Non-zero Sensor Aperture Effects in Signal Acquisition and Processing

Julius Foit, Jan Novák
Department of Microelectronics, Czech Technical University, Prague

Summary: Most natural signals (sensor outputs, for instance) have analog character. Quite frequently, sensor operation is based on some sort of finite non-zero aperture size, for instance in image data acquisition or magnetic data storage/retrieval. The non-zero aperture width has a decisive influence on the highest signal spectrum frequency component, or minimum size of original physical quantity detail that the system can process. Decreasing the aperture size improves the smallest recognizable detail, at the price of sensor sensitivity loss, however; even the sensor aperture reduction itself can present technological difficulties or hit the limits of fundamental physical laws (Rayleigh's principle etc.). A prominent part of the design problem is the task to obtain flat overall frequency response of the system, including the response to spatial frequency of non-electrical signals, while preserving the original complex signal waveform, i.e. a flat group-delay response. Current designs mostly fail to recognize the nature of the problem and its implications. This paper tries to make system designers aware of some simple ways how to deal with it.

1. Problem statement

When processing an analog signal, it happens now and then that it has to pass a non-minimum-phase system. In non-minimum phase systems there does not exist a mutually unambiguous analytical relation between the amplitude and phase frequency responses. The most frequently appearing case of a non-minimum phase system effect is the "slot function" distortion, appearing, for instance, in image ("pixel") acquisition or magnetic data storage/retrieval. The amplitude response of the slot function to spatial frequency is:

$$y = \frac{\sin x}{x}, \qquad (1)$$

where

$$x = \pi \frac{d}{\lambda}, \qquad (2)$$

where

λ is the signal spatial wavelength on the processing medium (image sensor pixel width or magnetic medium recorded signal spatial wavelength, for instance),

d is the processing slot width (image sensor pixel width, for instance).

Equation (2) assumes an ideal non-electrical signal transfer to the sensor (processing medium). Every real non-electrical signal transfer process has, however, its own critical spatial wavelength ψ (optical imaging system resolution, for instance) whose effect is almost identical to that of the slot function and is superimposed over it. As long as we can design the system in such a way that

$$\psi << d, \qquad (3)$$

we can neglect the critical wavelength of the non-electrical transfer system and limit the analysis to the effect of slot width *d* only. Both effects are usually present at the same time in real situations, superimposing over each other. However, as long as the overall system amplitude response is linear, both effects can be treated (and equalized) separately. Therefore, in the following we are going to discuss the situation for just one slot function as the only one present, assuming, for simplicity, that the condition stated in Equation (3) is satisfied.

2. Theory

A salient property of the slot function distortion is the fundamental difference from the frequency responses of minimum-phase systems, where the form of the phase vs. frequency response can always be unambiguously calculated from the amplitude vs. frequency response and vice versa. The situation is clearly shown by comparing Figs. 1 and 2.

Fig.1 shows the behavior of the slot-function characteristics; we can see that the slot-function phase response has a discontinuous character, with phase value φ jumping discontinuously between 0 and -π (or zero and -180 degrees, if you like), at the points where the signal spatial wavelength $n\lambda = d$ (for $n = 1, 2, 3, ...$ *etc.*). At these points, the amplitude response (signal transfer) is zero. The amplitude response *y* in Fig. 2 is shown in absolute values; sometimes it is drawn in true values as a periodical function with negative parts, corresponding to opposite phase in the intervals of $1 < (d/\lambda) < 2$, $3 < (d/\lambda) < 4$, ... etc. In practical applications, we usually limit the system operation range below the first zero crossing, i.e. $0 < (d/\lambda) < 1$, most frequently to the ratio value $(d/\lambda) \leq 0.95$. At $(d/\lambda) = 0.95$, the amplitude response $|y|$ of the slot function is more than 25 dB below its maximum (100%) value.

Fig. 2 shows the classical RC high-pass minimum-phase network behavior, featuring a 3 dB drop at the critical frequency $f_{cr} = 1/(2\pi RC)$, and a phase shift of $\pi/4$ at the same frequency, with a slope of the amplitude response equal to 20 dB/decade below it.

Drops in frequency response due to slot function effects as large as the one described above appear especially in more demanding image processing situations, as well as in high-density (or low speed, for that matter) magnetic recording, and they call for major high-frequency signal amplitude boost. In practically all common systems, this high-frequency boost is obtained by applying simple and straightforward minimum-phase networks of the RC or RLC type. The reason is obvious: they are inexpensive and easy to adjust. However, using this method for amplitude response correction leads to

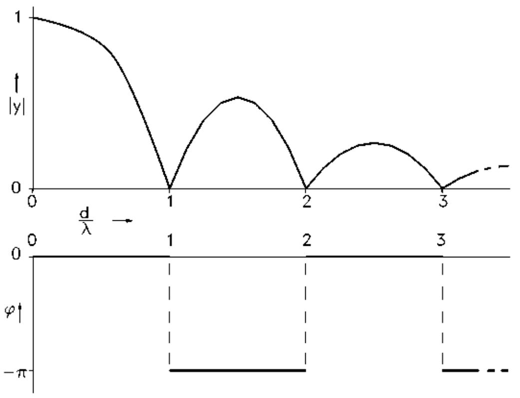

Fig. 1 – slot function response

errors in phase response at high frequencies, as can be easily understood by comparing Figs. 1 and 2. The effects of slot function compensation by minimum-phase circuits (errors in group delay characteristics) prevent making full use of the possibilities in sensor data acquisition for short signal spatial wavelengths.

In the particular case of analog signal magnetic storage the designers frequently try to adjust the system for an overall flat amplitude frequency response and do not care for phase errors. This, for instance, was the case of analog sound recording for many decades; the idea was based on the (wrong) assumption that in human hearing correct phase reproduction is of minor, if any, importance. Unfortunately, this is not exactly true: the human ear tolerates phase errors if they are not too large and the listening situation is limited to the ***A.C. periodical steady state only***. Once we start to deal with broad-spectrum transients, we are in deep trouble. Phase errors result in ***waveform distortion*** of such transients, regardless of perfect amplitude linearity of the whole system, no matter how well equalized is the amplitude frequency response.

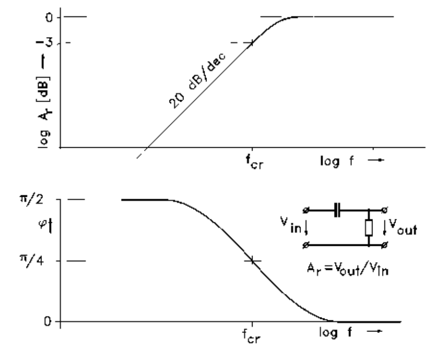

Fig.2 – responses of a minimum-phase network

This is the reason why many cheap video systems feature extremely poor resolution, much worse than could be expected from their sensor pixel count, and why acute-eared professional musicians despise low tape-speed recorders; usually they are unable to formulate their objections in clear expressions, nevertheless a closer examination always proves that the culprit is poor transient response of the system. A clear demonstration of the problem is the behavior of common analog tape recorders (including top-grade professional machines!) when trying to record and play back a simple rectangular waveform signal.

It is important to note that the popular belief that digital signal processing has done away with this problem is not exactly true since in almost all cases the initial input signal has the form of a weak analog physical quantity whose equivalent electrical signal must be amplified in an analog process prior to digitization. It is this analog pre-processing which can introduce phase errors into the signal ***before digitization***.

All this means that, in order to obtain a correct transient response (with maximum possible transition edge slope as limited by the system overall bandwidth), the system phase response should be flat. The problem can be approached in two ways. The first possible attitude is to use non-minimum-phase networks to correct the amplitude frequency response errors caused by the slot function. However, the necessary circuits are relatively complex and difficult to adjust. The second possible attitude makes use of the fact that since the phase distortion related to the standard R-(L)-C amplitude response correcting networks is a linear process, the system can be phase-equalized "after the fact", i.e. the acquired signal can be equalized later, after the “wrong-phase” frequency response correction.

Therefore the phase equalizing networks can be applied subsequently (but still before digitization), leaving the original signal-processing circuitry untouched.

3. Solution

The basic operating principle of phase-equalizing circuits is quite simple: the circuit must be an "all-pass", i.e. a circuit that has flat amplitude versus frequency response, with a properly frequency-dependent phase shift. Of course the phase shift characteristic should fit, at least approximately, to a form inverse to the phase error present in the signal that is being equalized.

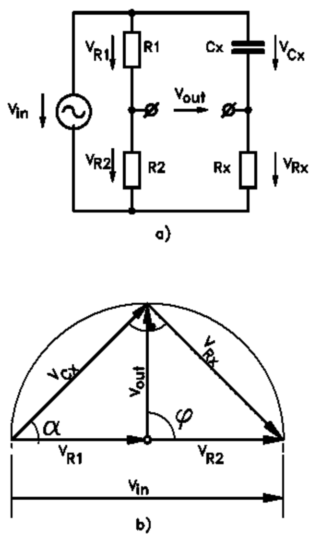

Fig.3 Phase shifter principle ($R_1 = R_2$)

The operating principle of a possible phase equalizer is based on the well known bridge-type all-pass phase shift network. Its principle is shown in Fig. 3a. The amplitude and phase frequency response can be read from the phasor diagram of the circuit shown in Fig. 3b. The output voltage V_{out} phasor has a frequency-independent amplitude and is phase-shifted by 90^0 ($\pi/2$) at the characteristic frequency

$$f_{cr} = \frac{1}{2\pi R_x C_x}, \qquad (4)$$

or, its phase shift versus frequency can be calculated as:

$$\varphi = 2\alpha = 2\tan^{-1}(\omega R_x C_x) \,. \qquad (5)$$

It can be shown [3] that a phase shifter of this type can equalize the phase shift of two identical independently cascaded simple RC networks like that in Fig. 1 in the whole 0 to ∞ frequency range, if its characteristic frequency is the same as that of the RC networks.

In real cases, the high-frequency boost does not start from zero, of course; typically, the high frequency boost seldom exceeds 30 dB. As a result, the phase equalization by a phase shifter according to Fig. 3 is not perfect in the whole frequency range; on the other hand, the boost frequency range, as well as the system overall frequency range are limited, permitting to adjust the phase equalization to fit the required form quite accurately.

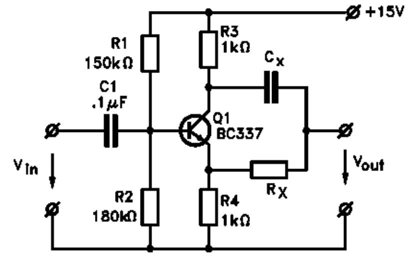

Fig. 4 Simple phase equalizer

An example of a simple practical phase shifter is shown in Fig. 4. The input signal is split in two halves by means of a split-load transistor amplifier. This circuit works quite well at low frequencies, yet it does have some drawbacks. First, it introduces a minor voltage transfer loss. Second, its output differential resistances in the points supplying input voltages to the RC phase shifting network are larger than negligible; moreover they do not have the same value.

Better performance can be obtained from a circuit connection shown in Fig. 5. Here, the input voltage split is obtained by using two operational amplifiers, one as a unity-gain inverter (OA1), the other as a voltage follower (OA2). OA3 serves as a buffer to separate the R_xC_x phase-shift network from external load. For flawless operation in the video signal frequency range, the operational amplifiers should have high input resistance and sufficient slew rate. A good choice can be fast JFET-input types.

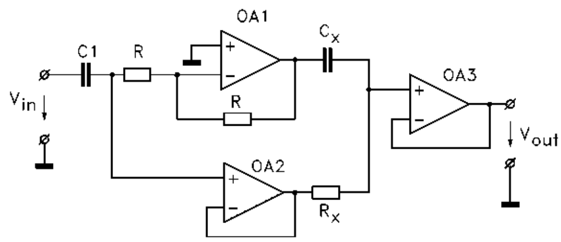

Fig. 5 Phase euqalizer with op. amps.

4. Practical results

To demonstrate the effects of proper phase response equalization we have selected an easily repeatable and very instructive example: the response of a slow-speed analog magnetic tape recorder. As a rule, these machines suffer heavily at the high end of their frequency response from slot function effects both due to a not negligible slot width of their record/playback heads, as well as due to operation near the critical wavelength of the recording medium. Since practically all these machines use frequency response equalizing networks of the minimum-phase type, their phase

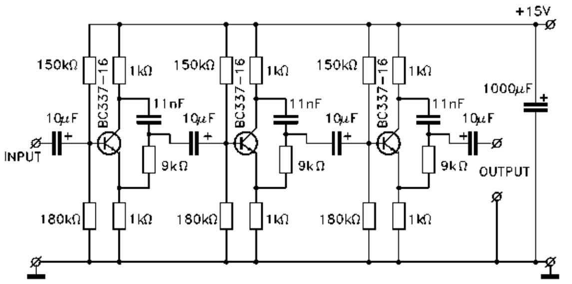

Fig 6. Phase equalizer network used for signals shown in Fig. 7

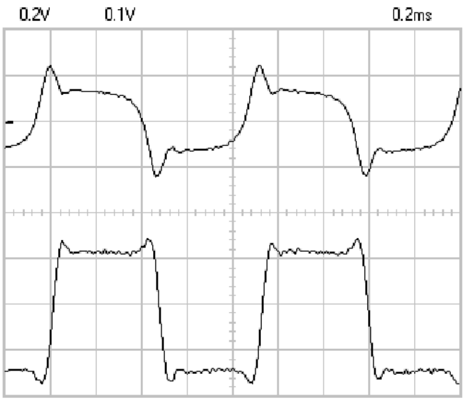

Fig. 7 Slow-speed playback signal before and after phase equalization

response is awful, resulting in highly objectionable waveform distortion of signal transient reproduction. A single phase equalizing stage can prove to be insufficient in similar cases; a single equalizer can have insufficient phase characteristic slope near the critical frequency. In such cases, several identical (or staggered) equalizer stages can be applied in cascade.

A practical case of phase equalization result is illustrated in Fig. 7. The top trace shows a typical playback output of a medium quality cassette tape recorder (Philips EL3302 cassette recorder, tape type TDK D-90, with h.f. recording bias optimized for a ±1.5 dB flat overall small-signal frequency response from 60 Hz to 11 kHz), of a 1.1 kHz 50% duty cycle rectangular waveform input signal (50 ns edge duration, recording level approximately 10 dB below maximum).

The bottom trace shows the same signal after being passed through a phase equalizer consisting of three cascaded identical equalizer stages (Fig. 6) using $R_x = 9\ k\Omega$, $C_x = 11{,}000\ pF$. Notice the slight "undershoot" preceding the equalized response transition edges. This is an inevitable result of bandwidth limitation in a system processing a waveform containing discontinuities. Those interested in the exact mathematics of this phenomenon should look for the "Gibbs effect" in books dealing with details of the Fourier transform theory [1].

It is important to note that the equalization was

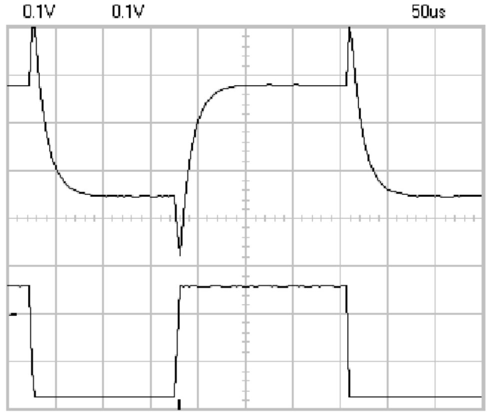

Fig. 8 Single-stage phase equalizer response

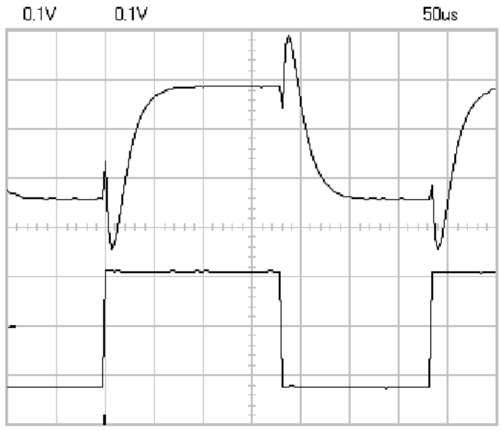

Fig. 9 Two-stage phase equalizer response

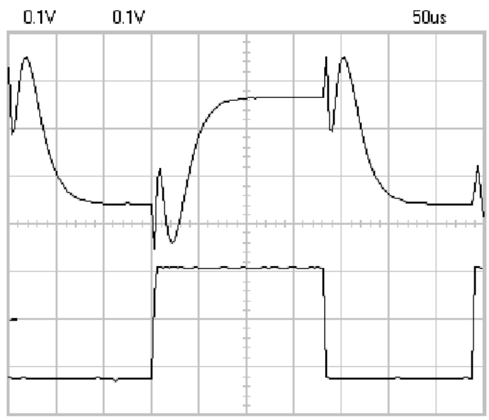

Fig. 10 Three-stage phase equalizer response

performed on the output signal of the machine externally, i.e. without any intervention in its internal circuitry.

To demonstrate the properties of the phasing network itself, see Figs. 8 to 10. They show the phasing network response alone to a rectangular waveform for one, two and three cascaded identical stages, respectively (see circuit in Fig.6, phasing resistances 1 kΩ, phasing capacitances 11,000 pF). Another important point regards noise: since the phase equalization is performed on a relatively high-level signal, and it does not change its amplitude frequency response, it does not appreciably deteriorate the noise figure of the whole system. These facts suggest an extremely interesting possibility for a marked improvement in reproduction quality of existing irreplaceable old recordings (for instance, during "re-mastering" for preparation of modern digital versions of old analog recordings). All previous considerations also apply to image processing and all similar data acquisition processes. Their performance capabilities can only be fully exploited in systems featuring correct phase characteristics.

References:

[1] Angot, A.: *Applied Mathematics for Electrical Engineers*. (pp. 97 – 99, Czech translation of the French original), SNTL, Prague, 1969

[2] Langer, O., **Foit, J**., Polej, B.: *Direkte auswertung von Spektren mittels modifizierten Fernsehanlagen*. Proceedings of XIV. Colloquium Spectroscopicum Internationale, pp. 799 – 807, Debrecen (Hungary) 1967 (in German)

[3] Foit. J.:*Step-response of tape recorders.* Acta Polytechnica III, ČVUT – SPN, Praha 1973 (in Czech)

[4] Foit, J.: *Calculation of amplitude and phase response from the step response by means of TI 58/59*. Acta Polytechnica III, 9, č. 7, SNTL – FEL, Praha 1984 (in Czech)

[5] Foit,J.: *Enhancement of recorded data retrieval* (Classifed Consultancy Report. Praha 2005., in Czech).