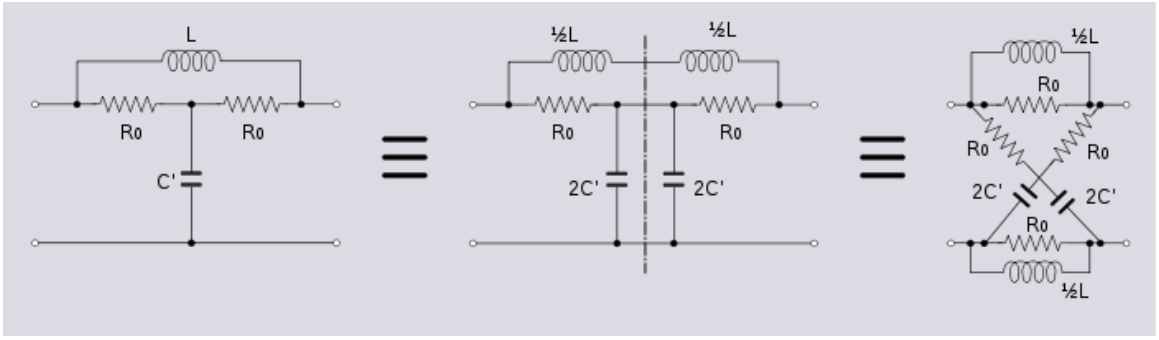
$$\frac{E}{I} = Z_{sc}$$

Recalling that these generator configurations are the precise way in which $Z_{oc}$ and $Z_{sc}$ were defined in the original two-port it is proved that the lattice is equivalent for those two cases. It is proved that this is so for all cases by considering that all other input and output conditions can be expressed as a linear superposition of the two cases already proved.

## *Examples*




Lattice equivalent of a T-section high-pass filter

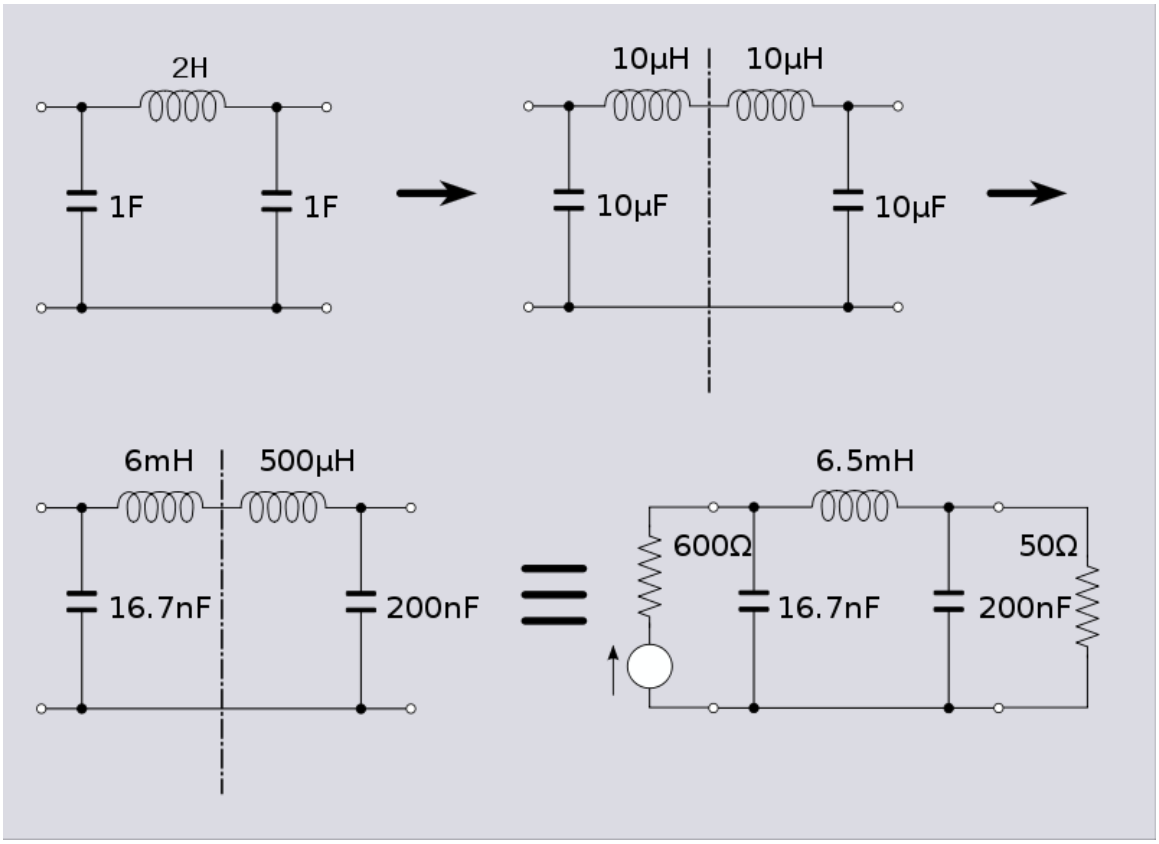



Lattice equivalent of a Zobel bridge-T low-pass filter

It is possible to use the Bartlett transformation in reverse; that is, to transform a symmetrical lattice network into some other symmetrical topology. The examples shown

above could just as equally have been shown in reverse. However, unlike the examples above, the result is not always physically realisable with linear passive components. This is because there is a possibility the reverse transform will generate components with negative values. Negative quantities can only be physically realised with active components present in the network.

## *Extension of the theorem*




Example of impedance and frequency scaling using a Π-section low-pass filter prototype. In the first transformation, the prototype is bisected and the cut-off frequency is rescaled from 1 rad/s to $10^5$ rad/s (15.9 kHz). In the second transformation, the bisected network is rescaled on the left side to operate at 600 Ω and on the right side to operate at 50 Ω.

There is an extension to Bartlett's theorem that allows a symmetrical filter network operating between equal input and output impedance terminations to be modified for unequal source and load impedances. This is an example of impedance scaling of a prototype filter. The symmetrical network is bisected along its plane of symmetry. One half is impedance-scaled to the input impedance and the other is scaled to the output impedance. The response shape of the filter remains the same. This does not amount to an impedance matching network, the impedances looking in to the network ports bear no relationship to the termination impedances. This means that a network designed by Bartlett's theorem, while having exactly the filter response predicted, also adds a constant

attenuation in addition to the filter response. In impedance matching networks, a usual design criteria is to maximise power transfer. The output response is "the same shape" relative to the voltage of the theoretical ideal generator driving the input. It is not the same relative to the actual input voltage which is delivered by the theoretical ideal generator via its load impedance.

The constant gain due to the difference in input and output impedances is given by;

$$A = \frac{V_2}{E} = \frac{2R_2}{R_1 + R_2}$$

Note that it is possible for this to be greater than unity, that is, a voltage gain is possible, but power is always lost.

Chapter 6

# Elliptic Filter

An **elliptic filter** (also known as a **Cauer filter**, named after Wilhelm Cauer) is a signal processing filter with equalized ripple (equiripple) behavior in both the passband and the stopband. The amount of ripple in each band is independently adjustable, and no other filter of equal order can have a faster transition in gain between the passband and the stopband, for the given values of ripple (whether the ripple is equalized or not). Alternatively, one may give up the ability to independently adjust the passband and stopband ripple, and instead design a filter which is maximally insensitive to component variations.

As the ripple in the stopband approaches zero, the filter becomes a type I Chebyshev filter. As the ripple in the passband approaches zero, the filter becomes a type II Chebyshev filter and finally, as both ripple values approach zero, the filter becomes a Butterworth filter.

The gain of a lowpass elliptic filter as a function of angular frequency ω is given by:

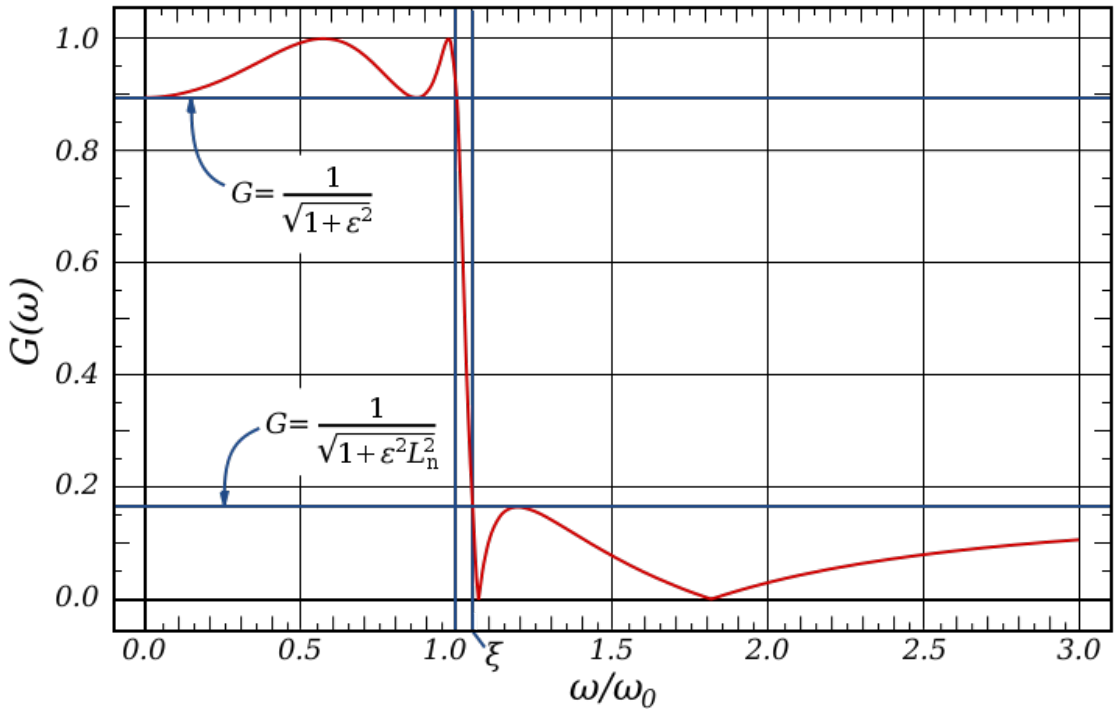
$$G_n(\omega) = \frac{1}{\sqrt{1 + \epsilon^2 R_n^2(\xi, \omega/\omega_0)}}$$

where $R_n$ is the *n*th-order elliptic rational function (sometimes known as a Chebyshev rational function) and

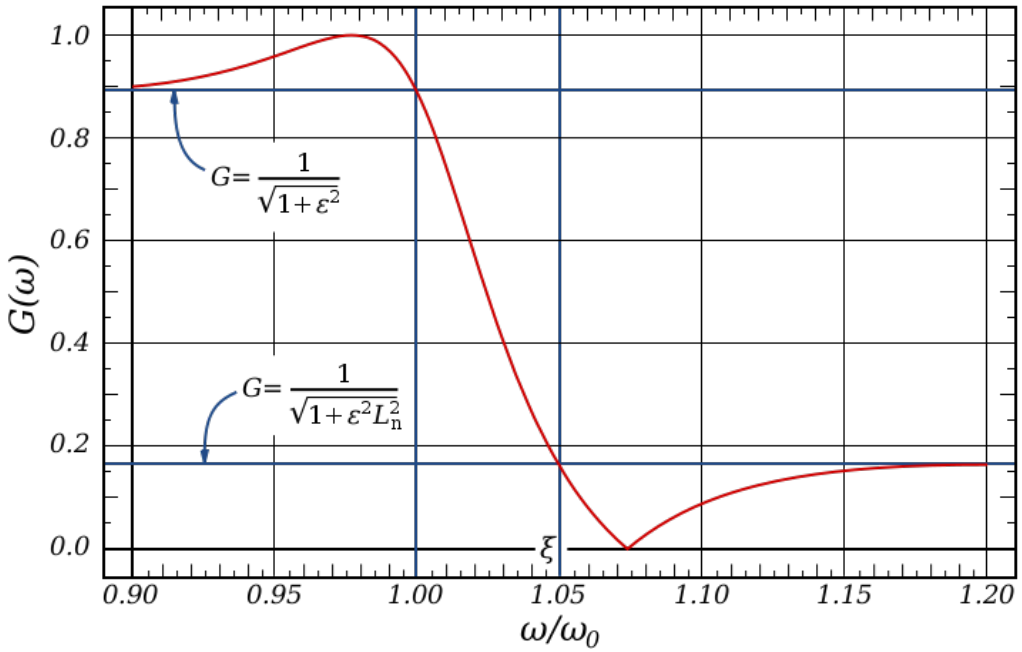
$\omega_0$ is the cutoff frequency
$\varepsilon$ is the ripple factor
$\xi$ is the selectivity factor

The value of the ripple factor specifies the passband ripple, while the combination of the ripple factor and the selectivity factor specify the stopband ripple.

## *Properties*




The frequency response of a fourth-order elliptic low-pass filter with ε=0.5 and ξ=1.05. Also shown are the minimum gain in the passband and the maximum gain in the stopband, and the transition region between normalized frequency 1 and ξ




A closeup of the transition region of the above plot.

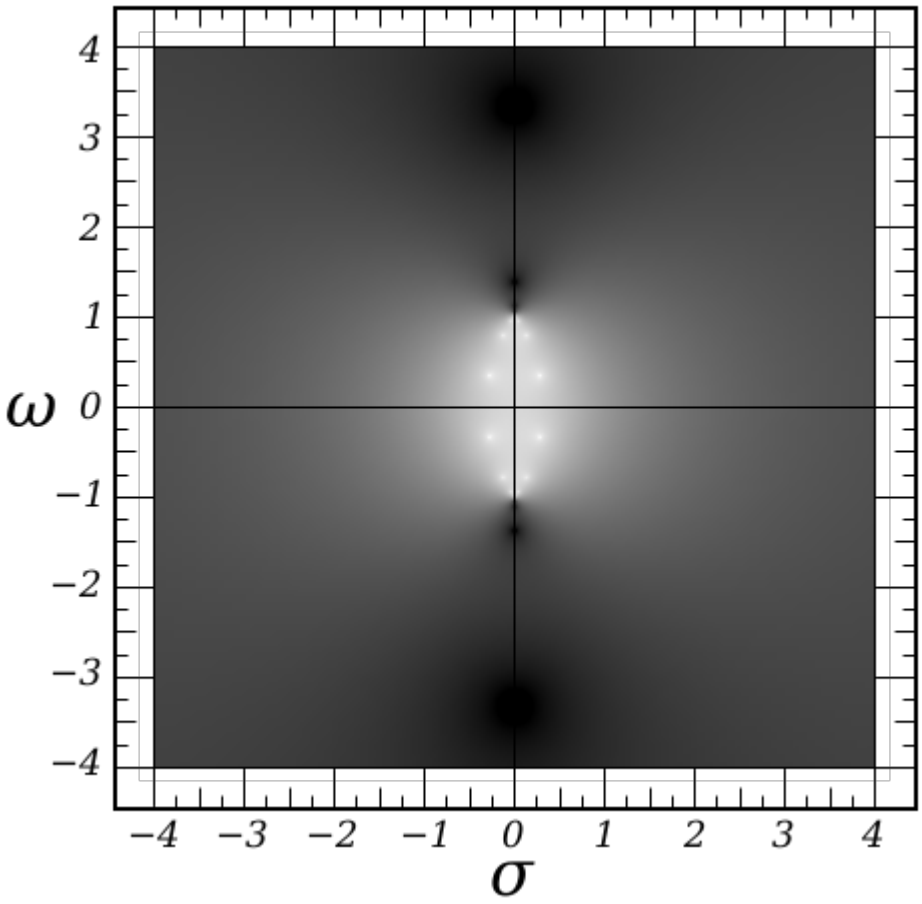
- In the passband, the elliptic rational function varies between zero and unity. The passband of the gain therefore will vary between 1 and $1/\sqrt{1+\epsilon^2}$.
- In the stopband, the elliptic rational function varies between infinity and the discrimination factor $L_n$ which is defined as:

  $$L_n = R_n(\xi, \xi)$$

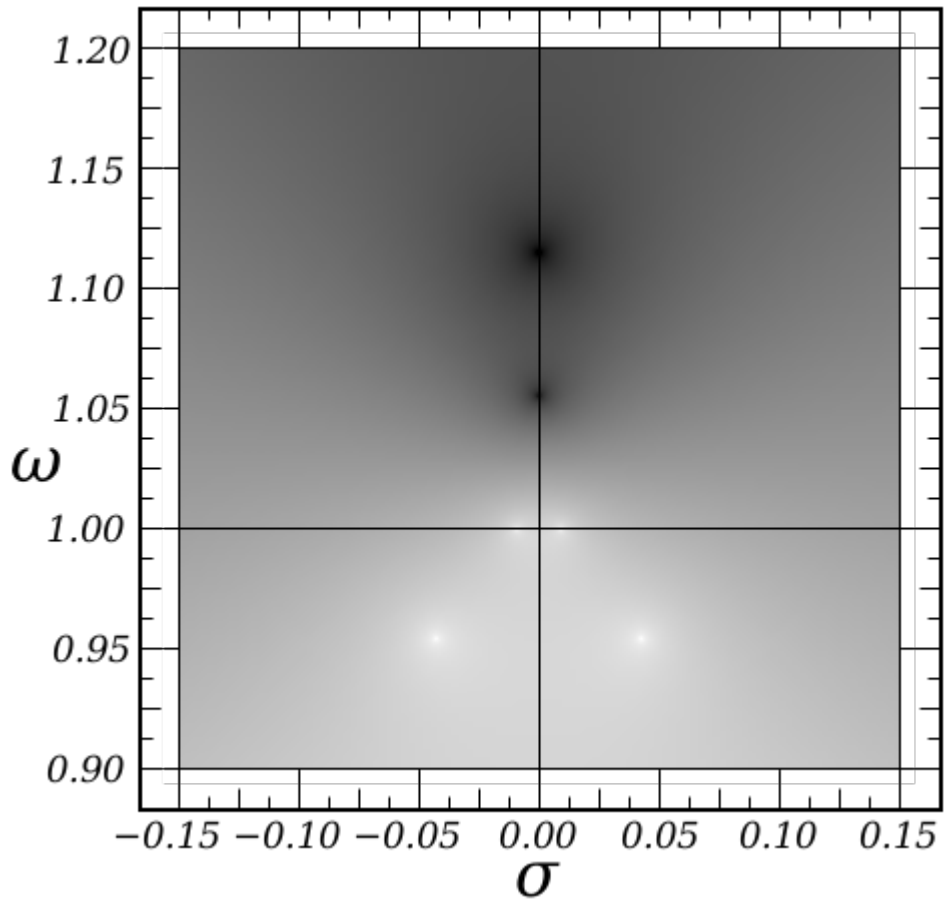
  The gain of the stopband therefore will vary between 0 and $1/\sqrt{1+\epsilon^2 L_n^2}$.
- In the limit of $\xi \to \infty$ the elliptic rational function becomes a Chebyshev polynomial, and therefore the filter becomes a Chebyshev type I filter, with ripple factor ε
- Since the Butterworth filter is a limiting form of the Chebyshev filter, it follows that in the limit of $\xi \to \infty$, $\omega_0 \to 0$ and $\epsilon \to 0$ such that $\epsilon\, R_n(\xi, 1/\omega_0) = 1$ the filter becomes a Butterworth filter
- In the limit of $\xi \to \infty$, $\epsilon \to 0$ and $\omega_0 \to 0$ such that $\xi\omega_0 = 1$ and $\varepsilon L_n = \alpha$, the filter becomes a Chebyshev type II filter with gain

  $$G(\omega) = \frac{1}{\sqrt{1 + \frac{1}{\alpha^2 T_n^2(1/\omega)}}}$$

### *Poles and zeroes*




Log of the absolute value of the gain of an 8th order elliptic filter in complex frequency space ($s=\sigma+j\omega$) with $\varepsilon=0.5$, $\xi=1.05$ and $\omega_0 = 1$. The white spots are poles and the black spots are zeroes. There are a total of 16 poles and 8 double zeroes. What appears to be a single pole and zero near the transition region is actually four poles and two double zeroes as shown in the expanded view below. In this image, black corresponds to a gain of 0.0001 or less and white corresponds to a gain of 10 or more.




An expanded view in the transition region of the above image, resolving the four poles and two double zeroes.

The poles of the gain of an elliptic filter may be derived in a manner very similar to the derivation of the poles of the gain of a type I Chebyshev filter. For simplicity, assume that the cutoff frequency is equal to unity. The poles ($\omega_{pm}$) of the gain of the elliptical filter will be the zeroes of the denominator of the gain. Using the complex frequency $s = \sigma + j\omega$ this means that:

$$1 + \epsilon^2 R_n^2(-js, \xi) = 0$$

Defining $-js = \mathrm{cd}(w, 1/\xi)$ where cd() is the Jacobi elliptic cosine function and using the definition of the elliptic rational functions yields:

$$1 + \epsilon^2 \mathrm{cd}^2\left(\frac{nwK_n}{K}, \frac{1}{L_n}\right) = 0$$

where $K = K(1/\xi)$ and $K_n = K(1/L_n)$. Solving for $w$

$$w = \frac{K}{nK_n}\mathrm{cd}^{-1}\left(\frac{\pm j}{\epsilon}, \frac{1}{L_n}\right) + \frac{mK}{n}$$

where the multiple values of the inverse cd() function are made explicit using the integer index *m*.

The poles of the elliptic gain function are then:

$$s_{pm} = i\,\mathrm{cd}(w, 1/\xi)$$

As is the case for the Chebyshev polynomials, this may be expressed in explicitly complex form (Lutovac & et al. 2001, § 12.8)

$$\begin{aligned} s_{pm} &= \frac{a + jb}{c} \\ a &= -\zeta_n\sqrt{1-\zeta_n^2}\sqrt{1-x_m^2}\sqrt{1-x_m^2/\xi^2} \\ b &= x_m\sqrt{1-\zeta_n^2(1-1/\xi^2)} \\ c &= 1 - \zeta_n^2 + x_i^2\zeta_n^2/\xi^2 \end{aligned}$$

where $\zeta_n$ is a function of $n$, $\epsilon$ and $\xi$ and $x_m$ are the zeroes of the elliptic rational function. $\zeta_n$ is expressible for all *n* in terms of Jacobi elliptic functions, or algebraically for some orders, especially orders 1,2, and 3. For orders 1 and 2 we have

$$\begin{aligned} \zeta_1 &= \frac{1}{\sqrt{1+\epsilon^2}} \\ \zeta_2 &= \frac{2}{(1+t)\sqrt{1+\epsilon^2} + \sqrt{(1-t)^2 + \epsilon^2(1+t)^2}} \end{aligned}$$

where

$$t = \sqrt{1 - 1/\xi^2}$$

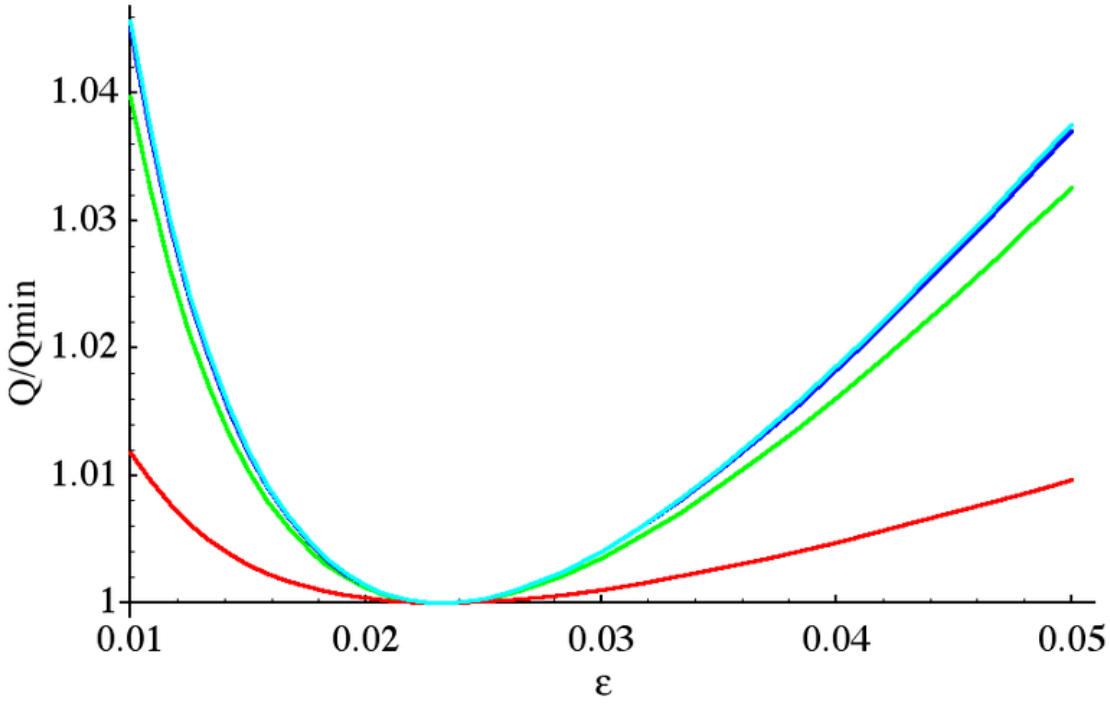
The algebraic expression for $\zeta_3$ is rather involved.

The nesting property of the elliptic rational functions can be used to build up higher order expressions for $\zeta_n$:

$$\zeta_{m\cdot n}(\xi, \epsilon) = \zeta_m\left(\xi, \sqrt{\frac{1}{\zeta_n^2(L_m, \epsilon)} - 1}\right)$$

where $L_m = R_m(\xi,\xi)$.

### *Minimum Q-factor elliptic filters*




The normalized Q-factors of the poles of an 8-th order elliptic filter with $\xi$=1.1 as a function of ripple factor ε. Each curve represents four poles, since complex conjugate pole pairs and positive-negative pole pairs have the same Q-factor. (The blue and cyan curves nearly coincide). The Q-factor of all poles are simultaneously minimized at $\varepsilon_{Qmin}=1/\sqrt{L_n}=0.02323...$

Elliptic filters are generally specified by requiring a particular value for the passband ripple, stopband ripple and the sharpness of the cutoff. This will generally specify a minimum value of the filter order which must be used. Another design consideration is the sensitivity of the gain function to the values of the electronic components used to build the filter. This sensitivity is inversely proportional to the quality factor (Q-factor) of the poles of the transfer function of the filter. The Q-factor of a pole is defined as:

$$Q = -\frac{|s_{pm}|}{2\mathrm{Re}(s_{pm})} = -\frac{1}{2\cos(\arg(s_{pm}))}$$

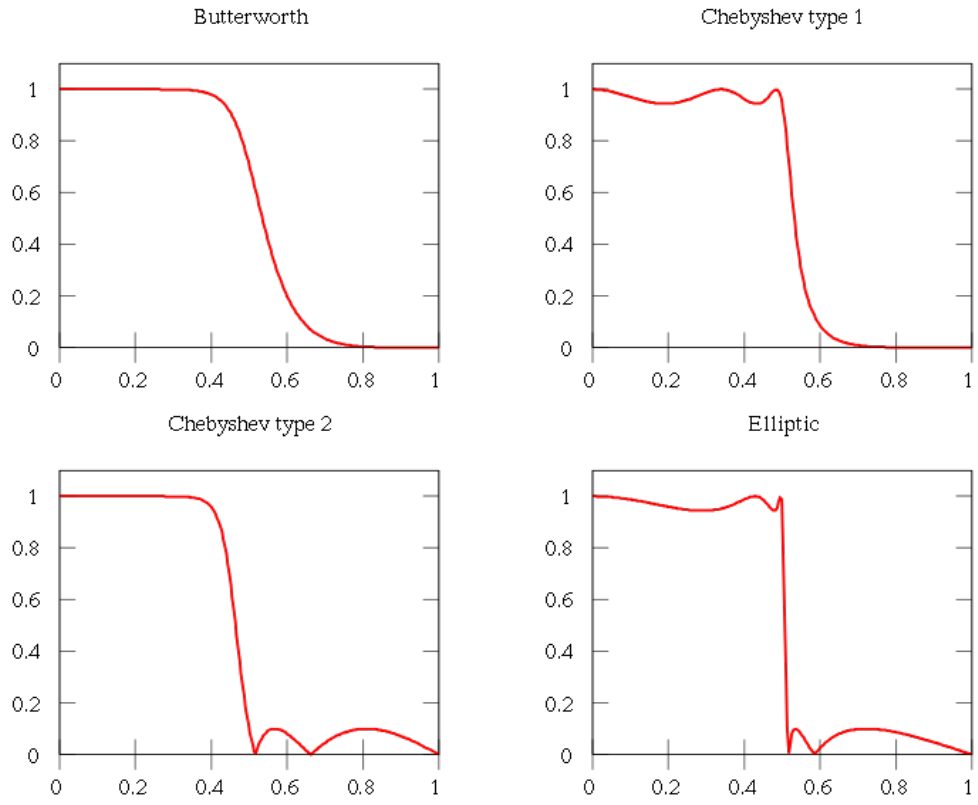
and is a measure of the influence of the pole on the gain function. For an elliptic filter, it happens that, for a given order, there exists a relationship between the ripple factor and selectivity factor which simultaneously minimizes the Q-factor of all poles in the transfer function:

$$\epsilon_{Qmin} = \frac{1}{\sqrt{L_n(\xi)}}$$

This results in a filter which is maximally insensitive to component variations, but the ability to independently specify the passband and stopband ripples will be lost. For such filters, as the order increases, the ripple in both bands will decrease and the rate of cutoff will increase. If one decides to use a minimum-Q elliptic filter in order to achieve a particular minimum ripple in the filter bands along with a particular rate of cutoff, the order needed will generally be greater than the order one would otherwise need without the minimum-Q restriction. An image of the absolute value of the gain will look very much like the image in the previous section, except that the poles are arranged in a circle rather than an ellipse. They will not be evenly spaced and there will be zeroes on the ω axis, unlike the Butterworth filter, whose poles are also arranged in a circle.

## *Comparison with other linear filters*

Here is an image showing the elliptic filter next to other common kind of filters obtained with the same number of coefficients:




As is clear from the image, elliptic filters are sharper than all the others, but they show ripples on the whole bandwidth.

Chapter 7

# Fast Kalman Filter and Filter Bank

## Fast Kalman Filter

The **fast Kalman filter (FKF)**, devised by Antti Lange (1941- ), is an extension of the Helmert-Wolf blocking (**HWB**) method from geodesy to real-time applications of Kalman filtering (KF) such as satellite imaging of the Earth. Kalman filters are an important software technique for building fault-tolerance into a wide range of systems, including real-time imaging.

### *Description*

The Fast Kalman filter applies only to systems with sparse matrices (Lange, 2001), since HWB is an inversion method to solve sparse linear equations (Wolf, 1978).

The ordinary Kalman filter is optimal for general systems. However, an **optimal** Kalman filter is probably stable only if Kalman's **observability** and **controllability** conditions are also satisfied (Kalman, 1960). These conditions are challenging to continuously maintain for a large system which means that even an optimal Kalman filter may diverge towards false solutions. Fortunately, the stability of an optimal Kalman filter can be controlled by monitoring its error variances if these can be reliably estimated. Their precise computation is, however, much more demanding than the optimal filtering itself but the FKF method may provide the required speed-up also in this respect.

### *Optimum calibration*

Calibration parameters are a typical example of those state parameters that may create serious observability problems if a narrow window of data (i.e. too few measurements) is continuously used by a Kalman filter (Lange, 1999). Observing instruments onboard orbiting satellites gives an example of **optimal** Kalman filtering where their calibration is done indirectly on ground (Olsson el al, 2001). There may also exist other state

parameters that are hardly or not at all observable (estimable) if too small samples of data are processed (analysed) at a time by any sort of a Kalman filter.

## Inverse problem

The computing load of the inverse problem of an ordinary **Kalman recursion** is roughly proportional to the cube of the number of the measurements processed simultaneously, which can always be set to 1 by processing each scalar measurement independently and (if necessary) performing a simple pre-filtering algorithm to de-correlate these measurements.

Even when many measurements are processed simultaneously, it is not unusual that the linear equation system is sparse, because some measurements turn out to be independent of some state or calibration parameters. In Satellite Geodesy problems (Brockmann, 1997), the computing load of the **HWB** (and **FKF**) method is only roughly proportional to the **square** of the number of the state parameters (and not of the measurements whose number may be billions).

## Reliable solution

Reliable operational Kalman filtering requires continuous fusion of data in real-time. Its optimality depends essentially on use of the error variances and covariances between all measurements and the estimated state and calibration parameters. This large error covariance matrix is obtained by matrix inversion from the respective system of **Normal Equations**. Its coefficient matrix is usually sparse and the exact solution of all estimated parameters can be computed by using the **HWB** method. The optimal solution may also be obtained by Gauss elimination using other sparse-matrix techniques or iterative methods based e.g. on Variational Calculus. However, these latter methods can solve the large matrix of all error variances and covariances only approximately and it would thus be impossible to do the data fusion in a strictly **optimal** fashion. Consequently, the filter's stability may become uncertain even if the observability and controllability conditions were satisfied.

The sparse coefficient matrix to be inverted may often have either a **bordered block- or band-diagonal** (BBD) structure. If it is band-diagonal it can be transformed into a block-diagonal form e.g. by means of a generalised Canonical Correlation Analysis (**gCCA**). The large matrix can thus be most effectively inverted in a blockwise manner by using the following **analytic inversion formula**:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{bmatrix}$$

of Frobenius where

$A$ = a large **block- or band-diagonal** (BD) matrix to be easily inverted, and,

$(D - CA^{-1}B)$ = a much smaller matrix called the Schur complement of $A$.

This is the **FKF** method that may make it computationally possible to estimate a much larger number of state and calibration parameters than an ordinary Kalman recursion can do. Their operational accuracies may also be reliably estimated from the theory of Minimum-Norm Quadratic Unbiased Estimation (MINQUE) of C. R. Rao (1920- ) and used for controlling the stability of optimal Kalman filtering.

### *Applications*

The **FKF** method extends the very high accuracies of Satellite Geodesy to Virtual Reference Station (**VRS**) Real Time Kinematic (**RTK**) surveying, mobile positioning and ultra-reliable navigation (Lange, 2003). First important applications will be real-time **optimum calibration** of global observing systems in Meteorology, Geophysics, Astronomy etc.

For example, a Numerical Weather Prediction (NWP) system can now forecast observations with confidence intervals and their operational quality control can thus be improved. A sudden increase of uncertainty in predicting observations would indicate that important observations were missing (observability problem) or an unpredictable change of weather is taking place (controllability problem). Remote sensing and imaging from satellites may partly be based on forecast information. Controlling stability of such **feedback** between the forecast and satellite data calls for the theory of optimal Kalman filtering. No suboptimal solution would do a proper job as public safety is usually at stake.

The computational advantage of **FKF** is marginal for applications using only small amounts of data in real-time data. Therefore improved built-in calibration and data communication infrastructures need to be developed first and introduced to public use before personal gadgets and machine-to-machine (M2M) devices can make the best out of **FKF**.

# Filter Bank

In signal processing, a **filter bank** is an array of band-pass filters that separates the input signal into multiple components, each one carrying a single frequency subband of the original signal. One application of a filter bank is a graphic equalizer, which can attenuate the components differently and recombine them into a modified version of the original signal. The process of decomposition performed by the filter bank is called *analysis* (meaning analysis of the signal in terms of its components in each sub-band); the output of analysis is referred to as a subband signal with as many subbands as there are filters in the filter bank. The reconstruction process is called *synthesis*, meaning reconstitution of a complete signal resulting from the filtering process.

In digital signal processing, the term **filter bank** is also commonly applied to a bank of receivers. The difference is that receivers also down-convert the subbands to a low center

frequency that can be re-sampled at a reduced rate. The same result can sometimes be achieved by undersampling the bandpass subbands.

Another application of filter banks is signal compression, when some frequencies are more important than others. After decomposition, the important frequencies can be coded with a fine resolution. Small differences at these frequencies are significant and a coding scheme that preserves these differences must be used. On the other hand, less important frequencies do not have to be exact. A coarser coding scheme can be used, even though some of the finer (but less important) details will be lost in the coding.

The vocoder uses a filter bank to determine the amplitude information of the subbands of a modulator signal (such as a voice) and uses them to control the amplitude of the subbands of a carrier signal (such as the output of a guitar or synthesizer), thus imposing the dynamic characteristics of the modulator on the carrier.

## *FFT filter banks*

A filter bank can be created by performing a sequence of FFTs on overlapping blocks of the input data. A weighting function is applied to each block to control the shape of the frequency responses of the filters. Instead of a conventional FFT window function, the weighting function is the impulse response of an FIR lowpass filter. The wider the shape of the frequency response**:**

1. the more often the FFTs have to be done to satisfy the Nyquist sampling criteria (which is what distinguishes a filter bank from a spectrum analyzer), and
2. the fewer filters that are needed to span the input bandwidth.

Eliminating unnecessary filters (i.e. decimation in frequency) can be accomplished most efficiently in the time-domain by summing subblocks of the weighted data-block, resulting in a smaller FFT size.

A special case occurs when, by design, the length of the subblocks is an integer multiple of the interval between FFTs. Then the FFT filter bank can be described in terms of one or more polyphase filter structures where the phases are recombined by an FFT instead of a simple summation. The number of subblocks is the impulse response length (or *depth*) of each filter.

## *Filter banks as time-frequency distributions*

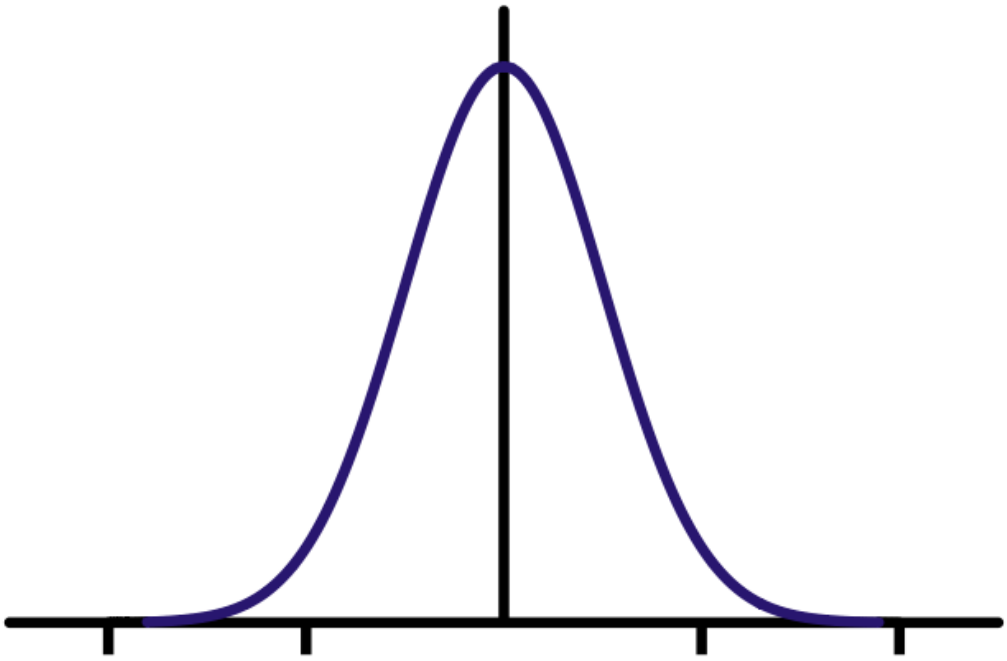
In time-frequency signal processing, a filter bank is a special quadratic time-frequency distribution (TFD) that represents the signal in a joint time-frequency domain. It is related to the Wigner-Ville distribution by a two-dimensional filtering that defines the class of quadratic (or bilinear) time-frequency distributions. The filter bank and the spectrogram are the two simplest ways of producing a quadratic TFD; they are in essence similar as one (the spectrogram) is obtained by dividing the time-domain in slices and then taking a

fourier transform, while the other (the filter bank) is obtained by dividing the frequency domain in slices forming bandpass filters that are excited by the signal under analysis.

Chapter 8

# Gaussian Filter and Digital Biquad Filter

## Gaussian Filter



Shape of a typical Gaussian filter

In electronics and signal processing, a **Gaussian filter** is a filter whose impulse response is a Gaussian function. Gaussian filters are designed to give no overshoot to a step

function input while minimizing the rise and fall time. This behavior is closely connected to the fact that the Gaussian filter has the minimum possible group delay. Mathematically, a Gaussian filter modifies the input signal by convolution with a Gaussian function; this transformation is also known as the Weierstrass transform.

## Definition

The one-dimensional Gaussian filter has an impulse response given by

$$g(x) = \sqrt{\frac{a}{\pi}} \cdot e^{-a \cdot x^2}$$

or with the standard deviation as parameter

$$g(x) = \frac{1}{\sqrt{2 \cdot \pi} \cdot \sigma} \cdot e^{-\frac{x^2}{2\sigma^2}}$$

In two dimensions, it is the product of two such Gaussians, one per direction:

$$g(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

where $x$ is the distance from the origin in the horizontal axis, $y$ is the distance from the origin in the vertical axis, and $\sigma$ is the standard deviation of the Gaussian distribution.

## Digital implementation

The Gaussian function is non-zero for $x \in [-\infty, \infty]$ and would theoretically require an infinite window length. However, since it decays rapidly, it is often reasonable to truncate the filter window and implement the filter directly for narrow windows, in effect by using a simple rectangular window function. In other cases, the truncation may introduce significant errors.

Filtering involves convolution. The filter function is said to be the kernel of an integral transform. The Gaussian kernel is continuous. Most commonly, the discrete equivalent is the sampled Gaussian kernel that is produced by sampling points from the continuous Gaussian. An alternate method is to use the discrete Gaussian kernel which has superior characteristics for some purposes. Unlike the sample Gaussian kernel, the discrete Gaussian kernel is the solution to the discrete diffusion equation.

Since the Fourier transform of the Gaussian function yields a Gaussian function, the signal (preferably after being divided into overlapping windowed blocks) can be transformed with a Fast Fourier transform, multiplied with a Gaussian function and transformed back. This is the standard procedure of applying an arbitrary finite impulse

response filter, with the only difference that the Fourier transform of the filter window is explicitly known.

Due to the central limit theorem, the Gaussian can be approximated by several runs of a very simple filter such as the moving average. The simple moving average corresponds to convolution with the constant B-spline, and, for example, four iterations of a moving average yields a cubic B-spline as filter window which approximates the Gaussian quite well.

Borrowing the terms from statistics, the standard deviation of a filter can be interpreted as a measure of its size. The cut-off frequency of the filter can be considered as the ratio between the sample rate $F_s$ and the standard deviation σ

$$f_c = \frac{F_s}{\sigma}$$

A simple moving average corresponds to a uniform probability distribution and thus its filter width of size *n* has standard deviation $\sqrt{(\sigma^2 - 1)/12}$. Thus *m* moving averages with sizes $\sigma_1, \ldots, \sigma_m$ yield a standard deviation of

$$\sigma = \sqrt{\frac{\sigma_1^2 + \cdots + \sigma_m^2 - m}{12}}.$$

(Note that standard deviations do not sum up, but variances do.)

When applied in two dimensions, this formula produces a Gaussian surface that has a maximum at the origin, whose contours are concentric circles with the origin as center. A two dimensional convolution matrix is precomputed from the formula and convolved with two dimensional data. Each element in the resultant matrix new value is set to a weighted average of that elements neighborhood. The focal element receives the heaviest weight (having the highest Gaussian value) and neighboring elements receive smaller weights as their distance to the focal element increases. In Image processing, each element in the matrix represents a pixel attribute such as brightness or a color intensity, and the overall effect is called Gaussian blur.

The Gaussian filter is non-causal which means the filter window is symmetric about the origin. This is usually of no consequence for most applications. In real-time systems, a delay is incurred because incoming samples need to fill the filter window before the filter can be applied to the signal.

# Digital Biquad Filter

In signal processing, a **digital biquad filter** is a second-order recursive linear filter, containing two poles and two zeros. "Biquad" is an abbreviation of "*biquadratic*", which refers to the fact that in the Z domain, its transfer function is the ratio of two quadratic functions:

$$H(z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}}$$

High-order recursive filters can be highly sensitive to quantization of their coefficients, and can easily become unstable. This is much less of a problem with first and second-order filters; therefore, higher-order filters are typically implemented as serially-cascaded biquad sections (and a first-order filter if necessary).

## *Implementation*

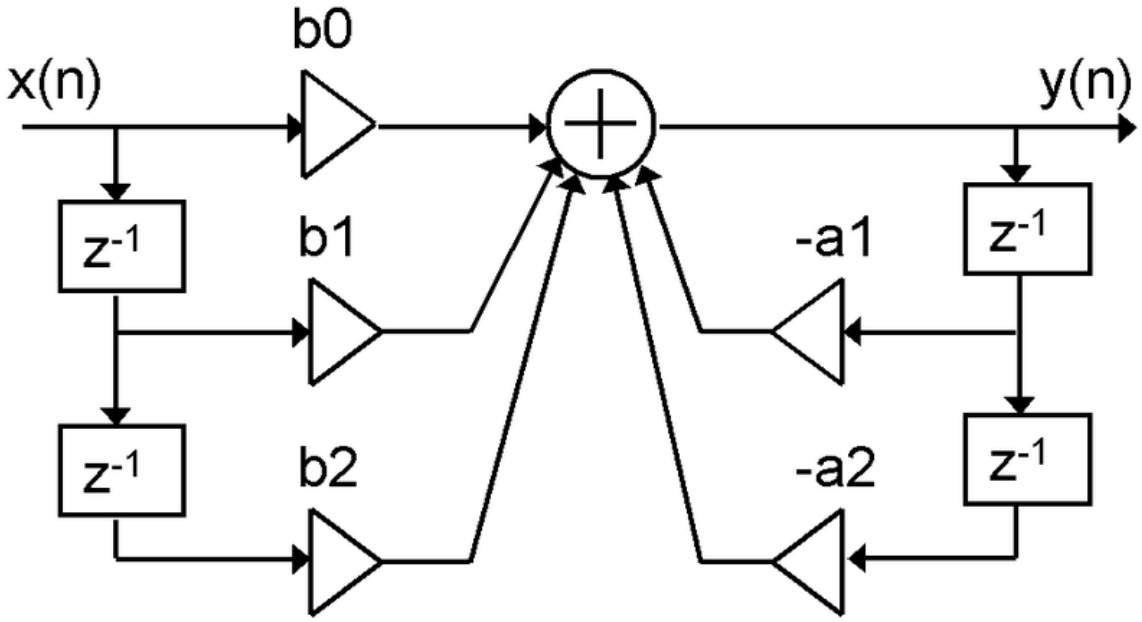
### Direct Form 1

The most straightforward implementation is the Direct Form 1, which has the following difference equation:

$$y(n) = b_0 x(n) + b_1 x(n-1) + b_2 x(n-2) - a_1 y(n-1) - a_2 y(n-2)$$

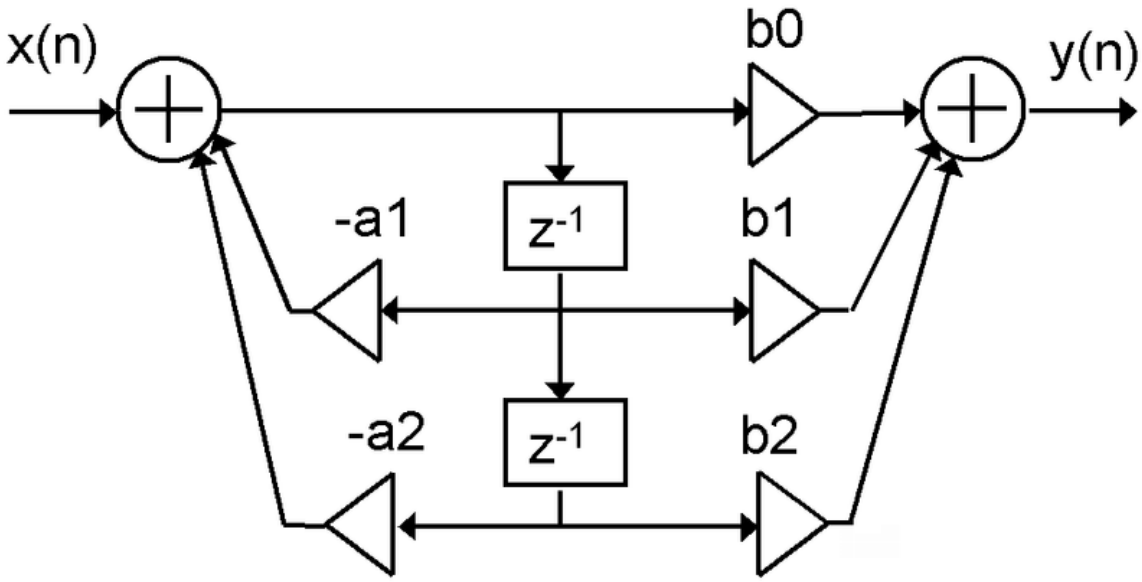
Here the $b_0$, $b_1$ and $b_2$ coefficients determine zeros, and $a_1$, $a_2$ determine the position of the poles.

Flow graph of biquad filter in Direct Form 1:




**Direct Form 2**

The Direct Form 1 implementation requires four delay registers. An equivalent circuit is the Direct Form 2 implementation, which requires only two delay registers:




The Direct Form 2 implementation is called the canonical form, because it uses the minimal amount of delays, adders and multipliers, yielding in the same transfer function as the Direct Form 1 implementation. The difference equations for DF2 are:

$$y(n) = b_0 w(n) + b_1 w(n-1) + b_2 w(n-2),$$

where

$$w(n) = x(n) - a_1 w(n-1) - a_2 w(n-2).$$

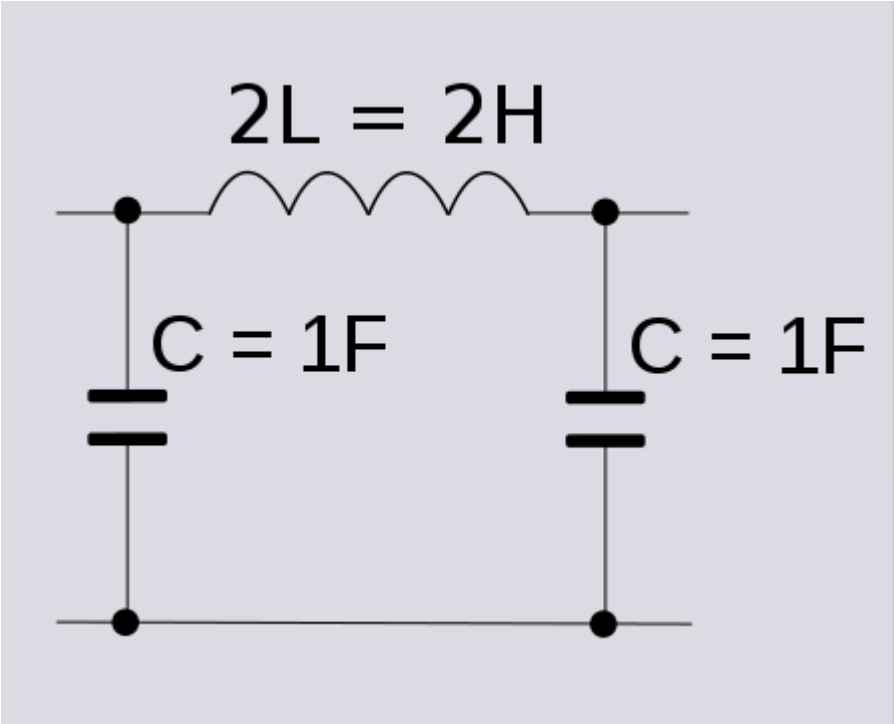
Chapter 9

# Prototype Filter

**Prototype filters** are electronic filter designs that are used as a template to produce a modified filter design for a particular application. They are an example of a nondimensionalised design from which the desired filter can be scaled or transformed. They are most often seen in regards to electronic filters and especially linear analogue passive filters. However, in principle, the method can be applied to any kind of linear filter or signal processing, including mechanical, acoustic and optical filters.

Filters are required to operate at many different frequencies, impedances and bandwidths. The utility of a prototype filter comes from the property that all these other filters can be derived from it by applying a scaling factor to the components of the prototype. The filter design need thus only be carried out once in full, with other filters being obtained by simply applying a scaling factor.

Especially useful is the ability to transform from one bandform to another. In this case, the transform is more than a simple scale factor. Bandform here is meant to indicate the category of passband that the filter possesses. The usual bandforms are lowpass, highpass, bandpass and bandstop, but others are possible. In particular, it is possible for a filter to have multiple passbands. In fact, in some treatments, the bandstop filter is considered to be a type of multiple passband filter having two passbands. Most commonly, the prototype filter is expressed as a lowpass filter, but other techniques are possible.




A low pass prototype constant k Π filter

## *Low-pass prototype*

The prototype is most often a low-pass filter with a 3dB corner frequency of angular frequency $\omega_c' = 1$ rad/s. Occasionally, frequency $f' = 1$ Hz is used instead of $\omega_c' = 1$. Likewise, the nominal or characteristic impedance of the filter is set to $R' = 1\ \Omega$.

In principle, any non-zero frequency point on the filter response could be used as a reference for the prototype design. For example, for filters with ripple in the passband the corner frequency is usually defined as the highest frequency at maximum ripple rather than 3dB. Another case is in image parameter filters (an older design method than the more modern network synthesis filters) which use the cut-off frequency rather than the 3dB point since cut-off is a well defined point in this types of filter.

The prototype filter can only be used to produce other filters of the same class and order. For instance, a fifth order Bessel filter prototype can be converted into any other fifth order Bessel filter, but it cannot be transformed into a third order Bessel filter or a fifth order Tchebyscheff filter.

## *Frequency scaling*

The prototype filter is scaled to the frequency required with the following transformation:

$$i\omega \rightarrow \left(\frac{\omega_c'}{\omega_c}\right) i\omega$$

where $\omega_c'$ is the value of the frequency parameter (e.g. cut-off frequency) for the prototype and $\omega_c$ is the desired value. So if $\omega_c' = 1$ then the transfer function of the filter is transformed as:

$$A(i\omega) \rightarrow A\left(i\frac{\omega}{\omega_c}\right)$$

It can readily be seen that to achieve this, the non-resistive components of the filter must be transformed by:

$$L \rightarrow \frac{\omega_c'}{\omega_c} L \quad \text{and,} \quad C \rightarrow \frac{\omega_c'}{\omega_c} C$$

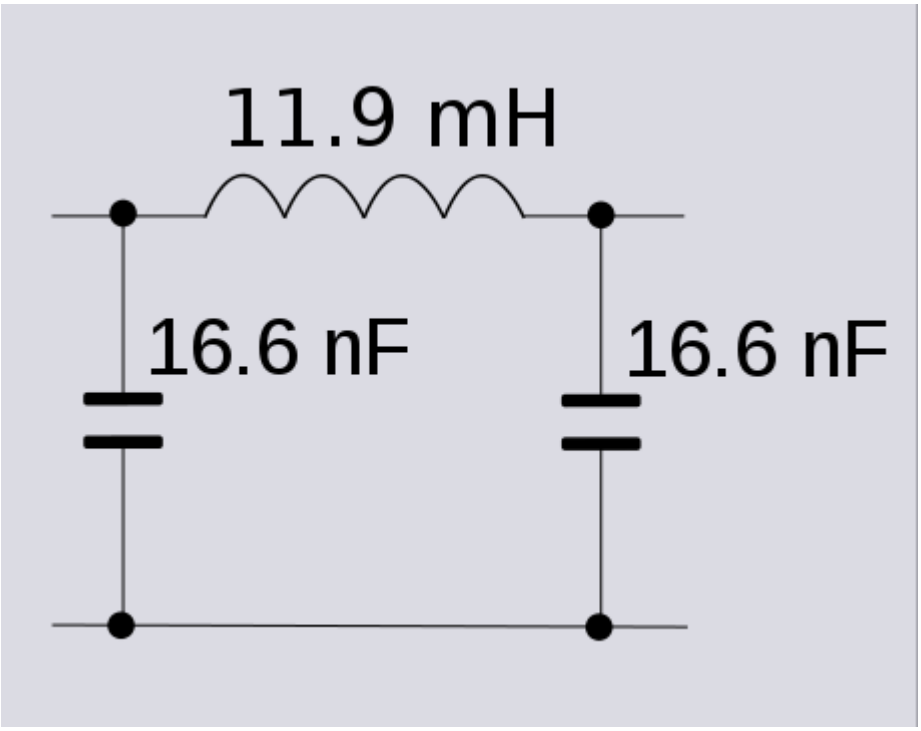
### *Impedance scaling*

Impedance scaling is invariably a scaling to a fixed resistance. This is because the terminations of the filter, at least nominally, are taken to be a fixed resistance. To carry out this scaling to a nominal impedance *R*, each impedance element of the filter is transformed by:

$$Z \rightarrow \frac{R}{R'} Z$$

It may be more convenient on some elements to scale the admittance instead:

$$Y \rightarrow \frac{R'}{R} Y$$




The prototype filter above, transformed to a 600Ω, 16kHz lowpass filter

It can readily be seen that to achieve this, the non-resistive components of the filter must be scaled as:

$$L \to \frac{R}{R'} L \quad \text{and,} \quad C \to \frac{R'}{R} C$$

Impedance scaling by itself has no effect on the transfer function of the filter (providing that the terminating impedances have the same scaling applied to them). However, it is usual to combine the frequency and impedance scaling into a single step:

$$L \to \frac{\omega_c'}{\omega_c} \frac{R}{R'} L \quad \text{and,} \quad C \to \frac{\omega_c'}{\omega_c} \frac{R'}{R} C$$

### *Bandform transformation*

In general, the bandform of a filter is transformed by replacing $i\omega$ where it occurs in the transfer function with a function of $i\omega$. This in turn leads to the transformation of the impedance components of the filter into some other component(s). The frequency scaling above is a trivial case of bandform transformation corresponding to a lowpass to lowpass transformation.

## Lowpass to highpass

The frequency transformation required in this case is:

$$\frac{i\omega}{\omega_c'} \rightarrow \frac{\omega_c}{i\omega}$$

where $\omega_c$ is the point on the highpass filter corresponding to $\omega_c'$ on the prototype. The transfer function then transforms as:

$$A(i\omega) \rightarrow A\left(\frac{\omega_c\,\omega_c'}{i\omega}\right)$$

Inductors are transformed into capacitors according to,

$$L' \rightarrow C = \frac{1}{\omega_c\,\omega_c'\,L'}$$

and capacitors are transformed into inductors,

$$C' \rightarrow L = \frac{1}{\omega_c\,\omega_c'\,C'}$$

the primed quantities being the component value in the prototype.

## Lowpass to bandpass

In this case, the required frequency transformation is:

$$\frac{i\omega}{\omega_c'} \rightarrow Q\left(\frac{i\omega}{\omega_0} + \frac{\omega_0}{i\omega}\right)$$

where $Q$ is the Q-factor and is equal to the inverse of the fractional bandwidth:

$$Q = \frac{\omega_0}{\Delta\omega}$$

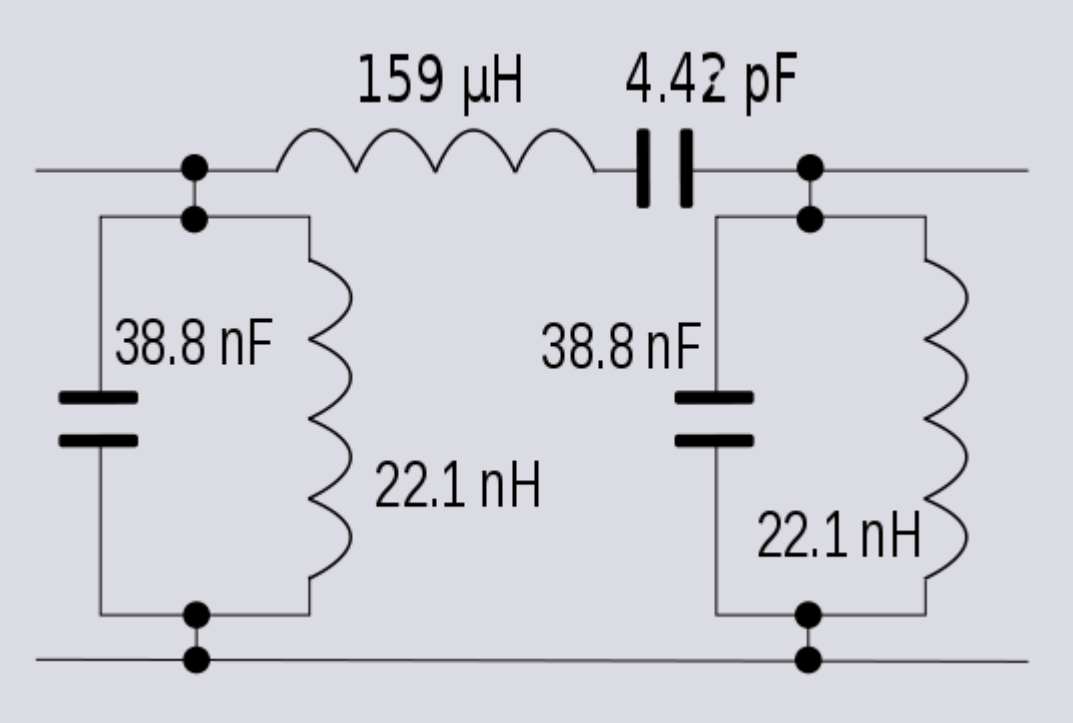
If $\omega_1$ and $\omega_2$ are the lower and upper frequency points (respectively) of the bandpass response corresponding to $\omega_c'$ of the prototype, then,

$\Delta\omega = \omega_2 - \omega_1$ and $\omega_0 = \sqrt{\omega_1\omega_2}$

$\Delta\omega$ is the absolute bandwidth, and $\omega_0$ is the resonant frequency of the resonators in the filter. Note that frequency scaling the prototype prior to lowpass to bandpass transformation does not affect the resonant frequency, but instead affects the final bandwidth of the filter.

The transfer function of the filter is transformed according to:

$$A(i\omega) \rightarrow A\left(\omega_c' Q\left[\frac{i\omega}{\omega_0} + \frac{\omega_0}{i\omega}\right]\right)$$




The prototype filter above, transformed to a 50Ω, 6MHz bandpass filter with 100kHz bandwidth

Inductors are transformed into series resonators,

$$L' \rightarrow L = \frac{\omega_c' Q}{\omega_0} L' \, , \, C = \frac{1}{\omega_0 \omega_c' Q} \frac{1}{L'}$$

and capacitors are transformed into parallel resonators,

$$C' \rightarrow C = \frac{\omega_c' Q}{\omega_0} C' \parallel L = \frac{1}{\omega_0 \omega_c' Q} \frac{1}{C'}$$

## Lowpass to bandstop

The required frequency transformation for lowpass to bandstop is:

$$\frac{\omega_c'}{i\omega} \to Q\left(\frac{i\omega}{\omega_0} + \frac{\omega_0}{i\omega}\right)$$

Inductors are transformed into parallel resonators,

$$L' \to L = \frac{\omega_c'}{\omega_0 Q} L' \parallel C = \frac{Q}{\omega_0 \omega_c'} \frac{1}{L'}$$

and capacitors are transformed into series resonators,

$$C' \to C = \frac{\omega_c'}{\omega_0 Q} C' \,,\; L = \frac{1}{\omega_0 Q \omega_c'} \frac{1}{C'}$$

## Lowpass to multi-band

Filters with multiple passbands may be obtained by applying the general transformation:

$$\frac{\omega_c'}{i\omega} \to \frac{1}{Q_1\left(\frac{i\omega}{\omega_{01}} + \frac{\omega_{01}}{i\omega}\right)} + \frac{1}{Q_2\left(\frac{i\omega}{\omega_{02}} + \frac{\omega_{02}}{i\omega}\right)} + \cdots$$

The number of resonators in the expression corresponds to the number of passbands required. Lowpass and highpass filters can be viewed as special cases of the resonator expression with one or the other of the terms becoming zero as appropriate. Bandstop filters can be regarded as a combination of a lowpass and a highpass filter. Multiple bandstop filters can always be expressed in terms of a multiple bandpass filter. In this way it, can be seen that this transformation represents the general case for any bandform, and all the other transformations are to be viewed as special cases of it.

The same response can equivalently be obtained, sometimes with a more convenient component topology, by transforming to multiple stopbands instead of multiple passbands. The required transformation in those cases is:

$$\frac{i\omega}{\omega_c'} \to \frac{1}{Q_1\left(\frac{i\omega}{\omega_{01}} + \frac{\omega_{01}}{i\omega}\right)} + \frac{1}{Q_2\left(\frac{i\omega}{\omega_{02}} + \frac{\omega_{02}}{i\omega}\right)} + \cdots$$

## *Alternative prototype*

In his treatment of image filters, Zobel provided an alternative basis for constructing a prototype which is not based in the frequency domain. The Zobel prototypes do not, therefore, correspond to any particular bandform, but they can be transformed into any of them. Not giving special significance to any one bandform makes the method more mathematically pleasing; however, it is not in common use.

The Zobel prototype considers filter sections, rather than components. That is, the transformation is carried out on a two-port network rather than a two-terminal inductor or capacitor. The transfer function is expressed in terms of the product of the series impedance, Z, and the shunt admittance Y of a filter half-section. This quantity is nondimensional, adding to the prototype's generality. Generally, ZY is a complex quantity,

$ZY = U + iV$ and as $U$ and $V$ are both, in general, functions of $\omega$ we should properly write,

$$ZY = U(\omega) + iV(\omega)$$

With image filters, it is possible to obtain filters of different classes from the constant k filter prototype by means of a different kind of transformation, constant k being those filters for which Z/Y is a constant. For this reason, filters of all classes are given in terms of $U(\omega)$ for a constant k, which is notated as,

$$ZY = U_k(\omega) + iV_k(\omega)$$

In the case of dissipationless networks, i.e. no resistors, the quantity $V(\omega)$ is zero and only $U(\omega)$ need be considered. $U_k(\omega)$ ranges from 0 at the centre of the passband to -1 at the cut-off frequency and then continues to increase negatively into the stopband regardless of the bandform of the filter being designed. To obtain the required bandform, the following transforms are used:

For a lowpass constant k prototype that is scaled:

$$R_0 = 1\,,\ \omega_c = 1$$

the independent variable of the response plot is,

$$U_k(\omega) = -\omega^2$$

The bandform transformations from this prototype are,

for lowpass, $$U_k(\omega) \rightarrow \left(\frac{i\omega}{\omega_c}\right)^2$$

for highpass, $$U_k(\omega) \to \left(\frac{\omega_c}{i\omega}\right)^2$$

and for bandpass, $$U_k(\omega) \to Q^2 \left(\frac{i\omega}{\omega_0} + \frac{\omega_0}{i\omega}\right)^2$$

Chapter 10

# Network Synthesis Filters

**Network synthesis** is a method of designing signal processing filters. It has produced several important classes of filter including the Butterworth filter, the Chebyshev filter and the Elliptic filter. It was originally intended to be applied to the design of passive linear analogue filters but its results can also be applied to implementations in active filters and digital filters. The essence of the method is to obtain the component values of the filter from a given mathematical polynomial ratio expression representing the desired transfer function.

## *Description of method*

The method can be viewed as the inverse problem of network analysis. Network analysis starts with a network and by applying the various electric circuit theorems predicts the response of the network. Network synthesis on the other hand, starts with a desired response and its methods produce a network that outputs, or approximates to, that response.

Network synthesis was originally intended to produce filters of the kind formerly described as "wave filters" but now usually just called filters. That is, filters whose purpose is to pass waves of certain wavelengths while rejecting waves of other wavelengths. Network synthesis starts out with a specification for the transfer function of the filter, H(s), as a function of complex frequency, s. This is used to generate an expression for the input impedance of the filter (the driving point impedance) which then, by a process of continued fraction or partial fraction expansions results in the required values of the filter components. In a digital implementation of a filter, H(s) can be implemented directly.

The advantages of the method are best understood by comparing it to the filter design methodology that was used before it, the image method. The image method considers the characteristics of an individual filter section in an infinite chain (ladder topology) of

identical sections. The filters produced by this method suffer from innaccuracies due to the theoretical termination impedance, the image impedance, not generally being equal to the actual termination impedance. This is not the case with network synthesis filters, the terminations are included in the design from the start. The image method also requires a certain amount of experience on the part of the designer. The designer must first decide how many sections and of what type should be used, and then after calculation, will obtain the transfer function of the filter. This may not be what is required and there can be a number of iterations. The network synthesis method, on the other hand, starts out with the required function and outputs the sections needed to build the corresponding filter.

In general, the sections of a network synthesis filter are identical topology (usually the simplest ladder type) but different component values are used in each section. By contrast, the structure of an image filter has identical values at each section - this is a consequence of the infinite chain approach - but may vary the topology from section to section to achieve various desirable characteristics. Both methods make use of low-pass prototype filters followed by frequency transformations and impedance scaling to arrive at the final desired filter.

## *Important filter classes*

The class of a filter refers to the class of polynomials from which the filter is mathematically derived. The order of the filter is the number of filter elements present in the filter's ladder implementation. Generally speaking, the higher the order of the filter, the steeper the cut-off transition between passband and stopband. Filters are often named after the mathematician or mathematics on which they are based rather than the discoverer or inventor of the filter.

### Butterworth filter

Butterworth filters are described as maximally flat, meaning that the response in the frequency domain is the smoothest possible curve of any class of filter of the equivalent order.

The Butterworth class of filter was first described in a 1930 paper by the British engineer Stephen Butterworth after whom it is named. The filter response is described by Butterworth polynomials, also due to Butterworth.

### Chebyshev filter

A Chebyshev filter has a faster cut-off transition than a Butterworth, but at the expense of there being ripples in the frequency response of the passband. There is a compromise to be had between the maximum allowed attenuation in the passband and the steepness of the cut-off response. This is also sometimes called a type I Chebyshev, the type 2 being a filter with no ripple in the passband but ripples in the stopband. The filter is named after

Pafnuty Chebyshev whose Chebyshev polynomials are used in the derivation of the transfer function.

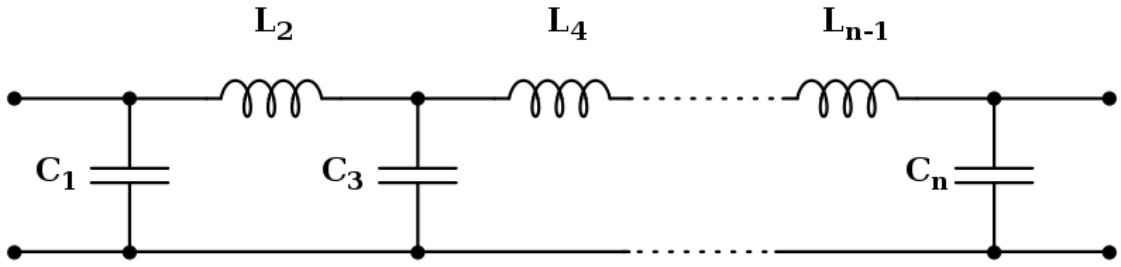
## Cauer filter

Cauer filters have equal maximum ripple in the passband and the stopband. The Cauer filter has a faster transition from the passband to the stopband than any other class of network synthesis filter. The term Cauer filter can be used interchangeably with elliptical filter, but the general case of elliptical filters can have unequal ripples in the passband and stopband. An elliptical filter in the limit of zero ripple in the passband is identical to a Chebyshev Type 1 filter. An elliptical filter in the limit of zero ripple in the stopband is identical to a Chebyshev Type 2 filter. An elliptical filter in the limit of zero ripple in both passbands is identical to a Butterworth filter. The filter is named after Wilhelm Cauer and the transfer function is based on elliptic rational functions.

## Bessel filter

- The Bessel filter has a maximally flat time-delay (group delay) over its passband. This gives the filter a linear phase response and results in it passing waveforms with minimal distortion. The Bessel filter has minimal distortion in the time domain due to the phase response with frequency as opposed to the Butterworth filter which has minimal distortion in the frequency domain due to the attenuation response with frequency. The Bessel filter is named after Friedrich Bessel and the transfer function is based on Bessel polynomials.

## *Driving point impedance*




Low-pass filter implemented as a ladder (Cauer) topology

The driving point impedance is a mathematical representation of the input impedance of a filter in the frequency domain using one of a number of notations such as Laplace transform (s-domain) or Fourier transform (jω-domain). Treating it as a one-port network, the expression is expanded using continued fraction or partial fraction expansions. The resulting expansion is transformed into a network (usually a ladder network) of electrical elements. Taking an output from the end of this network, so realised, will transform it into a two-port network filter with the desired transfer function.

Not every possible mathematical function for driving point impedance can be realised using real electrical components. Wilhelm Cauer (following on from R. M. Foster) did much of the early work on what mathematical functions could be realised and in which filter topologies. The ubiquitous ladder topology of filter design is named after Cauer.

There are a number of canonical forms of driving point impedance that can be used to express all (except the simplest) realisable impedances. The most well known ones are;

- Cauer's first form of driving point impedance consists of a ladder of shunt capacitors and series inductors and is most useful for low-pass filters.
- Cauer's second form of driving point impedance consists of a ladder of series capacitors and shunt inductors and is most useful for high-pass filters.
- Foster's first form of driving point impedance consists of parallel connected LC resonators and is most useful for band-pass filters.
- Foster's second form of driving point impedance consists of series connected LC anti-resonators and is most useful for band-stop filters.

## *Prototype filters*

Prototype filters are used to make the process of filter design less labour intensive. The prototype is usually designed to be a low-pass filter of unity nominal impedance and unity cut-off frequency, although other schemes are possible. The full design calculations from the relevant mathematical functions and polynomials are carried out only once. The actual filter required is obtained by a process of scaling and transforming the prototype.

Values of prototype elements are published in tables, one of the first being due to Sidney Darlington. Both modern computing power and the practice of directly implementing filter transfer functions in the digital domain have largely rendered this practice obsolete.

A different prototype is required for each order of filter in each class. For those classes in which there is attenuation ripple, a different prototype is required for each value of ripple. The same prototype may be used to produce filters which have a different bandform from the prototype. For instance low-pass, high-pass, band-pass and band-stop filters can all be produced from the same prototype.

Chapter 11

# Propagation Constant

The **propagation constant** of an electromagnetic wave is a measure of the change undergone by the amplitude of the wave as it propagates in a given direction. The quantity being measured can be the voltage or current in a circuit or a field vector such as electric field strength or flux density. The propagation constant itself measures change per metre but is otherwise dimensionless.

The propagation constant is expressed logarithmically, almost universally to the base e, rather than the more usual base 10 used in telecommunications in other situations. The quantity measured, such as voltage, is expressed as a sinusoidal phasor. The phase of the sinusoid varies with distance which results in the propagation constant being a complex number, the imaginary part being caused by the phase change.

## Alternative names

The term propagation constant is somewhat of a misnomer as it usually varies strongly with $\omega$. It is probably the most widely used term but there are a large variety of alternative names used by various authors for this quantity. These include, **transmission parameter**, **transmission function**, **propagation parameter**, **propagation coefficient** and **transmission constant**. In plural, it is usually implied that $\alpha$ and $\beta$ are being referenced separately but collectively as in **transmission parameters**, **propagation parameters**, **propagation coefficients**, **transmission constants** and **secondary coefficients**. This last occurs in transmission line theory, the term *secondary* being used to contrast to the *primary line coefficients*. The primary coefficients being the physical properties of the line; R,C,L and G, from which the secondary coefficients may be derived using the telegrapher's equation. Note that, at least in the field of transmission lines, the term transmission coefficient has a different meaning despite the similarity of name. Here it is the corollary of reflection coefficient.

## Definition

The propagation constant, symbol γ, for a given system is defined by the ratio of the amplitude at the source of the wave to the amplitude at some distance *x*, such that,

$$\frac{A_0}{A_x} = e^{\gamma x}$$

Since the propagation constant is a complex quantity we can write:

$$\gamma = \alpha + i\beta$$

where

α, the real part, is called the attenuation constant
β, the imaginary part, is called the phase constant

That β does indeed represent phase can be seen from Euler's formula;

$$e^{i\theta} = \cos\theta + i\sin\theta$$

which is a sinusoid which varies in phase as θ varies but does not vary in amplitude because;

$$\left|e^{i\theta}\right| = \sqrt{\cos^2\theta + \sin^2\theta} = 1$$

The reason for the use of base e is also now made clear. The imaginary phase constant, iβ, can be added directly to the attenuation constant, α, to form a single complex number that can be handled in one mathematical operation provided they are to the same base. Angles measured in radians require base e, so the attenuation is likewise in base e.

The propagation constant for copper (or any other conductor) lines can be calculated from the primary line coefficients by means of the relationship;

$$\gamma = \sqrt{ZY}$$

where;

$Z = R + i\omega L$, the series impedance of the line per metre and,
$Y = G + i\omega C$, the shunt admittance of the line per metre.

## *Attenuation constant*

In telecommunications, the term **attenuation constant**, also called **attenuation parameter** or **coefficient**, is the attenuation of an electromagnetic wave propagating through a medium per unit distance from the source. It is the real part of the propagation constant and is measured in nepers per metre. A neper is approximately 8.7dB. Attenuation constant can be defined by the amplitude ratio;

$$\left|\frac{A_0}{A_x}\right| = e^{\alpha x}$$

The propagation constant per unit length is defined as the natural logarithmic of ratio of the sending end current or voltage to the receiving end current or voltage.

### Copper lines

The attenuation constant for copper lines (or ones made of any other conductor) can be calculated from the primary line coefficients as shown above. For a line meeting the distortionless condition, with a conductance G in the insulator, the attenuation constant is given by;

$$\alpha = \sqrt{RG}$$

however, a real line is unlikely to meet this condition without the addition of loading coils and, furthermore, there are some frequency dependant effects operating on the primary "constants" which cause a frequency dependence of the loss. There are two main components to these losses, the metal loss and the dielectric loss.

The loss of most transmission lines are dominated by the metal loss, which causes a frequency dependency due to finite conductivity of metals, and the skin effect inside a conductor. The skin effect causes R along the conductor to be approximately dependent on frequency according to;

$$R \propto \sqrt{\omega}$$

Losses in the dielectric depend on the loss tangent (*tan*δ) of the material, which depends inversely on the wavelength of the signal and is directly proportional to the frequency.

$$\alpha_d = \frac{\pi\sqrt{\varepsilon_r}}{\lambda}\tan\delta$$

### Optical fibre

The attenuation constant for a particular propagation mode in an optical fiber, the real part of the axial propagation constant.

## Phase constant

In electromagnetic theory, the **phase constant**, also called **phase change constant**, **parameter** or **coefficient** is the imaginary component of the propagation constant for a plane wave. It represents the change in phase per metre along the path travelled by the wave at any instant and is equal to the angular wavenumber of the wave. It is represented by the symbol β and is measured in units of radians per metre.

From the definition of angular wavenumber;

$$k = \frac{2\pi}{\lambda} = \beta$$

This quantity is often (strictly speaking incorrectly) abbreviated to wavenumber. Properly, wavenumber is given by,

$$\tilde{\nu} = 1/\lambda$$

which differs from angular wavenumber only by a constant multiple of $2\pi$, in the same way that angular frequency differs from frequency.

For a transmission line, the Heaviside condition of the telegrapher's equation tells us that the wavenumber must be proportional to frequency for the transmission of the wave to be undistorted in the time domain. This includes, but is not limited to, the ideal case of a lossless line. The reason for this condition can be seen by considering that a useful signal is composed of many different wavelengths in the frequency domain. For there to be no distortion of the waveform, all these waves must travel at the same velocity so that they arrive at the far end of the line at the same time as a group. Since wave phase velocity is given by;

$$v_p = \frac{\lambda}{T} = \frac{f}{\tilde{\nu}} = \frac{\omega}{\beta}$$

it is proved that β is required to be proportional to ω. In terms of primary coefficients of the line, this yields from the telegrapher's equation for a distortionless line the condition;

$$\beta = \omega\sqrt{LC}$$

However, practical lines can only be expected to approximately meet this condition over a limited frequency band.

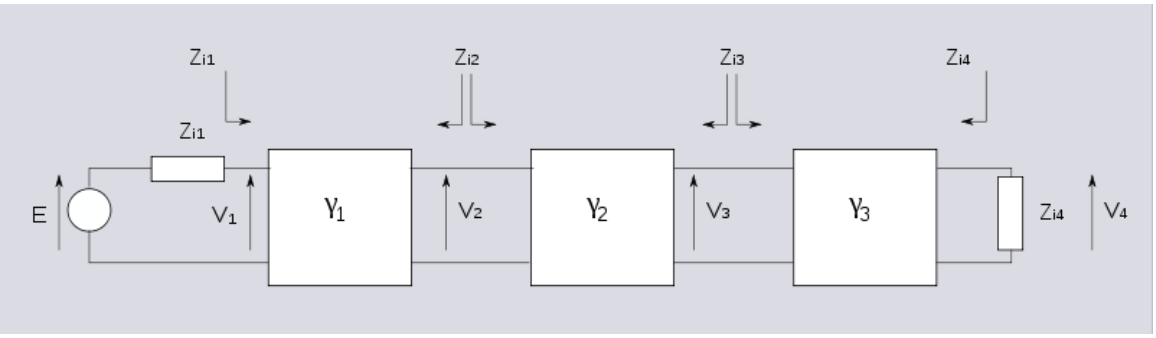
## Filters

The term propagation constant or propagation function is applied to filters and other two-port networks used for signal processing. In these cases, however, the attenuation and

phase coefficients are expressed in terms of nepers and radians per network section rather than per metre. Some authors make a distinction between per metre measures (for which "constant" is used) and per section measures (for which "function" is used).

The propagation constant is a useful concept in filter design which invariably uses a cascaded section topology. In a cascaded topology, the propagation constant, attenuation constant and phase constant of individual sections may be simply added to find the total propagation constant etc.

## Cascaded networks




Three networks with arbitrary propagation constants and impedances connected in cascade. The $Z_i$ terms represent image impedance and it is assumed that connections are between matching image impedances.

The ratio of output to input voltage for each network is given by,

$$\frac{V_1}{V_2} = \sqrt{\frac{Z_{I1}}{Z_{I2}}} e^{\gamma_1}$$

$$\frac{V_2}{V_3} = \sqrt{\frac{Z_{I2}}{Z_{I3}}} e^{\gamma_2}$$

$$\frac{V_3}{V_4} = \sqrt{\frac{Z_{I3}}{Z_{I4}}} e^{\gamma_3}$$

The terms $\sqrt{\frac{Z_{In}}{Z_{Im}}}$ are impedance scaling terms and their use is explained in the image impedance article.

The overall voltage ratio is given by,

$$\frac{V_1}{V_4} = \frac{V_1}{V_2} \cdot \frac{V_2}{V_3} \cdot \frac{V_3}{V_4} = \sqrt{\frac{Z_{I1}}{Z_{I4}}} e^{\gamma_1 + \gamma_2 + \gamma_3}$$

Thus for *n* cascaded sections all having matching impedances facing each other, the overall propagation constant is given by,

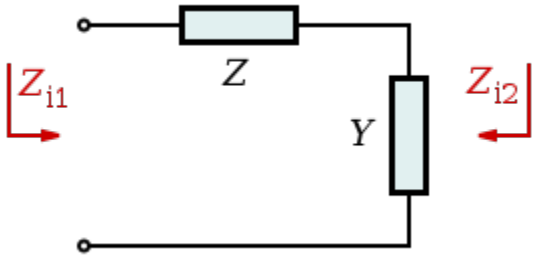
$$\gamma_{Tot} = \gamma_1 + \gamma_2 + \gamma_3 + \cdots + \gamma_n$$

Chapter 12

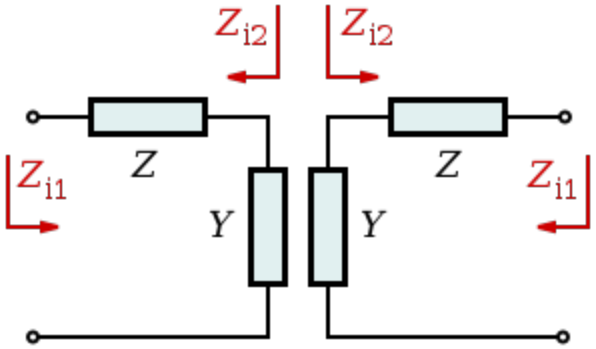
# Image Impedance

**Image impedance** is a concept used in electronic network design and analysis and most especially in filter design. The term image impedance applies to the impedance seen looking in to the ports of a network. Usually a two-port network is implied but the concept is capable of being extended to networks with more than two ports. The definition of image impedance for a two-port network is the impedance, $Z_{i\,1}$, seen looking in to port 1 when port 2 is terminated with the image impedance, $Z_{i\,2}$, for port 2. In general, the image impedances of ports 1 and 2 will not be equal unless the network is symmetrical (or anti-symmetrical) with respect to the ports.

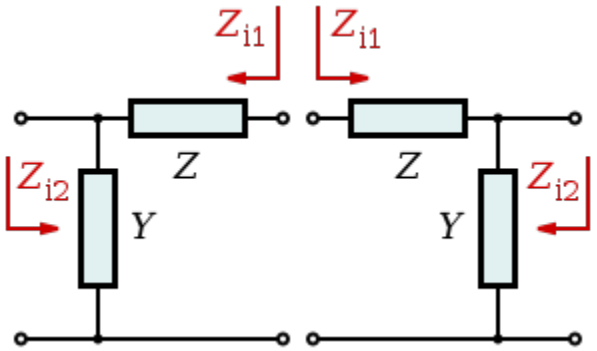
## *Derivation*




Simple 'L' network with series impedance Z and shunt admittance Y. Image impedances $Z_{i\,1}$ and $Z_{i\,2}$ are shown




Showing how a T section is made from two cascaded L half sections. $Z_{i\,2}$ is facing $Z_{i\,2}$ to provide matching impedances



Showing how a Π section is made from two cascaded L half sections. $Z_{i1}$ is facing $Z_{i1}$ to provide matching impedances

As an example, the derivation of the image impedances of a simple 'L' network is given below. The L network consists of a series impedance, *Z*, and a shunt admittance, *Y*.

The difficulty here is that in order to find $Z_{i1}$ it is first necessary to terminate port 2 with $Z_{i2}$. However, $Z_{i2}$ is also an unknown at this stage. The problem is solved by terminating port 2 with an identical network: port 2 of the second network is connected to port 2 of the first network and port 1 of the second network is terminated with $Z_{i1}$. The second network is terminating the first network in $Z_{i2}$ as required. Mathematically, this is equivalent to eliminating one variable from a set of simultaneous equations. The network can now be solved for $Z_{i1}$. Writing out the expression for input impedance gives;

$$Z_{i1} = Z + \frac{1}{2Y + \frac{1}{Z+Z_{i1}}}$$

and solving for $Z_{i1}$,

$$Z_{i1}^2 = Z^2 + \frac{Z}{Y}$$

$Z_{i2}$ is found by a similar process, but it is simpler to work in terms of the reciprocal, that is image admittance $Y_{i2}$,

$$Y_{i2}^2 = Y^2 + \frac{Y}{Z}$$

Also, it can be seen from these expressions that the two image impedances are related to each other by;

$$\frac{Z_{i1}}{Y_{i2}} = \frac{Z}{Y}$$

## Usage in filter design

When used in filter design, the 'L' network analysed above is usually referred to as a half section. Two half sections in cascade will make either a T section or a Π section depending on which port of the L section comes first. This leads to the terminology of $Z_{i\,T}$ to mean the $Z_{i\,1}$ in the above analysis and $Z_{i\,\Pi}$ to mean $Z_{i\,2}$.

## Relation to characteristic impedance

Image impedance is a similar concept to the characteristic impedance used in the analysis of transmission lines. In fact, in the limiting case of a chain of cascaded networks where the size of each single network is approaching an infinitesimally small element, the mathematical limit of the image impedance expression is the characteristic impedance of the chain. That is,

$$Z_i^2 \to \frac{Z}{Y}$$

The connection between the two can further be seen by noting an alternative, but equivalent, definition of image impedance. In this definition, the image impedance of a network is the input impedance of an infinitely long chain of cascaded identical networks (with the ports arranged so that like impedance faces like). This is directly analogous to the definition of characteristic impedance as the input impedance of an infinitely long line.

Conversely, it is possible to analyse a transmission line with lumped components, such as one utilising loading coils, in terms of an image impedance filter.

## Transfer function

The transfer function of the half section, like the image impedance, is calculated for a network terminated in its image impedances (or equivalently, as a section in an infinitely long chain of identical sections) and is given by,

$$A(i\omega) = \sqrt{\frac{Z_{I2}}{Z_{I1}}} e^{-\gamma}$$

where $\gamma$ is called the transmission function, propagation function or transmission parameter and is given by,

$$\gamma = \sinh^{-1}\sqrt{ZY}$$

The $\sqrt{\frac{Z_{I2}}{Z_{I1}}}$ term represents the voltage ratio that would be observed if the maximum available power was transferred from the source to the load. It would be possible to absorb this term into the definition of $\gamma$, and in some treatments this approach is taken. The purpose of separating it out is to scale γ so that $e^{\gamma}$ the actual voltage ratio may be. In the case of a network with symmetrical image impedances, such as a chain of an even number of identical L sections, the expression reduces to,

$$A(i\omega) = e^{-\gamma}$$

In general, $\gamma$ is a complex number such that,

$$\gamma = \alpha + i\beta$$

The real part of $\gamma$, represents an attenuation parameter, α in nepers and the imaginary part represents a phase change parameter, $\beta$ in radians. The transmission parameters for a chain of n half sections, provided that like impedance always faces like, is given by;

$$\gamma_n = n\gamma$$

As with the image impedance, the transmission parameters approach those of a transmission line as the filter section become infinitesimally small so that,

$$\gamma \to \sqrt{ZY}$$

with $\alpha$, $\beta$, $\gamma$, $Z$ and $Y$ all now being measured per metre instead of per half section.