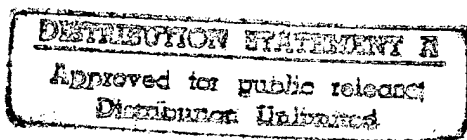


Serial Number 854,511
Filing Date 12 May 1997
Inventor Godi Fischer
 Alan J. Davis

NOTICE

The above identified patent application is available for licensing. Requests for information should be addressed to:

OFFICE OF NAVAL RESEARCH
DEPARTMENT OF THE NAVY
CODE OCCC
ARLINGTON VA 22217-5660



19970825 046

DTIC QUALITY INSPECTED 2

2
3 SIGMA-DELTA MODULATOR FOR WIDE BANDWIDTH APPLICATIONS

4
5 STATEMENT OF GOVERNMENT INTEREST

6 The invention described herein may be manufactured and used
7 by or for the Government of the United States of America for
8 governmental purposes without the payment of any royalties
9 thereon or therefore.

10
11 BACKGROUND OF THE INVENTION

12 (1) Field of the Invention

13 The present invention relates generally to an apparatus for
14 converting analog-to-digital or digital-to-analog signals, and
15 more particularly to a robust sigma-delta modulator for
16 implementing a high-precision analog-to-digital or digital-to-
17 analog converter.

18 (2) Description of the Prior Art

19 Sigma-delta modulators have become a popular means to
20 implement high-precision analog-to-digital converters (ADC's) or
21 digital-to-analog converters (DAC's) in that the technique
22 enables the realization of high-resolution data converters while
23 requiring only low to moderate precision analog components. FIG.
24 1 shows a detailed block diagram of a conventional sigma-delta

1 modulator 10. Modulator 10 consists of a summing junction 12, a
2 filter circuit 14, a quantizer 16 and a DAC 18 within a feedback
3 circuit. Sampling switch 20 provides analog input $X(z)$ to
4 summing junction 12 at the frequency of sampling clock 22. The
5 output 24 of summing junction 12 is provided to filter circuit 14
6 with the output 26 of filter circuit 14 provided to quantizer 16.
7 Quantizer 16 is also controlled by sampling clock 22 to convert
8 the analog signal produced by filter circuit 14 to an output
9 pulse density modulated (PDM) digital signal $Y(z)$. The output
10 $Y(z)$ is also fed back to DAC 18, which reconverts the digital
11 signal to an analog signal 28. Signal 28 is then applied to the
12 negative input of summing junction 12, such that, unless the
13 output $Y(z)$ is exactly the same as input $X(z)$, an error signal
14 will be developed by summing junction 12 which will then pass
15 through the loop to correct the output. This feedback loop
16 filters the difference between the previous output and the
17 current input sample. By means of algebraic decomposition and
18 linearity assumptions of quantizer 16, input signal $X(z)$ is
19 passed through a low pass filter that is wider than the desired
20 band, while the quantization error, i.e., the difference between
21 the value input to quantizer 16 and the value output from
22 quantizer 16, is filtered by a high pass filter with good stop
23 band suppression. Hence the signal is passed unsuppressed while
24 the quantization noise is attenuated in the band of interest.

1 Consequently, these devices have a relatively low conversion rate
2 when compared to other ADC's and DAC's due to the necessary high
3 OSR. Apart from limiting the frequency range, high OSR's
4 negatively impact power dissipation and the settling requirements
5 for the amplifiers employed in the discrete-time analog
6 integrators. It is therefor of great practical interest to find
7 alternative modulator topologies that can cope with low OSR's
8 while preserving the inherent insensitivity of the converter with
9 regard to its constituent analog components.

10 Three basic ways of reducing the OSR are known in the art.
11 The first method is to replace the typical single-bit quantizer
12 at the modulator output by a multi-bit quantizer as described by
13 Larson, L. E., Cataltepe, T. and Temes, G. C. in "Multi-bit
14 Oversampled Sigma-Delta A/D Converters With Digital Correction",
15 IEE Electronics Letters, vol. 24, pp. 1051-1052, Aug. 1989. This
16 method not only reduces the quantization noise thus increasing
17 the dynamic range of the converter, but also de-correlates the
18 quantization noise spectrum from the input signal. Such a system
19 is less likely to fall into a cyclic behavior which can give rise
20 to spurious tones in the passband. However, the major drawback
21 of this method is the extremely high linearity requirement for
22 the DAC in the feedback path of the modulator multi-loop
23 configuration. It is noted that the required accuracy of this
24 DAC must be greater than or equal to the quantization noise

1 suppression in the modulator pass band. For example, if the
2 modulator accuracy needs to be 16 bits, then the integral
3 linearity of the DAC must be no worse than 96dB or 0.00158%,
4 since DAC errors are added unfiltered to the input signal, as
5 shown in FIG. 1. Integral linearity errors in the DAC will
6 produce unwanted tones, i.e., harmonic distortion, which will
7 limit the system accuracy to that of the total harmonic
8 distortion. Unfortunately, the DAC integral linearity is
9 directly dependent upon the relative error in the ratios of
10 values derived from monolithic passive components. Thus, since
11 this ratio of values is 0.01% at best, the integral linearity
12 error can be kept to no better than 0.01%. making it difficult to
13 obtain greater than 60dB, or 10 bit accuracy due to the harmonic
14 distortion.

15 In a second approach, the single modulator loop is replaced
16 by a multi-loop configuration whereby the additional loop(s)
17 create an estimate of the quantization noise of the previous
18 loop(s). The noise estimate is digitally subtracted from the
19 previous loop output(s). The multi-loop solution whitens the
20 quantization noise and thus prevents the occurrence of spurious
21 tones. In theory, one can achieve arbitrarily good noise shaping
22 by cascading a sufficient number of stages. However, the
23 reduction of the quantization noise by signal subtraction
24 requires well matched capacitor ratios in the analog modulator

1 loops and extremely high op-amp open-loop gain values to minimize
2 integrator leakage. These requirements practically limit the
3 number of cascaded stages to two or three. Even with two or
4 three stages, 15 to 16 bit accuracy is difficult to achieve.

5 In the third approach, the order of the loop filter is
6 increased such that a more stable noise shaping filter is
7 possible. The multiple feedbacks of a higher order system tend
8 to de-correlate the quantization noise from the input signal. In
9 contrast to the multi-bit solutions, high-order single-bit
10 modulators preserve the insensitivity of single-bit low-order
11 circuits with respect to minor variations of the analog component
12 values. The major drawback of this approach is that there is a
13 progressive reduction in quantization noise suppression gains as
14 the OSR is lowered in order to maintain a stable operating point.

15 16 SUMMARY OF THE INVENTION

17 Accordingly, it is an object of the present invention to
18 provide a sigma-delta modulator topology that can cope with low
19 OSR's while preserving the inherent insensitivity of the
20 converter with regard to its constituent analog components.

21 Another object of the present invention is to provide a
22 sigma-delta modulator topology that does not require an extremely
23 high linearity for the DAC in the feedback path of the modulator
24 multi-loop configuration.

1 Still another object of the present invention is to provide
2 a sigma-delta modulator topology where the amplifier gain and
3 capacitor ratio matching requirements are kept reasonably low.

4 Other objects and advantages of the present invention will
5 become more obvious hereinafter in the specification and
6 drawings.

7 In accordance with the present invention, a wide-band sigma-
8 delta modulator is provided with a cascade of two modulators or
9 stages, the first stage being of third or greater order and the
10 second being of second or greater order. Both stages utilize
11 simple ternary quantizers to avoid linearity problems associated
12 with the DAC's in the feedback paths of the modulators and to
13 maintain stable operation. Reducing the quantization noise at
14 its source enables a more efficient noise shaping characteristic
15 by allowing the poles of the modulator loop filter to be placed
16 closer to the Nyquist frequency. Additional zeros are placed in
17 the noise transfer functions of third or higher order stages to
18 maximize the dynamic range. The new topology significantly
19 enhances the dynamic range of the system while hardly affecting
20 the circuit's sensitivity with regard to some non-idealities such
21 as finite amplifier open-loop gains and capacitor ratio
22 mismatches. The excess dynamic range gained by these additional
23 zeros is substantial and independent of OSR, hence, the present
24 invention is particularly useful in wide-band applications where

1 the OSR is inherently low. By employing only two cascaded
2 stages, the amplifier gain and capacitor ratio matching
3 requirements are kept reasonably low.
4

5 BRIEF DESCRIPTION OF THE DRAWINGS

6 A more complete understanding of the invention and many of
7 the attendant advantages thereto will be readily appreciated as
8 the same becomes better understood by reference to the following
9 detailed description when considered in conjunction with the
10 accompanying drawings wherein corresponding reference characters
11 indicate corresponding parts throughout the several views of the
12 drawings and wherein:

13 FIG. 1 is a detailed block diagram of a prior art sigma-
14 delta modulator;

15 FIG. 2 is a block diagram of a preferred embodiment of a
16 sigma-delta modulator of the present invention having a double
17 third-order cascade; and

18 FIG. 3 is a block diagram of another embodiment of a sigma-
19 delta modulator of the present invention having a third-order
20 second-order cascade.
21

22 DESCRIPTION OF THE PREFERRED EMBODIMENT

23 Referring now to FIG. 2, there is shown a block diagram of a
24 preferred embodiment of a sigma-delta modulator 50 of the present

1 invention having a double third-order cascade architecture. FIG.
2 2, as well as FIG. 3 to be described later, is based on the
3 inventors' papers "A Sigma-Delta Modulator Architecture For Wide
4 Bandwidth Applications", Proc. AISCAS-96, May 14 1996 and "A Two-
5 Stage Sixth-Order Sigma-Delta ADC With 16-bit Resolution Designed
6 For An Oversampling Ratio Of 16", Proc. MWSCAS-96, Aug 1996 which
7 are incorporated into this disclosure in their entirety by
8 reference. Dotted lines 52a and 52b generally denote the third
9 order modulators which make up modulator 50. Like components in
10 modulators 52a and 52b are described using identical numbering
11 for identical components such that the following description
12 applies to both modulators 52a and modulator 52b. Third order
13 modulator 52a has characteristics similar to modulator 10 of FIG.
14 1, in that input 54 is provided to summing junction 56 with
15 feedback signal 58 being applied to the negative input of summing
16 junction 56. However, filter circuit 14 of FIG. 1 has been split
17 into three single-order integrators 60, 62 and 64, integrators 60
18 and 64 being delaying integrators of the form $\frac{z^{-1}}{1-z^{-1}}$ and
19 integrator 62 being a non-delaying integrator of the form $\frac{1}{1-z^{-1}}$.
20 The integrators are joined by second and third summing junctions
21 66 and 68. Output 70 of summing junction 56 is provided to
22 integrator 60 which in turn provides output 72 to summing

1 junction 66. Output 74 of summing junction 66 is provided to
2 integrator 62 which in turn provides output 76 to summing
3 junction 68 with output 78 of summing junction 68 being provided
4 to integrator 64. As with the modulator of FIG. 1, the output 80
5 of integrator 64 is provided to ternary quantizer 82 with the
6 output of quantizer 82, designated as Y_1 for modulator 52a and Y_2
7 for modulator 52b, being fed back and applied to the negative
8 inputs of summing junctions 56, 66 and 68. The aspects of
9 modulators 52a and 52b just described are common to third-order
10 modulators well known in the art. Such prior art third-order
11 modulators are not normally designed in a cascade structure due
12 to their marginal stability. In order to achieve a noise shaping
13 characteristic with greater quantization noise suppression,
14 modulators 52a and 52b of the present invention have additional
15 finite zeros placed in the noise transfer function (NTF). The
16 implementation of the two zeros requires little analog circuitry
17 and is accomplished in the third-order modulators 52a and 52b of
18 the present invention by an additional feed back loop. Output 80
19 is fed back and applied, with the scaling factor $\delta_{61}/(bc)$ for
20 modulator 52a and scaling factor $\delta_{62}/(ef)$ for modulator 52b, to
21 the negative input of summing junction 66. This excess dynamic
22 range gained by these additional zeros is very substantial,
23 making this technique particularly useful in wide-band
24 applications where the OSR is inherently low, meaning that the

1 noise shaping filter stop band occupies more relative frequency
 2 from DC to π , i.e., half the sampling rate. In the preferred
 3 embodiment of the present invention, modulators 52a and 52b are
 4 cascaded, in that output 80 of integrator 64 in modulator 52a is
 5 also used as input to third-order modulator 52b, i.e., to summing
 6 junction 56 of modulator 52b. The digital noise cancellation
 7 (DNC) scheme of the present invention, denoted by dotted line 88,
 8 becomes more complex and requires two multiplier blocks, one to
 9 properly scale the first-stage quantization noise, i.e., the
 10 composite factor (abc) and one to implement the modified triple
 11 differentiation, i.e., the factor $(3 - \delta_{61})$, to correctly
 12 accommodate the NTF zero of the first stage. The final digital
 13 output, $Y(z)$, of DNC 88 is computed as follows:

$$14 \quad Y(z) = Y_1 z^{-2} + (1 - z^{-1})(1 - z^{-1}[2 - \delta_{61}] + z^{-2}) \left(\frac{1}{abc} Y_2 - Y_1 z^{-2} \right). \quad (1)$$

15 It is noted that the analog filter coefficients of integrators
 16 60, 62 and 64, denoted by a, b and c, respectively, can be
 17 selected such that the digital scaling factor $(1/abc)$ can be
 18 implemented by a simple shift operation, e.g., if the
 19 coefficients are chosen equal to 0.3, 0.5 and 0.8333,
 20 respectively, this multiplicand turns out to be 8, or 2^3 .

21 A quantitative expression for the excess noise suppression
 22 achieved by the finite zeros can be derived considering the

1 (ideal) numerator polynomial of the NTF of modulator 50 and can
2 be written as follows:

$$3 \quad N_{3+3}(z) = (1-z^{-1})^2(1-z^{-1}[2-\delta_{61}]+z^{-2})(1-z^{-1}[2-\delta_{62}]+z^{-2}). \quad (2)$$

4 Optimum values for the two finite zero loop gains, δ_{61} and δ_{62} ,
5 are chosen to minimize the 2-norm of $|N_{3+3}(z)|$, which is equivalent
6 to minimizing the noise power, i.e., the quantization error in
7 the desired band. The optimal values for δ_{61} and δ_{62} are thus:

$$8 \quad \delta_{61opt} = 0.8633 \frac{\pi^2}{OSR^2} \quad (3)$$

9 and

$$\delta_{62opt} = 0.4100 \frac{\pi^2}{OSR^2}. \quad (4)$$

10
11 Inserting an OSR of 16 yields $\delta_{61opt} = 0.01581$ and $\delta_{62opt} = 0.03328$.
12 It is noted that both values are very nearly equal to negative
13 powers of two. Thus the second digital scaling factor, i.e., the
14 difference $(3 - \delta_{61})$, can be realized by a simple shift and add
15 operation. Since the sensitivity of the signal to noise ratio
16 with respect to the two zero locations is inherently low, such an
17 approximation would have little effect on the dynamic range. The
18 implementation of DNC 88 as described above is shown in FIG. 2
19 with output Y_1 passing through first function 90, second function
20 92 and into summing junction 94. Scaling factor $(1/abc)$ is
21 applied to output Y_2 which is then passed through third function

1 96, fourth function 98 and is applied to the negative input of
2 summing junction 94. In addition, output 100 of first function
3 90 and output 102 of third function 96 are applied to the
4 positive and negative inputs of summing junction 104,
5 respectively. Scaling factor $(3 - \delta_{61})$ is applied to output 106
6 of summing junction 104, which is then split such that output
7 106a is input to summing junction 108 while output 106b first
8 passes through function 110 before being applied to the negative
9 input of summing junction 108. Output 112 of summing junction
10 108 is then applied to summing junction 94, the output of summing
11 junction 94 being the value of $Y(z)$.

12 The rms quantization error of the ternary quantizers are
13 denoted by e_{rms} . The influence of the loop filter poles can still
14 be approximately a gain factor despite the lower OSR such that
15 the resulting in-band quantization noise power of modulator 50
16 can be approximated by the following:

$$17 \quad n_{3+3}^2 \approx e_{rms}^2 6.5 \times 10^{-3} \frac{\pi^{12}}{13} OSR^{-13}. \quad (5)$$

18 The additional zeros thus improve the noise shaping performance,
19 independent of the OSR, by a factor approximately 153 or 22dB,
20 respectively. This corresponds to almost four extra bits of
21 resolution, well justifying the additional circuitry required in
22 DNC 88.

1 By employing only two cascaded stages and by limiting the
2 order of the loop filters to three, amplifier gain and C-ratio
3 matching requirements are kept reasonably low. In contrast to
4 single-loop circuits, cascaded modulators suffer significantly
5 from integrator leakage. Capacitor ratio mismatch errors, on the
6 other hand, result in an incomplete cancellation of the
7 quantization noise of the first stage. Denoting the order of the
8 second modulator loop by n , the amplifier gain demands of such a
9 two-stage cascade are shown to increase as $\left(\frac{OSR}{\pi}\right)^{n+1}$, while the C-
10 ratio matching conditions scale as $\left(\frac{OSR}{\pi}\right)^{-n}$. Referring now to the
11 embodiment of modulator 150 of FIG. 3, one can replace modulator
12 52b of FIG. 2 by a simpler second-order stage 152 in order to
13 relax the amplifier gain and C-ratio matching requirements.
14 However, reduction in the order of the second stage will lower
15 effective resolution since both a noise transfer function zero
16 and an integrator are traded off. As with modulator 52b of FIG.
17 2, the input to summing junction 154 of second-order stage 152,
18 is output 80 of integrator 64. Second-order stage 152 has
19 integrators 156 and 158. Integrator 156 is a non-delaying
20 integrator and integrator 158 is a delaying integrator. Output
21 160 of summing junction 154 is provided to integrator 156 which
22 in turn provides its output to summing junction 162 with the

1 output of summing junction 162 being provided to integrator 158.
 2 Output 164 of integrator 158 is provided to quantizer 166 with
 3 the output of quantizer 166 being the value Y_2 . Similar to
 4 output 58 of quantizer 16 of FIG. 2, output 164 of integrator 158
 5 is fed back to summing junctions 154 and 162. To accommodate the
 6 second-order configuration, output 164 is passed through function
 7 168 prior to being applied to the negative input of summing
 8 junction 154. Similarly, output 58 is passed through function
 9 170 prior to being applied to the negative input of summing
 10 junction 66 in modulator 52a. With the use of second-order
 11 modulator 152 only the NTF zero loop of output 80 of integrator
 12 64 to summing junction 66 is required. For the third-order,
 13 second-order modulator of FIG. 3, the output $Y(z)$ is computed as
 14 follows:

$$15 \quad Y(z) = Y_1 z^{-2} + (1 - z^{-1})(1 - z^{-1}[2 - \delta_{61}] + z^{-2}) \left(\frac{1}{abc} Y_2 - Y_1 z^{-2} \right). \quad (6)$$

16 The DNC scheme for the embodiment of FIG. 3, denoted by dotted
 17 line 172, therefor has a slightly different configuration than
 18 that of DNC 88 of FIG. 2 in that function 90 is replaced with
 19 function 174 and a scaling factor of $(3 - \delta_5)$ is applied to
 20 output 106 of summing junction 104.

21 The numerator polynomial for modulator 150 can be written as
 22 follows:

$$23 \quad N_{3+3}(z) = (1 - z^{-1})^3 (1 - z^{-1}[2 - \delta_5] + z^{-2}). \quad (7)$$

1 The optimum value of the loop gain of the first stage turns out
2 to be:

$$3 \quad \delta_{opt} = \frac{7 \pi^2}{9 OSR^2}. \quad (8)$$

4 Inserting this value into the noise power transfer function
5 yields the following approximate expression for the total output
6 noise power:

$$7 \quad n_{3+3}^2 \approx e_{rms}^2 \frac{4 \pi^{10}}{81 \cdot 11} OSR^{-11}. \quad (9)$$

8 The additional quantization noise suppression factor introduced
9 by this single zero is thus approximately equal to 20 or 13dB,
10 respectively. This still corresponds to more than two extra bits
11 of resolution. The additional multiplication factor $(3 - \delta_5)$ is
12 a comparatively small price to be paid for the substantially
13 improved dynamic range.

14 What has thus been described is a wide-band sigma-delta
15 modulator having a double third-order cascaded architecture.
16 Both third-order stages utilize simple ternary quantizers to
17 avoid linearity problems associated with the DAC's in the
18 feedback paths of the modulators. Additional finite NTF zeros
19 are added to the third-order stages by means of an additional
20 feed back loop between the output of the third integrator and the
21 second summing junction. The NTF zeros serve to maximize the
22 dynamic range. The excess dynamic range gained by these

1 additional zeros is substantial and independent of OSR, hence,
2 the present invention is particularly useful in wide-band
3 applications where the OSR is inherently low. The new topology
4 significantly enhances the dynamic range of the system while
5 hardly affecting the circuit's sensitivity with regard to some
6 non-idealities such as finite amplifier open-loop gains or
7 capacitor ratio mismatches. By employing only two cascaded
8 stages and by limiting the order of the loop filters to three,
9 amplifier gain and capacitor ratio matching requirements are kept
10 reasonably low. In addition, a second embodiment having a third-
11 order, second-order cascaded architecture is provided which
12 further relaxes the amplifier gain and C-ratio matching
13 requirements.

14 Thus, it will be understood that additional changes in the
15 details and arrangement of parts, or in the values for various
16 coefficients which have been herein described and illustrated in
17 order to explain the nature of the invention, may be made by
18 those skilled in the art within the principle and scope of the
19 invention as expressed in the appended claims. For example, by
20 implementing the sigma-delta modulator of the present invention
21 on a computer, the modulator can be utilized in a digital to
22 analog converter. Further, the modulator can be implemented as
23 discrete-time analog either by means of a switched-capacitor
24 filter/integrator or by means of a switched-current integrator.

1 It may also be implemented by means of an analog integrator where
2 phase is used to delay energy. Further, the architecture
3 described is applicable to any cascade of a third-order or higher
4 first stage and a second-order or higher second stage, e.g., a
5 fourth-order, fourth-order cascade, or a fifth-order, third-order
6 cascade.

7 In light of the above, it is therefore understood that
8 the invention may be
9 practiced otherwise than as specifically described.

1 Navy Case No. 78416

2

3 SIGMA-DELTA MODULATOR FOR WIDE BANDWIDTH APPLICATIONS

4

5 ABSTRACT OF THE DISCLOSURE

6 A wide-band sigma-delta modulator is disclosed having a cascade
7 of two modulators or stages, the first stage being of third or
8 greater order and the second being of second or greater order.
9 Both stages utilize simple ternary quantizers to avoid linearity
10 problems associated with the digital to analog converters in the
11 feedback paths of the modulators and to maintain stable
12 operation. Additional zeros are placed in the noise transfer
13 functions of third or higher order stages by means of additional
14 feed back loops between the output of the third integrator and
15 the second summing junction to maximize the dynamic range. The
16 new topology significantly enhances the dynamic range of the
17 system while hardly affecting the circuit's sensitivity with
18 regard to some non-idealities such as finite amplifier open-loop
19 gains or capacitor ratio mismatches. The excess dynamic range
20 gained by these additional zeros is substantial and independent
21 of OSR, hence, the present invention is particularly useful in
22 wide-band applications where the OSR is inherently low. By
23 employing only two cascaded stages, amplifier gain and capacitor
24 ratio matching requirements are kept reasonably low.

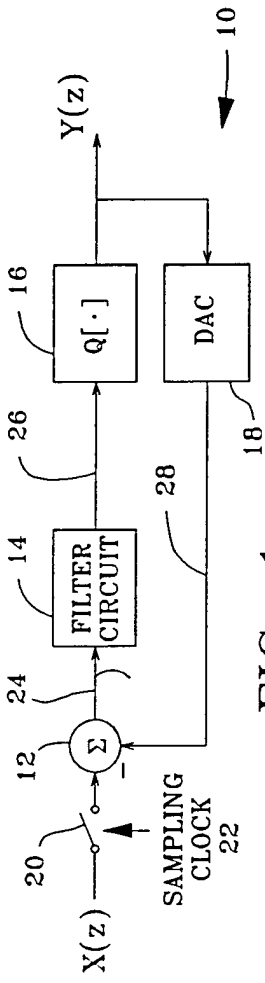


FIG. 1
(prior art)

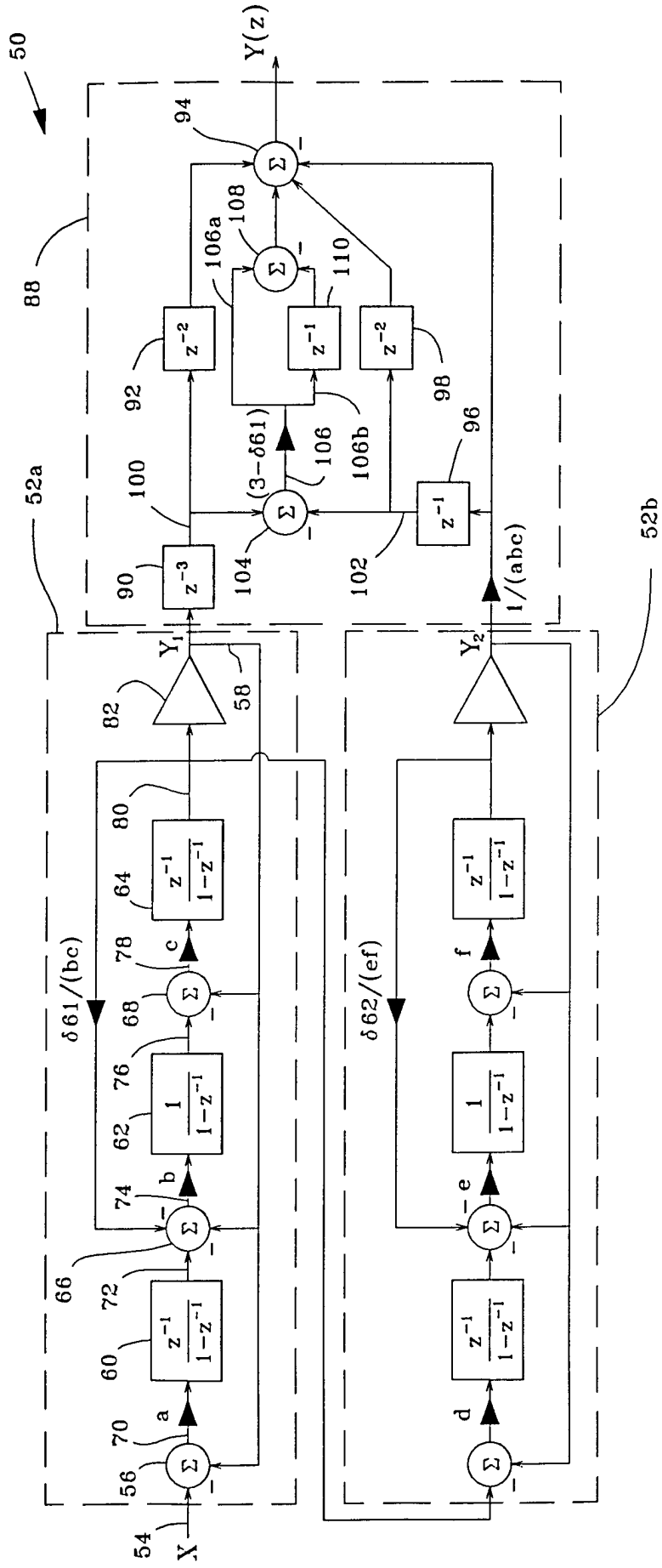


FIG. 2

150

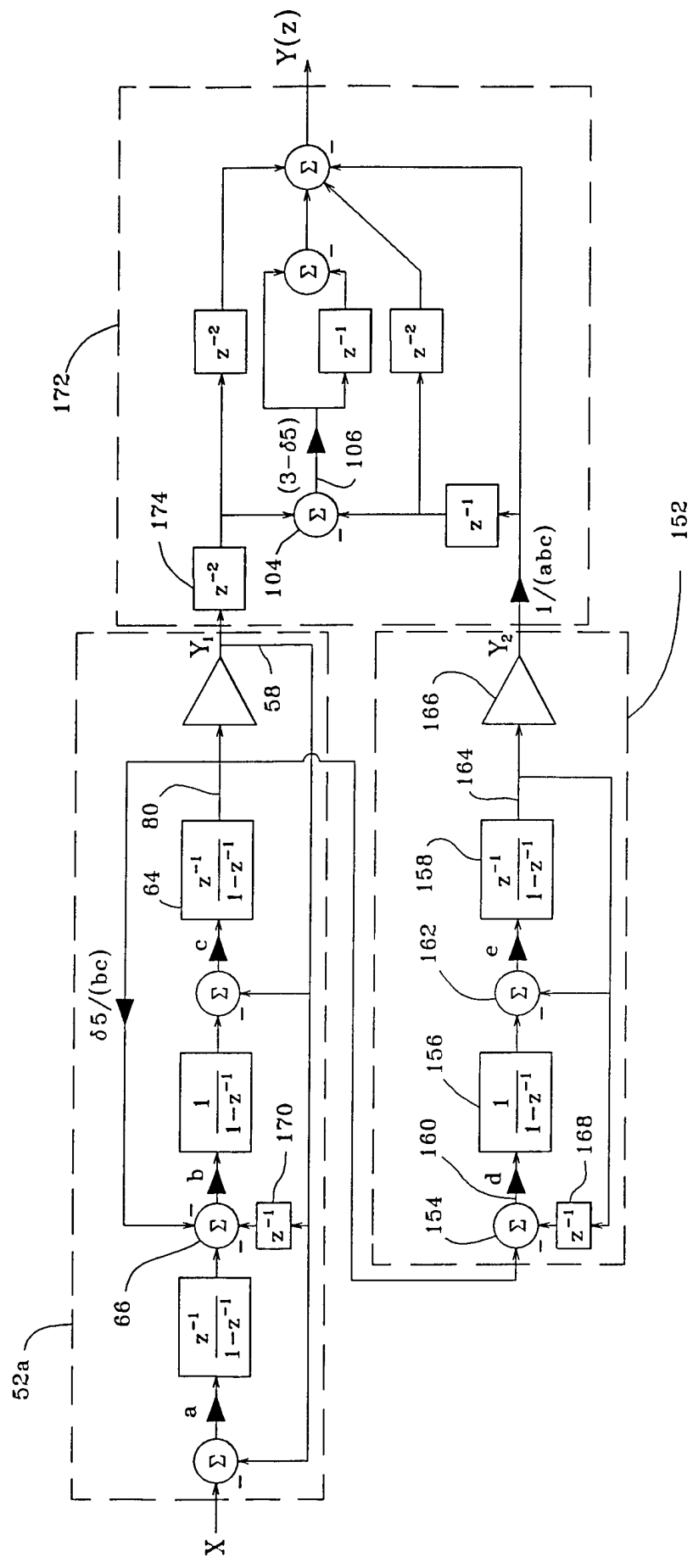


FIG. 3