computing@tanet.edu.te.ua
www.tanet.edu.te.ua/computing

ISSN 1727-6209
International Scientific
Journal of Computing

OPTIMAL POLAR QUANTIZATION OF COMPLEX VARIABLES WITH CIRCULARLY SYMMETRIC DENSITIES

Zoran Perić[1], **Srdjan Bogosavljević**[2], **Aleksandra Jovanović**[3]

1,3 "Faculty of Electronic Engineering", University of Nis, Serbia; e-mail: peric@elfak.ni.ac.yu
2 "Telecom Serbia", , 11000 Beograd, Serbia

Abstract: *In this paper we consider quantization of complex variables and mean-square error (MSE). The best polar quantizer is Wilson's unrestricted polar quantizer (UPQ) [1]. The MSE minimization is constrained only by the total number of quantization points, N. Our method is different from Wilson's algorithm [1] that has predetermined number of points M_i at each magnitude level i, $1\le i\le L$, which makes it impractical for large number of points. In our approach, we consider MSE as a function of the vector* $\mathbf{M}=(M_i)_{1\le i\le L}$ *whose elements are numbers of phase quantization levels at each magnitude level. The Wilson's method finds the optimal quantizer in such a way that the decision and reconstruction levels* **r**, **m** *are iterative found for each combination* **M**, *while the optimal combination is found by searching all combinations. Wilson's algorithm cannot be applied for middle and great N. The asymptotic analysis of the polar quantizers with circular symmetric densities is given in [2]. This analysis is approximate and cannot be applied for any number of points and for great N, which will be shown in this paper. We define the extension of the MSE over R^L (denoted by MSE(**P**)). We prove the convexity of this function and show an efficient way to find* $\mathbf{M}=(M_i)_{1\le i\le L}$ *by* $\mathbf{P}_{opt}$. *Our algorithm consists of two main iterative processes. The first iterative process finds* $\mathbf{P}_{opt}$, $\mathbf{r}_{opt}$, $\mathbf{m}_{opt}$ *with ε accuracy, while the second iterative process determines* $\mathbf{M}_{opt}$, $\mathbf{m}_{opt}$, $\mathbf{r}_{opt}$ *using* $\mathbf{P}_{opt}$ *as the starting value. This paper eliminates incompleteness from [1] and [2]. We also give an example of the quantizer construction for a Gaussian source. The authors see their work as a contribution in knowing the best possible solution in these classes of problems and also a possibility of applying the technique exposed inhere on other classes of problems and on larger dimensions.*

Keywords: *Quantization, polar quantization, iterative process, extended distortion function, convexity function, optimization, integer optimization .*

I. INTRODUCTION

Quantization is the heart of analog-to-digital conversion. Quantizers play an important role in the theory and practice of modern-day signal processing.The algorithm for the optimal scalar quantization has been introduced in literature since distant 1960 by Max. Extensive results have been developed on scalar quantization but more on vector quantization. The simplest vector quantization is polar quantization. Polar quantization techniques as well as their applications in areas such as computer holography, discrete Fourier transform encoding, image processing and communications have been studied extensively in the literature. However, the algorithm for the optimal polar quantization has not been found till now.

In these paper the existence of one minimum will be proved and we will give the algorithm after which we can effectively find the optimal solution. Problem of the integer optimization is present for larger dimensions than 2 when the lattice-encoders and geometrical source shapes are used [11].

This problem is effectively solved in our work (or, applied with small modifications it can be extended for problem [11]).Also, the compromise between the complexity and performance of the suggested solutions must always exist. In these sense we suggest an optimal quantizer with look-up table, which would memorize (**r**, **m**, **M**). Better performances obtained by noted technique can be gained using larger dimensions, but the complexity will arise consequently. Trellis coded quantization (TCQ) could be seen as better solution but its complexity is much larger and also the time-delay will be needed. Moreover, signal constellations mostly used in TCQ are two-dimensional, so the projecting method of these constellations and searchings to the nearest represent can be accomplished by techniques presented in this paper.

In this paper we consider a complex random variable $z = re^{j\phi}$ with circularly symmetric density function p*(z)=f(r)/(2π)*. Conventional polar quantization independently quantizes *r* and ϕ with *L* and *M* levels, respectively ($L\times M = N$). In other words, complex plane is partitioned into *L* concentric rings around $z=re^{j\phi}=0$, and each ring is divided angularly into *M* slices. In this paper we consider nonuniform polar quantizer, i.e. quantizer which consists of nonuniform magnitude and uniform phase quantizers. We use mean-square-error (MSE) distortion measure $D = E(\left|re^{j\phi} - me^{j\psi}\right|^2)$ where *m* and ψ represent the magnitude and phase reconstruction levels respectively.

The MSE quantization of the complex variables with circularly symmetric densities is considered by Bucklew and Gallagher [5], [6] where the minimization of the MSE is done assuming that the factorization of *N* into $L\times M$ is given. Wilson [1] showed that further MSE reduction could be achieved by allowing different numbers of points on different magnitude rings. These quantizers are called unrestricted polar quantizers (UPQs) because minimization of MSE is restricted only by total number of points, and complete freedom is allowed in distribution of numbers phase levels at different magnitude levels. Wilson also demonstrated vast improvement in performance over strictly polar quantizers for small *N* (*N<36*). He did not considered large values *N* due to the complexity of his design method. The minimization of the MSE over all combinations of number of points M_i (number of phase reconstruction levels) on the magnitude reconstruction levels *i*, $1\leq i\leq L$ becomes intractable for large *N*, ($N = \sum_{i=1}^{L} M_i$) since the number of these combinations grows exponentially with *N* (it is easy to show that there are 2^{N-1} such combinations). In the paper Peric and Stefanovic [7] analyses are given for optimal asymptotic uniform polar quantization. Optimal piecewise uniform product polar quantization was considered in the paper Peric and Milovic[10]. Swaszek and Ku gave [2] an asymptotic solution for this problem but without any mathematical proof of the optimality and using, sometimes, pretty hard approximations, which limit the application. We will point out some of the lacks of this work latter.

Instead of minimizing the MSE over all integer $\boldsymbol{M}_i$, $1\leq i\leq L$, they introduced (intuitively, without any proof!) the extension of the MSE function to the real field, i.e. they minimized MSE over M_i , $1\leq i\leq L$ assuming that M_i are real numbers (we will denote these real values by P_i). They approximated such extended real MSE function and found the approximate solutions by Langrange multipliers technique. They found approximations for: (a) optimal decision and reconstruction levels r_i and m_i of the magnitude quantizer, (b) the number of points M_i on each reconstruction level *i*, $1\leq i\leq L$, and (c) the resulting MSE. Their approach does not lead to the optimal solution, although they showed an improvement over quantization with same number of phase levels on each magnitude level. Swaszek and Ku find M_i by simple rounding of each P_i, $1\leq i\leq L$.

We define the extension of the MSE over R^L (denoted by MSE(***P***)). We prove the existence of one minimum and derive the expression for evaluating $\boldsymbol{P}_{opt}(\boldsymbol{r},\boldsymbol{m})$ for fixed values of representative levels, decision levels and number of levels L. We give the procedure (with proof) for evaluating $\boldsymbol{M}_{opt}(\boldsymbol{r},\boldsymbol{m})$ from $\boldsymbol{P}_{opt}(\boldsymbol{r},\boldsymbol{m})$. On the basis of these results, we give the iterative algorithm for determining $\boldsymbol{P}_{opt}(\boldsymbol{r}_{opt}, \boldsymbol{m}_{opt})$ and $\boldsymbol{M}_{opt}(\boldsymbol{r}_{opt}, \boldsymbol{m}_{opt})$ for optimal reconstruction and representative levels. The simpliest explanation of crossing from $\boldsymbol{P}_{opt}$ to $\boldsymbol{M}_{opt}$ comes from a proof that for the convex function with one minimum gradient method can be used for finding an optimal integer solution ($\boldsymbol{M}_{opt}$). Applying our algorithm, the predictions presented in [1] are realized and incompleteness from [2] is eliminated. In algorithm we use the combination of gradient and iterative methods because the system is not very compromised by the system of non-linear equations (the exact mathematical solution of optimization problem would come to the problem of solutioning system of non-linear equations for $\boldsymbol{r}_{opt}$,$\boldsymbol{m}_{opt}$, $\boldsymbol{M}_{opt}$, λ_{opt}.

Polar quantization techniques as well as their applications in areas such as computer holography, discrete Furrier transform encoding, image processing and communications have been studied extensively in the literature. Synthetic Aperture Radars (SARs) images can be represented in polar format (i.e., magnitude and phase components). The motivation behind this work is to maintain high accuracy of phase information that is required for some applications such as interferometry and polarimetry, without loosing massive amounts of magnitude information [3]. We also give an example of the quantizer construction for a Gaussian source. This case is of importance because using Gaussian quantizer on an arbitrary source we can take advantage of the central limit theorem and the known structure of an optimal scalar quantizer for a Gaussian random variable to code a general process by first filtering it to produce an approximately Gaussian density, scalar-quantizing the result, and then inverse-filtering to recover the original . Various processing techniques, when applied to non-Gaussian sources with memory, produce sequences, which are "approximately" independent, and Gaussian [16].

The paper is organized as follows. In Section II we introduce distortion function extended on set of reals, and its minimization. In this section we also describe the optimization procedure and prove its optimality. In Section III we apply the minimization procedure to the construction of the polar quantizers. We give step by step construction procedure, and present the Gaussian source example, where we derive the MSE in a closed form. Section IV gives an example of quantization of the Gaussian memoryless source by optimal non-uniform polar quantizers. Section V gives some conclusions.

2. POLAR QUANTIZATION : DESCRIPTION AND OPTIMIZATION

We assume that the quantizer input is a complex random variable with variance $2\sigma^2$ ($\sigma^2=1$) and circularly symmetric density function $f(r)$, and we consider non-uniform polar quantizer with L magnitude levels and M_i phase reconstruction points at magnitude reconstruction level m_i, $1\leq i\leq L$. In order to minimize the distortion we proceed as follows.

First we partition the magnitude range $[0,r_L]$ into magnitude rings by L decision levels r_i $1\leq i\leq L$ $(0=r_1<r_2<...<r_L<r_{L+1}=\infty)$ .The magnitude reconstruction levels obviously satisfy $(0<m_1<m_2<...<m_L$). Next we partition each magnitude ring into M_i phase subdivisions. Let $\phi_{i,j}$ and $\phi_{i,j+1}$ be two phase decision levels, and let $\psi_{i,j}$ be j-th phase reconstruction level for the i-th magnitude ring, $1\leq j\leq M_i$. Then $\phi_{i,j}=(j-1)2\pi/M_i$, and $\psi_{i,j}=(2j-1)\pi/M_i$. Total distortion D is a combination of granulation and overload distortions, $D=D_g+D_o$, wherein

$$D_g=\sum_{i=1}^{L-1}\sum_{j=1}^{M_i}\int_{\phi_{i,j}}^{\phi_{i,j+1}}\int_{r_i}^{r_{i+1}}[r^2+m_i^2-2rm_i\cos(\phi-\psi_{i,j})]\frac{f(r)}{2\pi}drd\phi$$
$$+\sum_{j=1}^{M_L}\int_{\phi_{L,j}}^{\phi_{L,j+1}}\int_{r_L}^{r_{\max}}[r^2+m_L^2-2rm_L\cos(\phi-\psi_{L,j})]\frac{f(r)}{2\pi}drd\phi \tag{1}$$

$$D_o=\sum_{j=1}^{M_L}\int_{\phi_{L,j}}^{\phi_{L,j+1}}\int_{r_{\max}}^{\infty}[r^2+m_L^2-2rm_L\cos(\phi-\psi_{L,j})]\frac{f(r)}{2\pi}drd\phi \tag{2}$$

Obviously, for fixed L, D_g is a function of $\boldsymbol{M}=(M_i)_{1\leq i\leq L}$, $D_g=\Delta(\boldsymbol{M})$, where Δ: $Z_+^L\to R$.
In order to be able to minimize D_g over integer vectors $\boldsymbol{M}=(M_i)_{1\leq i\leq L}$ we introduce an extension of the function Δ to the real field. We denote this function Δ^e: $R^L\to R$, and write $D_g^e=\Delta^e(\boldsymbol{P})$, where $\boldsymbol{P}=(P_i)_{1\leq i\leq L}\in R^L$ denotes the argument of extended function Δ_g^e.

Now, the decision and reconstruction levels of the uniform phase quantizer at i-th magnitude level can be written as

$$\begin{aligned}&\phi_{i,j}=(j-1)2\pi/P_i \quad 1\leq j\leq\lfloor P_i\rfloor+1; \quad \psi_{i,j}=(2j-1)\pi/P_i \quad 1\leq j\leq\lfloor P_i\rfloor\\&\phi_{i,P_i}=\phi_{i,\lfloor P_i\rfloor+1}+2\pi/P_i \quad \psi_{i,P_i}=\psi_{i,\lfloor P_i\rfloor}+2\pi/P_i\end{aligned}\,.$$

wherein $\lfloor P_i\rfloor$ and $\lceil P_i\rceil$ are the smallest and the largest integer respectively, such that $\lfloor P_i\rfloor\leq P_i\leq\lceil P_i\rceil$. Using this notation the extended granulation distortion function $D_g^e(\boldsymbol{P})$ can be written as

$$D_g^e=\sum_{i=1}^{L}\sum_{j=1}^{\lfloor P_i\rfloor}\int_{\phi_{i,j}}^{\phi_{i,j+1}}\int_{r_i}^{r_{i+1}}[r^2+m_i^2-2rm_i\cos(\phi-\psi_{i,j})]\frac{f(r)}{2\pi}drd\phi$$
$$+\sum_{i=1}^{L}(P_i-\lfloor P_i\rfloor)\int_{\phi_{i,\lfloor P_i\rfloor+1}}^{\phi_{i,P_i}}\int_{r_i}^{r_{i+1}}[r^2+m_i^2-2rm_i\cos(\phi-\psi_{i,P_i)})]\frac{f(r)}{2\pi}drd\phi \,. \tag{3}$$

After the integration over ϕ and the reordering, (1) and (3) become

$$D = \sum_{i=1}^{L} \int_{r_i}^{r_{i+1}} [r^2 + m_i^2 - 2rm_i \sin c(\frac{\pi}{M_i})] f(r) dr$$

$$D^e = \sum_{i=1}^{L} \int_{r_i}^{r_{i+1}} [r^2 + m_i^2 - 2rm_i \sin c(\frac{\pi}{P_i})] f(r) dr . \tag{4}$$

wherein *sinc(x)=sin(x)/x*. Implicit use has been made of the fact that in circularly symmetric densities the magnitude random variable with probability density *f(r)* is independent of the uniformly distributed *[−π,π]* phase random variable.

The minimization problem for restricted polar quantizer can be formulated in this way: for fixed N find $\boldsymbol{r} = (r_i)_{1\le i\le L}$, $\boldsymbol{m} = (m_i)_{1\le i\le L}$, L, and $\boldsymbol{M} = (M_i)_{1\le i\le L}$ for which D is minimal. The optimal decision and reconstruction level vectors $\boldsymbol{r}$ and $\boldsymbol{m}$ are obtained from the following set of equations

$$\frac{\partial D}{\partial m_k} = 0,\ 1 \le k \le L; \tag{5}$$

$$\frac{\partial D}{\partial r_k} = 0,\ 2 \le k \le L;\ r_1 = 0,\ r_{L+1} = \infty \tag{6}$$

Solution of (5) and (6) is

$$m_i = \sin c(\frac{\pi}{M_i}) \frac{\int_{r_i}^{r_{i+1}} rf(r) dr}{\int_{r_i}^{r_{i+1}} f(r) dr};\ 1 \le i \le L \tag{7}$$

$$r_i = \frac{m_i^2 - m_{i-1}^2}{2[m_i \sin c(\frac{\pi}{M_i}) - m_{i-1} \sin c(\frac{\pi}{M_{i-1}})]};\quad 2 \le i \le L \tag{8}$$

Substituting (7) into (4) we obtain

$$D = \sum_{i=1}^{L} \int_{r_i}^{r_{i+1}} r^2 f(r) dr - \sum_{i=1}^{L} (\sin c(\frac{\pi}{M_i}))^2 \frac{[\int_{r_i}^{r_{i+1}} rf(r) dr]^2}{\int_{r_i}^{r_{i+1}} f(r) dr}$$

$$D^e = \sum_{i=1}^{L} \int_{r_i}^{r_{i+1}} r^2 f(r) dr - \sum_{i=1}^{L} (\sin c(\frac{\pi}{P_i}))^2 \frac{[\int_{r_i}^{r_{i+1}} rf(r) dr]^2}{\int_{r_i}^{r_{i+1}} f(r) dr} . \tag{9}$$

Note that formula (7) is same as in [1], [5], [6], and that formula (8) can be obtained from [1] (by solving equation (8b) from [1]).

By substitution of (7) into (9) for UPQ ($r_{L+1} = \infty$) we have

$$D = 2\sigma^2 - \sum_{i=1}^{L} m_i^2 \int_{r_i}^{r_{i+1}} f(r) dr = 2\sigma^2 - 2(\sigma^*)^2 \tag{10}$$

where

$$2\sigma^2 = \sum_{i=1}^{L} \int_{r_i}^{r_{i+1}} r^2 f(r) dr$$

is the input signal variance, while

$$2(\sigma^*)^2 = \sum_{i=1}^{L} m_i^2 \int_{r_i}^{r_{i+1}} f(r) dr$$

is the output variance. Note that equation (10) corresponds to the description of the optimal quantizer given in [8, 9]. In addition to [8] we give the procedure for finding optimal distribution of phase quantization points on magnitude levels.

Now, instead of minimizing *D(**M**)* we will minimize extended function *D^e^(**P**)* and use the solution $\boldsymbol{P}_{opt}$ for finding the optimal integer vector $\boldsymbol{M}_{opt}$. The minimization of the extended function $D^e(\boldsymbol{P})$ for fixed number of magnitude levels L constrained by total number of reconstruction points N is formulated in this way: minimize $D^e(\boldsymbol{P})$ under the constraints

$$g_0(P_1, P_2, \cdots, P_L) = N - \sum_{i=1}^{L} P_i \ge 0$$

$$g_i(P_i) = P_i \ge 0;\quad 1 \le i \le L..$$

Before we describe the minimization procedure, we prove that the problem of minimization of the $D^e(\boldsymbol{P})$ is a convex programming problem. This follows directly from Lemma 1.

Lemma 1: Function $D^e(\boldsymbol{P})$ is convex and constraints $g_0(\boldsymbol{P})$ and $g_i(P_i)$ form the convex set (the proof is given in [3]).

The next theorem gives the properties of the minimum of the $D^e(\boldsymbol{P})$ with constraints $g_0(\boldsymbol{P})$ and $g_i(P_i)$.

Theorem 1: Global minimum $\boldsymbol{P}_{opt}$ of the function $D^e\,(\boldsymbol{P})$ constrained by

$$g_0(P_1,P_2,\cdots P_L)=N-\sum_{i=1}^{L}P_i\ge 0$$

$$g_i(P_i)=P_i\ge 0,\quad 1\le i\le L\,.$$

satisfies with accuracy δ (δ<< 1) the following equation (similary as in [3])

$$P_{iopt}=N\frac{\sqrt[3]{(1+\frac{\pi^2}{15(P_{iopt})^2})m_i^2\int_{r_i}^{r_{i+1}}f(r)dr}}{\sum_{j=1}^{L}\sqrt[3]{(1+\frac{\pi^2}{15(P_{jopt})^2})m_j^2\int_{r_j}^{r_{j+1}}f(r)dr}};\quad 1\le i\le L$$

.

In order to obtain $\boldsymbol{M}_{opt}(\boldsymbol{r,m})$ from $\boldsymbol{P}_{opt}$ we use the extended function J

$$J=\sum_{i=1}^{L}\int_{r_i}^{r_{i+1}}r^2f(r)dr-\sum_{i=1}^{L}(\sin c(\frac{\pi}{P_i}))^2\frac{[\int_{r_i}^{r_{i+1}}rf(r)dr]^2}{\int_{r_i}^{r_{i+1}}f(r)dr}+\lambda\sum_{i=1}^{L}P_i\,.$$

The optimal combination for fixed values of $\boldsymbol{r}$ and $\boldsymbol{m}$ is

$$\boldsymbol{M}_{opt}(\boldsymbol{r},\boldsymbol{m})=\arg\min_{\mathbf{M}\in Z_+^L}\left(D(\boldsymbol{M})-D^{\,e}(\boldsymbol{P}_{opt})\right)=\arg\min_{\mathbf{M}\in Z_+^L}\left(J(\boldsymbol{M},\lambda_{\boldsymbol{opt}})-J(\boldsymbol{P}_{opt},\lambda_{\boldsymbol{opt}})\right)=$$

$$=\arg\min_{\mathbf{M}\in Z_+^L}\left(J(\boldsymbol{M},\lambda_{\boldsymbol{opt}})-J(\boldsymbol{M}^{\boldsymbol{(0)}},\lambda_{\boldsymbol{opt}})+J(\boldsymbol{M}^{\boldsymbol{(0)}},\lambda_{\boldsymbol{opt}})-J(\boldsymbol{P}_{opt},\lambda_{\boldsymbol{opt}})\right)$$

where λ_{opt} is obtained from the stationary condition, while $\sum_{i=1}^{L}M_i=N$, $\sum_{i=1}^{L}M_i^{(0)}\ne N$. $\boldsymbol{M}^{(0)}$ is defined in the following way:

$$\boldsymbol{M}^{\boldsymbol{(0)}}=\arg\min_{\mathbf{M}^{(l)}\in Z_+^L}\left(J(\boldsymbol{M}^{\boldsymbol{(l)}},\lambda_{\boldsymbol{opt}})-J(\boldsymbol{P}_{opt},\lambda_{\boldsymbol{opt}})\right).$$

Now the optimal combination is satisfied by:

$$J(\boldsymbol{M}_{\boldsymbol{opt}},\lambda_{\boldsymbol{opt}})-J(\boldsymbol{P}_{opt},\lambda_{\boldsymbol{opt}})=\min_{\mathbf{M}\in Z_+^L}\left(J(\boldsymbol{M},\lambda_{\boldsymbol{opt}})-J(\boldsymbol{M}^{\boldsymbol{(0)}},\lambda_{\boldsymbol{opt}})\right)+J(\boldsymbol{M}^{\boldsymbol{(0)}},\lambda_{\boldsymbol{opt}})-J(\boldsymbol{P}_{opt},\lambda_{\boldsymbol{opt}})$$

i.e. $\boldsymbol{M}_{opt}(\boldsymbol{r},\boldsymbol{m})=\arg\min_{\mathbf{M}\in Z_+^L}\left(J(\boldsymbol{M},\lambda_{\boldsymbol{opt}})-J(\boldsymbol{M}^{\boldsymbol{(0)}},\lambda_{\boldsymbol{opt}})\right).$

The nearest integer combination $\boldsymbol{M}^{(0)}$ is determined using the procedure given in the following Lemma 2.

Lemma 2: Integer combination $\boldsymbol{M}^{(0)}$ that satisfies

$$\boldsymbol{M}^{\boldsymbol{(0)}}=\arg\min_{\mathbf{M}^{(l)}\in Z_+^L}\left(J(\boldsymbol{M}^{\boldsymbol{(l)}},\lambda_{\boldsymbol{opt}})-J(\boldsymbol{P}_{opt},\lambda_{\boldsymbol{opt}})\right)$$

is determined in the following way

$$M_i^{(0)}=\arg\min_{M_i^{(l)}\in Z_+^L}\left(\varDelta J_i(M_i^{(l)},\lambda_{\boldsymbol{opt}})\right)=\begin{cases}\lceil P_{iopt}\rceil, & if\quad \varDelta J_i(\lceil P_{iopt}\rceil,\lambda_{\boldsymbol{opt}})<\varDelta J_i(\lfloor P_{iopt}\rfloor,\lambda_{\boldsymbol{opt}})\\ \lfloor P_{iopt}\rfloor, & if\quad \varDelta J_i(\lfloor P_{iopt}\rfloor,\lambda_{\boldsymbol{opt}})<\varDelta J_i(\lceil P_{iopt}\rceil,\lambda_{\boldsymbol{opt}})\end{cases}.$$

The optimal combination $\boldsymbol{M}_{opt}(\boldsymbol{r,m})$ for fixed values of $\boldsymbol{r}$ and $\boldsymbol{m}$ is determined by applying the procedure given in Lemma 3.

Lemma3: The optimal combination $\boldsymbol{M}_{opt}(\boldsymbol{r,m})$ is determined from $\boldsymbol{M}^{(0)}$ in $\left|\sum_{i=1}^{L}M_i^{(0)}-N\right|$ successive steps. If $\sum_{i=1}^{L}M_i^{(0)}=N$, then follows that $\boldsymbol{M}_{opt}=\boldsymbol{M}^{(0)}$.

If $\sum_{i=1}^{L}M_i^{(0)}>N$, then in every step we decrease the number of points for one on that level where $\varDelta J_j=-\sum_{i=1}^{k}\frac{\partial J}{\partial(M_j^{(0)}-(i-1)/k)}\frac{1}{k},\quad 1\le j\le L$ is minimum. If $\sum_{i=1}^{L} M_i^{(0)} < N$, then in every step we increase the number of points for one on that level where

$\Delta J_j = \sum_{i=1}^{k} \frac{\partial J}{\partial(M_j^{(0)} + (i-1)/k)} \frac{1}{k}, \quad 1 \le j \le L$ is minimum.

Proof: The least change of the integer vector $\boldsymbol{M}^{(0)}$ is equal to 1, and the least number of the unit changes by which we obtain the vector $\boldsymbol{M}$ satisfying $\sum_{i=1}^{L} M_i = N$ is $n = \left|\sum_{i=1}^{L} M_i^{(0)} - N\right|$. The optimal combination satisfies

$$\boldsymbol{M}_{opt} = \arg\min_{\mathbf{M}\in Z_+^L}\left(J(\boldsymbol{M}, \lambda_{opt}) - J(\boldsymbol{M}^{(0)}, \lambda_{opt})\right) = \arg\min_{\mathbf{M}\in Z_+^L}(\Delta J)$$

$$\min(\Delta J) = \sum_{l=1}^{n} \min_j(\Delta J_j^{(l)})$$

If $\sum_{i=1}^{L} M_i^{(0)} > N$, using the result from Lemma 2 $\Delta J_j \approx \sum_{i=1}^{k} d_j e_j \rho \frac{\partial(J(\boldsymbol{P}_{opt} + (i-1)\rho \boldsymbol{d}))}{\partial(P_{jopt} + (i-1)d_j e_j \rho)}, \quad 1 \le j \le L$ in this case $d_j e_j = -1, \rho = \frac{1}{k}, \Delta J_j^{(l)} = -\sum_{i=1}^{k} \frac{\partial J}{\partial(M_j^{(0)} - (i-1)/k)} \frac{1}{k}, \quad 1 \le j \le L$. If $\sum_{i=1}^{L} M_i^{(0)} < N$, $d_j e_j = 1, \rho = \frac{1}{k}, \Delta J_j^{(l)} = \sum_{i=1}^{k} \frac{\partial J}{\partial(M_j^{(0)} + (i-1)/k)} \frac{1}{k}, \quad 1 \le j \le L$.

Using approximation of the function $D^e(L)$ given in [2]

$$D^e \approx \frac{1}{12L^2}\int_0^\infty \frac{f(r)}{[g'(r)]^2}dr + \frac{\pi^2 L^2}{3N^2}\frac{[\int_0^\infty r^{1/2} f^{1/2}(r)dr]^3}{[\int_0^\infty r^{-1/4} f^{1/4}(r)dr]^2}$$

it is easy to prove that D^e (as a function of number of magnitude levels, *L)* is a convex function (we denote this as $D^e(L)$). In the above formula *g(r)* is a compressing function given by

$$g(r) = \frac{\int_0^r s^{-1/4} f^{1/4}(s)ds}{\int_0^\infty s^{-1/4} f^{1/4}(s)ds}$$

Since $\frac{\partial^2 D^e}{\partial L^2} > 0$, it follows that D^e is indeed a convex function of L.

The optimal value for number of levels L can be found in the neighborhood of the estimation based on [2]. The number of points to be checked in order to find $\boldsymbol{M}_{opt}$ is reduced from 2^{N-1} to k (where k is a number of unit intervals in the neighborhood of the estimation for L).

3 ITERATIVE ALGORITHM FOR CONSTRUCTION OF POLAR QUANTIZERS:

In this Section we will give step by step procedure for constructing optimal polar quantizer. Then we will illustrate it on constructing the quantizer for Gaussian source. We start with the construction algorithm.

Step 1)

The preliminary numbers of magnitude levels, the "numbers" of reconstruction points, the reconstruction points and decision levels are calculated by using [2]

$$L = \frac{N^{1/2}}{\sqrt{2\pi}}\frac{\int_0^\infty s^{-1/4} f^{1/4}(s)ds}{(\int_0^\infty s^{1/2} f^{1/2}(s)ds)^{1/2}}$$

$$P_{iopt} = \frac{\sqrt{2\pi}N^{1/2} f^{1/4}(m_i)m_i^{3/4}}{(\int_0^\infty s^{1/2} f^{1/2}(s)ds)^{1/2}}$$

$$r_i = g^{-1}[(i-1)/L],\ 1 \le i \le L; \quad r_{L+1} = \infty$$

$$m_i = g^{-1}[(2i-1)/2L],\ 1 \le i \le L$$

Step 2)

We calculate $r_i^{(k+1)}$, $(r_1 = 0, r_{L+1} = \infty)$, $m_i^{(k+1)}$ (by using equations (7) and (8) where P_i substitutes

M$_i$) and calculate $P_i^{(k+1)}$,*1≤i≤L* (by using Theorem 1) as

$$r_i^{(k+1)} = \frac{(m_i^{(k)})^2 - (m_{i-1}^{(k)})^2}{2[m_i^{(k)}\sin c(\frac{\pi}{P_i^{(k)}}) - m_{i-1}^{(k)}\sin c(\frac{\pi}{P_{i-1}^{(k)}})]}; \quad 2 \le i \le L$$

$$m_i^{(k+1)} = \sin c(\frac{\pi}{P_i^{(k)}})\frac{\int_{r_i^{(k)}}^{r_{i+1}^{(k)}} rf(r)dr}{\int_{r_i^{(k)}}^{r_{i+1}^{(k)}} f(r)dr}; \quad 1 \le i \le L$$

$$P_i^{(k+1)} = N\frac{\sqrt[3]{(1+\frac{\pi^2}{15(P_i^{(k)})^2})(m_i^{(k)})^2\int_{r_i^{(k)}}^{r_{i+1}^{(k)}} f(r)dr}}{\sum_{j=1}^{L}\sqrt[3]{(1+\frac{\pi^2}{15(P_j^{(k)})^2})(m_j^{(k)})^2\int_{r_j^{(k)}}^{r_{j+1}^{(k)}} f(r)dr}}; \quad 1 \le i \le L$$

When we reach the fixed point of the iterative algorithm, we obtain $\boldsymbol{P}_{opt}$, i.e. we obtain the global minimum of $D^e(\boldsymbol{P})$.

Step 3)

The existence of the reconstruction point in the coordinate origin (*z=0*) should be also examined. So we put a reconstruction level to *z=0*, and optimize the polar quantizer whose *N-1* other reconstruction points are free variables. We find the $D^e(\boldsymbol{P})$ and compare it with a quantizer without reconstruction level at *z=0*. One with minimal $D^e(\boldsymbol{P})$ is a winner.

Step 4)

In order to find $\boldsymbol{M}_{opt}$ it is necessary to apply once more the iterative method where the number of points on levels is integer. The first integer combination is determined by

$$\boldsymbol{r}^{(1)} = \boldsymbol{r}_{opt}(\boldsymbol{P}_{opt}), \boldsymbol{m}^{(1)} = \boldsymbol{m}_{opt}(\boldsymbol{P}_{opt}), \lambda_{opt}^{(1)} = \frac{\boldsymbol{2\pi^2}}{\boldsymbol{3N^3}}(\sum_{j=1}^{L}\sqrt[3]{(m_j^{(1)})^2\int_{r_j^{(1)}}^{r_{j+1}^{(1)}} f(r)dr})^3$$

$$\boldsymbol{M}^{(1)} = \boldsymbol{M}_{opt}(\boldsymbol{r}_{opt}(\boldsymbol{P}_{opt}), \boldsymbol{m}_{opt}(\boldsymbol{P}_{opt})) = \underset{\mathbf{M}\in Z_+^L}{\arg\min}\left(J(\boldsymbol{M}, \lambda_{opt}^{(1)}) - J(\boldsymbol{P}_{opt}, \lambda_{opt}^{(1)})\right).$$

Lemma 2 and Lemma 3 determine the optimal combination. $\boldsymbol{M}^{(1)}$ is used as the starting value of more accurate iterative process for finding $\boldsymbol{M}_{opt}$. This iterative process is as follows:

For integer combination, we determine iterative $\boldsymbol{r_{opt}} = \boldsymbol{r}_{opt}(\boldsymbol{M}^{(k)}), \boldsymbol{m_{opt}} = \boldsymbol{m}_{opt}(\boldsymbol{M}^{(k)})$ (by using equations (7) and (8) where $M_i^{(k)}$ substitutes M_i).

Calculate $P_i^{(k+1)}$ *1≤i≤L* (by using Theorem 1 where $M_i^{(k)}$ substitutes $P_i^{(k)}$). The integer number of points is determined by applying Lemma 2 and Lemma 3

$$\boldsymbol{M}^{(k+1)} = \boldsymbol{M_{opt}}(\boldsymbol{r}_{opt}(\boldsymbol{M}^{(k)}), \boldsymbol{m}_{opt}(\boldsymbol{M}^{(k)})) = \underset{\mathbf{M}\in Z_+^L}{\arg\min}\left(J(\boldsymbol{M}, \lambda_{opt}^{(k)}) - J(\boldsymbol{M}^{\boldsymbol{(k,0)}}, \lambda_{opt}^{(k)})\right)$$

$$\boldsymbol{M}^{\boldsymbol{(k,0)}} = \underset{\mathbf{M}^{(k,l)}\in Z_+^L}{\arg\min}\left(J(\boldsymbol{M}^{\boldsymbol{(k,l)}}, \lambda_{opt}^{(k)}) - J(\boldsymbol{P}^{\boldsymbol{(k)}}, \lambda_{opt}^{(k)})\right), \quad \lambda_{opt}^{(k)} = \frac{\boldsymbol{2\pi^2}}{\boldsymbol{3N^3}}(\sum_{j=1}^{L}\sqrt[3]{(m_j^{(k)})^2\int_{r_j^{(k)}}^{r_{j+1}^{(k)}} f(r)dr})^3$$

The iterative procedure is terminated when $\boldsymbol{M}^{(k+1)} = \boldsymbol{M}^{(k)}$ the optimum is found and $\boldsymbol{M}_{opt} = \boldsymbol{M}_{opt}(\boldsymbol{r}_{opt}(\boldsymbol{M}^{\boldsymbol{(k)}}), \boldsymbol{m}_{opt}(\boldsymbol{M}^{\boldsymbol{(k)}})) = \boldsymbol{M}^{(k)}$, $\boldsymbol{r_{opt}} = \boldsymbol{r}_{opt}(\boldsymbol{M}^{(k)}), \boldsymbol{m_{opt}} = \boldsymbol{m}_{opt}(\boldsymbol{M}^{(k)})$.

Step 5)

It has been already explained in Section II, so we will complete this explanation on an example. If *P*$_1$*<4*, then we check other combinations for *P*$_2$*, P*$_3$*,* etc, and decrease the number of magnitude levels by one and repeat whole procedure. $D^e(L)$ is convex, and optimal value for *L* is in the neighborhood of the rough estimation of *L* given in [2].

Step 6)

Optimal value for r_{max} is obtained by repeating our optimization method for different r_{max} and choosing the values for which $D=D_g+D_o$ is minimal.

4. COMPLEX GAUSSIAN SOURCE EXAMPLE

AND COMPARISON WITH FORMER PAPERS

The number of combinations that is necessary to check to find the reconstruction magnitude and phase levels for non-extended MSE function (the MSE function of integer arguments) is *k*, instead of $\sum_{L=1}^{N} C_{N-1}^{L-1} = 2^{N-1}$ combinations in Wilsons's approach (*k* is a number of trials in the procedure of finding $\boldsymbol{L}_{opt}$ by using initial value obtained from [2]).

Wilson [1] in his paper finds the best combination by searching all possible combinations. His method cannot be applied for middle and great values of total number of points N. That is the reason why he considered the polar quantization of the Gaussian source for N<36. In addition, he concludes that it will be interesting to extend investigation for great values of N.

Moo and Neuhoff showed that UPQ could be very simply analyzed and optimized by using the vector quantization (VQ) version of Benett's integral [4]. They introduced the power law polar quantization (PLPQ) in which the number of phase levels is proportional to a power of the quantized magnitude $M_i \sim m_i^{\alpha}\sqrt{N}$ [4] (where is $0 \le \alpha \le 1$). However, since this relation is not completely accurate (although it gives an important insight into the connection between the power and number of phase quantization levels), PLPQ gives worse results than UPQ from [2]. In this paper we find optimal dependence between M_i and power α and show that result from [2] can be obtained from our result.

The paper [2] presents the optimization of approximate MSE by using compression function. Due to the approximations in calculating $\boldsymbol{P}_{opt}$ (made in that paper), there is great difference between their and our result independently of the value of N.

Method presented in the paper [2] can not be applied for some values of N and given number of level L. For number of level L, the total number of points is in the range [N1-N2), $N_1 = 2(round(L) - 0.5)^2$, $N_2 = 2(round(L) + 0.5)^2$. This follows from the fact that ***r*** and ***m*** are equal for any N in the range [N1-N2), and since $\boldsymbol{P}_{opt}$ is dependent on ***m***, N and introduced approximations then $\sum_{i=1}^{L} P_i = N$ will not be satisfied. In addition, for some values of N from the former range, we cannot reach $\sum_{i=1}^{L} M_i = N$ by setting $M_i = \lfloor P_i \rfloor$ or $M_i = \lceil P_i \rceil$ as it is said in the paper [2], (there is deviation between given and calculated number of points less or equal L).

On the difference of Wilson's algorithm, our algorithm is applicable for any number of points and is much simpler, and for N<36 results obtained by using both algorithms are identical. Our algorithm is more complex then the method presented in [2], but it gives the optimal results for any number of points N and there is not any constraints in application.

We consider the complex valued Gussian source with probability density function

$$p(re^{j\theta}) = \frac{1}{2\pi} \cdot re^{\frac{-r^2}{2}}$$

and apply the iterative procedure for construction of the polar quantizer with total number of reconstruction points *N*. Fig. 1 presents the quantization signal-to-noise ratio ($SNR = 10\log(\frac{2\sigma^2}{D})$) functional dependence on rate ($R = \log_2 N$).

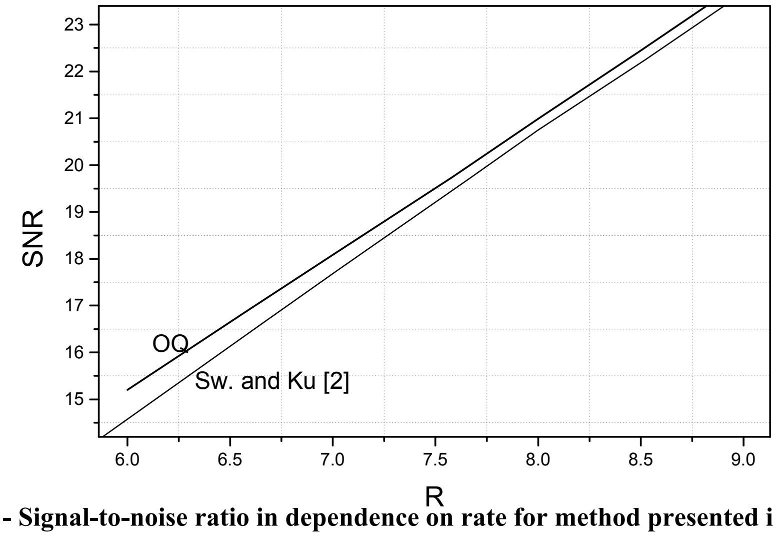

Fig. 1 - Signal-to-noise ratio in dependence on rate for method presented in [2] and our optimal quantization method for Gaussian source

With the iterative procedure for construction of the polar quantizer with the total number of reconstruction points *N=256* , the resulting MSE distortion is *D=0.0160776* . Note that the UPQ [2] has $D(=D_{[2]})=0.01683$, while optimal vector quantizer(OVQ) gives *D=0.01575*. OVQ gives the *0.28* of SNR gain relative to UPQ. Our method gives SNR which is only for 0.089 dB less than SNR of OVQ. The available design algorithms have very slow convergence unless the rate-dimension product is small [17], and implementation of the optimum vector quantizer is a computationally burdensome full search procedure. For great rates R there is deviation in the number of points on some levels (especially on the last one) and deviation in ***r*** and ***m*** among our results and results obtained by applying the algorithm from the paper [2], but the deviation in MSE is less then 0.08dB already for rates R>10.

Application of Optimal Polar Quantization:

Short-time pdf of speech segments are described by Gaussian pdf [13]. This paper addresses potential improvements achievable by means of joint quantization of two consecutive samples (x, y), referred to as two-dimensional quantization (2-D quantization), over the scalar quantization. Also a transform coding scheme known as spectral phase coding (SPC) is essentially a polar format representation of the discrete Fourier Transform (DFT) of a random phase time series. SPC utilises the discrete Fourier Transform and a two-dimensional quantizer to obtain its robust characteristics.

The design of optimal uniform polar quantization method is presented in image processing applying it on complex reflectivity function in Synthetic Aperture Radars (SARs) systems [14] (images can be represented in polar format i.e., magnitude and phase components). The motivation behind this work is to maintain high accuracy of phase information that is required for some applications such as interferometry and polarimetry, without loosing massive amounts of magnitude information. Also it may apply optimal polar quantization at Differential Pulse Code Modulation (DPCM) and Adaptive Differential Pulse Code Modulation (ADPCM). In DPCM and ADPCM systems it utilises uniform scalar quantization[12]. Optimal uniform scalar quantization for R=4 (bit/sample) has SNR=19.38dB [16] until optimal polar quantization has SNR=20.96. Optimal OPQ may achieve gain of about 1.58dB in regard to Optimal Scalar Quantization.

5 CONCLUSION

The solution given by Swaszek and Ku is the best one found by now but for big N. Wilson gives the way to find the optimal solution (full search of $\boldsymbol{M_{opt}}$) but his method can only be used for small values of N. Also, his method for finding $\boldsymbol{M_{opt}}$ and $\boldsymbol{r_{opt}}$ is more complex than search shown in our paper. Swaszek and Ku gave an asymptotic solution for this problem but without any mathematical proof of the optimality and using, sometimes, pretty hard approximations, which limit the application.

In this paper we prove the convexity of the function MSE in dependance on the disposition of number of points on levels under the constraint of the total number of points and derive the expression for determining the optimal number of points on levels for fixed values of decision thresholds and representation levels. We give the method for determining optimal integer combination and optimal real combination for the fixed values of the decision thresholds and representation levels. On the basis of these results the simple and complete iterative optimization procedure is given for constructing polar MSE quantizers for complex sources with circularly symmetric probability density. Two iterative processes are used. In the first iterative process we determine $\boldsymbol{P_{opt}}$, $\boldsymbol{r_{opt}(P_{opt})}$, $\boldsymbol{m_{opt}(P_{opt})}$ with ε accuracy. In the second one, which is more accurate, we determine $\boldsymbol{M_{opt}}$, $\boldsymbol{r_{opt}(M_{opt})}$, $\boldsymbol{m_{opt}(M_{opt})}$. We give the step by step optimization procedure (the algorithm for the optimal polar quantization) and demonstrate it on example of a Gaussian source. All incompleteness from the paper [1] and [2] are eliminated in this paper. Polar Quantization has a great application nowadays and we predict that it would have greater application in future.

REFERENCES

[1] S. G. Willson, "Magnitude/ phase quantization of independent Gaussian varieties*", IEEE Transaction on Communication*, vol. COM -28, pp. 1924-1929, Nov. 1980.

[2] P. F. Swaszek, T. W. Ku, "Asymptotic Performance of Unrestricted Polar Quantizer", *IEEE Transactions on Information Theory*, vol. 32, pp. 330-333, 1986.

[3]Z.H.Peric, B.V.Vasic, "Optimal Number Phase Divisions in Polar Quantization" Advances in Electrical and Computer Engineering, vol.3(10) Nr 1(19), pp. 5-11,2003

[4] P. W. Moo and D. L. Neuhoff "Polar Quantization revisited*", In Proc. IEEE International Symp. On Information Theory ISIT'97*, p. 60, Ulm, Germany, Jyly 1997,

[5] J. A. Bucklew and N.C. Gallagher, Jr., "Quantization schemes for bivariate Gaussian random variables", *IEEE Transactions on Information Theory*, vol. IT-25, pp. 537-543, Sept.

1979.

[6] J. A. Bucklew and N.C. Gallagher, Jr., "Two-dimensional quantization of bivariate circularly symmetric densities", *IEEE Transactions on Information Theory*, vol. IT-25, pp. 667-671, Nov. 1979.

[7] Z. Peric, M. Stefanovic, "Asymptotic Analysis of Optimal Uniform Polar Quantization", International Journal of Electronics and Communications (AEU), 56 (2002) No.5, pp.345-347.

[8] J. A. Bucklew and N. C. Gallagher, Jr., "A Note on Optimal Quantization", *IEEE Transactions on Information Theory*, vol. IT-25, No 3, pp. 365-366, May 1979.

[9] A. Gersho and R. M. Gray, *"Vector Quantization and signal Compression*", Kluwer Academ.Pub(1992).

[10] Z.H.Peric, D.Milovic "Piecewise uniform product polar quantization" International scientific journal Computing vo.2 No.3 pp. 144-152 2003

[11] T.R.Fischer, "Geometric source coding and vector quantization", *IEEE Trans. Inform. Theory*, vol.35 , pp.137-145,Jan.1989.

[12] D.Minoli, "Voice Over MPLS Planning and Designing Networks", McGraw-Hill.Pub(2002).

[13] N.S.Jayant and P.Noll, "DIGITAL CODING OF WAVEFORMS Principles and Applications to Speech and Video", Prentice-Hall, New Jersey (1984).

[14] Z.H.Peric, J.D.Jovkovic, *"Application of the Optimal Uniform Polar Quantization on Complex Reflectivity Function"* Advances in Electrical and Computer Engineering, vol.2 (9) Nr 1(17), pp. 80-85,2002.

[15] D.G.Jeong, J. Gibson" Uniform and Piecewise Uniform Lattice Vector Quantization for Memoryless Gaussian and Laplacian Sources", IEEE Trans., 1993,**IT-39**(3), pp. 786-804.

[16] K. Popat and K. Zeger, "Robust quantization of memoryless sources using dispersive FIR filters," *IEEE Trans.Commun*., vol. 40, pp. 1670-1674, Nov. 1992

[17] R.M.Gray and D.L.Neuhoff, "Quantization", *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2325-2384, October 1998.

***Zoran H. Peric** was born in Nis, Yugoslavia, in 1964. He received the B. Sc. degree in electronics and telecommunications from the Faculty of Electronic Engineering, Nis, Serbia, Yugoslavia, in 1989, and M. Sc. degree in telecommunications from the University of Nis, in 1994. He received the Ph. D. degree from the University of Nis, also, in 1999.*

He is currently Professor at the Department of Telecommunications, University of Nis, Yugoslavia. His current research interests include the information theory, source and channel coding and signal processing. He is particulary working on scalar and vector quantization techniques in image compression. He was author and coauthor in over 70 papers in digital communications. Dr Peric has been a Reviewer for IEEE Transactions on Information Theory. He is Vice-Dean of the Faculty of Electronic Engineering.

Zoran H. Peric is with the Faculty of Electronic Engineering, Aleksandra Medvedeva 14, 18000 Nis, Serbia, E-mail: peric@elfak.ni.ac.yu

***Srdjan M. Bogosavljevic** was born in Nis, Serbia, in 1967. He received the B. Sc. Degree in electronics and telecommunications from the Faculty of Electronic Engineering, Nis, Serbia, in 1992, and M. Sc. Degree in telecommunications from the Univeristy of Nis, in 1999. He has authored and coauthored 30 scientific papers. His current interests include the information theory, source coding, polar quantization.*

***Aleksandra Z. Jovanovic** was born in Nis, in 1971. She received the B. Sc. and M.Sc. degrees in electrical engineering from the Faculty of Electronic Engineering (Department of Telecommunications), University of Nis, Serbia, in 1995 and 1999, respectively. Her field of interest includes telecommunications theory, digital telecommunications, vector quantization, etc. She works as Teaching Assistant at the Faculty of Electronic Engineering.*