More on Digital Filters

© Malcolm Moore
23-Oct-2004

Upending the (First Order) EMA

When in school many of the maths problems were to “simplify” a particular equation or to ‘make “x” the subject’ of another equation. I never knew why they wanted us to do so many of these but in later life I realised that this was an adjunct to “thinking outside the square” and looking at a problem from another perspective.

In the case if the EMA it has till now always been done in terms of a constant, and that constant then related to another relationship involving periods – like days! The relationships between simple and moving averages is linked by the same area under both curves (a calculus thing) and the result is that:

Periods = (2/K) – 1 Where K is a constant between 0 and 1 (like a percentage)

Making K the subject then

K = 2 / (Periods +1)

And the EMA equation is initially given as:

EMA(0) = Today * K + EMA(1) * (1 – K)

(Where EMA(1) is the EMA for the last period)

Combining these two equations gives the following:

EMA(0) = Today * 2 / (Periods +1) + EMA(1) * (1 – 2 / (Periods +1))

= (2 * Today + (P – 1) * EMA(1)) / (P + 1)

Now that really simplifies the EMA, and that is why I said that the EMA is by far the easiest to calculate. Note that P does not have to be an integer (whole number)!

Higher Order Digital Filters

In analogue electronics, a higher order analogue filter usually consists of several components, in a ‘ladder’ formation (series, shunt, series, shunt…) from the input to the output, and the base number of reactive components (capacitors and coils (inductors)) totals up to tell you the ‘order’ of the filter. I found this to be a fascinating rule, and it worked with simplicity.

Most filter manufacturers hated coils because they were usually labour intensive to manufacture, difficult to get the right component parts, prone to assembly error and therefore expensive, and manufacturers would often do anything to either minimise the number of coils, and this led to a whole range of alternative filter designs, including the development of Digital Filters.

Such filter designs without coils included crystal and ceramic lattice filters for radio and communications purposes, surface wave filters in televisions, switched capacitor filters in telecommunications, followed by digital filters in CD players and Hi-Fi systems and some service equipment.

The beauty about digital filters was that they could be programmed into what is called a Digital Signal Processor (DSP), a small integrated circuit that was specifically designed to store and forward, multiply and divide with large digital words. Most DSP chips come with substantial analogue / digital conversion circuitry on board.

When delving into digital processors and digital filter technology with a view to using this technology for security price analysis, it suddenly struck that the clocking rate for digital filters well exceeds twice the maximum frequency (the Nyquist criteria) and more importantly that the number of stages in many filter algorithms can well exceed 100 stages.

Put into EOD data where the clocking rate is 24 hours per cycle, then if 200 cycles had to pass before any intelligence was to come out, then the delay would be in the order of a year – and this is a bit too slow! There had to be an alternate method.

Some studying of the IIR filter as per the start of this chapter clearly shows that the standard output from a unit step input is a piecewise stepped exponential 1^{st} order charge curve.

Further studies in several digital filter texts seem to show that higher order filters consist of the same 1^{st} order filter with added delays feeding back to a common point, and a large degree of input digital wave distorting (which is basically a 1^{st} order FIR filter) in itself.

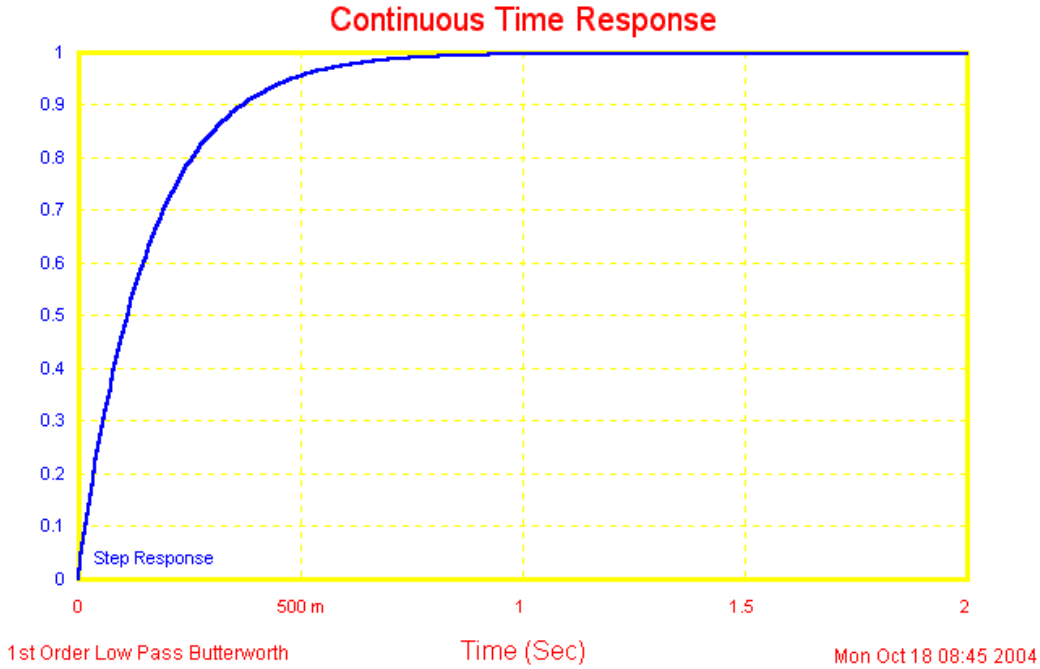
To synthesise equations, these filters are connected in series, parallel and/or latticed to 'make things work'! In my mind these were not right and I realised that filters above 1^{st} order needed another look at; as something is missing compared to lumped analogue filters.

When measuring the responses of analogue filters that are two simple physical domains that are commonly used. The first is the Frequency Domain and the second is the Time Domain. As it happens, Frequency is the reciprocal of Time, but each display shows up a filter in a very different light.

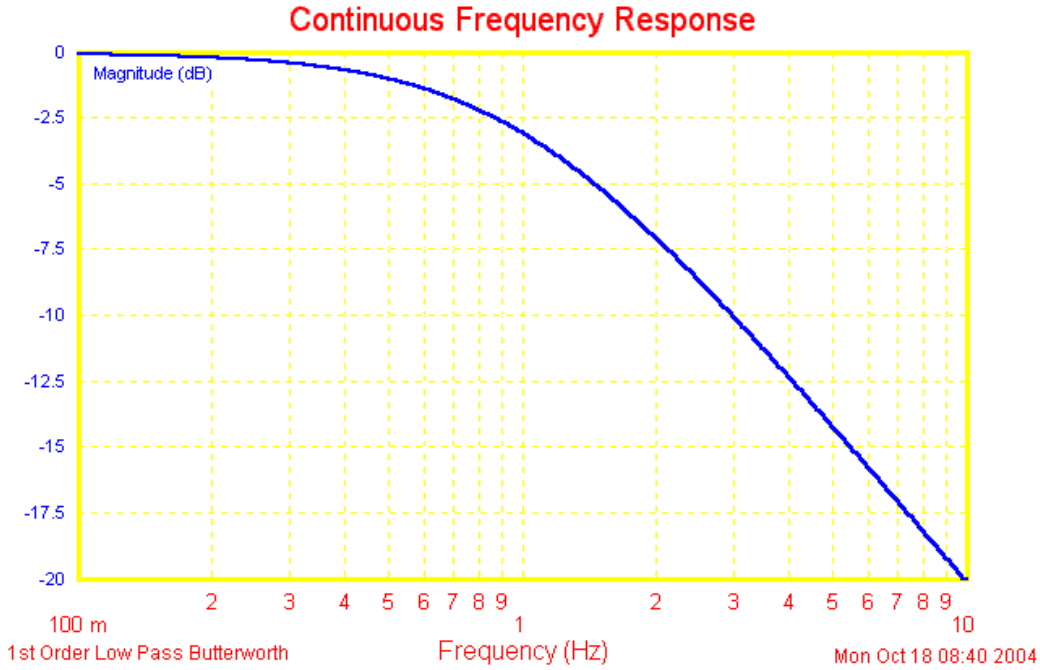
The Frequency response is critical on showing the cut-off point of a filter and how much the out of band response is attenuated, while the time response shows how the filter responds to a known excitation, and these responses can be calculated and measured, and the correlation between theory and reality is very close, meaning that the mathematical approximation used for analysis and design is very close to reality!

Cascading Filters

In the case of an EMA (which is a 1st order digital filter), the time response to a step input is a digital exponential charge curve[1]. It is unfortunately well documented as an analogue exponential charge curve and that has been 'done to death'!



In the frequency domain however, the response to a swept frequency has not been associated and it follows the graph below, in that it has a 3 dB (half power) frequency cut-off point and above that frequency the power asymptotically follows a –20 dB/decade line as shown on a logarithmic frequency graph below.



1 http://www.moore.org.au/sata/05_ema/20031214%20The%20Exponential%20Moving%20Average.pdf

The two graphs above show the time and frequency response of a 1st order filter with a frequency cut-off at 1 Hz. By direct relation with the continuous time curve above, the 80% mark crosses at about 250 msec, and we know that the 3 dB point is at 1 Hz, as shown in the right hand graph, and that 1 Hz has a cycle time of 1 sec which is 4 times 250 msec.

Using the SMA20 as the standard, the unit step time response moves from 0 to 100 % in 20 samples and reaches about 80% at about 16 samples, and an EMA20 crosses over at about the same (80%) point. So we have a direct correlation between the SMA and the EMA in the time domain and a similar frequency correlation point (3 dB) in the frequency domain.

In relating this; 16 samples divided by 4 times gives a normalised frequency of about 4 Hz, meaning that the normalised attenuation (in relation to 1 Hz as the normalised frequency reference) will be about 12.5 dB for a 1st order filter based on 20 days (EMA20).

An EMA40 filter would equate to 32 samples for the 80% transient point, and in normalised frequency terms that relates to about 8 Hz and the daily trade noise attenuation would be about 18 dB. This is why a longer EMA and/or SMA gives a smoother result than a shorter EMA or SMA, but a signal to noise ratio (SNR) less than 20 dB after filtering is hardly 'cleaning the water'!

To approach matters slightly differently; in Feedback Control Theory, the standard practice is to excite the system under test with a unit singularity step function and then measure the resultant rise time at the 10% and 90% points. In other words these are the 10% marker points from excitation and settling. This practice still comes out with a total of 80% but is centred on the overall movement, as opposed to concentrating on the end result only.

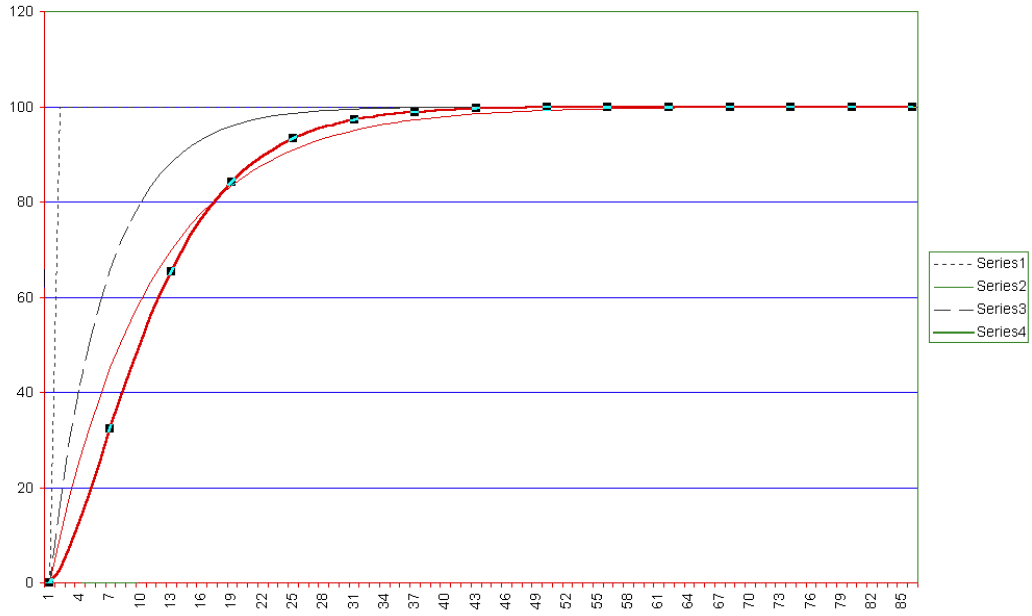
In Feedback Control Theory, 1st Order systems are rare, and for example the factor of 250 msec for 0 to 80% moves to about 285 msec for 10 % to 90% and that changes the reciprocal to about 3.5 instead of the original 4, but the approximation is close enough – for 1st Order filters!

So what would happen if we put a 1st order EMA followed by another 1st order EMA (in cascade)? In other words we would have the output of the first 1st order low (frequency) pass filter seeding directly into a second 1st order low (frequency) pass filter.

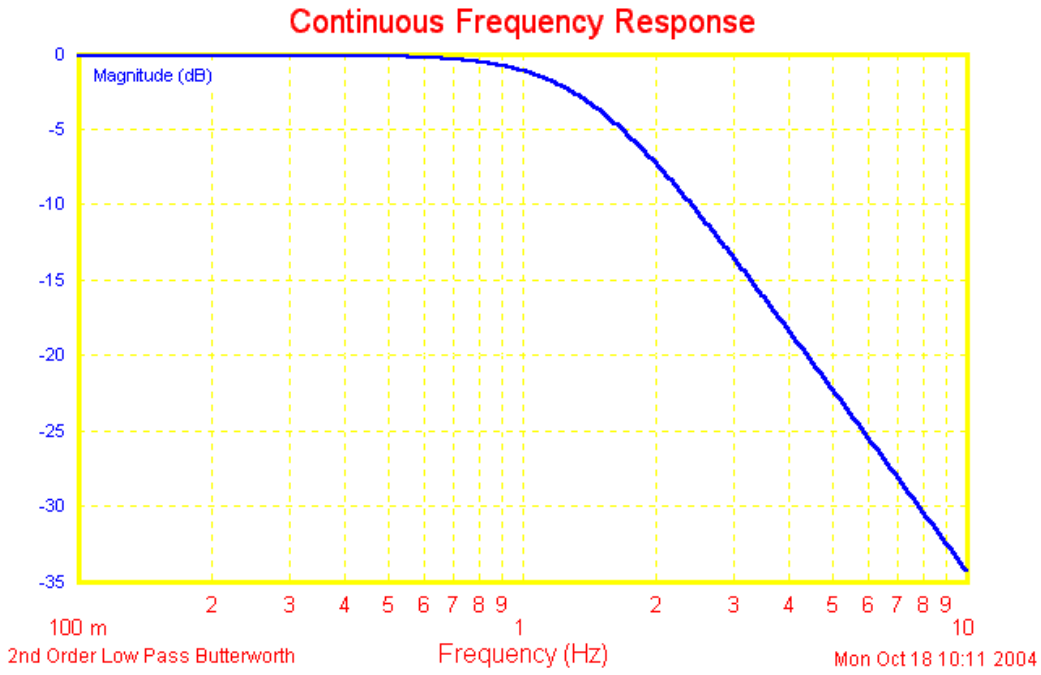
We could set up these filters so that the end time responses were approximately the same at the 80% mark, and the result would be an overdamped 2nd order filter, with an asymptotic frequency roll-off of -40 dB per decade as opposed to -20 dB per decade as before; that is twice as steep, but the cut-off frequency needs to be normalised down by root(2) or about 1.4 times to align the time response curve at the 80% mark as shown below on the next page:

Compare the dB scales on the right-hand graphs for Continuous Frequency Response and you will see that the frequency response is attenuated by about 35 db where the above graph the attenuation is only 20 dB. This is a big difference as this means that the trade noise can be substantially reduced!

In practice using two EMA11.2 cascaded, has an overall 3 dB cut off point very close to that of an EMA20 first order filter, and the resultant frequency response is very close to that on the right hand side graph, meaning the trading attenuation is about 18 dB down through this filter instead of 12.5 dB and if based on a 40EMA then the EOD noise would be attenuated by about 31 dB instead of 18 dB.



This 2nd Order EMA is substantially better than the 1st Order EMA and the time response curve; it approximates a sharper SMA time response curve and is sharper, possibly meaning that the cutover is sharper.



The immense problem about Technical Analysis is that 99.9% of all people who call themselves Technical Analysts, don't understand that the Transient Response is what they are actually looking at and making their 'decisions' on!

It takes quite some time to understand that the Transient Response of a First Order EMA (as used by almost all Technical Analysts) is clearly inferior to that of a First Order SMA or a Second Order EMA.

The problem is that the First Order EMA has an exponential attack/decay response that initially acts too quickly and then the tail acts far too slowly, and usually this long tail interferes with the next attack or decay so crossovers become more like tangents than intersections, making decisions rather vague and indecisive compared to crossovers when using SMAs.

In general 1st order EMAs are highly favoured by programmers because they are very easy to program in comparison to programming SMAs. As a consequence, most of the signalling indicators that are used by many Technical Analysts use a huge range of indicators that are unfortunately flawed in the first instance because most of these indicators are based on First Order EMAs for smoothing out the noisy data to generate their trading signals!

The original work shown here demonstrates how and where EMAs do not faithfully track the incoming signal nearly as well as SMAs do, and that there is a workaround of sorts in that a Second Order EMA (with a slight overshoot from a Step Excitation) will follow the SMA trajectory much closer than a First Order EMA, and the 2nd Order EMA can be programmed with about the same amount of simplicity as an SMA.

Copyright © Malcolm Moore, 2003, 2004, 2009.
Comments and Corrections are welcome