Joint Bit Allocation and Dimensions Optimization for Vector Transform Quantization

Vladimir Cuperman

Abstract—**In vector transform quantization (VTQ), vectors consisting of M consecutive samples of a waveform are transformed into a set of M coefficients that are quantized by $m \ll M$ vector quantizers. The bit allocation problem in the transform domain is considered for a memoryless stationary vector source encoded by a VTQ system. It is assumed that the vector quantizer parameters (dimension, codebook size) are subject to a complexity constraint. The vector quantization lower bound on the attainable distortion at a given (high) rate is used for deriving the bit allocation algorithm for given vector dimensions. Then, the joint optimization of vector dimensions and bit allocations is considered. Given a complexity constraint, the optimal dimensions depend on the bit allocation, which, in turn, depends on the dimensions. An iterative algorithm is proposed for solving this problem.**

Index Terms—**Source coding, vector quantization, transform coding, bit allocation.**

I. Introduction

The problem considered in this correspondence is the efficient coding of a memoryless vector source by using a transform coding approach. Traditionally, in such a system, the transform coefficients are quantized independently by a set of M scalar quantizers, where M is the transform size [2], [3].

It is easy to see that an M-dimensional vector quantizer (VQ) will achieve better performance than the scalar transform coder. Actually, a size M transform followed by M scalar quantizers is a particular case of an M-dimensional VQ, though not an optimal one. However, the implementation complexity of an M-dimensional VQ is of the order of 2^{Mr} multiplications/additions per sample where r is the rate in bits/sample. For typical values of interest such as $r = 1 - 2$ bits/sample and $M = 32 - 256$, the implementation complexity becomes untractable.

In vector transform quantization (VTQ), a size M transform is followed by a set of m VQ's, each having dimension $k_i, i = 1, 2, \cdots, m$ so that $\sum_{i=1}^{m} k_i = M$ [4]. In the system shown in Fig. 1, the transform S is applied to the input vector $\boldsymbol{x}$ to obtain the transformed vector $\boldsymbol{y}$. The components of the transformed vector $\boldsymbol{y}$ are regrouped in vectors $\boldsymbol{y}_i$, $i = 1, 2, \cdots, m$ and each vector $\boldsymbol{y}_i$ is quantized independently by the vector quantizer VQ_i. At the receiver, the vector $\boldsymbol{y}^*$ results by concatenating the reconstructed transformed vectors $\boldsymbol{y}_i^*$ obtained from the inverse vector quantizers IVQ_i by using the received indices in the VQ_i codebooks. Exhaustive (full) search is used to select the indices in the codebooks VQ_i. The reconstructed vector $\boldsymbol{x}^*$ is then obtained using the transform R, which in general may be different than the transform S employed in the encoder. The channel is assumed noiseless.

We assume that codebook sizes, N_i, and dimensions, k_i, are subject to a complexity constraint. The objective of this paper is to present a design procedure for VTQ which considers the joint optimization of bit allocations and VQ dimensions. The bit allocation design based on the VQ lower bound was introduced in [4] and applications to speech coding were discussed. The present paper extends the results in [4] by considering integer constraints on rate, joint optimization of bit allocations and dimensions, and evaluation of the rate-distortion performance for memoryless vector sources. The coding gain of the resulting VTQ configuration is also discussed.

Manuscript received February 2, 1989; revised August 6, 1991. This work was supported by the National Sciences and Engineering Research Council of Canada and by the Science Council of British Columbia. This work was presented in part at the IEEE International Symposium on Information Theory, Kobe City, Japan, June 19–24, 1988.

The author is with the School of Engineering Science, Simon Fraser University, Burnaby, BC, V5A 1S6 Canada. He also is with the Center for Information Processing Research, Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106.

IEEE Log Number 9203641.

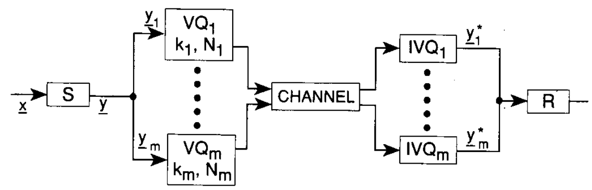

Fig. 1. Vector transform quantization (VTQ) block diagram.

II. Optimal Bit Allocation Based on the VQ Bound

Let $\boldsymbol{x}^{(j)}$ be a sequence of i.i.d. M-dimensional vectors having a jointly Gaussian distribution. The vectors $\boldsymbol{x}$ are transformed by a nonsingular decorrelating transform, S, into vectors $\boldsymbol{y}$ having independent components. (See Fig. 1.) The vectors $\boldsymbol{y}$ are encoded by a set of m VQ's, VQ_i, $i = 1, 2, \cdots, m$, having the dimensions k_i, such that $\sum_{i=1}^{m} k_i = M$. The objective is to minimize the average encoding distortion by adjusting the VQ_i bit allocations, $B_i = \log_2 N_i$, and the dimensions k_i, given a number of available bits per M-dimensional vector. The fidelity criterion used in this correspondence is the mean-squared error.

In this section, we will assume the number of VQ's m, and the dimensions, k_i, are fixed and will consider the following minimization problem.

Minimize

$$D = \frac{1}{M} E\left\{\|\boldsymbol{y} - \boldsymbol{y}^*\|^2\right\} \tag{1}$$

under the constraints

$$B_j \geq 0 \quad \text{and} \quad \sum_{j=1}^{m} B_j = B, \tag{2}$$

where B is the number of bits available per M-dimensional vector.

Let D_i be the distortion for VQ_i:

$$D_i = E\left\{\|\boldsymbol{y}_i - \boldsymbol{y}_i^*\|^2\right\} \tag{3}$$

then

$$D = \frac{1}{M} \sum_{i=1}^{m} D_i \tag{4}$$

Using the VQ bound (Gersho, [1]), the rate-distortion performance of an optionally chosen VQ_i can be approximated by

$$D_i = \gamma_i \hat{\sigma}_i^2 2^{-2B_i/k_i} \tag{5}$$

where $\hat{\sigma}_i^2 = (\det \Gamma_i)^{1/k_i}$, Γ_i is the covariance matrix of the vector $\boldsymbol{y}_i$,

$$\gamma_i = 2\pi c_i k_i \left(1 + \frac{2}{k_i}\right)^{k_i/2+1} \tag{6}$$

and c_i is the coefficient of quantization for the quantizer VQ_i [1]. Note that $\hat{\sigma}_i$ may be considered as an "equivalent" vector variance, and γ_i is actually a correction factor, depending on the coefficient of quantization for VQ_i and on the dimension k_i.

The VQ bound (derived first by Zador [7]) is actually an approximate formula for the least distortion of a VQ with a given dimension and a given large rate. For infinite dimension, this formula is equivalent to Shannon's lower bound [1]. The use of asymptotic high rate performance in the VTQ design is justified by the superior experimental results achieved by the resulting system design [4].

The rate-distortion performance (5) is a convex function of B_i. Hence, the solution to the minimization problem (1) can be found by following a procedure similar to that of Segall [3]. The resulting bit allocation is

$$B_i = \max\left(0, \frac{k_i}{2} \log_2 \frac{\hat{\sigma}_i^2 \gamma_i}{v}\right) \tag{7}$$

where the constant v may be found from the constraint on the total number of bits for encoding an M-dimensional vector:

$$\frac{1}{2} \sum_{j=1}^{m} {}^{*} k_j \log_2 \frac{\gamma_j \hat{\sigma}_j^2}{v} = B \tag{8}$$

Here, $\sum^*$ indicates summation only for the terms that obey

$$\gamma_j \hat{\sigma}_j^2 \geq v. \tag{9}$$

In the particular case $k_i = 1$, $i = 1, 2, \cdots, m$ and $m = M$, (7) gives a scalar bit assignment that minimizes the distortion D for a given number of bits per vector B.

The coefficient of quantization values are unknown except for dimensions one and two. However, a number of approximations based on lower or upper bounds can be used in conjunction with (7). Examples are the sphere lower bound [1], the Conway–Sloane lower bound [8], and the Voronoi lattice upper bound [8]. The results in this correspondence are based on using the values given by the Voronoi lattice upper bound.

Consider now the minimization of $E\{\|\boldsymbol{x} - \boldsymbol{x}^*\|^2\}$ under the constraints given in (2). Following the approach of [3], one can prove that the optimal transform coder should employ matrices R, S (see Fig. 1) such that $R = S^{-1}$ and S is given by the eigenvectors of the covariance matrix of the input vector, $\boldsymbol{x}$. Hence, S is the well-known Karhunen–Loeve transform (KLT) proven to be optimal for scalar transform coding systems under the assumptions of high rate and Gaussian input. The transform S decorrelates the components of the vector $\boldsymbol{y}$. This does not guarantee the components of $\boldsymbol{y}^*$ will be uncorrelated; however, the assumption that the components of $\boldsymbol{y}^*$ are uncorrelated is made here as an approximation in order to simplify the choice of matrix R to be $R = S^{-1}$. Under the assumptions previously discussed, the bit allocations for minimizing $E\{\|\boldsymbol{x} - \boldsymbol{x}^*\|^2\}$ are still given by (7).

III. The Coding Gain

An expression for the coding gain was derived in [4]. For completeness, it is briefly reviewed here. Let D_{PCM} be the PCM distortion evaluated by using the VQ bound (5) for $k_i = 1$. The VTQ gain over PCM is defined by

$$G_{\text{VTQ}} = \frac{D_{\text{PCM}}}{D} \tag{10}$$

Assuming that the components of the vector $\boldsymbol{y}$ are independent and using the allocation (7) and the relations (4), (5) for evaluating D, one can prove that [4]

$$G_{\text{VTQ}} = G_{\text{TC}} G_V \tag{11}$$

where G_{TC} is the scalar transform gain equal to the ratio of arithmetic and geometric mean of he transform coefficient variance. G_V is the additional gain due to the use of vector rather than scalar quantizers and

$$\log G_V = \frac{1}{M} \sum_{j=1}^{m} k_j \log G_j, \tag{12}$$

where G_j is the coding gain of VQ_j over PCM,

$$G_j = \frac{\gamma_{\text{PCM}}}{\gamma_j} \tag{13}$$

and γ_{PCM} is the value of γ_j for $k_j = 1$. To derive (12), the distortion D is evaluated by (4), (5) using the bit allocations (7), while the distortion D_{PCM} is obtained as the particular value of D for $m = M$ and $k_i = 1$, $i = 1, 2, \cdots m$.

For a Gaussian pdf, G_j estimated with the sphere lower bound used for the coefficients of quantization may be as high as 4.35 dB (asymptotically for large dimensions). A more conservative estimate of 3.2 dB is obtained using the Voronoi lattice upper bound for the coefficients of quantization. Expressions for γ_j for other dpf's and a discussion of the corresponding gain values can be found in [4].

VTQ has a couple of rather obvious properties which are reflected in (12), (13). First, VTQ is asymptotically optimal for increasing complexity (it reduces to a full search VQ which in its turn is asymptotically optimal). Second, for a given complexity constraint the performance degradation can be easily estimated, assuming the VQ gain for different dimensions is known.

IV. Joint Optimization of Dimensions and Bit Allocations

Relation (7) assumes that the bit assignment for each vector quantizer is a real positive number. In practical applications, the number of codewords in VQ_i, N_i, must be an integer. The constraint of the minimization problem (1) can be written for this case as

$$\sum_{j=1}^{m} \log_2 N_j = B \tag{14}$$

The problem of finding the minimum of D under the constraint (14) with N_j integers can be solved by marginal analysis [5]. Marginal analysis is a discrete optimization technique in which the codebook size allocations are found incrementally: the codebook size for the quantizer VQ_j is increased by one if the incremental return M_j for VQ_j is the largest. Here the incremental return criterion is the decrease of the squared mean error D_j when increasing N_j by one.

The distortion D_j can be written as a function of the number of codewords N_j:

$$D_j = \gamma_j \hat{\sigma}_j^2 N_j^{-2/k_j} \tag{15}$$

and taking into account the nonlinearity of the constraint (14), the incremental return criterion, M_j, is given by [5]

$$M_j = \frac{\gamma_j \hat{\sigma}_j^2 \left[N_j^{-2/k_j} - (N_j + 1)^{-2/k_j}\right]}{\log_2 (1 + 1/N_j)}. \tag{16}$$

The bit assignment algorithm based on marginal analysis consists of the following steps.

Step 1) Start with $N_i = 1$, $i = 1, 2, \cdots, m$.

Step 2) Find the index j for which M_j is maximum. Increase N_j by 1.

Step 3) If $\sum_{i=1}^{m} \log_2 N_j \geq B$ terminate. Else go to Step 2).

Now, we consider the joint optimization of vector dimensions and bit allocations for the encoding scheme given in Fig. 1 under the following complexity constraints:

$$\begin{aligned} N_j &\leq N_{\max} \\ k_j &\leq k_{\max} N_{\max}/N_j \end{aligned} \tag{17}$$

where $k_{\max}$ is the maximum dimension and $N_{\max}$ is the maximum number of codewords acceptable for VQ_j under the given complexity constraint. Note that a full search vector quantizer requires a memory of $N_j k_j$ words and the computational complexity is proportional to 2^{N_j}.

The optimization of vector dimensions will be based on maximizing the VTQ coding gain G_{VTQ}. The transform coding gain, G_{TC}, being independent of vector dimensions, (11) shows that the maximization of G_{VTQ} is equivalent to the maximization of G_V, the additional gain due to the use of vector rather than scalar quantizers.

Assume that the transform coefficients are ordered in the increasing order of their variances before defining the vectors $\boldsymbol{y}_i$, $i = 1, 2, \cdots, m$. Then based on (12), (13), it can be shown that the following procedure generates the dimensions which maximizes G_V under a given complexity constraint. First, $m = 1$, $k_1 = M$ is an optimal solution if it satisfies the complexity constraints. If for $m = 1$ the complexity constraint is not satisfied, $m = 2$, $k_1 + k_2 = M$, $k_1 \geq k_2$, is an optimal solution if k_1 has the largest value for which the complexity constraint is still satisfied. Finally, if for $m - 1$ the complexity constraint is not satisfied, $k_1, k_2, \cdots, k_m$ is an optimal solution if $k_1 \geq k_2 \geq \cdots \geq k_m$ and the dimensions $k_1, k_2, \cdots, k_m$ have the largest values for which the complexity constraint is still satisfied. Of course, for each set k_i, $i = 1, 2, \cdots, m$, checking the complexity constraints requires the computation of the bit allocation using marginal returns. Note that a given solution is considered here as "optimal" if it maximizes the gain G_V under a given complexity constraint.

The procedure previously defined requires an iterative algorithm for finding dimensions and bit allocations. The flowchart of the iterative algorithm is given in Fig. 2. The algorithm starts with an initial set of dimensions and computes the bit allocations using marginal analysis (16). Then, the dimensions are adjusted toward their optimal values, and the complexity constraints are tested. For practical (complexity) reasons, it is preferred to start with $m = M$ (scalar quantizers), rather than following the more "intuitive" approach and start with $m = 1$.

The main steps of the algorithm follow.

Step 1) Initialize the number of codebooks to $m = M$ and the dimensions to $k_i = 1$, $i = 1, 2, \cdots, m$. Set the iteration index to $j = 1$.

Step 2) Given dimensions k_i, $i = 1, 2, \cdots, m$, find the codebook sizes, N_i, using the marginal return algorithm.

Step 3) Check complexity constraints (17) for $i = 1, 2, \cdots, m$.

- IF (17) is not satisfied for $i = j$, decrease k_j by one, increase m by one, set $k_m = 1$, increase j by one, and GO TO Step 2).
- IF (17) is not satisfied for $i < j$, decrease k_i by one, increase m by one, set $k_m = 1$, set $j = i + 1$, and GO TO Step 2).
- Else, GO TO Step 4).

Step 4) IF $j = m$, STOP. Else, increase k_j by one, decrease m by one, and GO TO Step 2).

Note that the design algorithm starts with a scalar transform coder $(m = M)$ and then increases the dimension k_1 until the complexity constraint for VQ_1 is reached. The same procedure is then applied for k_2, k_3 and so on. Due to the fact that the bit assignment depends on all dimensions, the complexity constraints are checked for each VQ_i whenever one of the dimensions is incremented or decremented. In our experience, this algorithm converges in a few iterations.

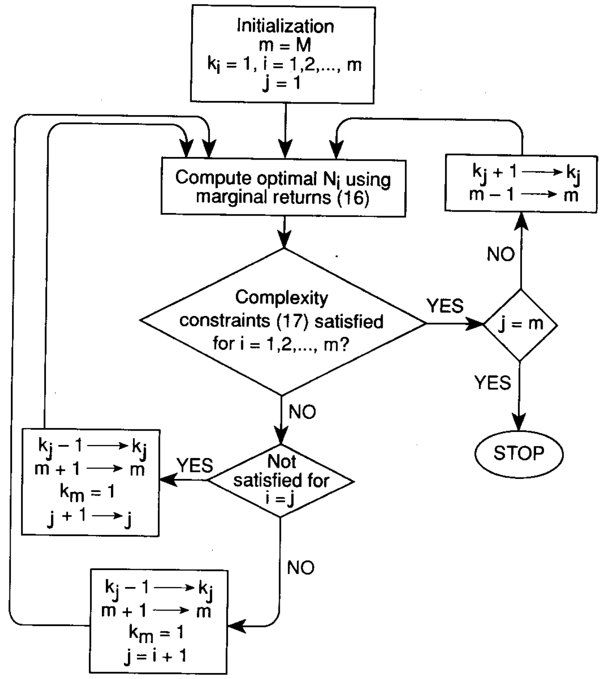

Fig. 2. Algorithm for joint optimization of bit allocations and dimensions.

V. Simulation Results for a Gauss–Markov Process

The joint optimization of bit allocations and dimensions was applied for designing vector transform quantizers (VTQ's) for a scalar Gauss–Markov source:

$$z(n) = u(n) + \rho z(n-1), \tag{18}$$

where $u(n)$ is a Gaussian process with zero-mean and unit variance. This process is often used for "benchmarking" encoding algorithms, because it is considered a useful mathematical model for some real sources (speech, images), and its performance bounds as given by the distortion-rate function are known. The source was encoded by "blocking" each M consecutive samples in an M-dimensional vector $\boldsymbol{x}$.

A file of 500 000 samples of the process (18) was used for training the vector quantizers and a different file of 60 000 samples for testing the encoders. The VTQ systems considered in the simulation use exhaustive search in each of the m codebooks. The complexity constraint used in the simulations was $N_{\max} = 512$ and $k_{\max} = 16$. The vector dimension was set at $M = 64$ and the rate was 1 bit/sample, i.e., $B = 64$ bits were used for and M-dimensional vector for all experiments.

Fig. 3 shows the signal-to-noise Ratio (SNR) achieved by the VTQ encoders versus the process parameter ρ. The rate-distortion function of this process and the best results achieved by an optimal DPCM encoder are plotted on the same figure [6]. The optimal DPCM system of [6] is a lower complexity system and the comparison is shown only for reference purposes.

Experimentally, it was found that constraining the number of codewords to be a power of two leads to only a slight performance degradation (less than 0.1 dB). In many practical applications, the number of codewords is constrained to be a power of two.

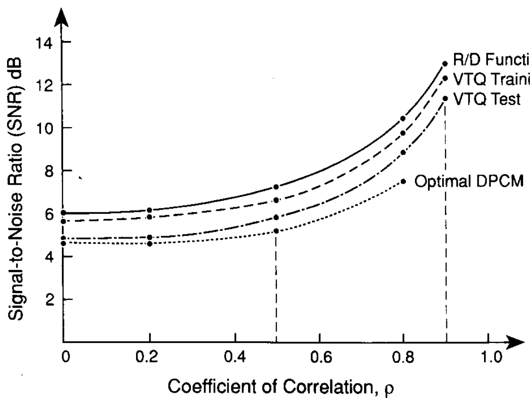

Fig. 3. Performance of vector transform quantization (VTQ) at a rate of 1 bit/sample for a Gauss–Markov source with correlation coefficient ρ. VTQ uses 64-dimensional vectors, $M = 64$, and the complexity constraints $N_{\max} = 512$, $k_{\max} = 16$.

For $\rho = 0.9$, the best performance achieved in training by a VTQ with $M = 128$ at a rate of 1 bit/sample was 12.7 dB. For comparison, comparable complexity systems based on finite state vector quantization achieved 12.1 dB, and vector predictive quantization achieved 11.6 dB [9]. The rate-distortion bound at a rate of 1 bit/sample for $\rho = 0.9$ is 13.2 dB.

REFERENCES

[1] A. Gersho, "Asymptotically optimal block quantization," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 373–380, July 1979.
[2] J. Y. Huang and P. M. Schulteiss, "Block quantization of correlated Gaussian random variables," *IEEE Trans. Commun. Syst.*, vol. CS-11, pp. 289–296, Sept. 1963.
[3] A. Segall, "Bit allocation and encoding for vector sources, *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 162–169, Mar. 1976.
[4] V. Cuperman, "On adaptive vector transform quantization for speech coding," *IEEE Trans. Commun.*, vol. 37, pp. 261–267, Mar. 1989.
[5] B. Fox, "Discrete optimization via marginal analysis," *Management Sci.*, vol. 13, no. 3, Nov. 1966.
[6] N. Favardin and J. W. Modestino, "DPCM for autoregressive sources," *IEEE Trans. Inform. Theory*, vol IT-31, pp. 402–418, May 1985.
[7] P. Zador, "Asymptotic quantization of continuous random variables," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 139–149, Mar. 1982.
[8] J. H. Conway and N. J. A. Sloane, "A lower bound on the average error of vector quantizers," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 106–109, Jan. 1985.
[9] R. M. Gray, "Vector quantization," *IEEE ASSP Mag.*, vol. 1, pp. 4–29, Apr. 1984.

Balanced Quadriphase Sequences with Optimal Periodic Correlation Properties Constructed by Real-Valued Bent Functions

Shinya Matsufuji and Kyoki Imamura

***Abstract*—The real-valued bent function was recently introduced by the present authors as a generalization of the usual p-ary bent function, p a prime, in such a way that the range of the function is the set of real numbers, i.e., not restricted to GF(p). The real-valued bent function was used to construct a family of $2^{n/2}$ balanced quadriphase sequences of period $2^n - 1$ with optimal periodic correlation properties, where n is a multiple of 4. A class of real-valued bent functions that map the set of all the m-tuples over GF(2) into the set $\{0, 1/2, 1, 3/2\}$ for an arbitrary m, is given. This is applied to generalize previous construction to the case where n is even, i.e., not restricted to a multiple of 4. It is also shown that the quadriphase sequences recently given by Novosad can be considered as one kind of sequences constructed by real-valued bent functions. Conditions are given for some families of the quadriphase sequences constructed by some real-valued bent functions to be balanced. The exact distributions of the periodic correlation values are derived for the families of the balanced quadriphase sequences.**

***Index Terms*— Real-valued bent function, sequence design, balanced quadriphase sequences, periodic correlation, spread spectrum communication.**

I. INTRODUCTION

Rothaus [1] defined a function from the set V_m of m-tuples over GF(2) into GF(2) to be bent if all the Fourier coefficients have unit magnitudes and studied some properties of the binary bent function with an even m. Olsen [2] showed that a family of 2^m balanced biphase sequences of period $2^{2m} - 1$ which have optimal correlation properties in the sense of Welch's lower bound [3] can be constructed by using the binary bent function on V_m with an even m, where a balanced sequence means that in one period -1 appears once more than $+1$. Kumar [4] defined a p-ary bent function from the set V_m of m-tuples over GF(p) into GF(p) and showed that there exists a bent function on V_m if m is odd and $p \neq 2$, and that a family of p^m p-phase sequences of period $p^{2m} - 1$ with optimal correlation properties can be constructed by the p-ary bent function in the similar manner to that by Olsen. We [5] showed that the p-phase sequences by Kumar are balanced sequences such that in one period each element $\exp(j2\pi i/p)$, $1 \leq i \leq p - 1$, appears once more than $+1$, and derived the exact distribution of the periodic correlation values. We call the sequence of period $P^{2m} - 1$ constructed by the p-ary bent function on V_m over GF(p) a "p-ary bent sequence."

The families of the quadriphase $\{\pm 1, \pm j\}$ sequences applied to the spread spectrum multiple access (SSMA) communication using the quadriphase shift keying (QPSK), where j denotes imaginary unit, are studied and constructed by Krone and Sarwate [6], but these do not have optimal correlation properties. Sole [7] showed that a family of $2^n + 1$ quadriphase sequences of period $2^n - 1$ with an odd n can be generated by using a unitary polynomial over the residues modulo 4. Recently, Udaya *et al.* [8] generalized the Sole's construction to an even n and Boztas *et al.* [9] also presented a family of $2^n + 1$ quadriphase sequences with near optimal correlation properties. We [10] gave a real-valued bent function extending the p-ary bent function such that it maps the set V_m with an even m over GF(p) to the real numbers, and a family of 2^m balanced quadriphase sequences of period $2^{2m} - 1$ with optimal correlation properties constructed by the real-valued bent function in the similar manner to that by Olsen and Kumar. Novosad [11] recently constructed a

Manuscript received August 30, 1991. This work was presented in part at the technical meeting of Spread Spectrum Communication Group of IEICE, March 1991.

S. Matsufuji is with the Department of Information Science, Saga University, Saga 840, Japan.

K. Imamura is with the Department of Computer Science and Electronics, Kyushu Institute of Technology, Iizuka, Fukuoka 820, Japan.

IEEE Log Number 9203031.