D10.9

Effects of FFT Coefficient Quantization on Sinusoidal Signal Detection

Tae H. Joo and Alan V. Oppenheim*

Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, MA 02139

Detection of a sinusoid of unknown frequency in wide band noise is performed efficiently by the FFT. The detector performs a hypothesis test on the magnitude of the FFT output. When the FFT is implemented, errors due to arithmetic roundoff and coefficient quantization limit the accuracy of the transform and degrade the detection performance. When the FFT is used as a detector of an unknown sinusoidal signal, the coefficient quantization error is significant and increases with the FFT length. We analyze the decimation in time, radix-2 FFT. The FFT output error is defined to be the maximum magnitude of the difference between the true FFT and the FFT computed with the quantized coefficients. An upper bound on the error is derived by a deterministic analysis and is verified to be close to the actually measured error. Using the functional form of the bound and scaling it to fit the measured error, an empirical formula for the error is derived. The probability of detection of the quantized-coefficient FFT is computed using the empirical error formula. The probability of detection curves are presented as a function of the FFT length. The simulations indicate that when a sufficient number of bits is used to quantize the coefficients, the probability of detection does not significantly degrade.

1 Introduction

The problem of detecting a weak sinusoid in wide band noise from a received signal of long duration arises in many different contexts, such as the detection of gravity waves[1] and the search for extraterrestrial intelligence[2]. The detection of an unknown complex sinusoid can be performed by applying hypothesis testing to the magnitude of the DFT. The maximum likelihood estimate of the unknown frequency corresponds to the frequency bin with the largest magnitude. Since the DFT effectively operates as a bank of matched filters in which the data length corresponds to the integration time, the probability of detection of the sinusoid increases with the transform length. In detecting gravity waves or searching for extraterrestrial intelligence, transform lengths of 2^{16} or longer are often required.

In implementing these transforms using the FFT, errors due to arithmetic roundoff and coefficient quantization limit the accuracy of the transform and degrade the detection performance. The effect of arithmetic roundoff on the FFT has been analyzed and is well documented[3,4]. In the context of detection of an unknown sinusoidal signal, arithmetic roundoff can be represented as additive white noise in the transform output. Analysis of coefficient quantization on detection performance is less straightforward and is analyzed in this paper. Specifically, we develop a deterministic bound on the error in the transform, which is then used in conjunction with experimental measurements of error to obtain an empirical formula for the error in the output of the FFT due to coefficient quantization.

A brief summary of detection of a complex sinusoidal signal using the DFT is presented in Section 2. Section 3 presents a deterministic, worst case analysis of the FFT output error due to coefficient quantization. We analyze two different implementations of the decimation in time, radix-2 FFT. The first implementation uses a table of precomputed coefficients and the second implementation computes the coefficients recursively. In Section 4, the degradation in probability of detection resulting from coefficient quantization is presented. In Section 5, we summarize our conclusions.

2 Detection of Complex Sinusoidal Signals using the DFT

In discussing the detection of a complex sinusoid in additive white Gaussian noise(WGN), the following model for the received signal, $x(n)$, is used: $x(n) = \begin{cases} Ae^{j\omega_0 n} + w(n) & \text{if signal exist} \\ w(n) & \text{if signal is absent} \end{cases}$ where A and ω_0 are unknown and $w(n)$ is zero mean WGN. For this signal model, the DFT performs as the matched filter for each discrete frequency[5]. The maximum likelihood estimate of ω_0 is the DFT bin with the largest magnitude, if ω_0 is a multiple of $\frac{2\pi}{N}$[6,7].

With $X(k)$ denoting the DFT of $x(n)$ so that

$$X(k) = \sum_{n=0}^{N-1} x(n) W_N^{kn} \tag{1}$$

where $W_N^{kn} = e^{-j\frac{2\pi}{N}kn}$ and with the assumption that $\omega_0 = \frac{2\pi}{N}k_0$, where k_0 is an integer, the probability density function for the magnitude of $X(k)$ is

$$p(z) = \begin{cases} \frac{z}{N\sigma^2} e^{-\frac{z^2}{2N\sigma^2}} u(z) & \text{if } k \neq k_0 \\ \frac{z}{N\sigma^2} I_0\left(\frac{zA}{\sigma^2}\right) e^{-\frac{z^2+(AN)^2}{2N\sigma^2}} u(z) & \text{if } k = k_0 \end{cases} \tag{2}$$

where $z = \sqrt{X_r^2 + X_i^2}$, $I_0(.)$ is the modified Bessel function of zeroth order and $u(z)$ is the unit step function. The probability density function of the magnitude is Rayleigh for $k \neq k_0$ and Rician for $k = k_0$[8].

*This work has been supported in part by the Advanced Research Projects Agency monitored by ONR under Contract No. N00014-81-K-0742, in part by the National Science Foundation under Grant ECS-8407285, in part by Sanders Associates, Inc., and in part by an Amoco Foundation Fellowship.

The density functions for $k = k_0$ and $k \neq k_0$ are increasingly disjoint as N increases and the probability of detection increases. The probability of detection, denoted P_D, and the probability of false alarm, denoted P_F, are derived using (2) and are given by

$$P_F = \int_{\eta}^{\infty} \frac{z}{N\sigma^2} e^{-\frac{z^2}{2N\sigma^2}} dz$$

and

$$P_D = \int_{\eta}^{\infty} \frac{z}{N\sigma^2} I_0(\frac{zA}{\sigma^2}) e^{-\frac{z^2+(AN)^2}{2N\sigma^2}} dz \quad (3)$$

where η is the threshold. For constant false alarm rate, the threshold value is $\eta = \sqrt{-2N\sigma^2 ln(P_F)}$. Therefore, when P_F is fixed, P_D increases as N increases. However, in the presence of coefficient quantization, the error in the transform output increases with increasing N. In the next section, we derive a bound on this error, which we then use in Section 4 to derive the probability of detection as a function of N with coefficient quantization.

3 Coefficient Quantization Noise

Analyses of coefficient quantization effects have been presented by other authors [9,10,11,12,13]. In the following discussion, we consider the decimation in time, radix-2 FFT algorithm. The required coefficients W_N^p, for $p = 0, 1, \cdots, \frac{N}{2} - 1$, can be either precomputed and stored in a table, or recursively computed at each stage of the FFT computation. In the recursive implementation, only $\log_2 N$ complex values must be stored for use as the initial values of the recursion. For large N, this results in significant savings in storage. However, as we show, using a table of $\frac{N}{2}$ precomputed coefficients is more accurate than using recursively computed coefficients.

In the decimation in time FFT, the coefficients W_N^{kn} in (1) are realized through combinations of coefficients associated with smaller length DFT. Specifically, it can be shown that for the decimation in time algorithm, (1) is effectively replaced by

$$X(k) = \sum_{n=0}^{N-1} x(n) W_N^{kb_{0,n}} W_N^{kb_{1,n}2} \cdots W_N^{kb_{M-1,n}2^{M-1}} \quad (4)$$

where $M = \log_2 N$, $n = b_{0,n} + b_{1,n}2 + \cdots + b_{M-1,n}2^{M-1}$, and $b_{i,n} = 0$ or 1 for $i = 0, 1, \cdots, M-1$ (*i.e.* $(b_{M-1,n} b_{M-2,n} \cdots b_{1,n} b_{0,n})$ is the binary representation of n). When the coefficients are quantized, (4) becomes

$$\hat{X}(k) = \sum_{n=0}^{N-1} x(n)(W_N^{kb_{0,n}} + \epsilon_{0,k}) \cdots (W_N^{kb_{M-1,n}2^{M-1}} + \epsilon_{M-1,k}) \quad (5)$$

where the difference between the true and the quantized coefficients is denoted by $\epsilon_{i,k}$. Because there is no quantization error in representing 1 or -1, $\epsilon_{M-1,k} = 0$

In analyzing the error, it is convenient to use matrix notation. Specifically, we express (1) as $\underline{X} = F\underline{x}$ where $\underline{x} = (x(0), x(1), \cdots, x(N-1))^T$, $\underline{X} = (X(0), X(1), \cdots, X(N-1))^T$ and F is the $N \times N$ matrix of coefficients with $(kn)^{th}$ element $f_{kn} = W_N^{kn}$. Correspondingly, (5) is written as $\hat{\underline{X}} = \hat{F}\underline{x}$ where $\hat{F}$ is formed by the quantized coefficients and an error vector is defined as $\underline{e} = \hat{\underline{X}} - \underline{X} = (\hat{F} - F)\underline{x}$. We choose the maximum FFT output error over all frequency bins as the measure for determining the degradation of the probability of detection of a sinusoid. Therefore, the error measure used is the infinity norm of $\underline{e}$:

$$\|\underline{e}\|_\infty = \max_k |e(k)| \quad (6)$$

where $e(k) = \hat{X}(k) - X(k)$. To derive an error measure which is independent of the input $\underline{x}$, we use the inequality

$$\|A\underline{v}\|_\infty \leq \|A\|_\infty \|\underline{v}\|_\infty \quad (7)$$

where A is a matrix, $\underline{v}$ is a vector, and the matrix norm is defined as $\|A\|_\infty = \max_k \sum_{n=0}^{N-1} |a_{kn}|$. Using (7), we then have $\max_k |e(k)| \leq \|\hat{F} - F\|_\infty \|\underline{x}\|_\infty$. The $(k,n)^{th}$ element of the difference matrix $(\hat{F} - F)_{kn}$ is

$$\begin{aligned}(\hat{F} - F)_{kn} &= (W_N^{kb_{0,n}} + \epsilon_{0,k}) \cdots (W_N^{kb_{M-1,n}2^{M-1}} + \epsilon_{M-1,k}) - \\ &\quad W_N^{kb_{0,n}} W_N^{kb_{1,n}2} \cdots W_N^{kb_{M-1,n}2^{M-1}} \end{aligned} \quad (8)$$

Precomputed coefficients: First, we consider using a table of $\frac{N}{2}$ precomputed coefficients. Each coefficient W_N^p is quantized such that $|\epsilon_{i,k}| \leq \sqrt{2}\Delta$ where Δ is $\frac{1}{2}$ of the quantizer step size. We assume that $|\epsilon_{i,k}|$ is small enough such that second and higher order error terms in (8) can be ignored. With this approximation, (8) becomes

$$\begin{aligned}(\hat{F} - F)_{kn} &\approx b_{0,n}\epsilon_{0,k}(W_N^{kb_{1,n}2} W_N^{kb_{2,n}4} \cdots W_N^{kb_{M-1,n}2^{M-1}}) \\ &\quad + b_{1,n}\epsilon_{1,k}(W_N^{kb_{0,n}} W_N^{kb_{2,n}4} \cdots W_N^{kb_{M-1,n}2^{M-1}}) + \cdots \\ &\quad + b_{M-1,n}\epsilon_{M-1,k}(W_N^{kb_{0,n}} \cdots W_N^{kb_{M-2,n}2^{M-2}}) \end{aligned} \quad (9)$$

Applying the triangle inequality to (9) and using the fact that $|\epsilon_{i,k}| \leq \sqrt{2}\Delta$ and $\epsilon_{M-1,k} = 0$, we can write that

$$|(\hat{F} - F)_{kn}| \leq \sqrt{2}\Delta(b_{0,n} + b_{1,n} + \cdots + b_{M-2,n})$$

Consequently,

$$\|\hat{F} - F\|_\infty = \max_k \sum_{n=0}^{N-1} |(\hat{F} - F)_{kn}| \leq \sqrt{2}\Delta \sum_{n=0}^{N-1} b_{0,n} + \cdots + b_{M-2,n}$$

As n ranges from 0 to $N-1$, $b_{i,n}$ for each i will be one for half the terms and zero for the remaining half. Consequently, since $M = \log_2 N$, $\|\hat{F} - F\|_\infty \leq \sqrt{2}\Delta\frac{N}{2}(\log_2 N - 1)$. Therefore, the maximum magnitude output error (6) of the quantized-coefficient FFT using a table of $\frac{N}{2}$ values is bounded by

$$\max_k |e(k)| \leq \sqrt{2}\Delta\frac{N}{2}(\log_2 N - 1)\|\underline{x}\|_\infty \quad (10)$$

Recursively computed coefficients: Next, we derive the error bound for the FFT output for which the coefficients are computed recursively using $\log_2 N$ stored initial values. At the i^{th} stage, the initial value $W_N^{2^{M-i}}$ is used to generate all the required coefficients for that stage. However, because $W_N^{2^{M-i}}$ is quantized, recursive computation increases the FFT output error. We assume that the quantization error for the initial value $W_N^{2^{M-i}}$ is $|\epsilon_{i,1}| \leq \sqrt{2}\Delta$. As the recursion is used to compute the next coefficient, the coefficient quantization error increases linearly. For example, assuming that $\|\epsilon_p^2\| \ll 1$ and using the triangular inequality, the error in computing W_N^{2p} is $\|(W_N^p + \epsilon_p)(W_N^p + \epsilon_p) - W_N^p W_N^p\| \approx \|2\epsilon_p W_N^p\| \leq 2\sqrt{2}\Delta$ If L terms of quantized W_N^ps are multiplied, the error is bounded approximately by $L\sqrt{2}\Delta$ because we assume that the quantization step

size is small enough to ignore non-linear error terms.

Again, applying the triangular inequality to (9) $\|\hat{F} - F\|_\infty = \max_k \sum_{n=0}^{N-1} b_{0,n}|\epsilon_{0,k}| + \cdots + b_{M-1,n}|\epsilon_{M-1,k}|$. The error $|\epsilon_{i,k}|$ obtains its maximum when $k = \frac{N}{2}-1$ such that $|\epsilon_{0,\frac{N}{2}-1}| \leq \sqrt{2}\Delta(\frac{N}{2}-1), |\epsilon_{1,\frac{N}{2}-1}| \leq \sqrt{2}\Delta(\frac{N}{4}-1)$, etc. Because $b_{i,n} = 1$ for $\frac{N}{2}$ terms only and $\log_2 -1$ terms are summed $\|\hat{F} - F\|_\infty \leq \sqrt{2}\Delta\frac{N}{2}((\frac{N}{2} + \frac{N}{4} + \cdots + 1) - M)$. Therefore, the maximum magnitude output error, (6), of the quantized-coefficient FFT using a recursion is bounded by

$$\max_k \|e(k)\| \leq \sqrt{2}\Delta\frac{N}{2}(N - 1 - \log_2 N)\|\underline{x}\|_\infty \qquad (11)$$

The error of the FFT output is proportional to $N\log_2 N$, as shown by (10), when the table of coefficients is used. The error of the FFT output is proportional to N^2, as shown by (11), when the recursion is used. Therefore, although more storage is required, using a precomputed table of coefficients proves to be more accurate.

To verify and measure the closeness of the above derived bounds to the exact FFT output error, (10) is checked by computing $\underline{e}$ explicitly for $\underline{x} = (1, e^{j\omega_0}, \cdots, e^{j\omega_0(N-1)})^T$. The result of the

$\log_2 N$	*MEASURED*	*BOUND*	*PREDICTED*
1	.0	.0	.086
2	.011	.011	.091
3	.022	.044	.107
4	.071	.132	.147
5	.201	.353	.249
6	.536	.883	.494
7	1.228	2.122	1.064
8	2.603	4.949	2.368
9	5.567	11.313	5.301
10	11.360	25.455	11.818
11	26.284	56.568	26.158
12	58.228	124.450	57.443

Table 1: $\|\underline{e}\|_\infty$ values

simulations employing an 8 bit uniform quantizer is shown in Table 1. The values under *MEASURED* are obtained by explicitly searching for the maximum error. The upper bound predicted by (10) is listed under *BOUND*. These simulations indicate that the predicted bound is approximately twice the actually measured values. This suggests that the bound can be scaled to predict the FFT output error. We use the functional form of the bound and incorporate the measurements to derive an empirical formula for the FFT output error. By minimizing the squared error, we solve for α and β to fit a linear model:

$$\alpha\,\Delta\frac{N}{2}(\log_2 N - 1) + \beta = measured\ \|\underline{e}\|_\infty \qquad (12)$$

The least squares solution is $\alpha = .65$ and $\beta = .08$. The values under *PREDICTED* are computed using (12). There is an excellent agreement between the predicted and the measured values particularly for larger value of N. To check the bound for even larger FFT, the error for an FFT of length 2^{16} is computed. Because the FFT length is rather large, our search for ω_0 was limited to a small frequency range. The search for the maximum resulted in $\|\underline{e}\|_\infty = 1356.65$. The upper bound given by (10) is 2715.3 and $\|\underline{e}\|_\infty$ is predicted to be 1251.45 by (12). Figure 1 plots the measured $\|\underline{e}\|_\infty$, the predicted (12), and the bound (10) for 8 bit uniform quantizer.

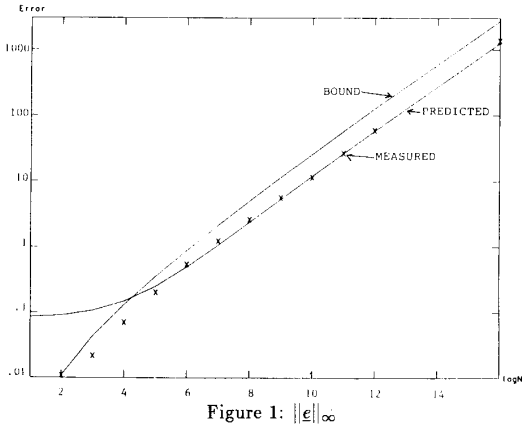

Figure 1: $\|\underline{e}\|_\infty$

4 Probability of Detection

The FFT output error due to coefficient quantization degrades the probability of detection. Each FFT bin has the probability of detection as given by (3). Because the FFT effectively implements a bank of matched filters, the definitions of the probability of detection and the probability of false alarm are modified. We define the probability of detection over all FFT bins, denoted P_d, as deciding that a signal exist at $k = k_0$. The probability of false alarm over all FFT bins, denoted P_f, is defined as deciding that a signal exist at $k \neq k_0$ where the frequency of the signal is $\omega_0 = \frac{2\pi}{N}k_0$.

The quantized-coefficient FFT for input $Ae^{j\omega_0 n} + w(n)$ is given by

$$\hat{X}(k) = \sum_{n=0}^{N-1} Ae^{j\omega_0 n}\widehat{W_N^{kn}} + \sum_{n=0}^{N-1} w(n)\widehat{W_N^{kn}}$$

where $\widehat{W_N^{kn}}$ denotes the quantized coefficients shown in (5). If the coefficients have no quantization error then $\sum_{n=0}^{N-1} Ae^{j\omega_0 n}W_N^{kn} = AN\delta(k-k_0)$. Let $\hat{z} = |\hat{X}(k)|$. For $k \neq k_0$, $|\sum_{n=0}^{N-1} Ae^{\omega_0 n}\widehat{W_N^{kn}}| \leq A\|\underline{e}\|_\infty$. We assume the equality for a conservative P_d estimation.

Therefore the probability density function of $\hat{z}$ is

$$p(\hat{z}) = \frac{\hat{z}}{N\sigma^2}I_0(\frac{\hat{z}}{N\sigma^2}A\|\underline{e}\|_\infty)\exp(-\frac{\hat{z}^2 + A^2\|\underline{e}\|_\infty^2}{2N\sigma^2})u(\hat{z}).$$

For $k = k_0$, $\|\sum_{n=0}^{N-1} Ae^{j\omega_0 n}\widehat{W_N^{kn}}\| \geq A(N - \|\underline{e}\|_\infty)$. Again we choose the equality for a conservative estimate of P_d. Therefore the probability density function of $\hat{z}$ becomes

$$p(\hat{z}) = \frac{\hat{z}}{N\sigma^2}I_0(\frac{\hat{z}}{N\sigma^2}A(N-\|\underline{e}\|_\infty))\exp(-\frac{\hat{z}^2 + A^2(N-\|\underline{e}\|_\infty)^2}{2N\sigma^2})u(\hat{z})$$

The probability of false alarm over all FFT bins is

$$\begin{aligned} P_f &= \int_\eta^\infty \frac{\hat{z}}{N\sigma^2}I_0(\frac{\hat{z}}{N\sigma^2}A\|\underline{e}\|_\infty)\exp(-\frac{\hat{z}^2 + A^2\|\underline{e}\|_\infty^2}{2N\sigma^2})d\hat{z} \\ &= Q(\frac{A\|\underline{e}\|_\infty}{\sqrt{N}\sigma}, \frac{\eta}{\sqrt{N}\sigma}) \end{aligned} \qquad (13)$$

where $Q(.)$ is the Marcum's Q-function[14]. The probability of detection over all FFT bins is

$$P_d = \int_\eta^\infty \frac{\hat{z}}{N\sigma^2} I_0\left(\frac{\hat{z}}{N\sigma^2} A(N - \|\underline{e}\|_\infty)\right) \cdot \exp\left(-\frac{\hat{z}^2 + A^2(N - \|\underline{e}\|_\infty)^2}{2N\sigma^2}\right) d\hat{z}$$

$$= Q\left(\frac{A(N - \|\underline{e}\|_\infty)}{\sqrt{N}\sigma}, \frac{\eta}{\sqrt{N}\sigma}\right) \quad (14)$$

We use the empirical formula (12) derived by assuming an 8 bit uniform quantizer in (13) and (14). Figure 2 shows P_d, given by (14), as a function of FFT length. It is generated using $A = 0.1$, $\sigma^2 = 1.0$, and a constant false alarm rate of $P_f = 0.01$, given by (13). This figure indicates that even though the error $\|\underline{e}\|_\infty$ increases as the data length increases, P_d improves also. The simulation shows that 8 bit quantization only slightly degrades the probability of detection.

5 Conclusions

We derived a bound for the FFT output error when the FFT coefficients are quantized. This deterministic analysis demonstrates that using a table of precomputed coefficients is more accurate than recursively computing the coefficients. Simulations were performed to verify the bound. We used the measured values to

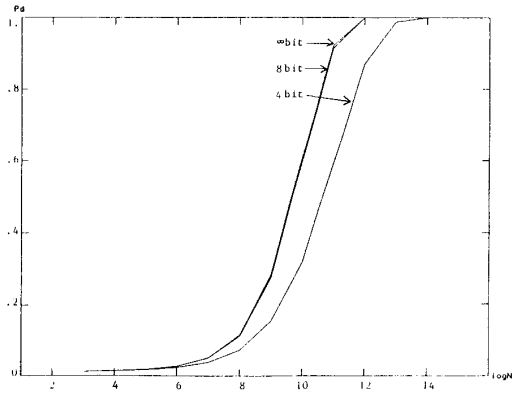

Figure 2: P_d vs N

scale the bound to to derive an FFT output error prediction formula. Using the error formula, the probability of detection and the probability of false alarm were derived. We showed that if the coefficients are quantized using a large number of bits, for instance 8 bits, then the degradation in the probability of detection is minimal.

References

[1] J. Livas, *Upper Limits for Gravitational Radiation from some Astrophysical Sources,* Ph.D Thesis, Massachusetts Institute of Technology, 1987.

[2] P. Horowitz, J. Forster, and I. Linscott, "The 8-Million Channel Narrowband Analyzer," pp. 361-371, *The Search for Extraterrestrial Life: Recent Developments,* M.D. Papagiannis(*ed.*), 1985.

[3] C. Weinstein and A. Oppenheim, "Effects of Finite Register Length in Digital Filtering and the Fast Fourier Transform," *Proceedings of IEEE,* vol.60, pp.957-976, 1972.

[4] A. Oppenheim and R. Schafer, *Digital Signal Processing,* McGraw-Hill Book Company, 1980.

[5] A. Viterbi, *Principles of Coherent Communication,* McGraw-Hill Book Company, 1966.

[6] D. Rife and R. Borystein, "Single-Tone Parameter Estimation from Discrete-Time Observations," *IEEE Trans. Info. Theory,* vol IT-20, no.5, pp.591-598, Sept., 1974.

[7] L. Palmer, "Coarse Frequency Estimation using the Discrete Fourier Transform," *IEEE Trans. Info. Theory,* vol IT-20, pp.104-109, Jan., 1974.

[8] A. Papoulis, *Probability, Random Variables, and Stochastic Processes,* McGraw-Hill Book Company, 1984.

[9] W. Knight and R. Kaiser, "A Simple Fixed-Point Error Bound for the Fast Fourier Transform," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-27, no.6, pp.615-620, Dec., 1979.

[10] D.James, "Quantization Errors in the Fast Fourier Transform," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-23, no.3, pp.227-283, June, 1975.

[11] D. Tufts, H. Hersey, and W. Mosier, "Effects of FFT Coefficient Quantization on Bin Frequency Response," *Proc. IEEE(Lett.),* vol. 60, pp.146-147, Jan. 1972.

[12] U. Heute, "Results of a Deterministic Analysis of FFT Coefficient Errors," *Signal Processing,* vol.3, pp.321-331, 1981.

[13] U. Heute and H. Schuessler, "FFT-Accuracy - New Insights and a New Point-of-View," *Proc. of ICASSP,* pp. 631-634, Boston, MA., 1983.

[14] H.L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I,* Wiley, 1968.