Anti-Aliasing Filters in Digital Photography

Douglas A. Kerr

Issue 1
April 19, 2025

ABSTRACT

Many digital cameras have, in front of the sensor array proper, what is variously described as a spatial low-pass filter, a blur filter, or an anti-aliasing filter. In this article we investigate the purpose and operation of such a filter.

1 PREFACE AND CAVEAT

A fairly complete draft of this article was written in 2011, and it somehow got trapped in some crack in my world and was not finished and published then. In 2025 I discovered it and completed it, still mostly based on the information I had in 2011.

Since then, there have been many advances in the implementation of anti-aliasing filters for digital cameras, and these have not been reflected in this article.

Nonetheless, the underlying theoretical concepts described here remain relevant to concept of aliasing, and to its avoidance.

2 INTRODUCTION

2.1 The anti-aliasing filter

Many digital cameras have, in front of the sensor array proper, what is variously described as a (spatial) low-pass filter, a blur filter or an anti-aliasing filter.

The purpose of this filter relates to the fact that a digital image, being composed of discrete pixels, is a “sampled” representation of the actual optical image. The need for a filter arises from this situation.

We will first explain the underlying theory in a different but parallel context: the digital representation of electrical waveforms (such as audio or video signals). The one-dimensional nature of a waveform (as contrasted to the two-dimensional nature of a camera image) allows the principles to be more clearly seen.

2.2 Sampling of an audio waveform

When we convert an electric audio signal to digital form, we usually begin by capturing the instantaneous value of the signal at regular intervals of time, a process called *sampling*. The remainder of the signal is ignored.

Nevertheless, if the rate at which we take samples is above twice the highest frequency contained in the signal, then the entire original waveform (including its values between the times of sampling) can be perfectly reconstructed from the series of sample values alone.[1] This is the principle expressed by the famous Nyquist-Shannon sampling theorem.[2]

Looking at it from the other direction, if we sample at a rate of f_s samples per second (*the sampling rate*), we will completely capture any signal containing component frequencies up to (but not including) $f_s/2$ Hz. That frequency is called the *Nyquist frequency*.

Suppose we present for sampling a signal containing components above the Nyquist frequency. Will they just be left behind in the sampling and reconstruction process? No, something more harmful occurs. Each such component will be replaced in the reconstructed signal by a spurious component. Its frequency lies below the Nyquist frequency by the same amount that the original "out of bounds" component was above it.

For example, suppose we have established a sampling rate of 8000 samples per second. This will accommodate (and tolerate) all frequency components below 4000 Hz. If the signal being sampled has a component (presumably one of no interest to us) at 4400 Hz, the reconstructed signal will contain nothing above 4000 Hz, but will have a spurious component at 3600 Hz.

In effect the maverick component travels under a false identity within the train of samples, having the same representation there as a component within the legitimate frequency range of the system. For this reason this phenomenon is called *aliasing*,[3] and its impact is called *aliasing distortion*.

[1] "Perfectly" assumes that each sample perfectly reflects the value of the signal at the instant of sampling. In an actual digital system, that is not precisely true.

[2] Harry Nyquist expressed it in 1928, and it was rigorously proven by Claude Shannon in 1959. Although Shannon has long been recognized for his gigantic contributions to the entire field of information theory, his role in the sampling theorem first set forth by Nyquist has only recently come to be widely mentioned.

[3] This term has unfortunately also come to be applied to other anomalies in signal and image processing to which it does not so aptly apply!

Appendix A gives a graphic illustration of the basic mechanism of aliasing.

2.3 The anti-aliasing filter

In most applications, we cannot just assume that a signal about to be sampled contains no components at or above the Nyquist frequency for the sampling rate we have adopted in the system. We must assure this by first passing the signal through a low-pass filter, one which removes any components at or above the Nyquist frequency. This is often called an *anti-aliasing filter*.

2.4 Isn't that counterproductive?

Many people, first learning about this, wonder if the inclusion of an anti-aliasing filter isn't counter-productive. Doesn't it artificially limit the range of signal frequencies accommodated by the digital signal?

If in fact the anti-aliasing filter cut off at a frequency just below the Nyquist frequency, then it would not limit the handling of any frequency component that could travel correctly through the system. The frequencies it would eliminate are only those that would show up in the reconstructed signal at the wrong frequency, and are thus of no use to us (but rather are troublemakers). (That's why we have the filter.)

If we actually wish to successfully transport signal components with frequencies above that limit, we must increase the sampling rate (and increase the cutoff frequency of the filter accordingly.)

3 NOW, DIGITAL PHOTOGRAPHY

3.1 The digital image as a sampled representation

We will for a moment consider a monochrome digital camera. While the principles we will discuss apply as well to a color camera, the details are a bit more difficult to grasp there.

In digital photography, the pattern of actual image illuminance along one row of pixel sensors[4] is equivalent to our audio signal, and the capturing of that pattern by regularly-spaced pixel sensors is an example of sampling. Here, the sampling rate is expressed not in samples per second but rather in terms of samples (pixels) per

[4] This would be equally true along one column of the sensors, or along any diagonal path through a set of sensors. In reality, the process proceeds two-dimensionally, but fully recognizing this introduces mathematical complications that would only interfere with our purpose right now.

millimeter (or some other unit of distance). The concept of the Nyquist frequency also pertains here, in this case also expressed in cycles per millimeter. (The frequencies in this case are not in terms of time—*temporal frequency*—as in the case of an electrical signal, but rather in terms of distance (space)—*spatial frequency*.)

If the image on the sensor array contains detail so fine that, looked at in terms of spatial frequency, it contains components at or above the Nyquist frequency corresponding to the pixel sensor spacing, then the image as reconstructed from the pixel data will contain spurious components—the result of aliasing. To prevent this corruption of the reconstructed image, we must filter out such overly-high-frequency components—overly fine detail—in the original image before it is sampled.

3.2 The impact of aliasing

Suppose that our digital camera has a pixel pitch of 100 px/mm. This corresponds to a sampling frequency of 100 samples/mm. The Nyquist frequency for this sampling would be 50 cycles per mm.

Suppose now that we have an image (formed on the sensor by the lens) consisting of alternating black and white vertical lines at a pitch of 36 lines per mm[5] (that is, with a spatial frequency of 18 cycles per mm). That is actually the *fundamental frequency* of the illuminance variation; unless the reflectance of the subject varied precisely as a sine function, there would also be frequency components at multiples at 18 cy/mm (for a symmetrical pattern, typically **odd** multiples). So let's assume that the frequencies contained in the variation were 18 cy/mm, 54 cy/mm, 90 cy/mm, 126 cy/mm and so forth. But for now I'll ignore all but the 18 cy/mm and 54 cy/mm components.

Then 54 cy/mm component is above the Nyquist limit, and cannot be (properly) captured by our sensor array. But it does have an impact on the set of sample values—exactly the same impact as would have been had by a component whose frequency was 46 cy/mm. That is, it would affect the train of samples exactly like a component whose frequency was as far below the Nyquist frequency as this component was above it.

When the image is later reconstructed (on a display screen, for example, the overall process will take that pattern of sample values as meaning a component in the illuminance variation with a frequency 46 cy/mm (the only "legitimate" meaning, in this system, of such a pattern within the series of sample values). So the reconstructed

[5] This notation counts black and white lines separately, as we do in video work.

image will be a pattern whose frequency components are 18 cy/mm and 46 cy/mm. Of course, 46 cy/mm is not an integer multiple of 18 cy/mm, so this will not look like a pattern of any kind of stripes at a rate of 36 stripes/mm (counting black and white separately). It will actually look mostly like a pattern of stripes at 36 stripes/mm whose "strength" varies at a pitch of 1.5 mm. (I'll spare you the computation of that).

The visible manifestation of such a "modulated" pattern of luminance[6] is what is often described as a *moiré pattern*.

3.3 Doctor, doctor!

> "Doctor, doctor, when I do this my elbow hurts."
>
> "Then don't do that."

We just saw that, in a system with sampling rate of 100 samples/mm, where the Nyquist frequency is 50 cy/mm, any components in the variation of illuminance across the image whose frequencies are at or above 50 cy/mm will not only not be included in the reconstructed signal but in fact will be replaced by spurious components.

How can we avert this? By making sure that there are no such components in the image that falls on the sensor.

Of course, in the electrical waveform case we say at the beginning, we do exactly that by interposing into the signal chain, before the point where the waveform is sampled a low-pass filter. Its response drops to essentially zero by the time we get to the Nyquist frequency. Thus there are (essentially) no components in the waveform presented for sampling which are at or above the Nyquist frequency.

Can we do the equivalent thing in the case of our sampled image? Conceptually yes, but, the execution of the concept is very challenging, and so what is typically done is a considerable compromise with the ideal concept.

4 SOME MORE BACKGROUND

Before we proceed, we need to review some concepts upon which we will draw shortly.

[6] Note that although it is the phenomenon of *illuminance* on the sensor to which the photodetectors respond, in a reconstructed image (as on a display screen), the varying phenomenon in the displayed reconstructed image is *luminance*.

4.1 The frequency response of an electrical filter

To establish the next principle in a familiar context, let us return to the world of audio signal waveforms.

A system handling an audio signal (such as an amplifier or a filter) has a *frequency response curve*. This is a plot of how much relative attenuation the system affords to signal components at different frequencies.[7] In an audio amplifier, we ordinarily seek to have this curve "flat" over the range of audio frequencies in which we are interested.

On the other hand, in a filter we have a response curve which is intentionally non-uniform. The "ideal" anti-aliasing filter we discussed above, for example, has a response curve that drops sharply as we approach the Nyquist frequency.

In Figure 1 we see the response curve of a "brick wall" low-pass filter, so-called because of its instantaneous decrease in response at its cutoff frequency.

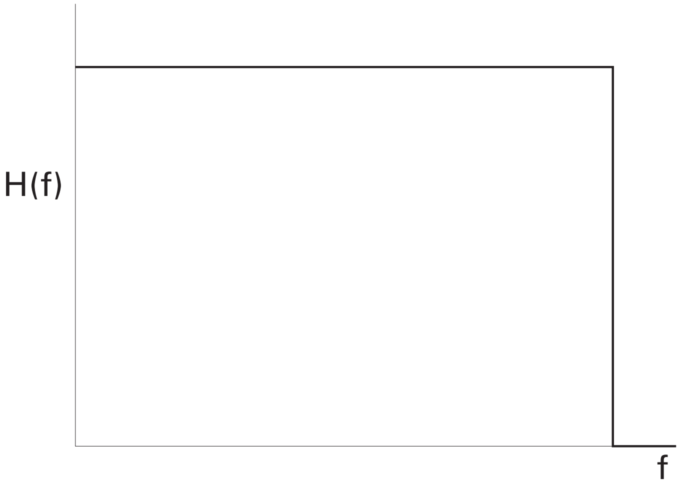

Figure 1. Frequency response of a "brick wall" low-pass filter

The horizontal axis is frequency, f. The labeling of the vertical axis as "H(f)" follows a convention in electrical engineering textbooks, where the response of a system, varying with frequency, is in general identified as the function "H(f)" (read "H of f", meaning that H is a function of f; that is, H varies with f).

[7] There is also a *phase response curve*. This tells, at any frequency, how the time reference of the output component at that frequency is shifted from the time reference of the original component. This is an important consideration, but I will nevertheless ignore it.

4.2 The modulation transfer function (MTF) of an optical "block"

An imaging system "block", such as a lens, has a comparable frequency response curve, called its modulation transfer function (MTF). This is a plot of how variations in the illuminance along a path across an image are attenuated, compared to the luminance variations in the scene itself, at different spatial frequencies. This approach treats the imaging system as a filter. [8] "Blurring", whether a result of imperfect focus, lens aberrations, diffraction effects, or the like, is a manifestation of decline in the MTF at higher spatial frequencies of interest to us.

4.3 The impulse response of an electrical filter

Suppose we have in the laboratory an electrical filter whose characteristics we wish to learn. We can send test signals at a large number of different frequencies through it, for each frequency measuring the relative amplitude (and perhaps phase) of the output signal.

But suppose we instead send through the filter a single pulse, of "very short" duration. The ideal theoretical concept would be a pulse of zero duration, although it would have to have an infinite voltage for it to contain any energy at all, and so we could not physically have such a thing. Nonetheless, such a pulse can be dealt with theoretically, and is called an *impulse*.

When we send this hypothetical impulse through the filter (once), a "one-shot" waveform comes out. This is called the *impulse response* of the filter.

If we capture that (perhaps on a recording oscilloscope), from it we can (through Fourier analysis) determine the entire frequency response curve of the filter (its complete phase response curve, too).

Note that if the filter actually had "flat" frequency and phase response curves (over all frequencies, to infinity), then impulse response would in fact just be an impulse; that is, what went in would come out, unchanged, just as we should expect.

[8] Note that it is the plot of MTF vs. **spatial frequency** that is comparable to the frequency response curve of a block in an electrical system, and we will be concerned with such plots here. However, in popular photographic work, it is common to show, for one or more specific spatial frequencies, a plot of the MTF vs. **locations across the image from the center outward**. It is important for the reader to understand that the MTF plots shown here do not correspond to the MTF presentation most often seen in connection with lenses and such.

In the case of a “brick wall” low-pass filter (whose frequency response we saw in Figure 1), the theoretical impulse response is shown in Figure 2.

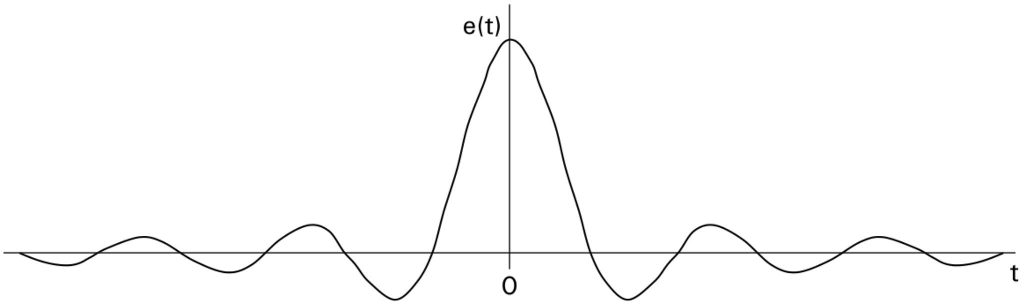

Figure 2. Impulse response of “brick wall” low-pass filter

We see here some interesting but vexing properties. The first is that the response begins before the time of the input impulse (t = 0). Of course that is an actual impossibility. Such a (theoretical) filter is called “prophetic”. I’ll come back to that in a little bit.

The second curiosity is that the response is infinite in time—that is, it begins an infinite time before the time of the input impulse, and lasts an infinite time after that (of course, I’ve only shown part of that). Filters whose impulse response are like that are in fact called infinite impulse response (IIR) filters.

In reality, an actual electrical filter of course can’t have an infinite impulse response (never mind the “prophetic” aspect). But that means it can’t have a “brick wall” frequency response—it means we cannot make such a filter in the real world.

Now, regarding the “prophetic” aspect. The symmetrical way in which I showed the output waveform implies that the phase response of the filter is zero for all frequencies. That is, no component of the output is “delayed”; thus the whole signal is not delayed compared to its “theoretical” manifestation).

But, once the duration of the output is no longer infinite (that is, we have given up our insistence on a “brick wall” response), we can imagine a phase response that in effect just says that the entire output is delayed from its theoretical schedule. If there is enough of that delay, then the output will not have to start before the time of the input pulse—that is, it will not have to be supernatural. And that is of course unavoidably the case for any “worldly” filter.

4.4 The point spread function of an optical block

Now imagine some "block" of an optical system (perhaps the lens of our camera). If we aim it at a tiny spot of light (a quasi-point source)[9], ideally it would generate a tiny spot (a quasi-point image) within the overall image.

But, because the block does not have a uniform (spatial) frequency response (MTF) over all spatial frequencies (to infinity), what comes out is not a point image, but some sort of blur figure. If we imagine the ideal but impractical case where the source is in fact a point source, then the result is called the *point spread function* of the optical block. It is described in terms of the illuminance of the pattern as we move along a radius from its center. If we plot that, it is exactly the equivalent of a plot of the impulse response of our electrical filter.

So, not surprisingly, if we know the point spread function of an optical block, we can (by Fourier analysis) determine its MTF vs. frequency "curve".

But there is a complication: the impulse response of an electrical filter is one dimensional (the voltage varies as a singe variable, time.) The point spread function of an optical block is two dimensional (the "blur figure" is two dimensional, as is the image of which it is a part).

It may well be "rotationally symmetrical", in which case the plot along a radius from the center describes the whole thing. But in the general case, it may not be rotationally-symmetrical. In that case, we would have to describe it with a separate plot of MTF vs. distance from the center along a radius at every possible orientation (or we of course could use a "three-dimensional" plot).

5 THE ELECTRICAL ANTI-ALIASING FILTER

We earlier introduced the concept of using a low-pass filter to strip from an electrical waveform all components whose frequencies were at or above the Nyquist frequency for the sampling situation we have.

What might such a filter be like (functionally)? It might seem that ideally it would have a "brick wall" response (as we saw in Figure 1). But as we heard, that is actually unattainable. (And if we even came close, there would be a number of troublesome side effects.)

[9] Conceptually, it would be ideal for this to have zero diameter, but of course then it would contain no light, so of course it would not even be visible.

Now, in practical systems, we do not seek to make such a utopian filter. Rather, we use a filter with a more practical “roll-off” characteristic. By doing so, we indeed limit the range of frequencies passed to less than that corresponding to the sampling rate of the system. For example, in standard basic digital telephony, we use a sampling rate of 8000 samples per second, which could theoretically cater to audio signal components at frequencies almost to 4000 Hz. But we only undertake to carry audio signal components up to about 3450 Hz, and use an anti-aliasing filter to match.

Such a filter might have the frequency response shown in Figure 3.

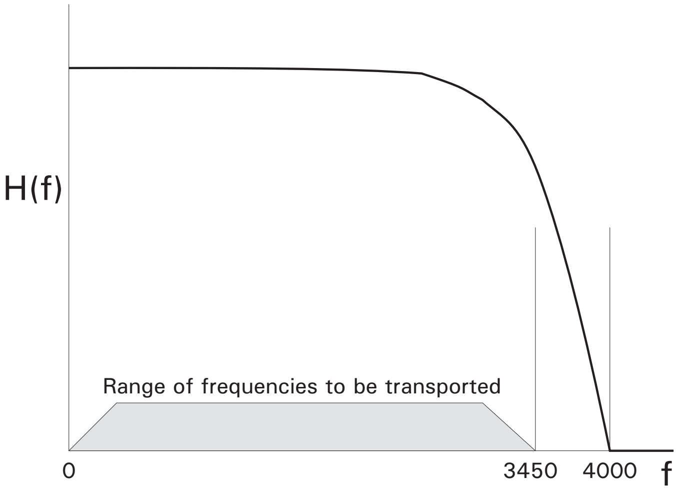

Figure 3. Frequency response of practical anti-aliasing low-pass filter

6 THE OPTICAL ANTI-ALIASING FILTER

We also earlier introduced the concept of using an “optical (spatial) low-pass filter” to strip from the pattern of variation of illuminance across the image all components whose frequencies were at or above the Nyquist frequency for the sampling situation we have.

What might such a filter be like (functionally)? Well, ideally its MTF would drop rapidly as the spatial frequency approached the Nyquist frequency. Suppose we somehow had been able to make such a thing. We might examine it in the optical laboratory by determining its point spread function.

By using inverse Fourier analysis, we can determine what that point spread function would need to be for the filter to exhibit that desired MTF response. And we find that it would be physically impossible. In particular, as we moved along the radius, the response would have to in some places be positive and in some places negative (just as we saw with the impulse response of the electrical low-pass filter).

But her this is a debilitating problem—we cannot have a negative Illuminance.

So in fact, we know already from a theoretical basis that we cannot make an optical low-pass filter with the spatial frequency response we think we would like to have.

So we have to be prepared to get some compromise spatial frequency response. And in fact, getting the some particular spatial frequency response (even one that is not theoretically unattainable) is no easy task.

So in actual digital camera design practice, at one time wide use was made of a certain type of filter that was (fairly) easy to make.

We will examine it soon.

6.1 Two-dimensionality considerations

Before we proceed with this matter, let me call attention again to a complication in the matter of anti-aliasing in a digital camera situation: the fact that the variation of illuminance we wish to "capture" by sampling is two-dimensional.

Again, lets assume that our sensor has its photodetectors spaced every 0.01 mm in a vertical/horizontal pattern. (The orientation of the pattern will now become critical.[10])

We said earlier that the Nyquist frequency (and we now have to say, with respect to frequency components of the variation of illuminance along a vertical or horizontal "track") is 50 cy/mm.

Suppose we consider variation of illuminance along a track at an angle of 45° to the horizontal. Along such a track, the spacing of the photodetectors is about 0.0141 mm. Accordingly, if we consider sampling of the illuminance pattern along a track oriented at 45° to the horizontal, we would have to consider the Nyquist frequency to be about 35.5 cy/mm. That is more stringent that we at first saw (where the Nyquist frequency seemed to be 50 cy/mm).

Wow! Does that mean that we must in fact strip from the overall pattern of illuminance all components whose spatial frequency is at or above 35.5 cy/mm? Or must we at least strip from the pattern of

[10] In some Fujifilm digital cameras, the grid of photodetectors is organized along 45° diagonal lines.

illuminance variation all components in a 45° diagonal direction whose frequencies are at or about 35.5 cy/mm?

Fortunately, neither of those. If in fact we limit the frequencies contained in the illuminance variation as seen along a horizontal or vertical track to be less than 50 cy/mm, then we are good to go. The image reconstructed from the collection of the sample values will not be "corrupted" by aliasing. It may however not be identical to the image delivered by the anti-aliasing filter onto the sensor array (which might have "forbidden" components in a diagonal direction.

6.2 The "four-spot" optical anti-aliasing filter

A bit ago, I pointed out that we could not, due to basic theoretical considerations, build an optical low-pass filter whose spatial frequency response (MTF) fell to zero "instantly" at a certain frequency (such as just below the Nyquist frequency for our sampling "setup"). And I also mentioned that if we aspired to have a certain spatial frequency response, one not precluded by those theoretical considerations, it is still very hard to do that.

In the wake of this, early workers in the field of anti-aliasing for digital cameras found a kind of filter that was (relatively) easy to make and whose spatial frequency response was "useful" in the war against aliasing. It is called the "birefringent four-spot filter".

In the original execution of this, we start with a plate made of a certain kind of quartz whose index of refraction is not the same in both directions (for example, vertical vs. horizontal, if we are thinking of a ray of light proceeding "away from us"). It is said to be "birefringent" ("bi-refracting"), thus that part of the name of the filter.

Placed in the proper context, such a plate, when receiving a cone of rays (from a "quasi-point source) that should form a tiny spot on the sensor, will divide the light into two emerging cones of rays, which would form two tiny spots on the sensor. (Oh, great, a double image!) How far apart they are depends on the thickness of the plate (that is, through what length of the material the rays are allowed to pass).

Now we place immediately behind this plate another one just like it, oriented at 90° to it. Each of the two cones of light emerging from the first plate is now divided and exits the second plate as two cones, which would form two tiny spots on the sensor. But because of the orientation change, the total output of the second plate is four cones, which will form four tiny spots on the sensor.

Think in terms of such a filter where the spot spacing is identical to the sample pitch (that is, the photodetector pitch), and the output spots are of infinitesimal diameter. If we now consider the MTF plot (spatial frequency response) of this filter (any optical block has one), and of course we now need to think of "with respect to the frequencies contained in a variation of illuminance along a track in one direction or another", we find that the MTF (with respect to either the vertical or horizontal direction) is approximately as shown in Figure 4.

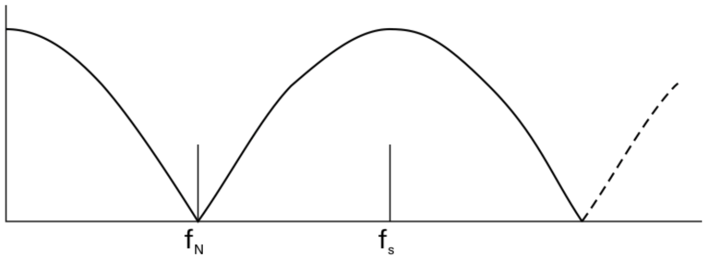

Figure 4. MTF of "four spot" filter

The frequency f_N is the Nyquist frequency of this system. The frequency f_s is 1/s, where s is the photodetector spacing (pitch).

In theory, the pattern we see would repeat to infinite frequency, but I have only shown it though a little past its "second null", the second frequency at which the response becomes zero.

Well, great! As a "low-pass filter" that is rather a flop. But it does have some appealing properties. Its response does in fact drop toward zero as we approach the Nyquist frequency (f_N), and is quite low immediately beyond it.

But as we continue beyond the Nyquist frequency, the response rises substantially, opening the door to components well above the Nyquist frequency, which of course would cause aliasing.

So why do we use such a filter. Well, because we can make it. How do we survive its fairly important shortcoming? Well, we rely on two other players in the drama to help it out. We'll get to that shortly.

In any case, we now understand why, in describing this kind of filter, we often hear, "its purpose is to make four point images from any point on the scene fall on four adjacent photodetectors". That does happen (if we assume a suitably located scene point), but that is not, *per se*, the objective. It is just a fact. And it makes possible a sometimes-seen "blackboard" demonstration of why aliasing (**for a component at exactly the Nyquist frequency**) is suppressed by the use of this filter. (I will spare the reader this slightly-misleading exercise.)

6.3 The role of the “photodetector MTF”

So far, I have discussed sampling of an image in terms of capturing the image illuminance at a large number of “points”. The implication is that it is in fact the illuminance **at a true point** that is captured; that is, the portion of the image to which an individual photodetector responds is very tiny—theoretically, of infinitesimal size. (It could not truly be of zero size, or the amount of light which the photodetector observed would always be zero.)

But in reality, we are not even close to this, by intent. In order to get the best noise performance of a sensor, it is desirable that each actual photodetector element catch as much light as possible, which means its “intake area” should be as large as possible. In fact, the fraction of the overall sensor area that is devoted to the collective intake areas of the photodetectors is a parameter that is often quoted when improved sensor designs are discussed. In the most recent sensor designs, the intake areas of the photodetectors have almost the same dimensions as the photodetector pitch.

Now for a moment consider a photodetector intake area that is circular (to finesse some complications), with a substantial diameter. Each photodetector thus regards a substantial region of the image (reporting approximately the average illuminance over it.

But a little thought will reveal that this is exactly the same as having photodetectors with “point” intake areas in front of which is an “blur filter”; that is, an optical low-pass filter.

This virtual filter of course has an MTF curve (its equivalent to the frequency response curve of an electrical filter).

Thus, in reality, in front of our sensor, in effect, is a pile of two filters, the actual anti-aliasing filter itself and the virtual filter recognizing the finite intake areas of the photodetectors.

If in fact the intake areas of the photodetectors were circular with a diameter equal to the photodetector pitch (so the circles nestled snugly across the sensor), and if all parts of the circle were equally “sensitive”, then the MTF plot of the virtual filter would be as shown in Figure 5.

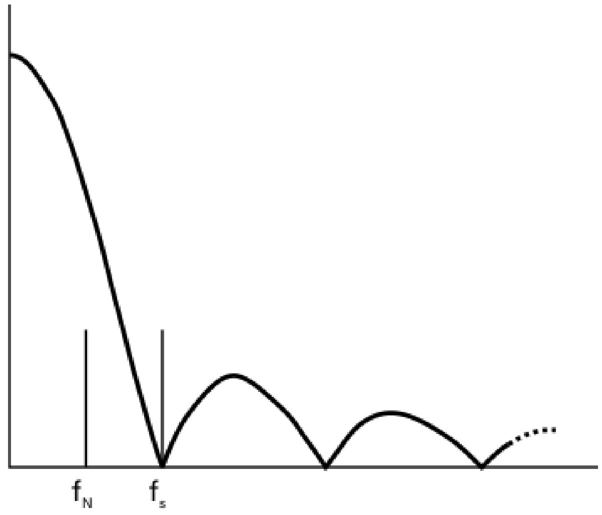

Figure 5. MTF of the photodetector

Again, here, f_s is 1/s, where s is the photodetector spacing (pitch), and f_N is our intended Nyquist frequency.

In theory, this pattern would extend to infinite frequency as a series of little hills of ever-decreasing height. Here, I have only shown the pattern to just past its “third null”.

Note that above the Nyquist frequency, f_N, its response is relatively small. Thus it provides considerable (but not complete) attenuation for components above the Nyquist frequency that are “given a pass” by the response of the four-spot filter.

In Figure 6, we see the combined MTF of our four-spot antialiasing filter and the photodetector itself (the horizontal scale of the photodetector MTF is different than in Figure 5).

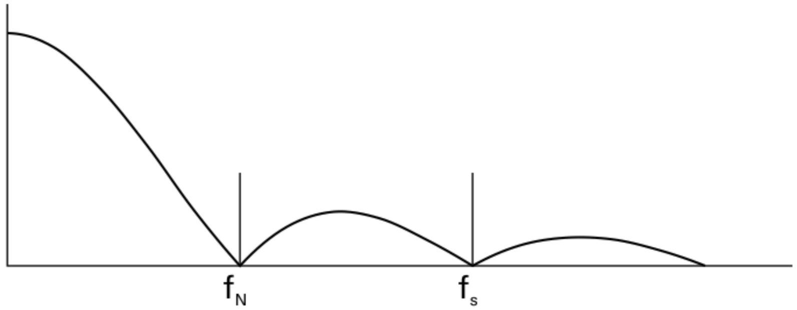

Figure 6. Combined MTF

We see that this much more like the MTF we really would like to have from an antialiasing standpoint.

6.4 The lens

Of course in real camera the lens does not have an MTF that is uniform out to an infinite spatial frequency. Rather, it rolls off, which is the source of the "finite resolution" of a real lens.

Thus the lens also serves to attenuate the higher frequency components in the image, including those that are above the Nyquist frequency. The combination of the MTF's of the three actors here—the lens, the four-spot anti-aliasing filter, and the virtual filter reflecting the structure of the photodetector array—will typically provide "sufficient" attenuation of "troublesome" components in the pattern of luminance variation of the scene (which is after all what we start with) so that aliasing is not "unacceptable".

Suppose we have just spent a zillion bucks on some lens with extremely high resolution? Does that screw up this story? Well, yes. We may find that with this lens in place we see artifacts of aliasing we did not see before.

You can't fool Mother Nature, even with lots of aftermarket money.

7 COLOR TIME

7.1 Entr'acte

So far I have insisted we think in terms of a monochrome camera, to avert contact with some very knotty complications. But in fact, few photographers have ever encountered a real monochrome digital camera.

So now that we have grasped the basic principle involved in the matter of aliasing, we will recognize the color camera.

7.2 Color camera basics

Most color cameras of interest to us use what is called a *tristimulus* sensor approach. Conceptually, at each point in the image we sample, we observe the light there with three photodetectors, which have different *spectral response curves*. This refers to the differing sensitivity they have for different wavelengths of light.[11]

In many cases, the three classes of photodetector have *spectral response curves* identified as "R", "G", and "B". The intent is to cause comfort among people somewhat familiar with the R, G, and B (red,

[11] This is wholly unrelated to the concept of spatial variation in light and spatial frequency.

green, blue) coordinates of an RGB-family color model or color space. But if we look at the actual spectral response curves of these three classes of photodetector in the typical camera, we find it a pretty big stretch to call them “R”, “G”, and “B”.

In any case, from the outputs of the three photodetectors that regard a certain point in the image (don’t worry yet about how can three of them fit in the same place) the system can make a reasonably-reliable estimate of the color of the light there. Not an exact determination? No. The reason is very complicated, and is a good subject for a different article.

Now to the matter of how can three photodetectors fit in the same spot so they can regard the same point in the image.

One way is used in “three-chip” cameras (an arrangement mostly used in video cameras at one time). There, beam splitters generate three copies of the image. Each one falls on a array of photodetectors, one for each sampling point (that is, one for each pixel of the delivered digital image). In one array, all the photodetectors have the “R” spectral response, in one array the “G” spectral response, and in one array the “B” spectral response.

Thus, each spot in the image is regarded by three photodetectors, one with each of the responses.

In digital still cameras using the Foveon-type sensor, at each sampling point (pixel location) there are in effect three photodetectors, one of each of the three classes, stacked one on top of the other. Each is partially transparent, so they all receive the light at that image point.

But in fact most of today’s digital still cameras use neither of these approaches.

7.3 The color filter array (CFA) sensor

Typical digital still cameras (and many video cameras) use what is spoken of as a “color filter array (CFA) sensor.

In a such a sensor we typically have three different classes of photodetector, identified as “R”, “G”, and “B”, having the frequency responses I spoke of earlier. They are planted across the scope of the sensor array in a repetitive “cluster” pattern (vertically and horizontally). Figure 7 shows the most common such pattern, known as the “Bayer” pattern.

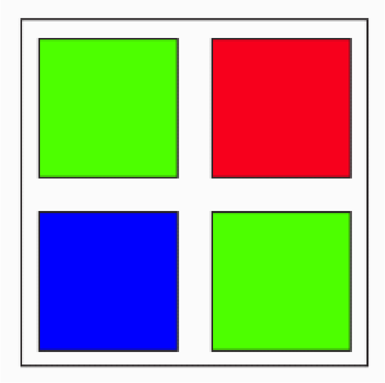

Figure 7. “Bayer” CFA photodetector cluster

We might wonder why each cluster contains two “G” photodetectors. The reason doesn’t affect our work here, but simplistically we can say that the overall sensitivity of the “G” photodetectors is less than that of the other two classes, and that is made up for by having twice as many of them

Now imagine that we consider the image to be made of three “layers”, conceptually containing the aspects of the light to which the three classes of photodetector respond. For a while, I will work from Figure 8.

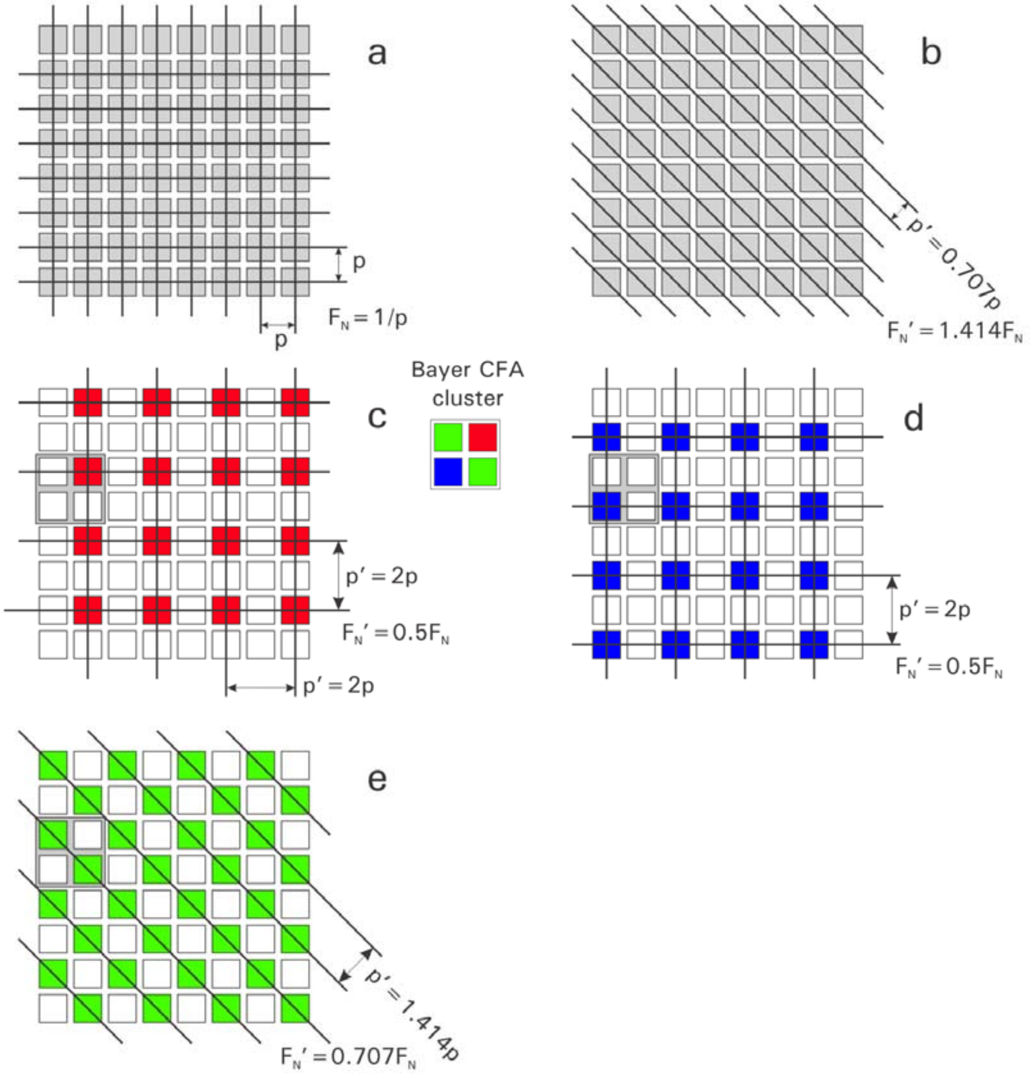

Figure 8. "Bayer" CFA sampling

Sections a and b show how various dimensions of eventual interest to us are defined. The following sections show how the three layers of the image are samples by the photodetectors of the three classes.

Because the Bayer photodetector pattern is cyclic, there is no unique way that we can consider four adjacent photodetectors to be one Bayer "cluster". That notwithstanding, on each of the later sections of this figure I have drawn a square around one arbitrary (but consistent) set of four adjacent photodetectors, which is certainly an illustrative Bayer "cluster".

The "R" layer is sampled, at regular spatial intervals, by the collection of "R" photodetectors (section c); the "B" layer is sampled, at regular spatial intervals, by the collection of "B" photodetectors (section d); and the "G" layer is sampled, at regular spatial intervals, by the collection of "G" photodetectors (section e).

If, for each of the layers, there are no components in its illuminance variation whose spatial frequency is at or above the Nyquist frequency for its set of photodetectors (determined by the spacing of the corresponding photodetectors), then the set of photodetector "readings" for that layer completely describes the entire illuminance variation for that layer.

Thus, the collection of all the photodetector "readings" completely described all three layers of the image. And from that suite of data, all three layers could be completely reconstructed at some distant point (or later time). And those three layers would (to the extent that this tristimulus scheme can represent color accurately) allow reconstruction of the **entire** image on a "color" basis.

Note that this does not mean the color at each photodetector location would be reconstructed, and in between we would have some kind of interpolated values. Shannon and Nyquist tell us that the entire image would be reconstructed.

But only if in the image we start with, in each of the three virtual layers of the image there are no components of illuminance variation whose frequencies are at or above the applicable Nyquist frequency.

Now lets consider the Nyquist frequency.

First recall that, given that the image is two-dimensional, we have to consider illuminance variations along two directions, normally horizontal and vertical. If we look at the pattern of "R" sampling points (section c of the figure), we can see that the spacing between sampling points—vertically and horizontally—is 2d (where d is the spacing between photodetectors), considering the whole collection, of all three classes).

Thus the sampling frequency is 1/(2d), and the Nyquist frequency (always half that) is 1/(4d).

It works that way for the "B" sampling points as well (section d).

Now for the "G" sampling points, things are a little different (section e). The easiest way to get at it is to say, "well for the 'G' world, the "tracks' we consider should be in the two **diagonal** directions." (Remember, the physics doesn't know what is vertical, what is horizontal, and what is diagonal.)

Now, the spacing between sampling points is about 1.41d; the sampling frequency is 1/(1.41d); and the Nyquist frequency (always half that) is about 1/(2.82d).

Does that mean we can allow greater frequencies in the “G” layer of our image than for the “R” and “B” layers? Well, conceptually, yes. But of course the three layers don’t really exist separately, so we might not really be able to take advantage of that notion. More on that later—we mustn’t brood about it now, or we will lose the momentum of the story.

Now suppose that we in fact arrange for an optical anti-aliasing filter that is actually a low-pass filter with a cutoff a little below the Nyquist frequency (for R and B): 1/4d. Then we should have aliasing-free operation.

But the resolution of the system is such that for, say, a 3000 × 2000 photodetector sensor, it would seem that we could really only deliver a 1500 px × 1000 px image.

But we all know how that’s not how it works in today’s cameras: a CFA camera with a 3000 × 2000 photodetector sensor can indeed delver a 3000 px × 2000 px image. How can that be?

The answer is an amazing process called *demosaicing*. The entire theory is very complex, and the details of actual practice relatively unknown, so I’ll describe it in a conceptual way.

If we want our sampled R layer to really participate in the sampled description of an image that is truly (for the example above) 3000 px × 2000 px, it must contain 3000 × 2000 pixel R values. But there are only 1500 ×1000 R photodetectors.

So the demosaicing algorithm puts in the missing samples, using a scheme of “smart interpolation”. Just simple interpolation between the adjacent R photodetector values won’t do it; the result would be able to become part of a 3000 px × 2000 px image, all right, but it wouldn’t have any more detail in it than the 1500 px ×1000 px image we thought of earlier.

The smart interpolation considers all the photodetector outputs from points of all flavors—R, G, and B—all around the “missing” point—maybe for quite a distance. Based on the realities of how the color in images vary, the result can be a set of quite good “estimated” values for all the “missing” samples of the R, G, and B layers.

In effect, we have for each layer samples (some direct, some concocted) at a regular interval (horizontal and vertical) of d. Now the Nyquist frequency is 1/2d. So we need to provide filtering to suppress components in the image at or above that. And we usually do that in the same way I most recently described for monochrome cameras, with a troika comprising:

- A birefringent four-spot filter, with an output point spacing of d (or thereabouts).
- The inherent MTF of the photodetectors, which these days have relatively large intake areas (compared to their pitch).
- (The dirty little secret) The MTF of whatever lens we find aboard.

7.4 With a Foveon-style filter

Foveon-style color camera sensors have a sensor with three layers, each having only one of the three primary color frequency responses at each pixel location.. So at each pixel location, we have a true "three-color" photodetector. Thus there is not any demosaicing required to "guess" as to the actual color at each pixel location.

Some colleagues have noted that in some or most cameras using a Foveon-style sensor,[12] there is no (overt) anti-aliasing filter. The accompanying comment is usually that "and no aliasing artifacts [notably, moiré patterns] are present".

From this, some conclude that the only (conceptual) need for an anti-aliasing filter is to prevent color moiré artifacts in the case of a CFA sensor.

Hopefully, the reader can see from the preceding discussions that aliasing is not only a creature of, and a problem with, the CFA sensor approach.

In any case, some workers have reported that indeed there are perceptible aliasing artifacts (often manifesting as color moiré patterns) in cameras with Foveon-style sensors.

7.5 In actual monochrome cameras

As further evidence of the claim that aliasing is only a creature of the CFA construct, some commenters note that in monochrome cameras (perhaps in particular video monochrome cameras, perhaps used for surveillance), there is often no (overt) anti-aliasing filter. What about that?

I'm not really familiar with these cameras. My guess is that they rely on both the photodetector MTF and the MTF of their (typically rather primitive) lenses to limit aliasing to a negligible degree.

[12] Recall that this information is basically from 2011.

7.6 Another explanation of the four-spot filter

A popular exploration of the four-spot filter in a CFA context is that it takes each point of light in the basic image and sends it to four adjacent photodetectors. Thus, in a sense, each spot is examined by R. G. and B photodetectors (twice for G, in fact). So it might seem that now we had a true tristimulus sensor (as in a three-chip camera or one with a Foveon sensor).

Now, goes the story, for the color of each pixel of our delivered image, we need only consider the outputs of the R, G, and B photodetectors “fed” by that pixel’s location on the image.

The fly in that ointment is that a certain “R” photodetector is “fed by” four pixel locations. So its output cannot be treated as the “R” coordinate for any pixel.

So the fact that the image light from each pixel location is split into four “spots”, each landing on a different photodetector, one “R”, two “G”, and one “B”, is just a result of how the four-spot filter gets the MTF seen in Figure 4.

8 A CLOSING REMINDER

Just a reminder that, as I advised in the Preface and Caveat, the information on implementation of anti-aliasing filters is per my best knowledge as of 2011 (when the manuscript for this article inadvertently went into a “deep sleep”). There have been substantial advances in that technology since, and as well advances in other ways to mitigate the aliasing phenomenon.

However, the descriptions of the theoretical basis of aliasing, and the conceptual way in which aliasing can me mitigated by a low-pass filter, remain as valid and pertinent as ever.

9 ACKNOWLEDGEMENT

Great thanks to my wife, Carla Kerr, for her insightful editing of an earlier draft of this difficult manuscript. But any residual editorial errors, and any technical errors, are solely my responsibility.

-#-

APPENDIX A

Aliasing in action

In this appendix, we will see a "blackboard" illustration of the phenomenon of aliasing. The context is a system in which an electrical waveform is to be captured by sampling and then reconstructed from the train of samples (as in a basic digital telephone transmission system—the context in which I first learned this theory, so I often resort to it in explaining the theory). The action is seen on Figure 9. I'll discuss its features as I go.

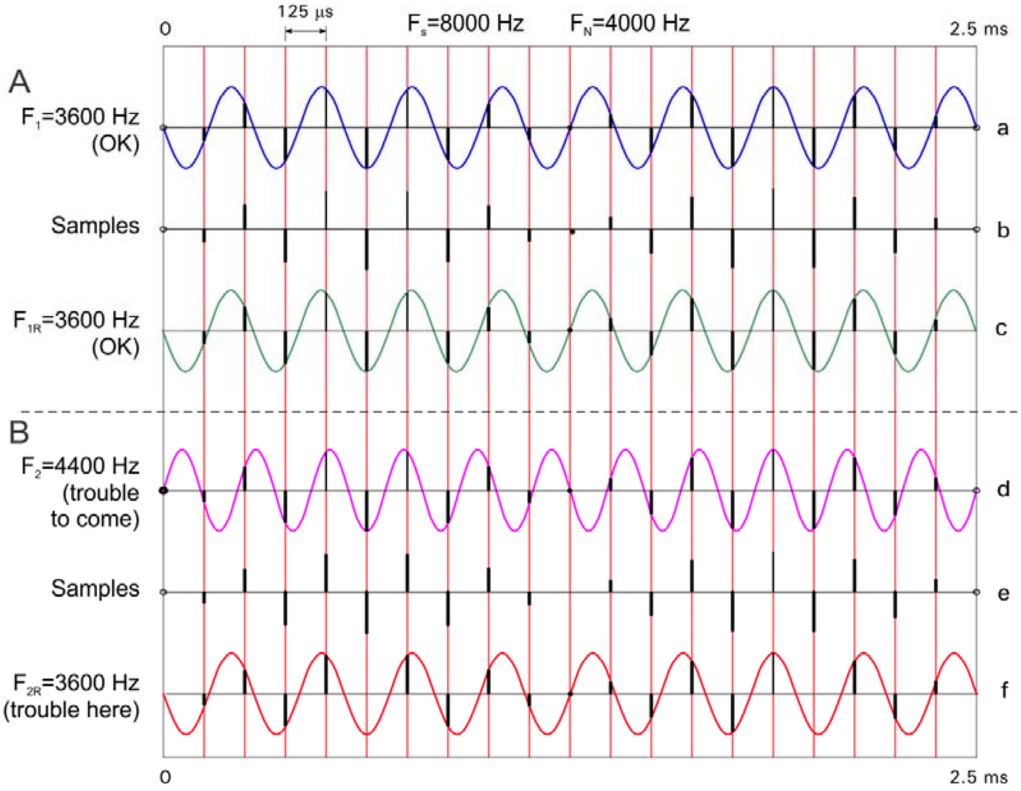

Figure 9. Aliasing in action

We assume that our need here is to transport, by sampling, audio signals with frequencies up to 3700 Hz, which would require a sampling rate of over 7400 Hz. To give ourselves a little leeway, we choose a sampling frequency (sampling rate) of 8000 Hz. According to Messrs. Shannon and Nyquist, we should be able to transport a signal all of whose frequency components have frequencies less than 4000 Hz, the Nyquist frequency.

We see a total period of 2.5 ms. The red lines on the figure represent the sampling instants, 125 μs apart (a sampling rate of 8000 Hz).

In section A, on line **a** we have applied a test signal, a single-frequency (sine wave) waveform with frequency 3600 Hz (within the Nyquist frequency limit).

The little vertical bars at the sampling instants point out the instantaneous voltage of the wave at those instants. These of course become the samples. Where the instantaneous voltage (and thus the sample value) is zero, I use a little dot rather than the bar.

We see the samples isolated on line **b** as if they were voltage pulses (which in fact they actually are immediately after the sampling).

We imagine the value of each of these pulses being put into digital form, transmitted to the "receiving" end, and used to reconstitute that pulse there (I do not show this).

On line **c**, we see the waveform the "reconstruction organ" would reconstruct from that train of (reconstituted) sample values. It is of course identical to the waveform of line a. Thus our original signal has been correctly (in fact, "precisely") reconstructed from its digital representation.[13]

In section B, on line **d**, we apply another test signal, this time a sine wave with frequency 4400 Hz. This is of course above the Nyquist frequency limit, and we should not expect it to play nicely. Again, we see the samples as they are taken, and then isolated (as if "pulses") on line **e**.

Son of a gun! This is the identical sequence of sample values we saw on line **b**.

On line **f**, we see what the reconstruction organ would make of that series of sample values. Not surprisingly, it makes a waveform just like the last time we fed it that same sequence of sample values, which is the same as the one we see on line **c**.

Thus what goes in as a 4400 Hz sine wave comes out as a 3600 Hz sine wave, a manifestation of *aliasing*.

Of course if this 4400 Hz waveform (which never should have been let into the sampler) was a component of a more complex original waveform, the overall waveform coming out of the reconstruction

[13] This presumes that the reconstituted sample pulses have "exactly" the same values as the original sample pulses, not exactly true in reality, owing to the discrete nature of their digital representation. But this is a separate matter, not a weakness in the concept of reconstruction from samples.

organ would be wholly different than the original one, an example of *aliasing distortion*.

Now, the cynical reader might say:

"But Kerr, you arranged the phase of the 4400 Hz waveform on line **d** so that the samples taken from it would be the same as the samples taken from the 3600 Hz waveform on line a."

Well, yes I did.

"So is it possible that aliasing only can be shown in such a contrived situation?"

No. Imagine that I did the first stage again with a 3600 Hz waveform of a different phase. Then the sequence of sample values on line **b** would be different. Now from that sequence the reconstruction organ would again reconstruct a 3600 Hz waveform, of course with a different phase (since the original waveform had a different phase—the whole process properly conserves the phase of the source waveform).

But I could then test with a 4400 Hz waveform, at a different phase as needed, and we would get that same sequence of sample values from it. And of course, the reconstruction organ would from it make a 3600 Hz waveform (at the same phase as this run's 3600 Hz source waveform).

Said in the other direction, regardless of the phase of the "unqualified" 4400 Hz waveform, the sequence of samples derived from it could also be derived by sampling a 3600 Hz waveform at some phase, and that is what the reconstruction organ would make from that sequence of sample values.

How does the reconstruction organ know that the series of samples seen on line **b** (and again on line **e**) represent a 3600 Hz waveform and not (maybe in both cases) a 4400 Hz waveform?

Well, the actual technical details are beyond this article, but we can say that the reconstruction organ was designed with a sampling rate of 8000 Hz in mind, and accordingly it is compelled to only generate frequency components in its output with frequencies less than 4000 Hz, the Nyquist frequency.

How is that constraint applied? In fact, in practice, the almost last thing in the reconstruction organ is a low-pass filter with a cutoff of a little less than 4000 Hz. And that is not just there for "enforcement" or "safety" purposes. It is what actually does the reconstruction! Now, just how it does that is the subject of a different article.