On the Uniform ADC Bit Precision and Clip Level Computation for a Gaussian Signal

Naofal Al-Dhahir *, Member, IEEE, and John M. Cioffi, Senior Member, IEEE
Stanford University, Stanford CA 94305†

Abstract

The problem of computing the required bit precision of analog–to–digital converters is revisited with emphasis on Gaussian signals. We present two methods of analysis. The first method fixes the probability of overload and sets the dynamic range of the quantizer to accommodate the worst–case signal–to–quantization noise ratio. The second method sets the clipping level of the quantizer to meet a desired overload distortion level, using knowledge of the input probability density function. New closed–form expressions relating the distortion–minimizing clip level of the uniform quantizer and the input bit rate are derived and shown to give remarkably close results to the optimum ones obtained using numerical iterative procedures in [6, 9, 10].

I. Introduction

Amplitude quantization refers to the transformation of an analog sample that can take a continuum of values into one that can only assume a finite set of levels. Inherent in the quantization process is the introduction of quantization noise that has two components: *granular* noise and *overload* distortion due to quantization errors that arise when the input signal amplitude lies within or outside the maximum input range of the quantizer, respectively.

In this correspondence, the focus on Gaussian input signals is motivated by our interest in **multi-carrier modulation** which has emerged over the past few years as an excellent modulation scheme that achieves high performance at a practical implementation complexity [1]. Briefly, the multicarrier signal is the sum of N (a large number typically 256 or 512) independent low–rate Quadrature Amplitude Modulated (QAM) subsignals (or sometimes called subchannels). Although the number of bits

---

*N. Al-Dhahir is now with GE R & D Center, Schenectady, NY 12301. Tel : (518) 387-7092, FAX : (518) 387-4042, Email : aldhahir@birch.crd.ge.com

†This work was supported by NSF under Contract NCR-9203131, JSEP under Contract DAAL 03-91-C-0010, and NASA under Contract NAG2-842.

and the input power allocated to each subchannel can be different in order to maximize the overall data rate, we shall assume identical QAM–constellation sizes and a flat power distribution across all subchannels.[1] Direct application of the *Central Limit Theorem* implies that the probability density function (pdf) of a multicarrier signal can be considered for all practical purposes to be approximately **Gaussian**. In other words,

$$P_{v_{in}}(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \ , \tag{1}$$

where, without loss of generality, we have assumed that the input signal has unity power.

In Section II, we present two methods for computing the bit precision of the uniform quantizer for a multicarrier input signal. The problem of optimizing the clip level (to minimize the total quantization distortion) of the uniform quantizer was first solved in [6] by resort to iterative numerical procedures. In Section III, under the common assumption of uniformly–distributed granular noise, we present new closed–form analytical expressions that demonstrate very accurately the complicated functional interdependence between the optimum clip level and the input bit rate (in bits/symbol). Details of the derivation are given in Appendices A and B.

II. Bit Precision Analysis

Three sources contribute to the required number of the analog–to–digital converter (ADC) bits for a multicarrier signal. The first source is information bits that achieve a desired probability of error level in the presence of additive white Gaussian noise (AWGN). Secondly, extra quantization bits are needed to keep the quantization noise at a factor of α ($\alpha > 1$) times smaller than the AWGN level. Finally, additional bits are added to accommodate the high peak–to–average ratio (PAR) of the multicarrier input signal.

II-A. First Method

For an M-point QAM signal constellation, from which a multicarrier subsignal is derived, the signal–to–noise ratio (SNR) required to support $\overline{b}_{info}$ information bits per dimension ($M = 2^{2\overline{b}_{info}}$) at a probability of error, P_e, is given by [5]

$$SNR \approx \frac{(2^{2\overline{b}_{info}} - 1)}{3} (Q^{-1}(\frac{P_e}{4(1 - 2^{-\overline{b}_{info}})}))^2 \ , \tag{2}$$

where $Q(x) \stackrel{def}{=} \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{v^2}{2}} dv$.

[1]It was shown in [7] that this assumption simplifies the analysis significantly and serves as a good approximation to the general case of non–uniform bit and power distributions across the subchannels.

For the mid–rise uniform quantizer shown in Figure 1, if we assume $\overline{b}$ output bits per dimension, then it is a simple matter to show that the SNR at the output of the uniform quantizer is given by [4]

$$SNR_q = 6\overline{b} + 4.77 - 20\log_{10}(K_{OL}) \quad : \quad \text{in dB} \; . \tag{3}$$

Using this method, the PAR factor, K_{OL}, is set large enough such that the probability that the magnitude of the input sample exceeds K_{OL}, known as the probability of overload P_{OL}, is very small (should be less than P_e) [2].

Finally, we request that the overall introduced quantization noise level is at a level of α ($\alpha > 1$) times smaller than all other noise (or less). This ensures that quantization effects do not degrade the SNR by more than $10log_{10}(1+\frac{1}{\alpha})$ dB. Putting it all together, the minimum total number of ADC bits, $\overline{b}^{(1)}_{min}$, must satisfy

$$6\overline{b}^{(1)}_{min} + 4.77 = 10\log_{10}(SNR) + 10log_{10}(\alpha) + 20log_{10}(K_{OL}) \; . \tag{4}$$

Example 1 *If we restrict the maximum number of information bits per dimension to be 4 for each subchannel at $P_e = 10^{-7}$ then $SNR \approx 34$ dB. Moreover, for a probability of overload $P_{OL} = 2Q(K_{OL}) = 10^{-9}$, we have PAR=$20\log_{10}(K_{OL}) = 15.72$ dB. In this case, and under a 0.1 dB maximum SNR degradation due to overload distortion ($\alpha = 40$), we get*

$$\overline{b}^{(1)}_{min} = \frac{1}{6}(34 + 16 + 15.72 - 4.77) = 10.16 \;\; bits \; .$$

II-B. Second Method

The second approach uses knowledge of the pdf of the input signal to set the overload distortion power, denoted by σ^2_{OL}, exactly at a pre–determined level. More precisely, the desired clip level, c, is determined as a function of σ^2_{OL} by solving the integral equation

$$\begin{aligned} \sigma^2_{OL} &= 2\int_c^{\infty}(u - y_{max})^2 P_{v_{in}}(u)du \\ \Rightarrow \sigma^2_{OL}\;(dB) &= 3 + 10\log_{10}\left(\frac{1}{\sqrt{2\pi}}\int_c^{\infty}(u - y_{max})^2 e^{-\frac{u^2}{2}}du\right) \; , \end{aligned} \tag{5}$$

where y_{max} is the maximum output level of the uniform quantizer and is related to c by (c.f. Figure 1)

$$y_{max} = c - \frac{c}{2^{\overline{b}}} = c(1 - 2^{-\overline{b}}) \; . \tag{6}$$

In Appendix A, we solve (5) analytically and obtain

$$\sigma^2_{OL}\;(dB) = 3 + 10\log_{10}\{(c^2(1 - 2^{-\overline{b}})^2 + 1)Q(c) - 2ce^{-\frac{c^2}{2}}(\sqrt{2}\pi(1 - 2^{-\overline{b}}) - \frac{1}{\sqrt{2\pi}})\} \; . \tag{7}$$

Figure 2 depicts the variation of the overload distortion power as a function of the clip level of the uniform quantizer for various input bit rates (in bits/symbol). In order to keep the total quantization noise power level $10\log_{10}(\alpha)$ dB below the SNR at a desired P_e, we need $\overline{b}_q$ bits per dimension determined from

$$6\overline{b}_q + 4.77 = 10\log_{10}(\alpha) + 20\log_{10}(Q^{-1}(P_e)) \; , \tag{8}$$

where the second term is the SNR required to maintain *binary* transmission at a probability of error P_e. Therefore, the minimum number of ADC bits required using this method is given by

$$\overline{b}^{(2)}_{min} = \overline{b}_{info} + \overline{b}_q + log_2(c) \; . \tag{9}$$

To summarize, the parameters P_e, α, and $\overline{b}_{info}$ are common to both methods. The first method assumes a certain value for the PAR of the multicarrier input signal (which fixes P_{OL}) while the second method assumes a certain value for the overload distortion power (which fixes the clip level).

Example 2 *Using the same parameters of Example* 1*, we get (using (8))*

$$6\overline{b}_q + 4.77 = 16 + 14.32 \Rightarrow \overline{b}_q = 4.26.$$

In setting the clip level, σ^2_{OL} should be made low enough such that the probability of clipping is less than P_e. *Using Figure* 2*, for* $\sigma^2_{OL} = -80$ *dB,* $c \approx 5.57$ *and* $\overline{b}^{(2)}_{min} = 4 + 4.26 + log_2(5.57) = 10.74$ *bits, which is a more conservative design than that of Example* 1.

Figure 3 depicts the variation of $\Delta\overline{b} \stackrel{def}{=} \overline{b}^{(1)}_{min} - \overline{b}^{(2)}_{min}$ with the probability of clipping (at $P_e = 10^{-7}$ and $\overline{b}_{info} = 4$), while the PAR is varied between 14 and 17 dB.

III. Clip Level Optimization

Although the problem of optimizing the clip level (or equivalently the step size) of the uniform quantizer to minimize the overall quantization noise distortion has been analyzed before for uniform and non–uniform quantizers and for a number of input distributions in [6, 9], *no closed–form analytical solutions were given (except in very special cases).* Instead, the solutions were obtained using a computationally-intensive iterative numerical procedure and the optimum step size together with the corresponding minimum quantization distortion were tabulated versus the input bit rate. In [10], Roe derived a very accurate analytical expression relating the number of output levels of the uniform quantizer to the input bit rate through the use of Taylor series approximations and the Euler–McClaurin summation formula and assuming a large number of output levels. Unfortunately, Roe's expressions were not presented in a useful form for direct evaluation. We are also aware of some recent work on the ADC bit requirements for multicarrier signals in Asymmetric Digital Subscriber Line (ADSL) applications [7]. There, the authors set the clip level such that the same SNR is obtained as when clipping is not applied.

Under the *common assumption of uniformly–distributed granular noise* [4], we show in Appendix B that the optimum clip level, $c_{opt.}$, of the uniform quantizer is related to the number of quantizer output levels, $2^{\overline{b}}$, by the quadratic equation

$$A_1(2^{\overline{b}})^2 + A_2(2^{\overline{b}}) + A_3 = 0 \ , \tag{10}$$

where the A_is are completely determined by $c_{opt.}$ (see Appendix B, Equations (18)-(20)) and $\overline{b}$ is calculated from the *unique* physically meaningful solution of (10).[2]

Figure 4 depicts $c_{opt.}$ as a function of $\overline{b}$ as obtained from (10); also shown on the same plot are Roe's approximate results and the optimum clip level calculated numerically in [6]. It is clear that all three curves are remarkably close even for low $\overline{b}$. It is worth mentioning that we had to perform extensive computations to simplify and evaluate Roe's expressions to be able to compare them with our closed–form expressions. This comparison revealed a remarkable closeness between the two expressions although they were derived using different approaches.

Once $c_{opt.}$ is found, the optimum step size $\Delta_{opt.}$ is calculated using

$$\Delta_{opt.} = \frac{c_{opt.}}{2^{\overline{b}-1}} \ . \tag{11}$$

Table I lists the step sizes and the maximum signal–to–quantization noise ratios (SQNR) achievable for $\overline{b}$ varying from 1 to 8. Also listed are the optimum values determined using the iterative numerical techniques of [6]. It is clear that our analytical approach yields a very accurate estimate of the optimum step size for all input bit rates. It is also worth mentioning that our assumption of a uniformly–distributed granular noise (c.f. (15)) results in a slightly optimistic estimate of SQNR for high input bit rates ($\overline{b} = 7, 8$ in Table I).

We conclude this section by emphasizing that while the resulting probability of clipping (when the clip level is chosen to minimize the total quantization distortion) might be acceptable for speech applications, it is unacceptable for data transmission applications where $P_e < 10^{-6}$. Figure 5 clearly illustrates this fact.

IV. Conclusions

Two approaches for computing the bit precision of the uniform quantizer for a multicarrier input signal were presented and compared. Under the well–justified assumption of uniformly–distributed granular noise, new analytic expressions relating the distortion–minimizing clip level of the uniform quantizer to the input bit rate (in bits/symbol) were derived for a multicarrier input signal (Gaussian

[2] It can be checked that for the range of clip levels (or equivalently probabilities of clipping) considered in this paper, $\frac{A_3}{A_1} < 0$ which implies that one of the roots of (10) will be negative and should be discarded (since $2^{\overline{b}} > 0$).

pdf). The high accuracy of the derived expressions was verified by comparing them with the optimum solutions obtained through iterative numerical techniques.

For data transmission applications that require high reliability levels, we showed that minimizing the overall quantization distortion is not an appropriate criterion by itself for selecting the clip level; the resulting probability of clipping must also be taken into account.

Appendix

A. Evaluating the Integral

In this appendix, we derive Equation (7). Towards this end, consider the integral

$$\begin{aligned} I &\stackrel{def}{=} \frac{1}{\sqrt{2\pi}} \int_c^\infty (u - y_{max})^2 e^{-\frac{u^2}{2}} du \quad : \text{where } y_{max} = c - \frac{c}{2^{\bar{b}}} \\ &= \frac{y_{max}^2}{\sqrt{2\pi}} \int_c^\infty e^{-\frac{u^2}{2}} du - \frac{2y_{max}}{\sqrt{2\pi}} \int_c^\infty u e^{-\frac{u^2}{2}} du + \frac{1}{\sqrt{2\pi}} \int_c^\infty u^2 e^{-\frac{u^2}{2}} du \\ &\stackrel{def}{=} I_1 - I_2 + I_3 \ . \end{aligned}$$

$$\text{Now, } \quad I_1 = \frac{y_{max}^2}{\sqrt{2\pi}} \int_c^\infty e^{-\frac{u^2}{2}} du \stackrel{def}{=} c^2(1 - 2^{-\bar{b}})^2 Q(c) \ . \tag{12}$$

$$I_2 = \frac{2y_{max}}{\sqrt{2\pi}} \int_c^\infty u e^{-\frac{u^2}{2}} du = \frac{2y_{max}}{\sqrt{2\pi}} \int_{\frac{c^2}{2}}^\infty e^{-v} dv = \sqrt{\frac{2}{\pi}} c(1 - 2^{-\bar{b}}) e^{-\frac{c^2}{2}} \ . \tag{13}$$

For the evaluation of I_3 we use the integral identity (A8) in [8]

$$\int u^k e^{-u^2} du = \frac{e^{-u^2}}{2} \sum_{j=0}^{r-1} \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k+1}{2} - j)} u^{k-2j-1} + (1-s)\Gamma(r + \frac{1}{2})(\frac{1}{2} - Q(\sqrt{2}u)) \ ,$$

where $k = 2r - s$, $s \in \{0, 1\}$, and $\Gamma(n) \stackrel{def}{=} \int_0^\infty x^{n-1} e^{-x} dx$. For our case, $k = 2$, $r = 1$, and $s = 0$.

After some algebraic manipulation, we get

$$I_3 = \frac{c e^{-\frac{c^2}{2}}}{\sqrt{2\pi}} + Q(c) \ . \tag{14}$$

From (12),(13), and (14) we have

$$I = (c^2(1 - 2^{-\bar{b}})^2 + 1) Q(c) - c e^{-\frac{c^2}{2}} (\sqrt{\frac{2}{\pi}} (1 - 2^{-\bar{b}}) - \frac{1}{\sqrt{2\pi}}) \ ,$$

and (7) follows immediately.

B. Optimizing the Clip Level

The total quantization noise variance consists of two components

- Granular noise (assumed uniformly–distributed) whose variance *increases* with *increasing* clip level and is given by [4]

$$\sigma_g^2 = \frac{c^2}{3} 4^{-\overline{b}} \ . \tag{15}$$

- Overload distortion whose variance *decreases* with *increasing* clip level and is given by

$$\sigma_{OL}^2 = 2 \int_c^{\infty} \frac{1}{\sqrt{2\pi}} (u - c(1 - 2^{-\overline{b}}))^2 e^{-\frac{u^2}{2}} du \ .$$

Using the results of Appendix A, the total quantization noise variance is given by

$$\sigma_q^2 = \sigma_g^2 + \sigma_{OL}^2 = \frac{c^2}{3} 4^{-\overline{b}} + 2(c^2(1 - 2^{-\overline{b}})^2 + 1) Q(c) - 2ce^{-\frac{c^2}{2}} \left(\sqrt{\frac{2}{\pi}}(1 - 2^{-\overline{b}}) - \frac{1}{\sqrt{2\pi}}\right) . \tag{16}$$

Differentiating σ_q^2 with respect [3] to c and setting the result to zero we obtain after some rearrangement

$$K_1 c_{opt.} + K_2 c_{opt.} Q(c_{opt.}) + K_3 c_{opt.}^2 e^{-\frac{c_{opt.}^2}{2}} + K_4 e^{-\frac{c_{opt.}^2}{2}} = 0 \ , \tag{17}$$

where,

$$K_1 = \frac{2}{3} 4^{-\overline{b}} \quad ; \quad K_2 = 4(1 - 2^{-\overline{b}})^2 \ .$$

$$K_3 = -\sqrt{\frac{2}{\pi}} (1 + (1 - 2^{-\overline{b}})^2 - 2(1 - 2^{-\overline{b}})) = -\sqrt{\frac{2}{\pi}} 4^{-\overline{b}} \quad ; \quad K_4 = -\sqrt{\frac{8}{\pi}} (1 - 2^{-\overline{b}}) \ .$$

Instead of solving (17) numerically (using very accurate analytical approximations of the $Q(.)$ function reported in [3]), it can be easily checked that (17) can be recast as a quadratic equation in $2^{\overline{b}}$ in the form of (10) where

$$A_1 = 4c_{opt.} Q(c_{opt.}) - \sqrt{\frac{8}{\pi}} e^{-\frac{c_{opt.}^2}{2}} \ . \tag{18}$$

$$A_2 = \sqrt{\frac{8}{\pi}} e^{-\frac{c_{opt.}^2}{2}} - 8c_{opt.} Q(c_{opt.}) \ . \tag{19}$$

$$A_3 = \frac{2}{3} c_{opt.} + 4c_{opt.} Q(c_{opt.}) - \sqrt{\frac{2}{\pi}} c_{opt.}^2 e^{-\frac{c_{opt.}^2}{2}} \ . \tag{20}$$

---

[3] Note that

$$\frac{d}{dc} Q(c) \stackrel{def}{=} \frac{d}{dc} \int_c^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du = \frac{-1}{\sqrt{2\pi}} e^{-\frac{c^2}{2}} \ .$$

These expressions are identical to those of Roe [10] *except* for a slight deviation in the equation for [4] A_3, which becomes negligible as $\overline{b}$ becomes large (since the assumption of uniformly–distributed granular noise becomes more accurate then).

For the special case of a high input bit rate ($2^{-\overline{b}} \ll 1$), (17) reduces to

$$(\frac{2}{3}4^{-\overline{b}})c_{opt.} + 4c_{opt.}Q(c_{opt.}) - \sqrt{\frac{8}{\pi}}e^{-\frac{c^2}{2}} = 0 \ ,$$

which implies that

$$\overline{b} = -\frac{1}{2}\log_2(3\sqrt{\frac{2}{\pi}}\frac{e^{-\frac{c^2_{opt.}}{2}}}{c_{opt.}} - 6Q(c_{opt.}))\ .$$

References

[1] J.A.C. Bingham. Multicarrier Modulation for Data Transmission: An Idea Whose Time Has Come. *IEEE Communications Magazine*, pages 5–14, May 1990.

[2] J.A.C. Bingham. Private Communication. Amati Communications Corporation, September 1992.

[3] P.O. Borjesson and C.E. Sundberg. Simple Approximations of the Error Function $Q(x)$ for Communications Applications. *IEEE Transactions on Communications*, COM-27(3):639–643, March 1979.

[4] N.S. Jayant and P. Noll. *Digital Coding of Waveforms.* Prentice-Hall, 1984.

[5] E. Lee and D. Messerschmitt. *Digital Communication.* Kluwer Academic Publishers, 1988.

[6] J. Max. Quantizing for Minimum Distortion. *IRE Transactions on Information Theory*, IT-6:7–12, March 1960.

[7] D. Mestdagh, P. Spruyt, and B. Biran. Analysis of Clipping Effect in DMT–based ADSL Systems. In *International Conference on Communications*, pages 293–300, May 1994.

[8] E.W. Ng and M. Geller. A Table of Integrals of the Error Functions. *Journal of Research, National Bureau of Standards*, 73B:1–20, January-March 1969.

[9] M.D. Paez and T. H. Glisson. Minimum Mean–Squared–Error Quantization in Speech PCM and DPCM Systems. *IEEE Transactions on Communications*, pages 225–230, April 1972.

[10] G. M. Roe. Quantizing for Minimum Distortion. *IEEE Transactions on Information Theory*, IT-10:384–385, October 1964.

[4] The second and third terms in (20) are multiplied by a factor of $\frac{2}{3}$ in Roe's expression.

List of Figures

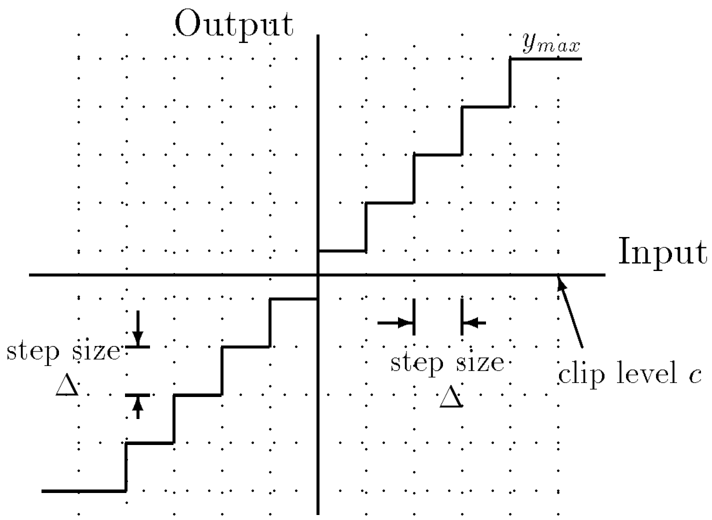

Fig. 1. Input-output characteristics of a mid-riser uniform quantizer

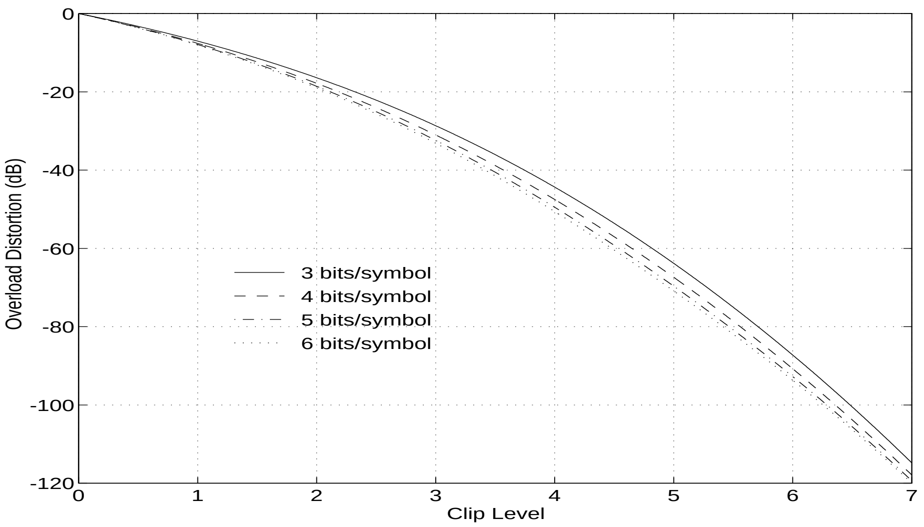

Fig. 2. Overload distortion vs. clip level for a unit–power Gaussian input signal

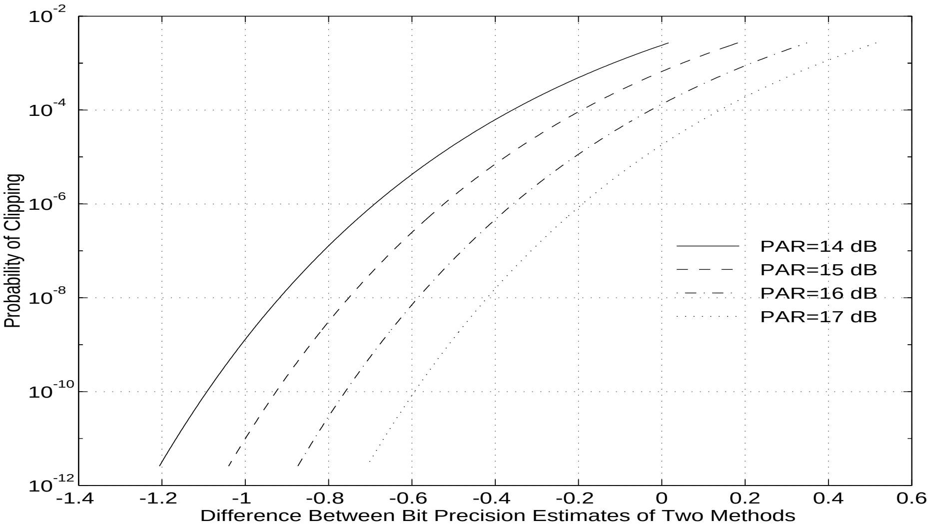

Fig. 3. Difference between the bit precision estimates of Methods 1 and 2

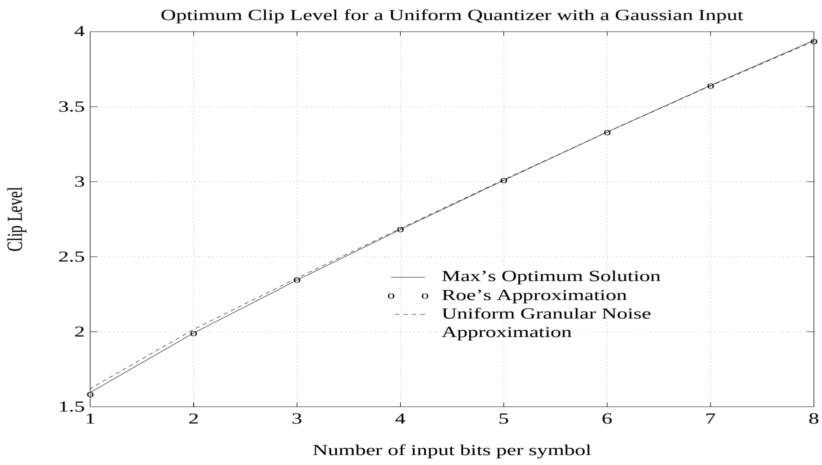

Fig. 4. Optimum clip level of uniform quantizer vs. bit rate (in bits/symbol)

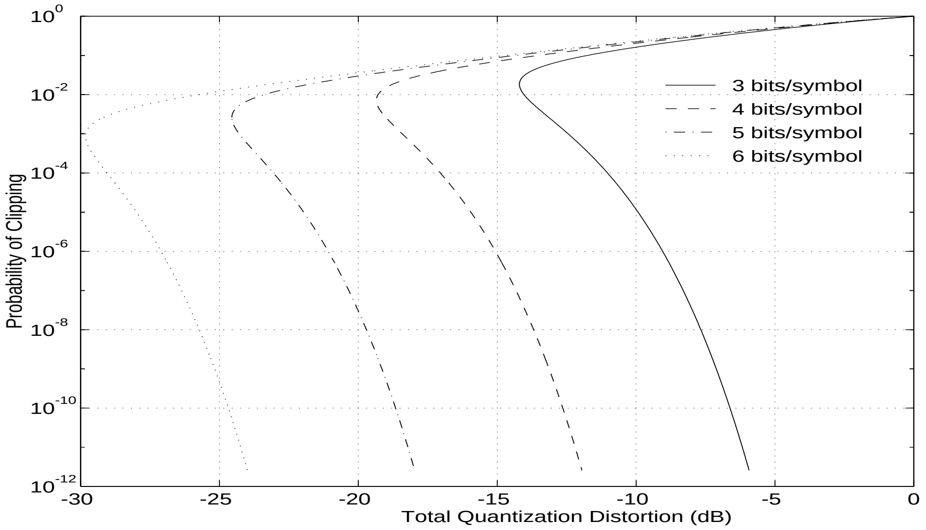

Fig. 5. Variation of the probability of clipping with total quantization distortion

$\overline{b}$ (bits/symbol)	Δ from (11)	$\Delta_{opt.}$ from [6]	$SQNR$ (dB) from (16)	$SQNR_{opt.}$ (dB) from [6]
1	1.621	1.5956	4.0564	4.4
2	1.008	.9957	9.0939	9.25
3	0.5895	.586	14.2	14.27
4	0.336	.3352	19.35	19.38
5	0.1883	.1881	24.554	24.57
6	0.1041	.1041	29.821	29.83
7	.0569	.0569	35.152	35.13
8	0.0308	.0308	40.542	40.34

Table I. Comparison between our approximate analytical solution and the optimum one for a Gaussian input signal