

AN OPTIMAL BIT ALLOCATION ALGORITHM FOR SUB-BAND CODING

P.H. Westerink, J. Biemond and D.E. Boeke

*Delft University of Technology
Department of Electrical Engineering
Information Theory Group
P.O. Box 5031, 2600 GA Delft
The Netherlands*

ABSTRACT

Bit allocation is the problem of optimally assigning quantizers to a number of sources that are to be coded at a given total bit rate. Traditional solutions to this problem are known to be sub-optimal or are only applicable to coding schemes with quantizers that have an integer number of bits. This paper presents an optimal bit allocation algorithm that is suitable for all practical situations. Each source to be coded is assumed to have its own set of admissible quantizers (which can be either scalar or vector quantizers) which do not need to have integer bit rates. Also, the distortion versus rate characteristic of each quantizer set may have an arbitrary shape. The algorithm is very simple in structure and can be applied to any practical coding scheme (such as a sub-band coder) that needs dynamic bit allocation.

1. INTRODUCTION

In bit allocation, as used e.g. in sub-band coding, the objective is to optimally assign to each sub-band its own quantizer, in such a manner that the total number of bits equals a given quota and the overall distortion is minimal.

Among the first researchers to recognize the bit allocation problem were Huang and Schultheiss [1], who derived an analytical expression for the bit allocation in the case of a number of Gaussian sources. Their derivation, however, was based on a mathematical relation $d(r)$ between the quantizer distortion d and the corresponding bit rate r , which is only accurate for large bit rates. Although the expression is easy to evaluate it has a sub-optimal solution (due to the approximation of $d(r)$) and furthermore, negative bit assignments may occur that would require "trial and error" to find a practical bit assignment. The latter problem can be avoided by using special procedures to evaluate the bit allocation expression (Segall [2], Ramstadt [3]). Most of the existing literature on bit allocation (see, for instance for sub-band coding, Woods and O'Neil [4]) uses a similar analytic expression for the bit allocation, based on an approximate relation $d(r)$.

Another viewpoint on the bit allocation problem is taken by Trushkin [5]. He proposes an algorithm for optimal bit allocation that does not use any

mathematical relation between d and r . Instead, he assumes that all quantizers have already been designed and consequently, that all possible pairs of d and r are known. The algorithm described then finds the best combination of existing quantizers that will add up to a desired total number of bits at minimum distortion. The algorithm, however, is only suitable for integer values of the bit rate r , which is a major drawback in many practical situations. Another bit allocation algorithm, based on the same assumptions, was derived by Shoham and Gersho [6], who considered the problem of bit allocation in the case of coding a set of sources with vector quantizers (also having an integer bit rate).

In this paper we shall derive a generalization of Trushkin's algorithm. The major extension of the algorithm lies in allowing also real values for the bit rate r , in contrast to the special case considered in [5] (and [6]) where r can only take on integer values. This new algorithm therefore fits more to practical situations, where the number of bits per quantizer is not necessarily an integer value (e.g. when entropy coding is used).

In the sequel we shall often refer to bit allocation in sub-band coders, but the algorithm can be applied in general to all situations where a vector of sources needs to be quantized.

2. THE BIT ALLOCATION PROBLEM

The bit allocation problem that we meet in sub-band coding can in general be formulated as the problem to minimize the overall quantization error

$$D = \sum_{n=1}^N d_n, \quad (1a)$$

under the constraint of a given total bit rate

$$R = \sum_{n=1}^N r_n, \quad (1b)$$

by assigning to each sub-band n the appropriate quantizer having a distortion d_n and rate r_n . Note that the only assumption we make about the overall distortion measure is that it can be written as the sum of the separate distortions d_n , where each d_n can be the result of any arbitrary distortion measure (e.g. not just only the mean squared

error). In addition no restrictions are imposed on the possible choice of quantizers, in fact, each sub-band can have its own set of admissible quantizers (among which can be vector quantizers). To visualize the situation at hand, we consider 4 sub-bands that each have a set of 4 admissible quantizers and calculate the R and D values for all possible combinations of quantizers. As a result, as shown in Figure 1, we obtain $4^4 = 256$ different realizations of R and D pairs, each representing one particular bit allocation. So the bit

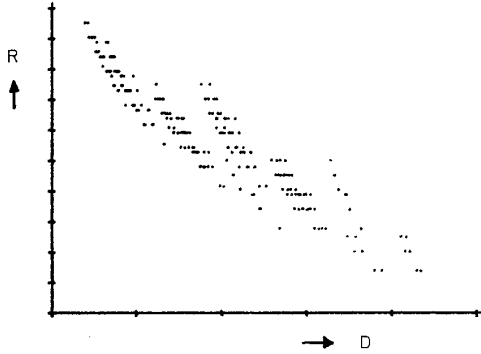


Figure 1. R(D) plot of all possible bit allocations in the case we have 4 sub-bands that each have 4 admissible quantizers.

allocation problem we have here, is to choose the best combination of quantizers from the R(D) plot that is close to the desired total bit rate. The obvious question that arises is how to define a certain bit allocation as being optimal. By analogy to rate-distortion theory (Berger, [7]), where R(D) curves are known to be convex, we here define the optimal bit allocations as those points that lie on the lower convex hull of the plot of all possible bit allocations. In Figure 2 the lower convex hull is plotted for the same example as was shown in Figure 1.

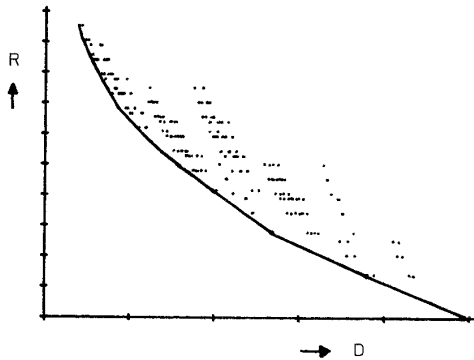


Figure 2. Optimal bit allocations defined as the lower convex hull of the R(D) plot of Figure 1.

Finding a convex hull of a set of points has been thoroughly investigated in literature and many fast algorithms have been designed to solve the problem (Toussaint, [8]). In our situation, however, the number of points is usually so large that we cannot reasonably use conventional convex hull algorithms (N sub-bands with each M quantizers yield M^N possible bit allocations). We therefore have to use another algorithm that does not require the calculation of all possible bit allocations, but finds the convex hull in a limited number of computations. In the following we shall derive such an algorithm, based on eqs. (1) and a property of a convex hull.

Each point (bit allocation) in the R(D) plot is the result of a certain choice of the quantizer for each sub-band. If we denote the assignment of quantizer v to sub-band n by v_n , then each possible bit allocation is described by a vector $\underline{v} = (v_1, \dots, v_N)$. Rewriting eqs. (1a) and (1b) we then denote an arbitrary bit allocation $\underline{v}$ as a point in the R(D) plot with coordinates

$$D(\underline{v}) = \sum_{n=1}^N d_n(v_n), \quad (2a)$$

$$R(\underline{v}) = \sum_{n=1}^N r_n(v_n). \quad (2b)$$

A property of a convex hull of a set of points which is useful for our purpose is based on the observation that if we draw a line through two neighboring points that lie on the convex hull, then all other points of the set must lie on or at one side of this line [8]. This is illustrated in Figure 3, where according to the definition of $\underline{v}$ in eq. (2), we denote the points on the convex hull by the vectors $\underline{l}$ and $\underline{k}$ and the line through these points by $L(\underline{l}, \underline{k})$. The definition of a convex hull

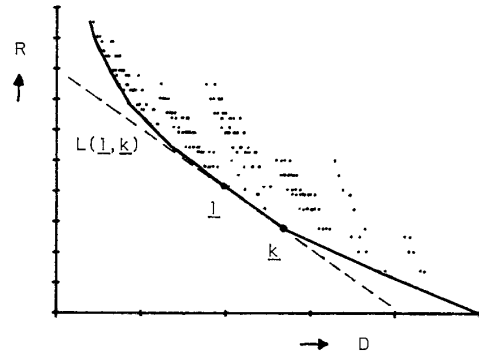


Figure 3. Property of two neighboring points $\underline{l}$ and $\underline{k}$ on the the convex hull of a set: all points of the set must lie on or at one side of the line $L(\underline{l}, \underline{k})$.

can thus be formulated using the equation of the straight line through $\underline{l}$ and $\underline{k}$:

$$[R(\underline{v}) - R(\underline{k})] \geq S(\underline{l}, \underline{k}) \cdot [D(\underline{v}) - D(\underline{k})] \quad \forall \underline{v}, \quad (3)$$

where $S(\underline{l}, \underline{k})$ is the slope of the line $L(\underline{l}, \underline{k})$:

$$S(\underline{l}, \underline{k}) = \frac{R(\underline{l}) - R(\underline{k})}{D(\underline{l}) - D(\underline{k})}. \quad (4)$$

Note that the equality sign in eq. (3) is for those points $\underline{v}$ that lie on the line $S(\underline{l}, \underline{k})$. Substitution of eqs. (2a) and (2b) into eq. (3) yields that for every possible vector $\underline{v}$ it must hold that:

$$\sum_{n=1}^N [r_n(\underline{v}_n) - r_n(\underline{k}_n)] \geq S(\underline{l}, \underline{k}) \sum_{n=1}^N [d_n(\underline{v}_n) - d_n(\underline{k}_n)]. \quad (5)$$

Next we bring $S(\underline{l}, \underline{k})$ under the summation sign and bring everything to the left side of the inequality sign:

$$\sum_{n=1}^N \left\{ [r_n(\underline{v}_n) - r_n(\underline{k}_n)] - S(\underline{l}, \underline{k}) [d_n(\underline{v}_n) - d_n(\underline{k}_n)] \right\} \geq 0. \quad (6)$$

This inequality now reflects the relationship between two neighboring points on the convex hull. In the case that we already know one of the two points we can find the other point by solving eq. (6) (which in fact will yield two solutions since a point will always have two neighbours). Using the new point again, but now as the point that is known, it is then possible to iteratively find all points of the convex hull. Finding the optimal bit allocations is therefore equal to repeatedly finding a vector $\underline{l}$ that satisfies eq. (6) for every vector $\underline{v}$, where an initial vector $\underline{k}$ is given. The iteration can be stopped when the total bit rate R is sufficiently close to the desired bit rate. It is also possible to stop the iteration when a certain desired distortion is reached.

3. ALGORITHM DERIVATION

The sum in eq. (6) will be larger than or equal to zero if all the elements in the summation are larger than or equal to zero, i.e. we must find a vector $\underline{l}$ such that for every $\underline{v}_n$

$$[r_n(\underline{v}_n) - r_n(\underline{k}_n)] - S(\underline{l}, \underline{k}) [d_n(\underline{v}_n) - d_n(\underline{k}_n)] \geq 0. \quad (7)$$

The solution for $\underline{l}$ in $S(\underline{l}, \underline{k})$ can be split into two different situations, each resulting in one of the neighbours of $\underline{k}$:

$$1. d_n(\underline{v}_n) - d_n(\underline{k}_n) < 0 \Rightarrow S(\underline{l}, \underline{k}) \geq \frac{r_n(\underline{v}_n) - r_n(\underline{k}_n)}{d_n(\underline{v}_n) - d_n(\underline{k}_n)} \quad \forall \underline{v}_n, \quad (8a)$$

$$2. d_n(\underline{v}_n) - d_n(\underline{k}_n) > 0 \Rightarrow S(\underline{l}, \underline{k}) \leq \frac{r_n(\underline{v}_n) - r_n(\underline{k}_n)}{d_n(\underline{v}_n) - d_n(\underline{k}_n)} \quad \forall \underline{v}_n. \quad (8b)$$

Without loss of generality we next concentrate on the first case. The solution to eq. (8a) is:

$$S(\underline{l}, \underline{k}) = \max_{n=1}^N s(\underline{v}_n, \underline{k}_n) \quad \forall \underline{v}_n, \quad (9)$$

where, by analogy to $S(\underline{l}, \underline{k})$, $s(\underline{v}_n, \underline{k}_n)$ is defined as

$$s(\underline{v}_n, \underline{k}_n) = \frac{r_n(\underline{v}_n) - r_n(\underline{k}_n)}{d_n(\underline{v}_n) - d_n(\underline{k}_n)}. \quad (10)$$

Substitution of eqs. (2a) and (2b) into the definition of $S(\underline{l}, \underline{k})$ in eq. (4), and using the definition of $s(\underline{v}_n, \underline{k}_n)$ in eq. (10) yields:

$$S(\underline{l}, \underline{k}) = \frac{\sum_{n=1}^N [r_n(\underline{l}_n) - r_n(\underline{k}_n)]}{\sum_{n=1}^N [d_n(\underline{l}_n) - d_n(\underline{k}_n)]} = \frac{\sum_{n=1}^N s(\underline{l}_n, \underline{k}_n) [d_n(\underline{l}_n) - d_n(\underline{k}_n)]}{\sum_{n=1}^N [d_n(\underline{l}_n) - d_n(\underline{k}_n)]}, \quad (11)$$

and by combining eqs. (9) and (11) we finally obtain the equation from which we can solve the vector $\underline{l}$, given the vector $\underline{k}$:

$$\sum_{n=1}^N [s(\underline{l}_n, \underline{k}_n) - \max_{m=1}^N s(\underline{v}_m, \underline{k}_m)] [d_n(\underline{l}_n) - d_n(\underline{k}_n)] = 0. \quad \forall \underline{v}_n \quad (12)$$

Excluding the trivial solution $\underline{l} = \underline{k}$, the solution to eq. (12) is:

$$\begin{cases} \underline{l}_n = \underline{v}_n & \text{for that } n \text{ for which } s(\underline{v}_n, \underline{k}_n) \\ & \text{has its maximum, and} \\ \underline{l}_n = \underline{k}_n & \text{for all other } n. \end{cases} \quad (13)$$

The $\underline{v}_n$ that may be considered when determining the maximum in eq. (13) must satisfy the constraint as formulated in eq. (8a). This means that we may only consider quantizers for the new bit allocation $\underline{l}_n$ that have a smaller distortion than $\underline{k}_n$. Because the total distortion D is the sum of the individual distortions, in this case we are calculating the points of the convex hull in the direction of decreasing distortion D . Eq. (8b) has the same solution as the solution to eq. (8a) just derived, except that we now have to find the minimum $s(\underline{v}_n, \underline{k}_n)$. According to the constraint in eq. (8b) we are then searching in the direction of increasing distortion D . The initial point for the algorithm can easily be found in both cases. If we search in the direction of decreasing distortion D , we take the point that is determined by assigning to each sub-band the quantizer with the lowest bit rate that is admissible for that sub-band. Since the total bit rate will then be minimal this point will always lie on the convex hull. In the same manner the initial point for searching in the direction of increasing D is determined by assigning to each sub-band its quantizer with the lowest distortion.

Finally the algorithm for searching in the direction of decreasing distortion D can be

described as follows:

1. Determine the initial bit allocation by assigning to each sub-band an admissible quantizer that has the lowest bit rate.
2. Calculate for each sub-band n all possible values of $s(v_n, k_n)$ with $d_n(v_n) < d_n(k_n)$ and find for each sub-band the quantizer for which $s(v_n, k_n)$ is maximal.
3. Determine the sub-band for which the maximum $s(v_n, k_n)$ is the largest and assign to this sub-band the quantizer for which this maximum is obtained.
4. Calculate the new rate R and the new distortion D and check if the desired bit rate or the desired distortion is approximated sufficiently close; if so, then stop the algorithm.
5. Repeat 2, 3 and 4, but do 2 only for the sub-band to which the new quantizer is assigned (the other values have not changed).

4. COMPARISON TO A SUB-OPTIMAL SOLUTION FOR SUB-BAND CODING OF IMAGES

To show the importance of our approach, in this section we compare the optimal algorithm with the bit allocation procedure as is described by Woods and O'Neil in [4], by applying both methods to the sub-band coding of images. Following [4] we employ a 16 band decomposition of the image frequency band using a tree structure of two-dimensional separable quadrature mirror filters and decimators. Each sub-band is coded separately using DPCM. The admissible set of quantizers for the 4 sub-bands containing the lowest frequencies consists of 1, 2, 4, 8, 16 and 32 level quantizers, while the other sub-bands use a set that consists of 1, 3, 5, 8, 16 and 32 level quantizers. All quantizers are based on a Laplacian probability density function with a mean squared error (MSE) criterion. Table 1 shows the quantizer assignments (indicated by the number of levels) for both methods at 0.75 bits per pixel. As can be seen, the sub-optimal solution of [4]

sb	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
a	16	8	4	4	1	3	1	1	1	1	1	1	1	1	1	1
b	16	4	2	2	3	3	1	3	1	1	1	1	1	1	1	1

Table 1. Quantizer assignments per sub-band (sb) at 0.75 bits/pixel for the optimal bit allocation algorithm (a) and the bit allocation method of Woods and O'Neil (b).

slightly favors the lower bit rate quantizers. This undoubtedly is the consequence of the approximation for low bit rates of the relation $d(r)$ on which the sub-optimal bit allocation procedure is based. Figure 4 shows a comparison of the signal-to-noise ratio (SNR) versus the bit rate for both bit allocation methods. It can be seen that the gain obtained by applying the optimal algorithm over the sub-optimal solution is nearly 1.0 dB in terms of the SNR.

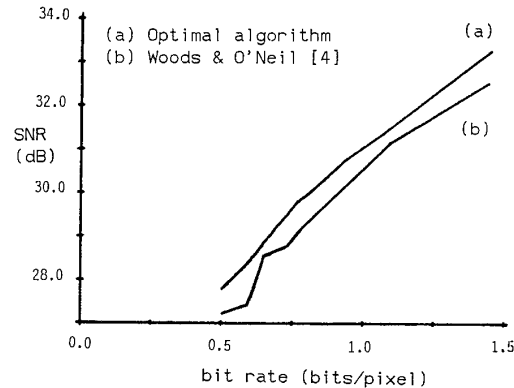


Figure 4. SNR versus bit rate comparison of the optimal bit allocation algorithm (a) and the sub-optimal procedure of [4] (b).

5. CONCLUSIONS

In this paper we have described an optimal bit allocation algorithm that is applicable to all practical situations. The algorithm is simple in structure and requires just the rates and distortions of the quantizers that are to be used. When applied to sub-band coding of images the performance of the coder is higher when compared to using a sub-optimal bit allocation procedure.

REFERENCES

- [1] J.J.Y. Huang and P.M. Schultheiss, "Block Quantization of Correlated Gaussian Random Variables", IEEE Trans. on Comm. Syst., Vol. 11, No. 3, Sept. 1963, pp. 289-296.
- [2] A. Segall, "Bit Allocation and Encoding for Vector Sources", IEEE Trans. on Inf. Theory, Vol. IT-22, No. 2, March 1976, pp. 162-169.
- [3] T.A. Ramstadt, "Considerations on Quantization and Dynamic Bit Allocation in Subband Coding", Proc. ICASSP 1986, Tokyo, Japan, April 1986, pp. 841-844.
- [4] J.W. Woods and S.D. O'Neil, "Sub-band Coding of Images", IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-34, No. 5, Oct. 1986, pp. 1278-1288.
- [5] A.V. Trushkin, "Optimal Bit Allocation Algorithm For Quantizing a Random Vector", Translated from Russian in: Probl. Inf. Transmission, Vol. 17, No. 3, July-Sept. 1981, pp. 156-161.
- [6] Y. Shoham and A. Gersho, "Efficient Codebook Allocation for an Arbitrary set of Vector Quantizers", Proc. ICASSP 1985, Tampa, Florida, USA, March 1985, pp. 1696-1699.
- [7] T. Berger, *Rate Distortion Theory*, Prentice Hall, Inc. Englewood Cliffs, New Jersey, 1971.
- [8] G.T. Toussaint, "Pattern Recognition and Geometrical Complexity", Proc. of the 5th Int. Conf. on Pattern Recognition, Miami Beach, Florida, USA, December 1980, pp. 1324-1347.