THE INFLUENCE OF QUANTIZATION AND FIXED POINT ARITHMETIC UPON THE BER PERFORMANCE OF TURBO CODES

Yufei Wu, Brian D. Woerner

Mobile and Portable Radio Research Group
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061-0350, USA
e-mail: yufei@vt.edu

Abstract – The majority of the performance studies of turbo codes presented in the literature to date have assumed the use of floating point arithmetic. However, if fixed point arithmetic is employed, a corresponding degradation in the BER performance of the turbo decoding algorithm is expected. In this paper [1] , the effects of quantization and fixed point arithmetic upon the Log-MAP and APRI-SOVA decoding algorithms for a BPSK communication system are quantified via simulation. It is shown that, with proper scaling of the signal prior to quantization, no degradation of the BER performance is incurred with eight-bit quantization, and even four-bit quantization can provide acceptable BER performance.

I. INTRODUCTION

After turbo codes were first introduced in [1], researchers have turned their attention to the design, performance and application of turbo codes. Two major classes of soft-input/soft-output decoding algorithms exist: (1) the maximum *a posteriori* (MAP) estimator and its suboptimal forms (Log-MAP and Max-Log-MAP) [2] [3], and (2) the *a priori* soft output Viterbi algorithm (APRI-SOVA) and its variants [3] [4]. To the authors' knowledge, the analysis of these algorithms in the published literature has primarily assumed the use of floating point arithmetic.

However, in a real-time decoder, it is likely that the decoding algorithms would need to be implemented using fixed point arithmetic. The amplitude of both signals and coefficients in the decoder is discrete. Fixed point arithmetic and quantization inevitably add noise to the system. Adequate signal-to-noise ratio (SNR) at the quantizer and throughout the whole process are essential for decoder implementation. A simple solution is to increase the signal levels since the rounding noise level is fixed for a given structure. However, the signal level cannot be increased too much, otherwise the dynamic range of the quantizer and the fixed point arithmetic will be exceeded and overflow follows. Thus a balanced scaling factor, or optimal gain, is to be found for the signal before it is fed into the decoding processor.

Motivated by the above considerations, we investigated the dynamic range adjustment, and the influence of quantization and fixed point arithmetic for the implementation of the turbo decoder. Two typical decoding algorithms were experimented with: Log-MAP and APRI-SOVA. The remainder of the paper is organized as follows. In Section II, the system model and the simulation configuration are introduced. In Section III, the optimal gain for a quantizer of a certain resolution is obtained for low to medium E_b/N_0. In Section IV, the simulation results using the Log-MAP and APRI-SOVA algorithms are presented. Conclusions are presented in section V.

II. SYSTEM MODEL

A block diagram of the communication system considered in this paper is shown in Figure 1. Information bits $d_k = 0$ or $d_k = 1$ with equal probability. The turbo encoder consists of two identical parallel recursive systematic convolutional (RSC) encoders having rate 1/2, constraint length 3, and octal generators 7 (feedback) and 5 (feed forward). d_k is grouped into frames of length 1022 bits for encoding.

[1]This work has been supported by the MPRG Industrial Affiliates Foundation and the Defense Advanced Research Projects Agency (DARPA) through the GloMo program.

Two tail bits are added at the end of each frame to return the first RSC encoder to all-zero state, while leaving the state of the second encoder open. A random interleaver is used in between the two constituent encoders. The two parity bit streams are alternatively punctured [1] to increase the overall code rate to 1/2 before BPSK modulation.

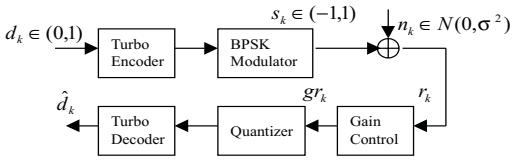

Figure 1: System model.

The coded bits are BPSK modulated into bipolar signal $(-1, +1)$ and transmitted through the channel. Additive white Gaussian noise (AWGN) noise $n_k \in N(0, \sigma^2)$ is added to the transimitted signal by the channel. After the BPSK demodulation, the received bits are scaled by a factor g before they are digitized by the quantizer.

In simulation of the quantization, the continous signal within the voltage range (V_{low}, V_{hi}) is mapped to integer numbers between $(-2^{n-1}, 2^{n-1} - 1)$, where n is the number of bits of resolution of the quantizer [5]. Usually $V_{low} = -V_{hi}$, and 0 is the center of a bin. The quantization is realized by dividing the region between the voltages (V_{low}, V_{hi}) into 2^n evenly spaced bins. These bins are numbered between -2^{n-1} and $2^{n-1}-1$, inclusive. The bin width is $\delta = (V_{hi} - V_{low})/2^n$, and the bin boundaries are at $\{-\infty; V_{low} + \delta/2; \cdots; V_{low} + (2m-1)\delta/2; \cdots; V_{hi} - 3\delta/2; +\infty\}$, $m = 1, ..., 2^n - 1$. For each continuous input sample, a search is performed to identify the bin in which the sample lies, and the corresponding integer number will be used as the quantized value of the input.

The phenomenon modeled in this paper is the number of quantization bits available. Thus there was no attempt to optimize the number of bits available in the internal data path, which is usually higher than the number of quantization bits. All the intermediate results are integers which have length less than 32 bits.

Two turbo decoding algorithms, Log-MAP and APRI-SOVA, are used to decode the quantized signal. In the Log-MAP algorithm, the computations are performed in log arithmetic and the Jacobian logarithm is used [2]:

$$\ln(e^{\delta_1} + e^{\delta_2}) = \max(\delta_1, \delta_2) + f_c(|\delta_1 - \delta_2|)$$

where $f_c(x) = \ln(1 + e^{-x})$ is a nonlinear correction function. All the other compuations with Log-MAP can be performed in the same manner as with the floating point arithmetic, except $f_c(\cdot)$. Due to the use of fixed point arithmetic, the quantities δ_1 and δ_2 in the Log-MAP algorithm can be thought of as having been scaled from their floating point counterparts by a factor of approximately $2^n/(V_{hi} - V_{low})$. In our simulation, the floating point $f_c(\cdot)$ was implemented using the fixed point counterpart $f'_c(\cdot)$. This necessitated the quantity $|\delta_1 - \delta_2|$ to be first scaled to $|\delta'_1 - \delta'_2|$ as follows:

$$|\delta'_1 - \delta'_2| = \frac{|\delta_1 - \delta_2|(V_{hi} - V_{low})}{2^n\, g}, \qquad (1)$$

where g is the scaling factor. The look-up table for $f'_c(\cdot)$ was constructed by computing the nonlinear correction function using the value of $|\delta'_1 - \delta'_2|$ as follows:

$$f'_c(|\delta_1 - \delta_2|) = \text{Int}\left[\frac{2^n\, g\, f_c(|\delta'_1 - \delta'_2|)}{V_{hi} - V_{low}}\right], \qquad (2)$$

where Int$[\cdot]$ stands for the operation of rounding to nearest integer.

In APRI-SOVA, all the computations are linear. Therefore, there is no need to adjust the operations of the algorithm to accomodate the fixed point arithmetic.

III. OPTIMAL GAIN

For the system in Figure 1, to achieve the best signal-to-distortion ratio at the quantizer, the amplitude of the signal is adjusted by the gain control. For a given input distribution, there exists an optimal scaling factor which would minimize the distortion, or maximize the signal-to-distortion ratio, of the quantizer.

Here the optimal gain is derived for BPSK signals passing through an AWGN channel. Assume the BPSK signal constellation is $(+1, -1)$ with equal probability, the pdf (probability distribution function) of the signal s is: $p_s(s) = 0.5[\delta(s+1)+\delta(s-1)]$. The Gaussian noise n has zero mean, σ^2 variance and pdf: $p_n(n) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{n^2}{2\sigma^2}\right)$. The received signal r is the summation of the pure signal and the additive channel noise: $r = s + n$. Assume that s and n are independent, the pdf of r is the convolution of

$p_s(s)$ and $p_n(n)$, or:

$$p_r(r) = \int_{-\infty}^{\infty} p_n(r-y)p_s(y)dy \qquad (3)$$
$$= \frac{1}{2\sqrt{2\pi}\sigma}\left(\exp\left(-\frac{(r-1)^2}{2\sigma^2}\right)+\exp\left(-\frac{(r+1)^2}{2\sigma^2}\right)\right)$$

After been scaled by a factor g, the signal becomes $x = gr$. g is assumed to be a constant for transmitting a frame. The scaled signal x has pdf:

$$p(x) = \frac{1}{|g|}p_r\left(\frac{x}{g}\right) \qquad (g>0) \qquad (4)$$
$$= \frac{1}{2\sqrt{2\pi}g\sigma}\left(\exp\left(-\frac{(x-g)^2}{2g^2\sigma^2}\right)+\exp\left(-\frac{(x+g)^2}{2g^2\sigma^2}\right)\right)$$

Or, x's distribution is equivalent to the summation of two Guassian distributions: $N(g, g\sigma)$ and $N(-g, g\sigma)$.

Assume the quantization levels are $\tilde{x}_k$, and the quantization boundaries are (x_{k-1}, x_k), where $k = 1, ..., L$, and $L = 2^n$ is the number of quantization levels. The distortion function is thus:

$$D = \sum_{k=1}^{L}\int_{x_{k-1}}^{x_k}(x-\tilde{x}_k)^2 p(x)dx = A + \sum_{k=1}^{L}\left[-2\tilde{x}_k B_k + \tilde{x}_k^2 C_k\right] \qquad (5)$$

where $A = \int_{-\infty}^{\infty} x^2 p(x)dx$, $B_k = \int_{x_{k-1}}^{x_k} xp(x)dx$, $C_k = \int_{x_{k-1}}^{x_k} p(x)dx$, and $p(x)$ is as in (4). The following relations are derived, where $Q(\cdot)$ is the commonly used Q function:

$$A = g^2(\sigma^2+1)$$
$$B_k = \frac{g\sigma}{2\sqrt{2\pi}}\left[\exp\left(\frac{-(x_{k-1}-g)^2}{2g^2\sigma^2}\right)-\exp\left(\frac{-(x_k-g)^2}{2g^2\sigma^2}\right)\right.$$
$$\left.+\exp\left(\frac{-(x_{k-1}+g)^2}{2g^2\sigma^2}\right)-\exp\left(\frac{-(x_k+g)^2}{2g^2\sigma^2}\right)\right]$$
$$+\frac{g}{2}\left[Q\left(\frac{x_{k-1}-g}{g\sigma}\right)-Q\left(\frac{x_k-g}{g\sigma}\right)\right.$$
$$\left.-Q\left(\frac{x_{k-1}+g}{g\sigma}\right)+Q\left(\frac{x_k+g}{g\sigma}\right)\right]$$
$$C_k = \frac{1}{2}\left[Q\left(\frac{x_{k-1}-g}{g\sigma}\right)-Q\left(\frac{x_k-g}{g\sigma}\right)\right.$$
$$\left.+Q\left(\frac{x_{k-1}+g}{g\sigma}\right)-Q\left(\frac{x_k+g}{g\sigma}\right)\right]$$

From the above it is easy to see that the signal power is equal to A: $S = \int_{-\infty}^{\infty} x^2 p(x)dx = A = g^2(\sigma^2+1)$. And the signal-to-distortion ratio (SDR) is: $S/D = A/(A + \sum_{k=1}^{L}\left[-2\tilde{x}_k B_k + \tilde{x}_k^2 C_k\right])$. To find the optimal scaling factor, first the curve of SDR versus g is plotted. Then the optimal g is found to be the value corresponding to the maximum SDR.

The SDR plot for a four-bit, range $(-1, 1)$ quantizer is shown in Figure 2. The curves of the optimal gain when the quantizer range is $(-1, 1)$ are plotted in Figure 3. Let the optimal gain for quantizer range $(-1, 1)$ be g_1 at a certain E_b/N_0. When the dynamic range of the quantizer is $(-v, v)$, the optimal gain is $g_v = vg_1$.

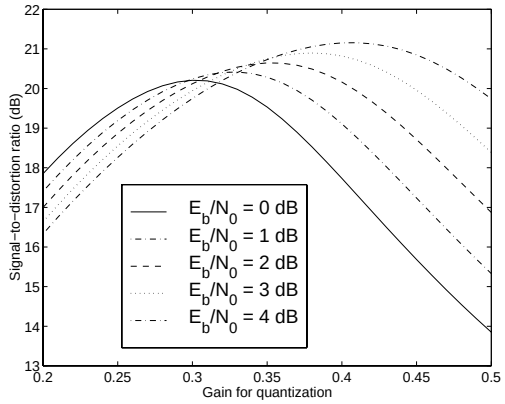

Figure 2: SDR for a $n = 4$, range $(-1, 1)$ quantizer.

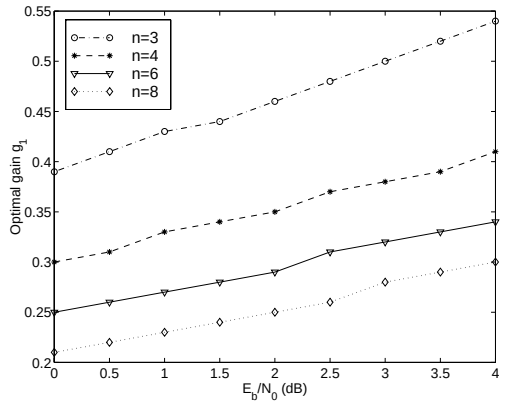

Figure 3: Optimal scaling factor for n-bit quantizer.

The following conclusions can be made about the scaling factor for an n-bit quantizer:

- There is an optimal gain to scale the received signal before it is fed into the quantizer. When the gain is too small, quantization noise will distort the signal; when the gain is too large, saturation will occur. Both cases result in the undesirable degradation of SDR.
- Figure 2 shows that larger values of n allow

for higher SDR. However, with proper scaling, a lower resolution quantizer can yield a higher SDR than that of a higher resolution quantizer, if the input of the latter is not properly scaled.

- It can be seen from Figure 3 that, over the E_b/N_0 range that is examined, for a given n, the optimal gain increases approximately linearly with the E_b/N_0 (dB) of the signal at the input of the quantizer. Also, for a given E_b/N_0, a higher resolution (larger n) quantizer needs a smaller gain to maximize SDR.

IV. SIMULATION RESULTS

The noise affecting the fixed point arithmetic decoding procedure is composed of two parts: the quantization error of the quantizer at the input of the decoder, and the accumulated errors resulting from the rounding or truncation of multiplication products inside the decoder. Rounding error at each step propagates through the whole decoding process. Since the decoder is a complicated system with feedback, analysis of rounding error effect is hard to derive analytically. Simulation is used instead to study the effect.

In the simulations, three different quantizer resolution, $n = 3, 4, 8$-bits, were examined. Perfect knowlege of the channel was assumed at the decoder. Two quantizer ranges were considered: $(-8, 8)$ and $(-0.5, 0.5)$. Range $(-0.5, 0.5)$ was chosen to illustrate the overflow problem with the quantizer, and range $(-8, 8)$ was chosen to show the rounding error problem. The optimal gain obtained in Section III was used in contrast to the case without scaling, and the case using floating point arithmetic. After eight iterations, the decoder estimates $\hat{d}_k$ were compared with the information bits d_k to determine BER.

In Figures 4 through 7, the BER curves are plotted versus E_b/N_0. Figures 4 and 5 are for a quantizer of range $(-0.5, 0.5)$. Figures 6 and 7 are for a quantizer of range $(-8, 8)$. The following observations are made from the plots.

- The effects of quantization are more evident at higher E_b/N_0, and the effects of AWGN tend to dominate at lower E_b/N_0. This can be seen from the fact that all curves for the optimally scaled cases tend to converge at lower E_b/N_0.

- As expected, higher n provides better performance for both decoding algorithms when the distribution of the signal is fixed. The most significant performance increases were observed when the main fixed point error was contributed by rounding (no scaling with quantizer having a range of $(-8, 8)$). However, in all cases, the improvement followed a law of diminishing returns. Since the difference between eight-bit quantization and floating point was negligible, no more than eight-bit quantization is required for accurate decoding of turbo codes.

- In cases of severe overflow (no scaling with quantizer having a range of $(-0.5, 0.5)$), both decoding algorithms exhibit marked inability to correct errors. However, the performance of APRI-SOVA did improve with increasing either n or E_b/N_0, in contrast to the Log-MAP algorithm, implying that the APRI-SOVA algorithm is more computationally stable.

- With range $(-8, 8)$, the BER for three-bit quantization with optimal gain control was lower than that of four-bit quantization without gain control. This indicates that, for a low resolution quantizer, it was crucial to adjust the gain so that the received signals fit in the dynamic range of the quantizer properly.

- For all cases in which the received signal is optimally scaled, in the region of $E_b/N_0 > 1$ dB, the coding gain of the Log-MAP algorithm is about 0.5 dB greater than that of the APRI-SOVA slgorithm for the same value of n.

V. CONCLUSION

In this paper, the influence of quantization and fixed point arithmetic upon the BER performance of turbo decoders was discussed. The following points were shown:

- Scaling the received signal by an optimal gain leads to excellent turbo decoder performance, even with very few bits ($n = 4$).

- APRI-SOVA involves no extra computation with fixed point arithmetic because of its linearity. It is also more computationally stable. However, Log-MAP generally performs better than APRI-SOVA, except when the overflow problem is severe.

- Overflow is a more severe problem than the rounding noise. It cannot be solved by increasing the number of quantization bits. Adjusting

the received signal to fit it into the full scale range of the quantizer is important.

Further investigations include finding the optimal gain for the Rayleigh fading channel, investigating the nonuniform quantization, and estimating the channel characteristic.

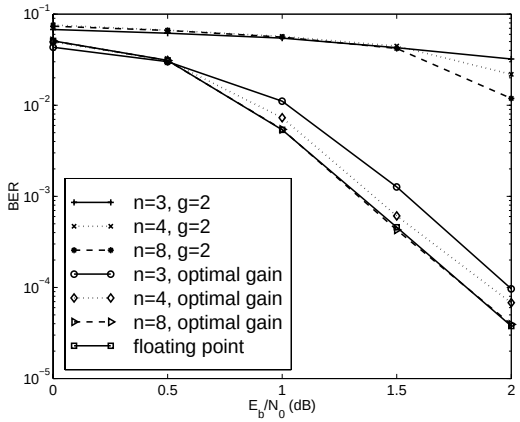

Figure 4: BER using Log-MAP for various resolution quantizers with range $(-0.5, 0.5)$.

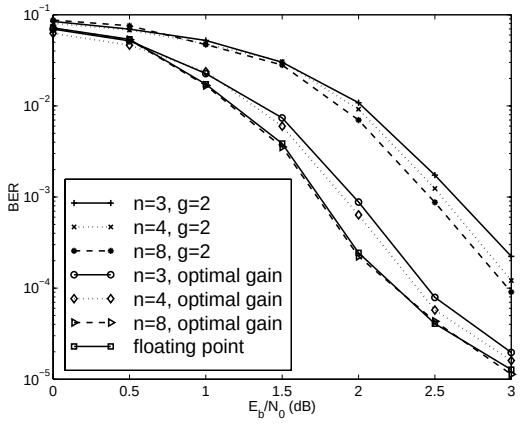

Figure 5: BER using APRI-SOVA for various resolution quantizers with range $(-0.5, 0.5)$.

REFERENCES

(1) C. Berrou, A. Glavieux, and P. Thitimasjshima, "Near Shannon limit error-correcting coding and decoding: turbo-codes (1)," in *Proc., IEEE Int. Conf. on Commun.*, (Geneva, Switzerland), pp. 1064–1070, May 1993.

(2) P. Robertson, P. Hoeher, and E. Villebrun, "Optimal and sub-optimal maximum a posteriori algorithms suitable for turbo decoding," *European Trans. on Telecommun.*, vol. 8, pp. 119–125, Mar./Apr. 1997.

(3) J. Hagenauer, P. Robertson, and L. Papke, "Iterative (turbo) decoding of systematic convolutional codes with the MAP and SOVA algorithms," in *Proc., ITG Conf.*, (Munchen, Germany), pp. 21–29, Sept. 1994.

(4) L. Papke, P. Robertson, and E. Villebrun, "Improved decoding with the SOVA in a parallel concatenated (turbo-code) scheme," in *Proc., IEEE Int. Conf. on Commun.*, pp. 102–106, 1996.

(5) T. K. Blankenship, "Design and implementation of a pilot signal scanning receiver for CDMA personal communication services systems," Master's thesis, Virginia Tech, Apr. 1998.

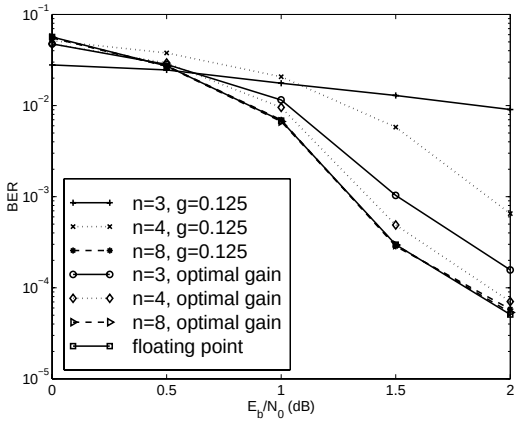

Figure 6: BER using Log-MAP for various resolution quantizers with range $(-8, 8)$.

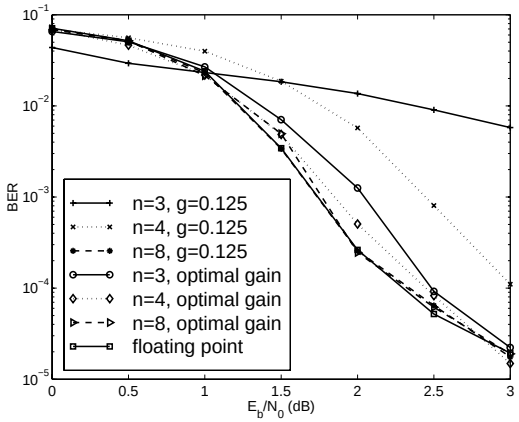

Figure 7: BER using APRI-SOVA for various resolution quantizers with range $(-8, 8)$.