Digital Dither: Decreasing Round-off Errors in Digital Signal Processing

Péter Csordás , András Mersich, István Kollár

Budapest University of Technology and Economics
Department of Measurement and Information Systems

cp333@hszk.bme.hu , ma351@hszk.bme.hu , kollar@mit.bme.hu

Abstract *– Dithering is a known method for increasing the precision of analog-to-digital conversion. In this paper the basic problems of quantization and the principles of analog dithering are described. The benefits of dithering are demonstrated through practical applications. The possibility and the conditions of using digital dither in practical applications are examined. The aim is to find an optimal digital dither to minimize the round-off error of digital signal processing. The effects on bias and variance using uniform and triangularly distributed digital dithers are investigated and compared. Some DSP implementations are discussed.*

Keywords *– digital dither, round-off error, bias, variance*

I. STATE-OF-THE ART IN THE FIELD

In DSP applications data must be quantized before storing them in the memory. This happens after importing analog signals (A/D conversion, input quantization), and after every arithmetic operation (round-off).

Input quantization is extensively studied in the literature [1, 6]. In accordance with the theory described, although some information is always lost, the moments of the signal can be exactly calculated from the quantized samples if the Quantizing Theorem is fulfilled (QT I or QT II, [1]). The problem is however, that in real cases this theorem can never be totally fulfilled, and even for a reasonable approximation an analog-digital converter with high resolution may be needed. This problem has been previously analyzed and it was shown that the use of a dither is a good solution: adding a carefully designed random or pseudo-random noise to the signal before quantization, the bias of the moments can be decreased or eliminated. Dithers are usually characterized by their probability density function (PDF) and by the characteristic function (CF, the inverse Fourier-transform of the PDF).

While the theory of input quantization is well elaborated, much less is known about the quantization which happens after each arithmetic operation (round- off error). The situation is quite different from the above one, since the result of any arithmetic operation is already discrete in amplitude, even its distribution is usually non-standard. Therefore, quantization of quantized signals needs to be analyzed. This cannot be bound to a certain processing unit in the DSP, but happens simply during operation and/or during storage of the result into memory cells. Moreover, since the signal to be quantized is in digital form, only digital dither can be applied, and only if additional bits are available to make it possible to use. The method has already been discussed for example in [4], but exact analysis has not been published yet. The goal of this paper is to contribute to the filling of this gap. We will deal here with fixed-point number representation only, the results can be extended to floating-point later.

II. SOME PRACTICAL APPLICATIONS USING DITHER

The dither effect is already present in some industrial applications. An example is the precision weigher: if a weight is measured in the stationary state of the instrument, an error of one LSB has to be taken into account. However, rapid and precise measurement is possible knowing the characteristic of the weigher and making the most of the dither effect of the transient, by making an LS fit to the transient response [9].

In audio techniques, a triangularly distributed dither can be used to minimize the "fade-out noise" (described in [3]).

In medical applications, motion analysis is an important tool for example for early detection of Parkinsons disease. In this method, the movement is examined by means of following some markers fixed to given anatomical points of the patient. The precise detection of the marker centers is a key task. We know the form of the marker (circle in the image), but we are only interested in the position of it. The several-pixel marker picture acts as a dither, allowing more precise determination of the center than measurement of a

0-7803-7864-4/03/$17.00 ©2003 IEEE.

single point would allow. In addition, trembling also acts as dither, further improving the possibility of precision.

III. PROBLEMS TO SOLVE, MODEL FOR EXAMINATION

Theoretically speaking the digital dither – like the analog one – should be applied to the input signal before the AD conversion, because the information lost during quantization is lost forever, unless quantization errors can be played out against each other in some kind of averaging. This is why a high-bit resolution accumulator is to be used, which allows that the result of an operation (multiplication, addition, trigonometric function) is not immediately quantized by the ALU, but the long result can be manipulated by adding the dither *before* storage.

In this paper we estimate the upper limit of the deviation of the moments with the forms given in [2] and with a numerical calculation in MATLAB. After that we examine the effects of using dither and give the parameters of an optimal dither.

IV. DITHER FORMS AND DSP IMPLEMENTATIONS

Digital dither means that the dither is implemented in the form of finite bit length numbers. It is a reasonable approach that we try to approximate suggested analog dithers by numbers, that is, by discrete distributions. Two examples are shown in Fig. 1. These approximate the two dithers suggested for use in digital signal processing: uniform and triangularly distributed dithers. To implement a symmetric uniform dither with N bits, N+1 bits are needed, because the half an LSB shifting relative to zero must be realized, to avoid nonzero dither mean (Fig. 1a). Another possibility would be to use a non-symmetric dither at a price of small systematic errors.

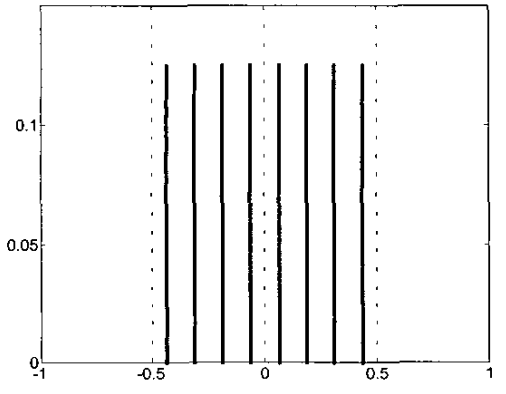

Fig. 1a: PDF of a symmetric uniformly distributed dither. Number of dither bits: N=3

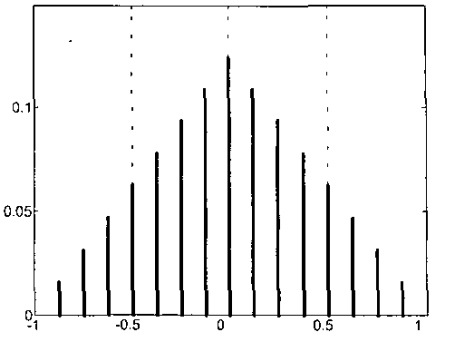

Fig 1b: PDF of a triangularly distributed dither, derived as the convolution of two uniform dithers.

In number representation (storage) and the calculation of different arithmetic operations, the quantum size is the LSB in number representation. Therefore, the dither amplitude needs to be in this order of magnitude, and we need to add the dither *before* truncating the result. This is only possible if the numbers to be quantized (the *precise* or more or less precise results) are at hand. Therefore, we need the numbers with sub-LSB precision. We see two options for this.

- In many DSP's, the results are generated in the accumulator, which has an extended precision. If the numbers are properly scaled, the sub-LSB bits appear at the end of the accumulator, and the dither can be added to it before storage. The maximum reasonable bit number of the dithers is determined by the number of excess bits in the accumulator. Our aim can be to verify if this bit number is enough, and/or determine the necessary bit number if this is smaller.
- If the accumulator has no excess bits, then we have to choose another way to determine the sub-LSB bits of the result. This may need special programming: e.g. in multiplication, separation of the two terms into upper and lower bit segments, calculation of the four products, and use these for determining the result and its sub-LSB bits. Since this is to be done for each operation, the effectiveness depends on the necessary operations – it is well possible that selecting a DSP with more bits is the cheaper solution.

V. CALCULATIONS

We need to evaluate the effect of the above dithers in terms of mean value and variance of the result. Let us denote the quantum size (the LSB in number representation) by q.

The CFs of the above dither functions are:

For discrete uniform dither:

$$\Phi_{d\Pi} = \frac{\sin\left(\frac{qu}{2}\right)}{2^N \cdot \sin\left(\frac{qu}{2^{N+1}}\right)} \tag{1a}$$

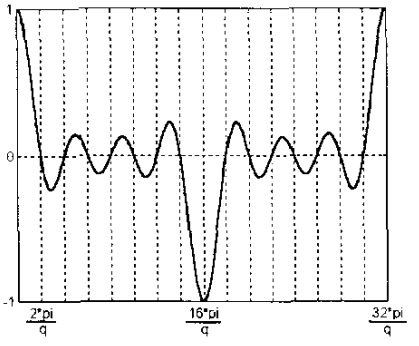

Fig. 2a: $\Phi_{d\Pi}(u)$, discrete sinc for N=3

For discrete triangular dither:

$$\Phi_{d\Delta} = \left(\frac{1}{2^N}\frac{\sin\left(u\frac{q}{2}\right)}{\sin\left(u\frac{q}{2^{N+1}}\right)}\right)^2 \tag{1b}$$

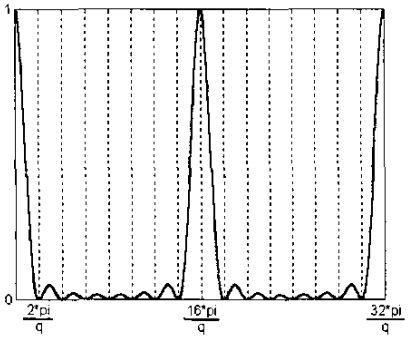

Fig. 2b: $\Phi_{d\Delta}(u)$, for discrete sinc^2 for N=3

The moments of signal x are to calculate with the equation:

$$E\{x^n\} = \frac{1}{j^n}\cdot\frac{\partial^n \Phi_x(u)}{\partial u^n}_{|u=0} \tag{2a}$$

This form is modified by quantizing and dithering:

- Quantization means the convolution of the PDF with a rectangular window, and so the multiplication of the CF with a *sinc* function
- The windowed PDF is sampled. This causes the periodic repetition of the CF
- Adding a dither means multiplication with the dithers CF

The result is:

$$E\{x_q^{\,n}\} = \frac{1}{j^n}\sum_{k=-\infty}^{\infty}\left[\frac{\partial^n}{\partial u^n}\left(\Phi_x(u)\cdot\Phi_d(u)\cdot sinc\left(\frac{qu}{2}\right)\right)\right]_{u=\frac{2k\pi}{q}} \tag{2b}$$

This means for the bias and for the second moment:

$$E\{x_q\} = \frac{1}{j}\sum_{k=-\infty}^{\infty}\left[sinc\cdot\left(\Phi_x\dot{\Phi}_d + \dot{\Phi}_x\Phi_d\right) + si\dot{n}c\cdot\Phi_x\Phi_d\right]_{u=\frac{2k\pi}{q}} \tag{3a}$$

$$E\{x_q^{\,2}\} = -\sum_{k=-\infty}^{\infty}\left[sinc\cdot\left(\Phi_x\ddot{\Phi}_d + 2\dot{\Phi}_x\dot{\Phi}_d + \ddot{\Phi}_x\Phi_d\right)\right]_{u=\frac{2k\pi}{q}} - \tag{3b}$$
$$-\sum_{k=-\infty}^{\infty}\left[2\cdot si\dot{n}c\cdot\left(\Phi_x\dot{\Phi}_d + \dot{\Phi}_x\Phi_d\right) + si\ddot{n}c\cdot\Phi_x\Phi_d\right]_{u=\frac{2k\pi}{q}}$$

Where the arguments of *sinc* and Φ are the same as in (2b)

In the tables below, some often used, important forms are given. These are needed to extract the equations above.

Table 1.: *sinc* function

Form	Value for $u=\frac{2k\pi}{q}$	
	$k=0$	$k\neq 0$
$sinc = \frac{\sin\left(\frac{qu}{2}\right)}{qu/2}$	1	0
$si\dot{n}c = \frac{\cos\left(\frac{qu}{2}\right) - sinc}{u}$	0	$\frac{q}{k\pi}\frac{(-1)^k}{2}$
$si\ddot{n}c = \left(\frac{2}{u^2} - \frac{q^2}{4}\right)\cdot sinc - \frac{2\cos\left(\frac{qu}{2}\right)}{u^2}$	$-\frac{q^2}{12}$	$\left(\frac{q}{k\pi}\right)^2\frac{(-1)^{k+1}}{2}$

Table 2.: dither characteristics

Form	Value for $u=\frac{2k\pi}{q}$	
	$k=2^N\cdot l$	$k\neq 2^N\cdot l$
$\Phi_{d\Pi}=\frac{\sin\left(\frac{qu}{2}\right)}{2^N\cdot\sin\left(\frac{qu}{2^{N+1}}\right)}$	$(-1)^l$	0
$\dot{\Phi}_{d\Pi}=\frac{q}{2^{N+1}\sin\left(\frac{qu}{2^{N+1}}\right)}\cdot\left[\cos\left(\frac{qu}{2}\right)-\Phi_{d\Pi}\cos\left(\frac{qu}{2^{N+1}}\right)\right]$	0	$\frac{q(-1)^k}{2^{N+1}\sin\left(\frac{k\pi}{2^N}\right)}$
$\ddot{\Phi}_{d\Pi}=-\Phi_{d\Pi}\frac{q^2}{4}\left(1-\frac{1}{2^{2N}}\right)-\dot{\Phi}_{d\Pi}\frac{q\cos\left(\frac{qu}{2^{N+1}}\right)}{2^N\sin\left(\frac{qu}{2^{N+1}}\right)}$	$-\frac{q^2(-1)^l}{12}\left(1-\frac{1}{2^{2N}}\right)$	$\frac{q^2(-1)^{k+1}\cos\left(\frac{k\pi}{2^N}\right)}{2^{2N+1}\sin^2\left(\frac{k\pi}{2^N}\right)}$
$\Phi_{d\Delta}=\Phi_{d\Pi}^2$	1	0
$\dot{\Phi}_{d\Delta}=2\cdot\Phi_{d\Pi}\cdot\dot{\Phi}_{d\Pi}$	0	0
$\ddot{\Phi}_{d\Delta}=2\cdot\Phi_{d\Pi}^2+2\cdot\Phi_{d\Pi}\ddot{\Phi}_{d\Pi}$	$-\frac{q^2}{12}\left(2-\frac{2}{2^{2N}}\right)$	$\frac{q^2}{2^{2N+1}\sin^2\left(\frac{k\pi}{2^N}\right)}$

By means of these expressions we obtain the following equations:

The mean value of an input signal (without dither) can be expressed as:

$$E\{x_q\}=E\{x\}+\frac{1}{j}\sum_{\substack{k=-\infty\\k\neq 0}}^{\infty}\Phi_x\left(\frac{2k\pi}{q}\right)\frac{q(-1)^k}{2k\pi} \tag{4a}$$

Noticing that the dither CF's take the values 0, ±1 at the desired points, we obtain for discrete uniform dither:

$$E\{x_q\}=E\{x\}+\frac{1}{j}\sum_{\substack{k=-\infty\\k\neq 0}}^{\infty}\Phi_x\left(\frac{2^{N+1}k\pi}{q}\right)\frac{1}{2^N}\frac{q(-1)^k}{2k\pi} \tag{4b}$$

and for discrete triangular dither:

$$E\{x_q\}=E\{x\}+\frac{1}{j}\sum_{\substack{k=-\infty\\k\neq 0}}^{\infty}\Phi_x\left(\frac{2^{N+1}k\pi}{q}\right)\frac{1}{2^N}\frac{q}{2k\pi} \tag{4c}$$

For the investigation we take the worst case, which is a constant input signal with amplitude y. The PDF and the CF are:

$$f_x(x)=\delta(x-y),\quad \Phi_x(u)=e^{juy} \tag{5}$$

Substituting this in equations (4a)-(4b)-(4c) we obtain the distortion of the first moment (the sums in each equation) dependent of the input y, which is illustrated in Fig. 3.

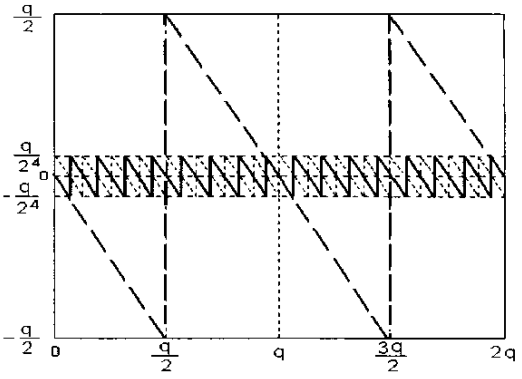

Fig. 3 Effect of various dithers on a constant input signal (N=3). Without dither: dashed, with uniform dither: solid, with triangular dither: dotted.

Heuristically, we could argue that we can measure the mean value of the input signal at best with the resolution of the dither. This is why the maximum bias of the mean value as represented in Fig. 3 is the same for the two dither signals, and is equal to 0.5/2N LSB in both cases.

Let us examine the second moment with the same method. Similarly to (4a):

$$E\{x_q^2\}=E\{x^2\}+\frac{q^2}{12}+\frac{q^2}{2\pi^2}\sum_{\substack{k=-\infty\\k\neq 0}}^{\infty}\Phi_x\left(\frac{2k\pi}{q}\right)\frac{(-1)^k}{k^2}+\frac{q}{\pi}\sum_{\substack{k=-\infty\\k\neq 0}}^{\infty}\dot{\Phi}_x\left(\frac{2k\pi}{q}\right)\frac{(-1)^{k+1}}{k} \tag{6a}$$

The conversion is distortion-free on condition that

$$\Phi_x\left(\frac{2k\pi}{q}\right)=\dot{\Phi}_x\left(\frac{2k\pi}{q}\right)=0 \qquad k=\pm1,\pm2\ldots$$

In this case, the result is the second Sheppard correction:

$$E\{x^2\}=E\{x_q^2\}-\frac{q^2}{12}.$$

Generally, the distortion of the second moment is a function of the input represented by Φ_x. The effects of using dither are shown below.

For uniform dither, similarly to (4b):

$$E\{x_q^2\} = E\{x^2\} + \frac{q^2}{12} + \frac{q^2}{12}\left(1 - \frac{1}{2^{2N}}\right) + \frac{q^2}{2\pi^2}\sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \Phi_x\left(\frac{2^{(N+1)}k\pi}{q}\right)\frac{(-1)^k}{(2^N k)^2} + \frac{q}{\pi}\sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \Phi_x\left(\frac{2^{(N+1)}k\pi}{q}\right)\frac{(-1)^{k+1}}{2^N k} - \frac{q}{\pi}\sum_{\substack{k=-\infty \\ k\neq 2^N l}}^{\infty} \Phi_x\left(\frac{2k\pi}{q}\right)\frac{1}{2^N}\frac{q}{2\cdot\sin\left(\frac{k\pi}{2^N}\right)\cdot k} \tag{6b}$$

For triangular dither, similarly to (4c):

$$E\{x_q^2\} = E\{x^2\} + \frac{q^2}{12} + \frac{q^2}{12}\left(2 - \frac{2}{2^{2N}}\right) + \frac{q^2}{2\pi^2}\sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \Phi_x\left(\frac{2^{(N+1)}k\pi}{q}\right)\frac{1}{(2^N k)^2} - \frac{q}{\pi}\sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \Phi_x\left(\frac{2^{(N+1)}k\pi}{q}\right)\frac{1}{2^N k} \tag{6c}$$

The constants $\frac{q^2}{12}\left(1-\frac{1}{2^{2N}}\right)$ and $\frac{q^2}{12}\left(2-\frac{2}{2^{2N}}\right)$ are coused by the variance of the uniform and the triangular dither.

Equations (6b) and (6c) prove the benefits of the triangularly distributed dither. Namely, the last sum in (6b) disappears for $N\to\infty$, while in (5c) it remains.

Knowing the first and the second moment, the variance can be calculated as:

$$\mathrm{var}\{x_q\} = E\{x_q^2\} - E^2\{x_q\} \tag{7}$$

After the substitution $\Phi_x(u) = e^{jux}$ we obtain Fig. 4. The goal is to achive a constant, input independent variance. In case of audio applications, for example, this means the elimination of the fade-out noise, because the variance correspondes to the sound power.

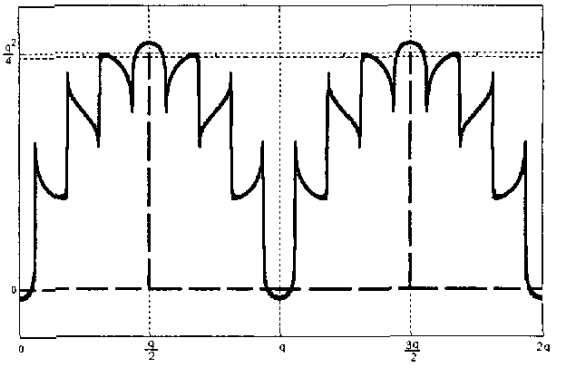

Fig. 4 Effects of various dithers on the variance of the constant analog input. Without dither: dashed, with uniform dither: solid, with triangular dither: dotted.

Without dither the variance is independent of the input apart from the q/2 impulses at the half quantum steps. Using a uniform dither the dependence is *much* greater. (The curve approximates with $N\to\infty$ the analogue, parabolic function.) Benefit of the triangular dither is obvius. At cost of a higher constant variance, the impulses are smaller.

Fig 5. illustrates how this effect varies with the bit-length of the dither. Already one bit increasion couses a distinct improvement.

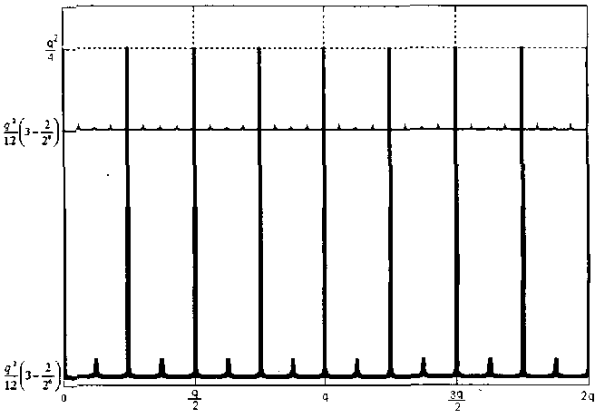

Fig. 5: Variance of a constant analog input signal using triangularly distributed digital dithers.
N=3 : fat, N=4 : thin.
With growing N the variance is less dependent of the input.

When one considers that 8 free bits in accumulator are accessible in many often-used DSP architectures, this result is useful for practical purposes. For example the Motorola DSP560xx series has an accumulator-length of 56 bits and stores the results in 48 bits. The Texas Instruments TMS320C6000 processor family however has a 40/32 bit architecture. Both types have 8 bit overflow protection, which can be used as dither. In other DSP architectures, where these ‚guardig' bits are not avilable, the virtual enlargement of the accumulator has to be done.

Using the above results, if we know the PDF of the input signal, we can give the type and bit-length of a necessary dither to assure the minimal bias for first-order and second-order moments.

VI. SIGNIFICANCE OF RESULTS

We studied analytically and with numerical calculations (with MATLAB) the effects of uniform and triangular discrete dithers in case of different constant input signals. Both types of dithers decrease the bias the same way, but in case of variance the triangularly distributed one is a better choice. We have found that digital triangular dither can be the effective solution to eliminate many undesirable effects, such as round-off errors.

REFERENCES

[1] [1] B. Widrow, I. Kollár and M.-C. Liu, "Statistical Theory of Quantization", IEEE Trans. on Instrumentation and Measurement, Vol. 45 No. 2, pp. 353-61, Apr. 1996.

[2] [2] Sripad, A. B. and Snyder, D. L., "A Necessary and Sufficient Condition for Quantization Errors to Be Uniform and White." IEEE Trans. on Acoustics, Speech and Signal Processing, Vol: ASSP-25, No. 5 pp. 442-448. 1977.

[3] [3] J. Vanderkooy and S. P. Lipshitz, "Dither in Digital Audio", Journal of Audio Engineering Society, Vol. 35, No 12, pp 966-975, December 1987.

[4] [4] S. P. Lipshitz and R. A. Wannamaker, "Quantization and dither: a theoretical survey", Journal of the Audio Engineering Society, Vol. 40, No 5 pp. 335-375, May 1992

[5] [5] I. Kollár, "Bias of Mean Value and Mean Square Value Measurements Based on Quantized Data", IEEE Trans. on Instrumentation and Measurement, Vol. 43, No 5, October 1994.

[6] [6] B. Widrow, "A study of rough amplitude quantization by Means of Nyquist Sampling Theory", IRE. Trans. Circuit Theory, Vol. 3, No. 4, pp 226-276, December 1956

[7] [7] D. T. Sherwood, "Some theorems on quantization and an example using dither", in Conference Record of the 19th Asilomar Conference on Circuits, Systems & Computers, Pacific Grove, CA, Nov. 6-8, 1986, 86CH2331-7, 1986, pp. 207-12.

[8] [8] R. Dunay and I. Kollár, „MATLAB-Based Analysis of Round-off Noise." Periodica Polytechnica Ser. El. Eng., Vol. 43, No. 1, pp. 53-64. 1999.

[9] [9] I. Kollár, "Steady-State Value Measurements with Quantized Data", 2nd IMEKO TC8 Symposium "Theoretical and Practical Limits of Measurement Accuracy", Budapest, May 10-12, 1983, pp. 92-101