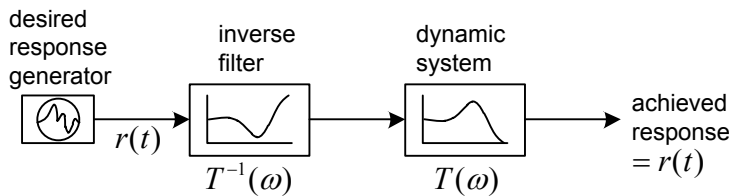


Causal Inverse Control

Brad Thoen, Staff Engineer
MTS Systems Corporation
Rev. 18-Aug-14

Inverse Control

A common form of inverse control involves processing the desired response signal with a filter whose frequency response is the inverse of the dynamic system to be controlled. This is shown schematically below. The overall frequency response between the desired response generator and the dynamic system output is $T^{-1}(\omega) \cdot T(\omega)$, which has unity magnitude and zero phase, so the achieved response is an exact replica of the desired response.



The inverse filter is typically a finite impulse response (FIR) filter, whose output $y(k)$ at sample instant k is computed from N present and past inputs $x(k-i)$, weighted by N coefficients h_i , where $0 \leq i \leq N-1$:

$$y(k) = h_0 x(k) + h_1 x(k-1) + \cdots + h_i x(k-i) + \cdots + h_{N-1} x(k-N+1)$$

What gives this filter its name is the fact that an impulse applied to the input shifts through the filter in N sample periods, after which its influence on the output is gone, i.e., its impulse response has a finite duration.

The inverse filter coefficients h_i are determined by measuring the dynamic system frequency response $T(\omega)$, inverting it to obtain $T^{-1}(\omega)$, then inverse Fourier transforming $T^{-1}(\omega)$ to obtain the inverse impulse response. The inverse impulse response samples are then assigned to the inverse FIR filter coefficients, taking advantage of the fact that an FIR filter's coefficients are equivalent to its impulse response. It is important to note that the number of filter coefficients N must be at least as great as the number of inverse impulse response samples, otherwise important inverse dynamic content will be lost. In other words, the duration of the inverse filter's impulse response must be at least as long as the dynamic system's inverse impulse response.

Causality

A causal system is one whose outputs depend on current or past inputs but not future inputs, i.e., outputs occur *after* inputs and thus the principle of cause and effect holds. All real world dynamic systems are causal. By contrast, a noncausal system is one whose outputs are influenced by future inputs, i.e., outputs can occur *before* inputs. This is the equivalent of predicting the future. Noncausal systems are a mathematical abstraction and do not exist in the real world.

A causal system has positive delay between input and output. Its impulse response is nonzero for $t \geq 0$, and is zero for $t < 0$. A noncausal system has negative delay (also called *anticipation*) between input and output, and some or all of its impulse response is nonzero for $t < 0$.

Inverse Control and Causality

Causality is of concern in inverse control because of the following fact: *the inverse of a real world dynamic system is always noncausal*. Intuitively this makes sense because all real world systems exhibit delay, and the inverse of delay is anticipation, i.e., noncausality. Because the inverse is noncausal, it cannot be directly implemented with a realtime digital filter, which itself is a causal device.

The causality dilemma can be solved by altering one of the design goals of inverse control: retain the goal of unity magnitude, but allow nonzero phase, i.e., delay. It is the insistence on zero phase between desired and achieved responses that results in a noncausal inverse. By relaxing this constraint, the inverse can be made causal. This is accomplished by shifting the noncausal inverse impulse response to the right in time, i.e. delaying it, until it becomes causal.

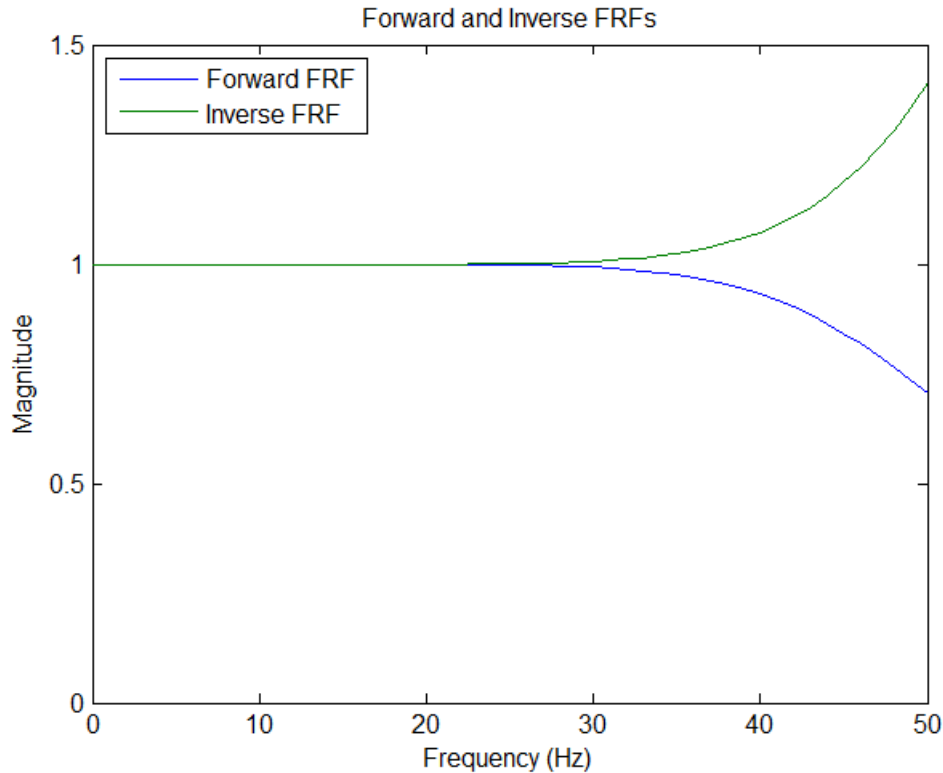
Example

In the following sections, the concepts of causality and inverse control will be demonstrated using an example dynamic system in the form of a fourth-order Butterworth filter with a -3dB frequency of 50 Hz and a sample rate of 512 Hz. The inverse filter will initially have an impulse response duration of 0.5 second, which will then be increased to 1 second for improved results.

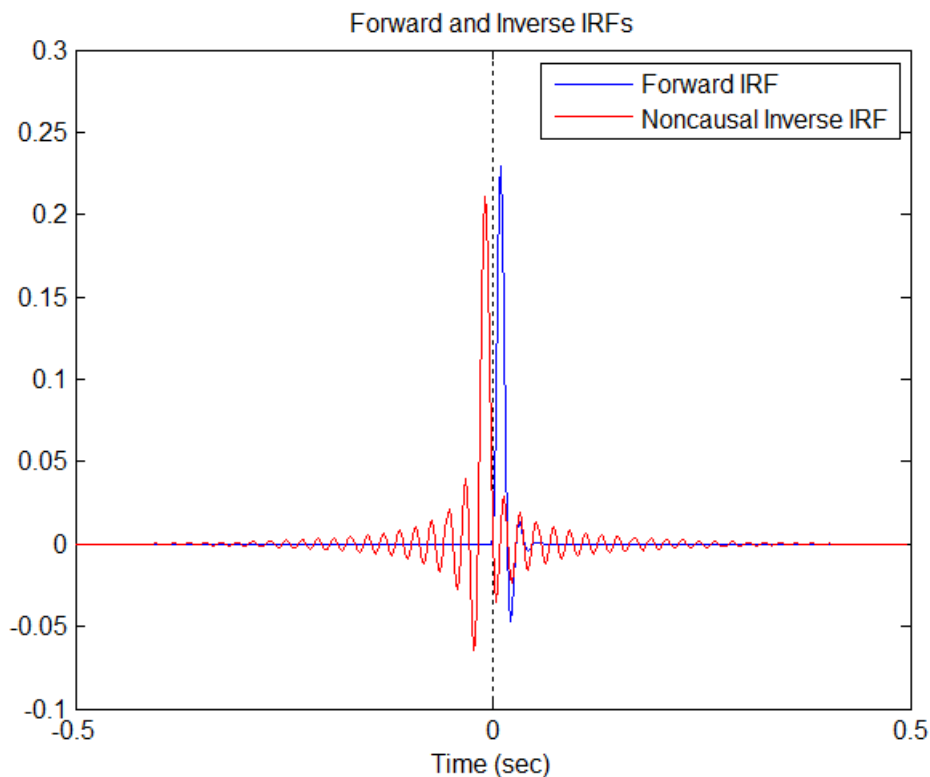
Notes on nomenclature:

- 1) FRF: Frequency Response Function, which is another expression for “frequency response”.
- 2) IRF: Impulse Response Function, which is another expression for “impulse response”.
- 3) Forward FRF or IRF: The frequency or impulse response of the dynamic system, to distinguish it from that of the inverse system.

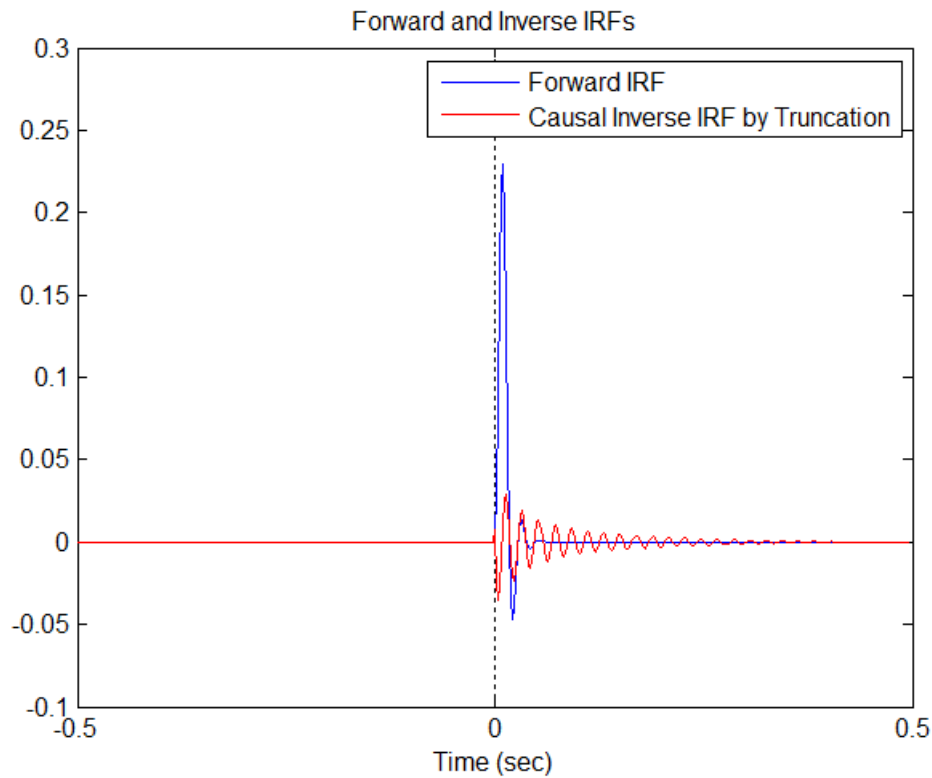
The forward and inverse FRFs of the example dynamic system are shown below. Note that inversion is typically done only over a limited frequency range, in this case, up to 50 Hz, even though this system has response up to the Nyquist frequency of 256 Hz. This bandlimiting process is done to prevent low response or low coherence frequency regions from creating high level incoherent system excitation when inverted.



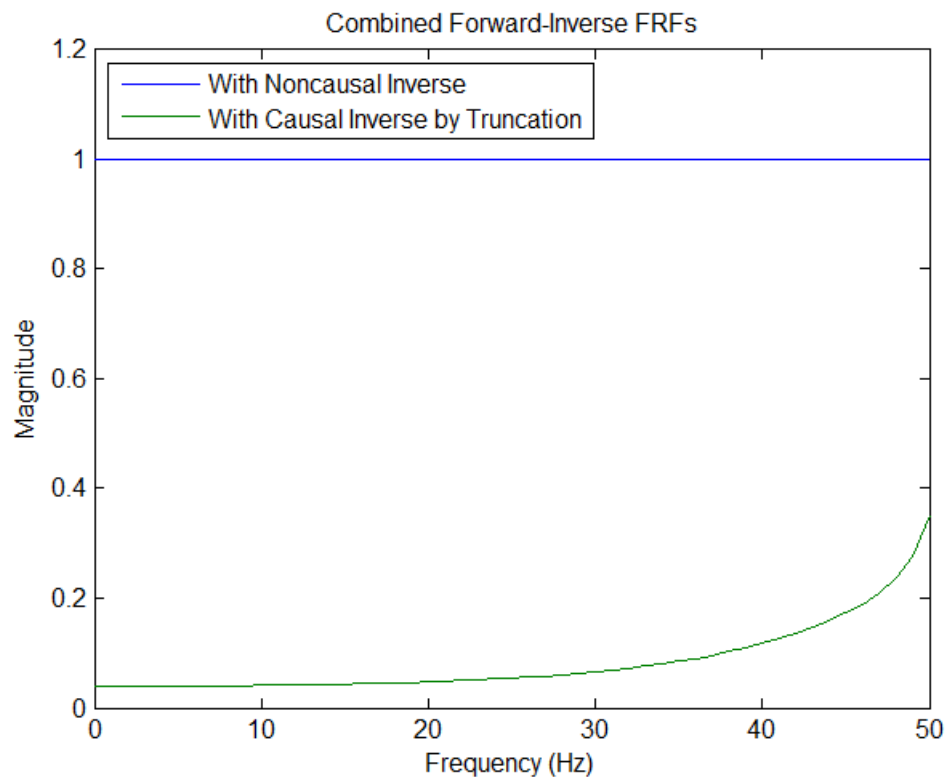
The forward and inverse IRFs obtained by inverse Fourier transforming the FRFs are shown below. Note that the forward IRF is causal, i.e., nonzero for $t \geq 0$ and zero for $t < 0$. The inverse FRF has its main response for $t < 0$, making it noncausal, but also has response for $t \geq 0$ as well. Also note that the duration of inverse IRF far greater than that of the forward IRF.



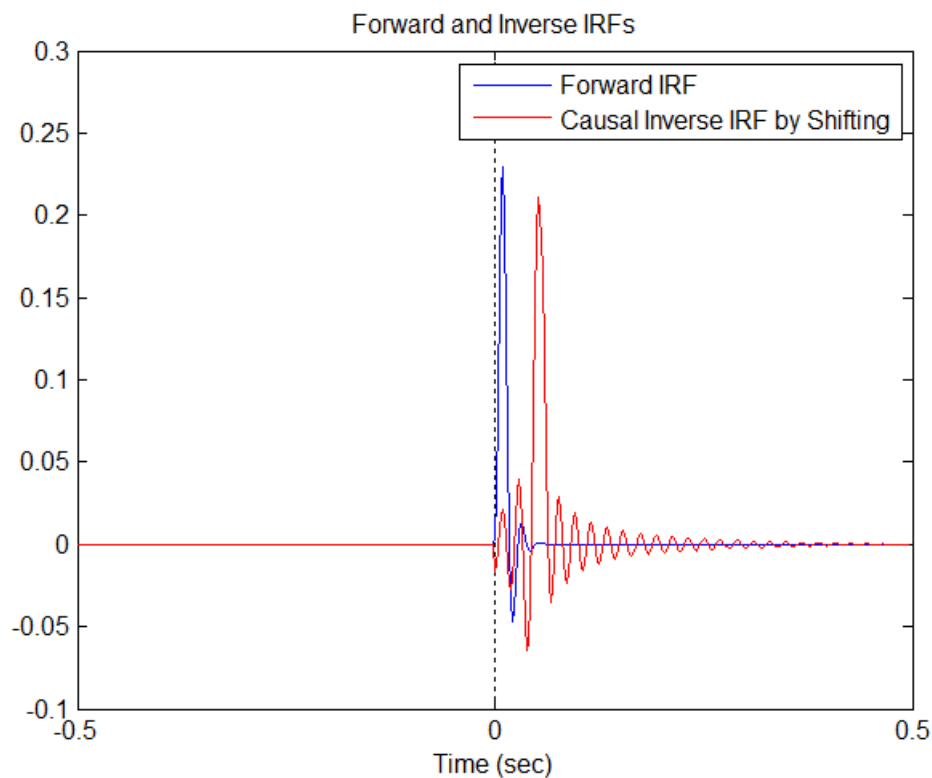
One way to make the inverse IRF causal is to truncate everything to the left of $t = 0$:



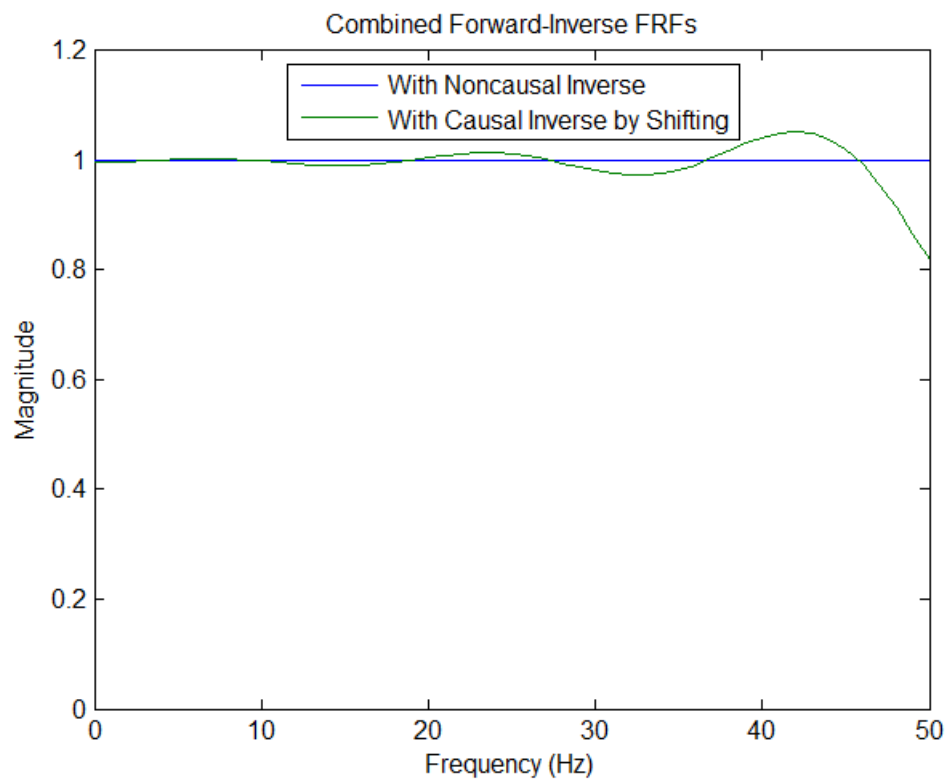
However, this radical truncation of inverse dynamic content results in a poor inverse FRF, which when combined with the forward FRF results in a combined FRF that is far from the ideal unity magnitude:



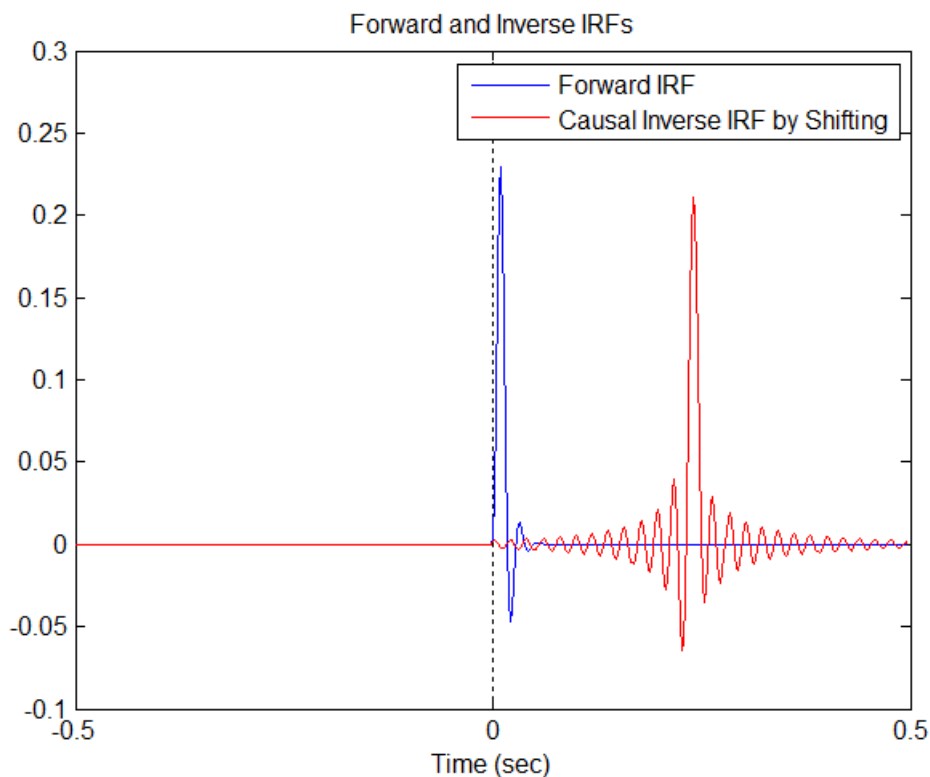
Better results can be obtained if the inverse IRF is shifted to the right before the truncation that makes it causal. Here, the shift is 0.0625 second, which brings the main response impulse to the right of $t = 0$:



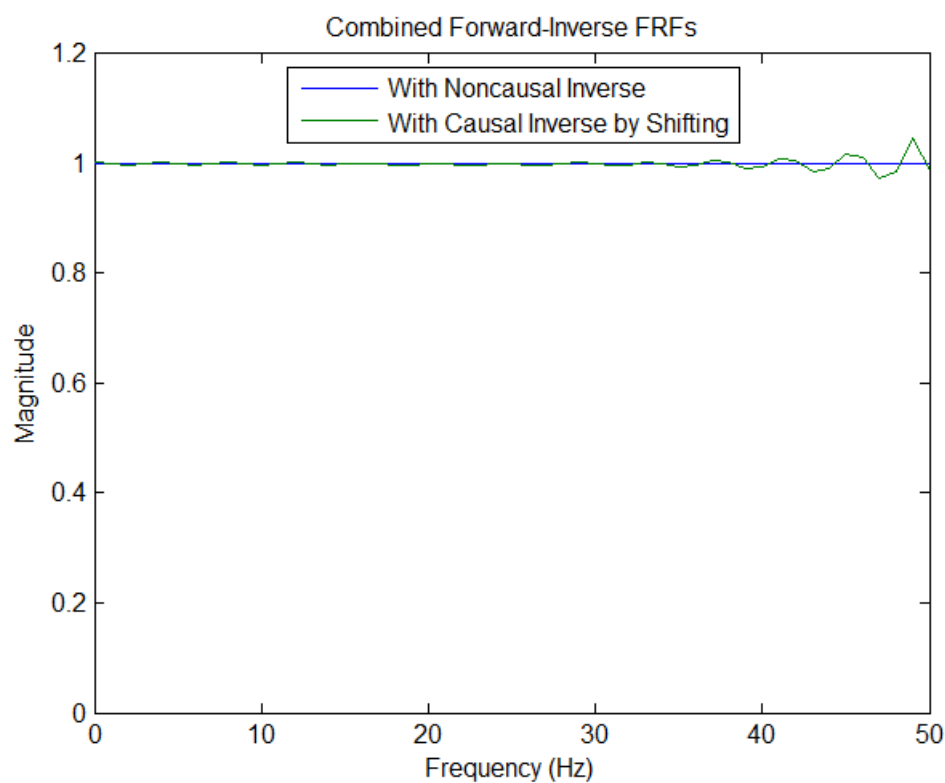
The combined forward-inverse FRF is much better, but is still considerably different from unity because too much of the front end of the inverse IRF is still being truncated:



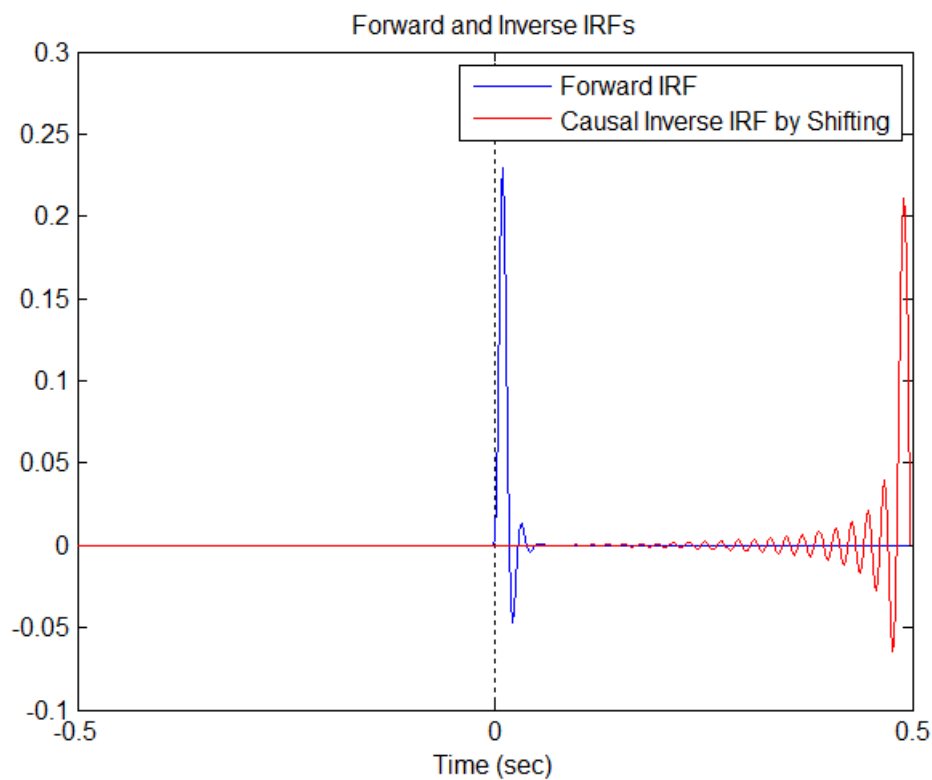
Continue to shift the inverse IRF to the right, in this case 0.25 second. Most of the inverse impulse response is now contained within the inverse filter's 0.5 second impulse response duration, therefore little inverse dynamic content is truncated on either the front or the tail end:



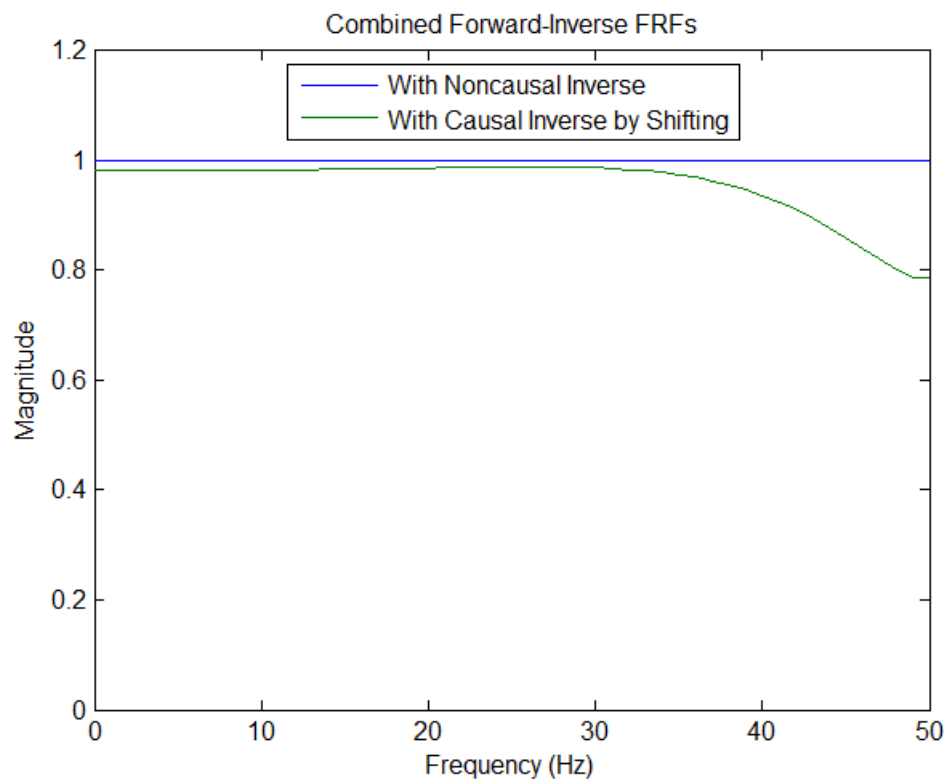
This results in a combined forward-inverse FRF that is quite good because much of the inverse IRF has been preserved:



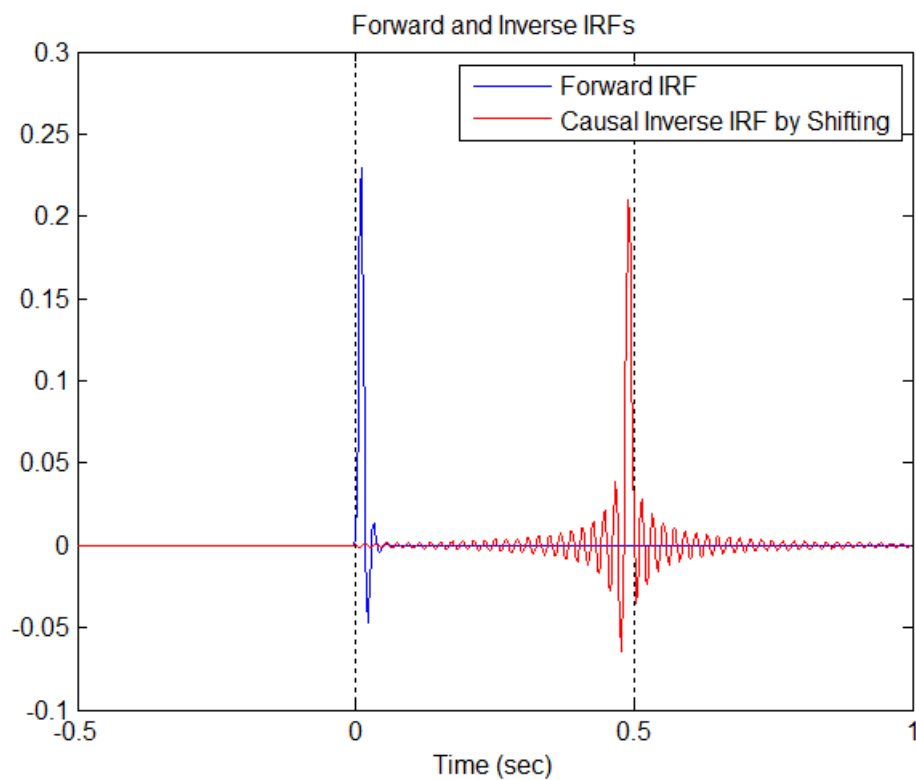
Even more of the inverse FRF's front end can be preserved by shifting it further to the right, in this case 0.5 second:



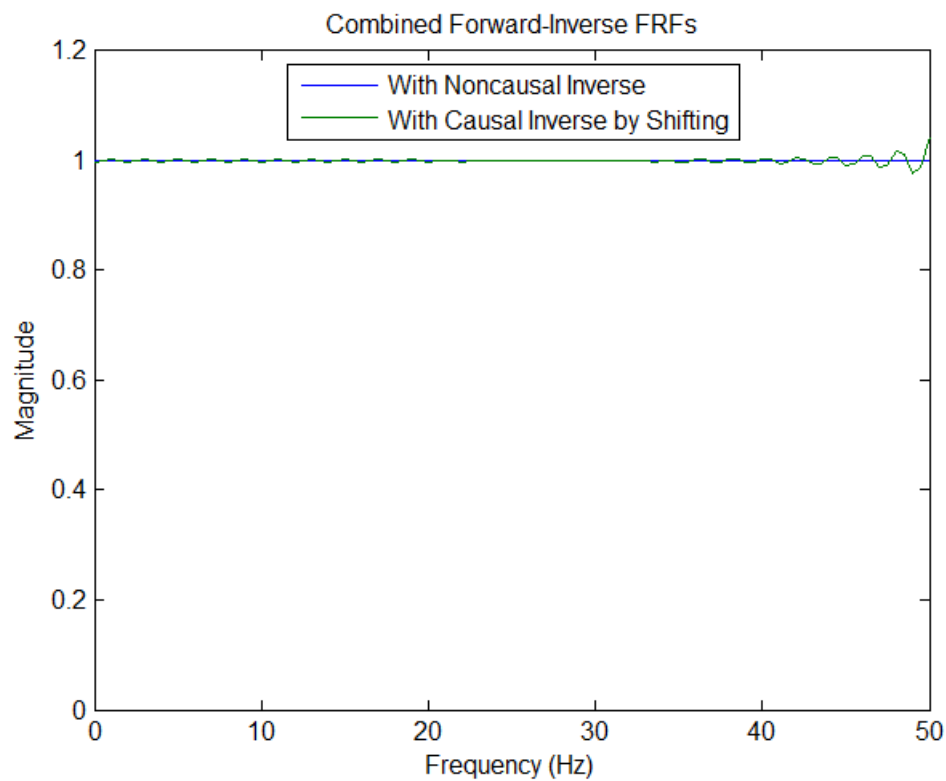
However, the combined forward-inverse FRF is worse, because the entire tail end of the inverse FRF is truncated by fitting it within inverse filter's 0.5 second impulse response duration:



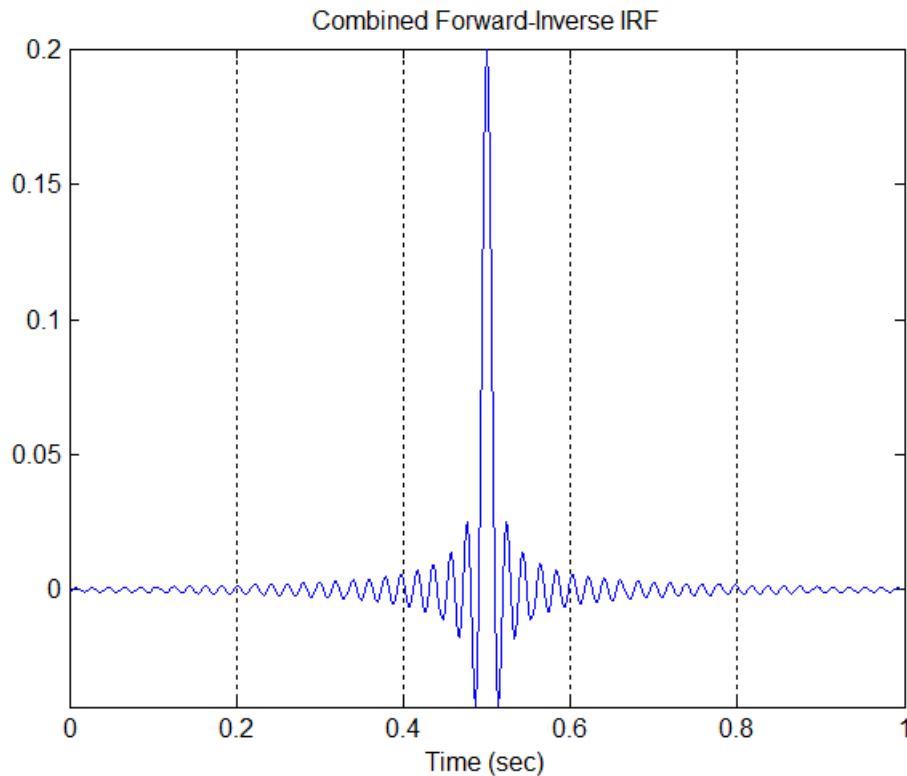
Preserve the tail end of the inverse IRF by doubling the inverse filter's 0.5 second impulse response duration to 1 second:



The combined forward-inverse FRF is excellent, because most of the inverse IRF fits within the confines of the inverse filter, and very little content is truncated:



The combined forward-inverse IRF clearly shows an overall input-output delay equal to the applied shift:



Note that the combined forward-inverse IRF is not a pure impulse as one might expect. This is because inversion was done only over a limited bandwidth, resulting in a combined forward-inverse system that is a nearly perfect lowpass filter, whose IRF is a sinc function ($\text{sinc}(x) = \sin(x)/x$) like that seen in the above plot. Had inversion been done over the entire bandwidth, the combined forward-inverse IRF would indeed be a pure impulse.

Tuning Strategy for Causal Inverse Control

To construct a causal inverse filter for inverse control purposes, follow these steps:

- 1) Increase the *anticipation* delay to shift the inverse IRF to the right until the front end of the inverse IRF is mostly to the right of $t = 0$.
- 2) Increase the *impulse response length* until the tail end of the inverse FRF is mostly to the left of $t = \text{impulse response length}$.

Notes:

- 1) Inverse IRFs are not always symmetric as it was in the example, where by coincidence it happened that $\text{anticipation} = \text{impulse response length} / 2$. Often the front end of the inverse IRF will have a different duration than the tail end. That is why there are two inverse filter adjustments: *anticipation* for the front end, and *impulse response length* for the tail end.
- 2) The *impulse response length* is determined by the duration of the inverse IRF, which in general is far greater than the duration of forward IRF.
- 3) The inverse FRF is infinite in duration, so some truncation must occur. The goal is to truncate as little of it as possible.