Several methods for approximating the exact calculation of the magnitude of quadrature components are faster, and require smaller programs, than the exact methods themselves.

Magnitude Approximations for Microprocessor Implementation

W. Thomas Adams* and John Brady

Applied Research Laboratories
University of Texas at Austin

Many data processing systems require computation of the magnitude of vector quantities expressed in rectangular coordinates. All the various applications, such as the amplitude computation of quadrature samples of the magnitude portion of a rectangular-to-polar conversion, require the operation $R = \sqrt{I^2+Q^2}$. Here, we will review several popular forms of piece-wise approximation and will demonstrate their advantages. We will give examples of typical implementations.

Motivation

The need to compute the magnitude of vector components occurs repeatedly in graphics and signal processing computations. As more of this processing is done digitally and as system data rates increase, less time is available for each computation. Previously, in systems where the data rate permitted, excellent approximations to $\sqrt{I^2+Q^2}$ could be achieved through Cordic routines and through successive approximation techniques over large dynamic ranges. For systems in which vector components are expressed in eight or fewer bits, exact methods such as look-up tables or programmed logic arrays have been used. For time-constrained systems of larger dynamic range, the exact computation has been replaced by piece-wise linear approximations.

Although the approximations typically introduce error in the result, they reduce processing time. We will discuss ease of implementation, amount of error introduced, and computation speed for several approximations.

Background

An excellent survey of piece-wise linear methods is found in a letter by A. E. Filip to the editor of the *IEEE Transactions on Audio and Electroacoustics*.[1] Filip approaches the problem of error in approximating $\sqrt{I^2+Q^2}$ by constraining the error to an equiripple case. He then compares his equiripple approach to several other approximations (see "Rationale for magnitude approximation" on page 28). The approximations that we will examine here consist of the ones discussed in Filip's article

*Now with IBM, Entry Systems Division, Austin, Texas

DOCKET ALARM
Find authenticated court documents without watermarks at docketalarm.com.

and some other methods. All these methods consist of a linear approximation to the function $\sqrt{I^2+Q^2}$ and require that the magnitude of I and Q be compared and that the largest and smallest of the two be defined. We will employ Filip's notation to avoid confusion.

Rationale for magnitude approximation

Several different approaches can be used to obtain the values for a linear estimate of the magnitude of a vector. The approach depends on the factors most important to the user. Factors include mean error output, standard deviation of error, maximum error, and computational complexity and speed. Filip[1] wanted the maximum error to be small and evenly distributed over the range of inputs. To achieve his goals, he selected an equiripple criterion. This means that the peak error was equal to the error at the extremes of the ratios of the inputs. Using this criterion, Filip obtained an erorr of less than one percent at the expense of producing coefficients not as easily implemented as those produced specifically for speed of execution or ease of implementation.

Many signal processing applications can tolerate more than a one-percent error and have highly constrained throughput requirements. The need for high processing throughput and ease of implementation led us to the selection of coefficients which can be expressed as fractions whose denominator is a power of 2, so that shifting can be used for division. Since one set of coefficients cannot always satisfy error constraints, two sets of coefficients must sometimes be selected for different ranges of input. Methods which use one set of coefficients are called one-line approximations, and methods which use two sets of coefficients are called two-line approximations.

Table 1.
One-line approximations for $\sqrt{I^2+Q^2}$, with computed errors.*

	VALUE SET						
	1	2	3	4	5	6	7
COEFFICIENT a	0.961	1.000	1.000	0.969 (31/32)	0.948	1	1
COEFFICIENT b	0.398	0.267	0.500	0.375	0.393	0.375	0.25
PEAK ERROR $\lvert e_{max}\rvert$ (percent)	3.95	10.4	11.8	4.97	5.19	6.8	11.6
MEAN ERROR $\bar{e}$ (percent)	1.30	0.	8.7	1.20	0	4.0	0.656
STANDARD DEVIATION OF THE ERROR σ_e (percent)	2.70	3.87	9.21	2.70	2.33	2.56	4.11

Therefore, we define

$$x = \max(|I|,|Q|)$$
$$y = \min(|I|,|Q|).$$

The reason for the selection of the largest and smallest of I and Q may not be obvious unless one considers that if either I or Q is much larger than the other it will tend to dominate the result of $\sqrt{I^2+Q^2}$. By selecting the largest of I and Q, we can find appropriate coefficients or multipliers, called a and b, for x and y such that the amplitude—the resultant $R = \sqrt{I^2+Q^2}$—is approximated by $R = ax+by$. In the one-line approximations, one set of coefficients holds for all ratios of y/x or, alternately stated,

$$a = a_1$$
$$b = b_1$$

for $\theta = 0 \rightarrow 45°$, where $\theta = \tan^{-1} y/x$. For the two-line approximations, an angle θ_o is defined such that for $\theta < \theta_o$,

$$a = a_1$$
$$b = b_1$$

and for $\theta > \theta_o$,

$$a = a_2$$
$$b = b_2.$$

The values for a, b, and θ_o must be chosen to minimize cost, size, or error and/or to maximize speed. Filip presents several sets of values for a and b and computes the mean, the standard deviation, and the peak error for both one-line and two-line approximations. Table 1 and 2 give the values of a and b and the mean, standard, and peak errors for several value sets in addition to Filip's value sets; all computations have been verified. Region I is the portion of the approximation in which the coefficients (a_1,b_1) are used. Figures 1 and 2 are plots of the errors introduced by the different value sets. Figure 1 shows a comparison of the one-line methods presented in Table 1, while Figure 2 shows a comparison of the two-line methods from Table 2. Generally, the two-line methods require more execution time, but their errors are lower than those of the one-line methods. Our experience has been that the one-line approximations are adequate for many signal processing tasks but also that the two-line approximations are not too difficult to program.

Analysis

Of the value sets shown in Figure 1, the ones that minimize processing time are those for which the a and b coefficients can be expressed as a quotient whose denominator is a power of 2, such as 1/2, 3/8, or 3/4.

DOCKET ALARM
Find authenticated court documents without watermarks at docketalarm.com.

can be used to accomplish the multiplication. Of the one-line approximations, the value sets which fit the power-of-2 quotient criterion are sets 3, 4, 6, and 7, which are plotted in Figure 3. It is easy to see that value set 3 $(x+y/2)$ provides the simplest and fastest implementation, since only one shift operation and one addition are required in addition to the comparison. Value set 4 is much more difficult to implement due to the 31/32 coefficient, and value sets 6 and 7 lie between value sets 3 and 4 in difficulty, with value set 7 having the smallest mean error. Notice that the coefficients for value set 7 $(x+y/4)$ are nearly the same as those for value set 2 $(x+0.267y)$ and that the coefficients for value set 6 are very close to those for value sets 1, 4, and 5. It may be concluded that value sets 3, 6, and 7 are the easiest to implement and that all have less than 3/4-dB mean error and less than 1-dB peak error. These errors are entirely acceptable for many systems.

If peak errors of less than 1/2 dB are required, then a two-line method will be needed. Two-line methods require more execution time and are slightly more difficult to implement. In order to implement a two-line method, one must separate the regions by measuring the magnitude of the difference between the quadrature components to determine whether one component is larger than the other by some preset amount. The magnitude of the difference determines which of the regions is chosen for the approximation. The implementation is made simpler if this difference is a power of 2. Table 2 shows all the two-line methods we examined. The methods with coefficients that are powers of 2 and whose difference magnitude is 2 are value sets 4-9. Figure 4 is a plot of value set 4 and Figure 5 is a plot of value set 7; value sets 4 and 7 were chosen since they have the lowest errors of value sets 4-9. The peak errors are less than 0.26 dB and the mean errors are 0.05 dB and 0.012 dB, respectively. This reduction in error from the one-line methods may be significant enough in some applications to justify the more extensive software needed for two-line methods.

If there is a criterion other than ease of implementa-

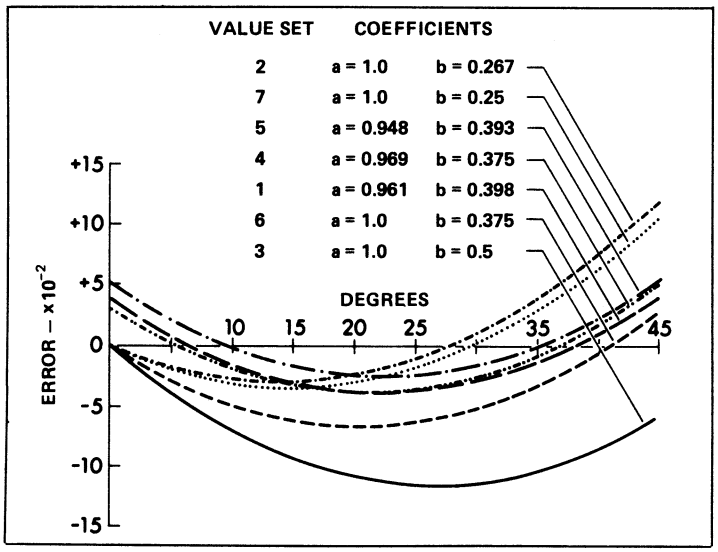


Figure 1. Comparison of one-line approximations to R = $\sqrt{I^2+Q^2}$.

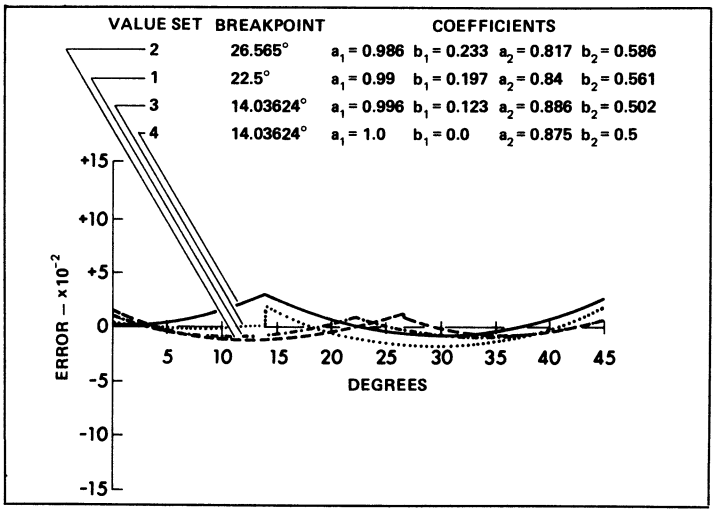


Figure 2. Comparison of two-line approximations to R = $\sqrt{I^2+Q^2}$.

Table 2.
Two-line approximations for $\sqrt{I^2+Q^2}$, with computed errors.*

		VALUE SET 1	2	3	4	5	6	7	8	9
	BREAKPOINT	$\theta_0=\pi/8$	$\theta_0=\tan^{-1}(1/2)$	$\theta_0=\tan^{-1}(1/4)$	$\theta_0=\tan^{-1}(1/4)$	$\theta_0=\tan^{-1}(1/4)$	$\theta_0=\tan^{-1}(1/4)$	$\theta_0=\tan^{-1}(1/2)$	$\theta_0=\tan^{-1}(1/2)$	$\theta_0=\tan^{-1}(1/2)$
COEFFICIENT a	REGION I	0.990	0.986	0.996	1.0	1.0	1.0	1.0	1.0	1.0
	REGION II	0.840	0.817	0.886	0.875	0.875	1.0	0.875	0.875	1.0
COEFFICIENT b	REGION I	0.197	0.233	0.123	0	0.125	0	0.125	0	0
	REGION II	0.561	0.586	0.502	0.5	0.5	0.5	0.5	0.5	0.5
PEAK ERROR $\lvert e_{max}\rvert$ (percent)	REGION I	0.970	1.36	0.376	2.98	0.778	2.97	4.92	10.5	10.5
	REGION II	0.970	0.650	1.84	2.98	2.95	11.8	2.77	2.77	11.8
MEAN ERROR $\bar{e}$ (percent)		0.323	0.354	0.461	0.617	0.143	6.76	0.499	2.17	1.96
STANDARD DEVIATION σ_e (percent)		0.644	0.765	1.05	1.23	0.021	5.4	1.45	2.99	7.11

*Portions from Filip's "Linear Approximations . . ."[1]

DOCKET ALARM
Find authenticated court documents without watermarks at docketalarm.com.

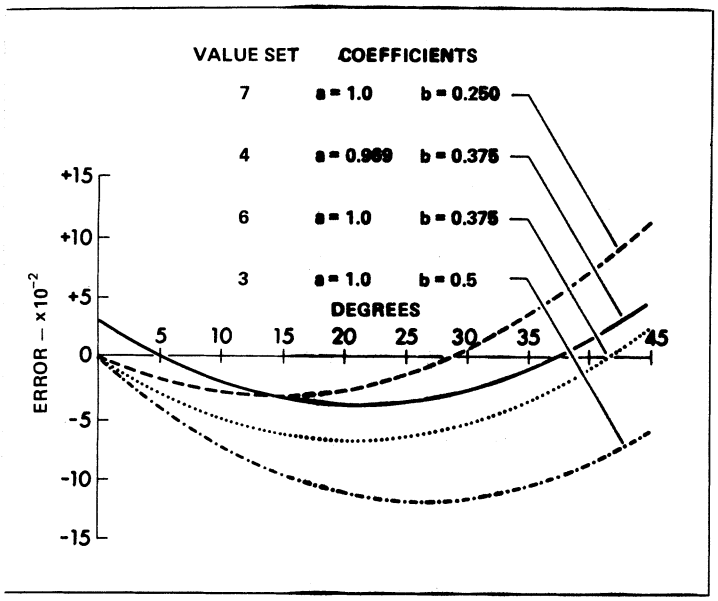


Figure 3. One-line methods with power-of-2 coefficients.

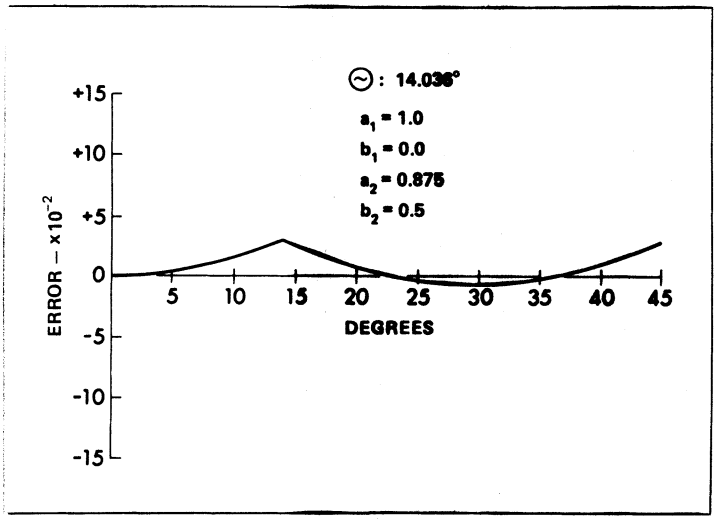


Figure 4. Plot of the "fast" two-line method.

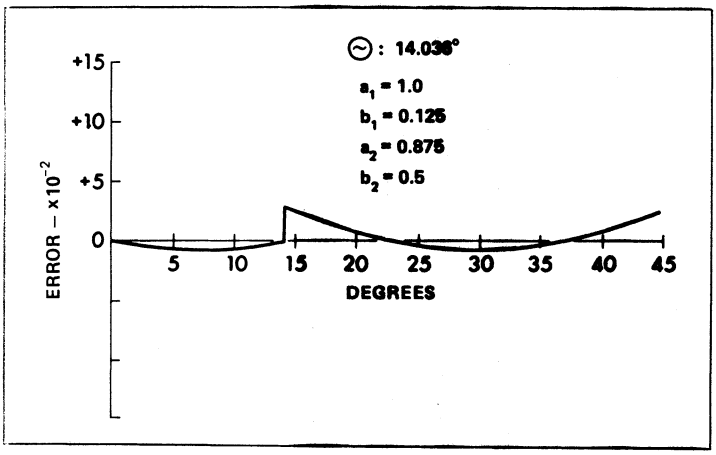


Figure 5. Plot of the two-line approximation, value set 7.

tion, then peak error, mean error, or standard deviation may be considered. The tables give the values of the errors so that reasonable design selections can be made. In general, a random distribution of the input values is assumed.

Implementations

The software realization for one one-line method and one two-line method will be described here in order to illustrate the relative complexity of the various methods.

One-line method. The one-line method chosen for implementation is a = 1, B = 1/4, since it produces a low mean error (0.656 dB). Figure 6 is a flowchart of this method. Since a comparison of the magnitude of these components is needed, the absolute value of the quadrature components is computed first. Then the magnitudes are compared, and the larger of the quadrature components is placed in a register called X. The smaller of the two is placed in a register called Y. The contents of the Y register are shifted two times to the right, which approximates division by 4. There is some error involved in this method of division, but the average error for a 12-bit number is less than 0.5 percent. The resulting contents of the Y register are added to the contents of the X register to obtain the final result. Note that if an approximation such as $x+3y/8$ were to be implemented, the Y register would be shifted three times for division by 8, and the resulting contents of Y would be added to the X register three times to implement the multiplication.

Two-line method. The two-line method chosen for implementation is Region I, a = 1.0, b = 0, and Region II, a = 7/8, b = 1/2, $\theta_o = \tan^{-1}(b/a)$. This method was chosen because it has a low mean error (0.61 percent) and is representative of all two-line methods. Figure 7 is a flowchart of this method. As in the one-line method, the absolute values of the I and Q components are computed first. Then these absolute values are compared. The comparison must determine whether either the x or y component is four times as large as the other. Depending on the outcome of this comparison, one of two methods is chosen. If one component is at least four times larger than the other, the angle is in Region I and the largest number is chosen. The approximation is then complete. If one component is *not* four times larger than the other, the angle is in Region II and the approximation must be computed with a = 7/8 and b = 1/2. The X and Y registers holding the absolute values of the coefficient must be shifted the appropriate number of places in order to approximate divisions by 8 and by 2. The registers must be summed to an accumulator and the X register must be added to itself seven times; the Y register must be added to the total once. The operation is then complete.

These one- and two-line approximation methods are not difficult to implement, result in faster operation, and

DOCKET ALARM
Find authenticated court documents without watermarks at docketalarm.com.

Several methods can be used to approximate the exact calculation of the magnitude of quadrature components. A number of one-line methods give faster computation, with less software, than exact methods. For even lower errors, two-line approximations can be easily implemented, with even higher-speed results. ■

Acknowledgment

The authors wish to thank Patti Hall for the generation of the computer plots. This work was sponsored under Naval Sea Systems Command Contract N00024-76-C-6022.

Reference

1. A. E. Filip, "Linear Approximations to $\sqrt{x^2+y^2}$ Having Equiripple Characteristics," *IEEE Trans. Audio and Electroacoustics,* Vol. AU-21, No. 3, 1973, pp. 554-556.

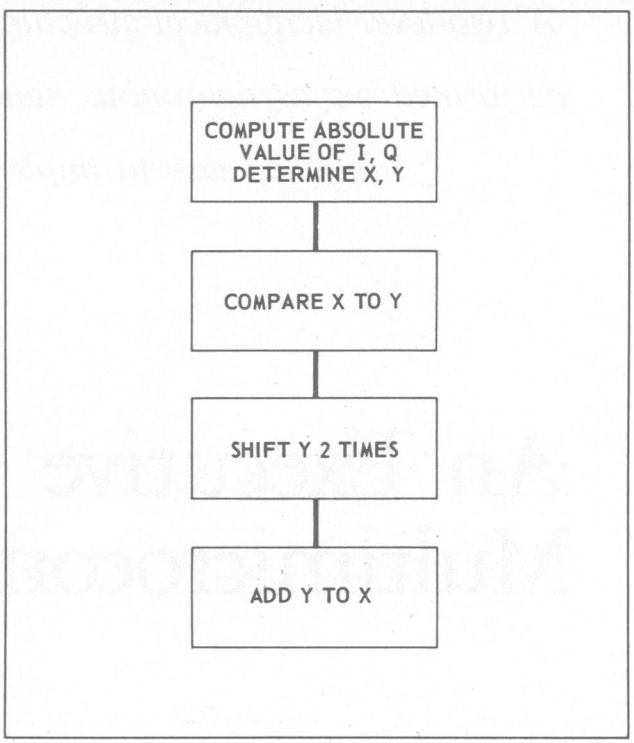
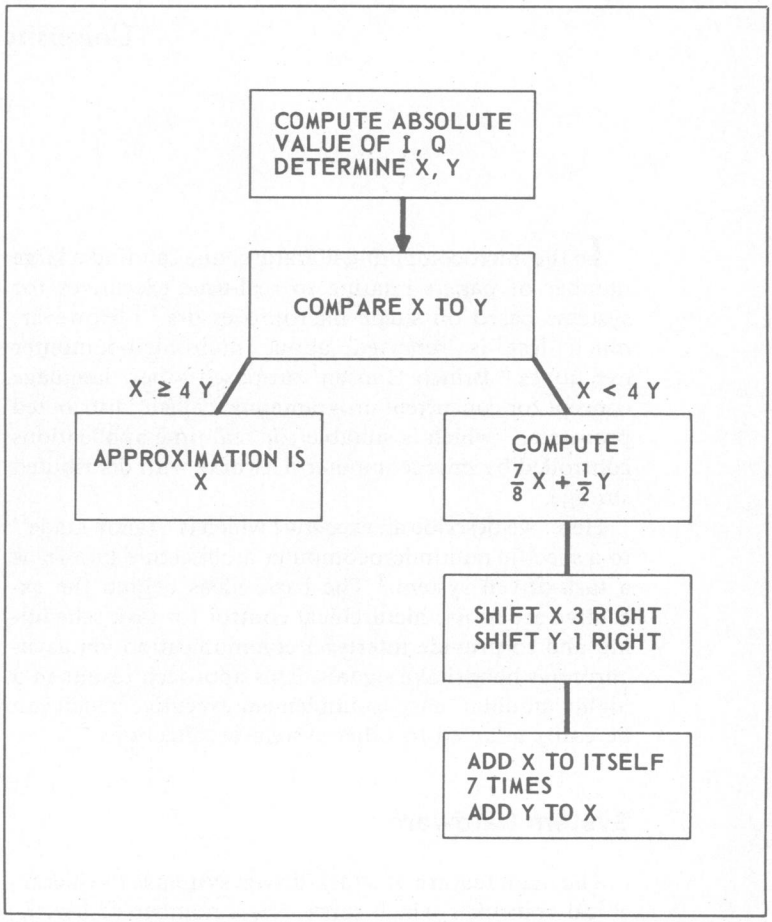
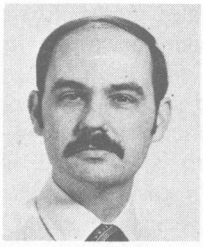


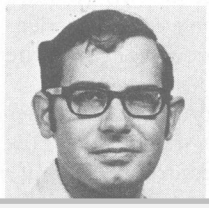
Figure 6. Flowchart for the one-line approximation, value set 7: a + (1/4) b.





W. Thomas Adams has been an engineer with IBM Corporation, Entry Systems Division, Austin, Texas, since 1978. His responsibilities include new product development. He was a research engineer with the Applied Research Laboratories of the University of Texas at Austin from 1969 until 1977. A member of the IEEE, Adams was chairman of the Central Texas Chapter of the Computer Society for 1982-83. He received a BS in electrical engineering in 1969 and an MS in electrical engineering in 1974, both from the University of Texas at Austin.

Adams' address is IBM Corporation, F25/045, 11400 Burnet Rd., Austin, TX 78758.



John Brady is an engineer at the Applied Research Laboratories of the University of Texas at Austin, where he is involved in sonar signal processing applications. He received a BS in 1972 and an MS in 1978, both in electrical engineering and both from the University of Texas at Austin.

Brady's address is Applied Research

DOCKET ALARM Find authenticated court documents without watermarks at docketalarm.com.