QUANTIZATION AND SAMPLING OF NOT NECESSARILY BAND-LIMITED SIGNALS

Milan S. Derpich, Daniel E. Quevedo*, Graham C. Goodwin* and Arie Feuer†*

*School of Electrical Engineering and Computer Science, The University of Newcastle, Australia
†Department of Electrical Engineering, The Technion-Israel Institute of Technology, Israel.

ABSTRACT

This paper presents novel results on the joint problem of sampling and quantization of non bandlimited signals. Existing literature typically focuses either on sampling in the absence of quantization, or, conversely, studies quantization for already sampled signals. Our emphasis here is on the issues that arise at the intersection of these two design problems. We argue that the joint problem can be formulated and solved to any desired level of accuracy, using moving horizon optimization methods. We present several examples which show that consideration of the combined sampling and quantization problem gives important performance gains, relative to strategies which don't specifically address the interaction between these two problems.

1. INTRODUCTION

In many DSP applications, one needs to convert continuous time, continuous valued signals, to sampled ones, which have a finite bitrate. This leads to an important set of questions regarding the best way to represent a signal by a sequence of sampled and quantized values, such that the information loss inherent in the sampling process is minimized in some sense. The above problems lead to important questions such as:" If one wants to represent a not bandlimited signal using a fixed number of bits, then how should we quantize and sample the signal to obtain the lowest possible reconstruction distortion?"

Our approach to this joint design problem is to convert it into a sampled data moving horizon optimization problem with quantized decision variables.

Background to the work described here arises from four distinct streams. The first of these is associated with the problem of sampling in the absence of quantization [1][2]. Recent work has placed these earlier results in a modern Hilbert space framework and has established connections to splines and wavelets [3][4][5]. Also, there has been work on the approximation error arising from the sampling of not bandlimited signals [6][7][3].

The second related field of research is concerned with quantization of signals where the sampling strategy has been pre ordained [8]. Recent work on this problem includes [9][10][11].

The third stream of prior work arises in the area of sampled data control theory. Here, the emphasis has typically been on regulation (zero reference) problems with unconstrained decision variables [12] [13]. We will utilize related ideas here. However, we will need to extend the concepts to account for non zero reference signals and quantized decision variables.

E-mails: milan.derpich@studentmail.newcastle.edu.au; daniel.quevedo@newcastle.edu.au; graham.goodwin@newcastle.edu.au; feuer@ee.technion.ac.il

The fourth stream of prior work includes alternative methodologies for addressing the joint design problem. These include approximating quantization as an additive noise source and using Hilbert-space methods [14] [15] and other non optimization based approaches [16].

Our approach differs from the work described above by virtue of the fact that we design the joint optimal *sampler and quantizer* using sampled data quantized moving horizon optimization. We show in section 4 that our algorithm leads to significant performance gains, compared with alternative approaches which do not take account of the interaction between sampling and quantization.

2. PROBLEM FORMULATION

As foreshadowed in the introduction, we focus our attention to the sampling-quantization structure depicted in figure 1.

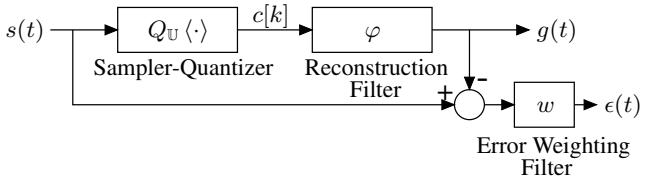

Fig. 1: Block diagram of the system

The following symbols and notation are used in figure 1 and in the remainder of this paper:

- $s(t) \in L_2$ is the input signal, not necessarily band-limitted.
- T is the sampling period. Sampling is regularly spaced.
- $f_s \triangleq 1/T$ is the sampling frequency.
- $\varphi(t)$ is the impulse response of the reconstruction filter. It is a design choice.
- $g(t)$ is the output of the system.
- $w(t)$ is the impulse response of the weighting filter. It is assumed given.
- $\mathbb{U}$ is the set of quantization levels. It is assumed given.
- The Fourier transform of a continuous time signal $f(t)$ is written $\widehat{f}(\Omega)$. For a discrete time signal $d[k]$, its discrete Fourier transform is written $\widehat{d}(e^{j\omega})$.
- The continuous time convolution of two functions $f_1(t)$, $f_2(t)$ is denoted by $(f_1 * f_2)(t)$.
- The symbol $'$ denotes the time reversion of a signal, i.e.,

$$f'(t) = f(-t)$$

.

Our goal is to find, for a given sampling rate and reconstruction filter, the pre-processing that minimizes the $\mathbf{L_2}$ norm of the frequency weighted reconstruction error

$$\|\epsilon(\cdot)\|^2_{\mathbf{L_2}} = \int_{-\infty}^{\infty} \epsilon^2(t)dt \tag{1}$$

where

$$\epsilon(t) \triangleq w(t) * [s(t) - g(t)]\,; \quad t \in \Re \tag{2}$$

and

$$g(t) = \sum_{k\in\mathbb{Z}} c[k]\varphi(t-k); \quad t \in \Re \tag{3}$$

When there are no quantization constraints, the optimal pre-processor is known to be a linear time invariant filter matched to φ, see [3]. On the other hand, when quantization is introduced, the problem must be re stated as finding the *optimal sampler-quantizer* $Q^{\circ}_{\mathbb{U}}\langle\cdot\rangle$

$$Q^{\circ}_{\mathbb{U}}\langle\cdot\rangle : s(t) \xrightarrow{Q^{\circ}_{\mathbb{U}}\langle\cdot\rangle} c^*[\cdot] \tag{4}$$

such that $\{c^{\circ}[k]\}$ is the sequence that minimizes the error

$$\|\epsilon(\cdot)\|^2_{\mathbf{L_2}} = \int_{-\infty}^{\infty} \left[v(t) - \sum_{k\in\mathbb{Z}} c[k]f(t-k)\right]^2 dt \tag{5}$$

where $v(t) \triangleq (s*w)(t)$ and $f(t) \triangleq (\varphi * w)(t)$. The quantization constraint is formulated by restricting

$$c[k] \in \mathbb{U},\ \forall k \in \mathbb{Z} \tag{6}$$

Note that the operator $Q^{\circ}_{\mathbb{U}}\langle\cdot\rangle$ is time-variant and non-linear.

We will first show that the $\mathbf{L_2}$ (continuous time) optimization problem with sampling in (5) is equivalent to an $\mathbf{l_2}$ (discrete time) optimization problem, where the weighting values are time varying and depend, inter-alia, on the continuous time signal in the time intervals between samples. This is established in the following lemma.

Lemma 1 *Define $F[k,n]$ and $Y[k]$, $k,n \in \mathbb{Z}$ via*

$$Y[k] \triangleq \int_{-\infty}^{\infty} v(t)f(t-kT)dt \tag{7}$$

$$F[k,n] \triangleq \int_{-\infty}^{\infty} f(t-kT)f(t-nT)dt \tag{8}$$

Then

$$\|\epsilon(\cdot)\|^2_{\mathbf{L_2}} = \int_{-\infty}^{\infty} v^2(t)dt - 2\sum_{k\in\mathbb{Z}} c[k]Y[k] + \sum_{k\in\mathbb{Z}}\sum_{n\in\mathbb{Z}} c[k]c[n]F[k,n] \tag{9}$$

Proof 1 *The result is obtained by Expanding (5) and exchanging the order of sum and integral.*

Remark 2 *The above result can be regarded as a more general form of sampled data control techniques.*

Remark 3 *If there was no quantization, then the optimization becomes a convex quadratic problem, and $c^*[k]$ can be found by differentiating (9) and then solving*

$$0 = \sum_{n\in\mathbb{Z}} c[n]F[k,n] - Y[k]; \quad \forall k \in \mathbb{Z} \tag{10}$$

Note that $Y[k] = (v * f')(kT)$ and $F[k,n] = (f * f')([n-k]T)$. This turns the sum in (10) into a discrete time convolution. If, in addition. we choose $\widehat{w}(\Omega) = 1$, it can be shown that $c^{\circ}[k] = \int s(\xi)\tilde{\varphi}_d\left(\frac{\xi}{T} - j\right)d\frac{\xi}{T}$, where $\tilde{\varphi}_d(\Omega)$ is the dual of φ, just as in [17].

When the coefficients $c[k]$ are constrained to belong to a finite set of scalars, finding $c^{\circ}[k]$ requires the knowledge of the entire input signal (which is unsuited for on-line applications). Furthermore, the number of required calculations grows exponentially with the length of the input sequence. Thus, $Q^{\circ}_{\mathbb{U}}\langle\cdot\rangle$ cannot be implemented online. This issue is addressed below.

3. THE MOVING HORIZON SAMPLER-QUANTIZER

The difficulties mentionned above can be avoided as follows:

At a given instant, rather than taking into account $t \in \Re$, we consider only a fixed time interval and optimize the coefficients contained therein. We then shift the horizon ahead by T and repeat the calculations, in an iterative process which yields a near-optimal result [18]. To be more precise, at iteration ℓ, the horizon of coefficients to be optimized extends from $c[\ell]$ to $c[\ell+N-1]$, $N \in \mathbb{Z}^+$. The time interval over which the error is to be minimized by these coefficients ranges from $(\ell - B)T$ to $(\ell+N)T$, $B \in \mathbb{Z}^+$. Once the optimal combination of coefficients is found, only the first of them, $c[\ell]$, is sent to the output of the sampler-quantizer. Then, the horizon is shifted forward T, and iteration $\ell+1$ begins. N can be considered the *forward horizon length*. It contains the coefficients to be optimized. On the other hand, B can be viewed as the *backward horizon length*, which accounts for non-causality of φ.

In each iteration, the cost function to be minimized with respect to $c[\ell], \ldots c[\ell+N-1]$ is defined as

$$\tilde{J}_l \triangleq \int_{(\ell-B)T}^{(\ell+N)T} \tilde{\epsilon}^2(t)dt \tag{11}$$

Compare to (1). In (11),

$$\tilde{\epsilon}(t) \triangleq v(t) - \sum_{k=-\infty}^{\ell-1} c[k]f(t-kT) - \sum_{k=\ell}^{\ell+N-1} c[k]f(t-kT) \tag{12}$$

is the filtered error excluding the effect of the coefficients beyond the horizon. The second term on the right hand side of (12) captures the effect of decisions which have already been made [1], whilst the third term of (12) captures the effects of the current optimization variables. If $f(t)$ is non-causal, the future coefficients (yet unknown) will also affect the output within the present horizon, thus unavoidably degrading the quality of the optimization. This effect is mitigated by the use of a long optimization horizon. Indeed, the horizon lengths N and B are system design parameters, which allow one to manage the trade-off between computational complexity and performance. Fortunately, our experience with related schemes shows

[1] Notice that the second term on the right hand side of (12) can be conveniently captured by the use of a state space representation for φ and w

that near optimal performance can often be obtained for rather small values of N [9][10][11]. Thus, excellent performance can often be achieved in on-line applications

If we define the difference between the filtered input signal and the part of the filtered output due to past coefficients in (12) as

$$y_\ell(t) \triangleq v(t) - \sum_{k=-\infty}^{\ell-1} c[k] f(t-kT) \tag{13}$$

and substitute (12) into (11), then the cost function for the horizon at iteration ℓ can be written in matrix form as

$$\tilde{J}_\ell = \vec{\mathbf{c}}_\ell^T \mathbf{F}_N \vec{\mathbf{c}}_\ell - 2\mathbf{Y}_\ell^T \vec{\mathbf{c}}_\ell + \int_{(\ell-B)T}^{(\ell+N)T} y^2(t)dt \tag{14}$$

where

$$\vec{\mathbf{c}}_\ell \triangleq (c[\ell], \ldots, c[\ell+N-1]) \tag{15}$$

The vector $\mathbf{Y}_\ell \in \Re^N$ and the symmetric matrix $\mathbf{F_N} \in \Re^{N\times N}$ are defined as

$$Y_\ell[j] \triangleq \int_{-BT}^{NT} y_\ell(t+\ell T) f(t-jT)dt \tag{16}$$

$$F_N[j,k] \triangleq \int_{-BT}^{NT} f(t-jT) f(t-kT)dt \tag{17}$$

We will denote $\vec{\mathbf{c}}_\ell^*$ the sequence of coeficients that minimizes (14).

The proposed algorithm, beginning at instant ℓT, can be stated as follows:

Step 1.- Calculate the matrix $\mathbf{F}_N$ in (17)
Step 2.- Calculate $\mathbf{Y}_\ell$
Step 3.- Find the optimizer $\vec{\mathbf{c}}_\ell^*$ by minimizing (14)
Step 4.- Output $c_\ell^*[\ell]$, the first element of $\vec{\mathbf{c}}_\ell^*$ (see (15))
Step 5.- Increment ℓ by 1 and go to Step 2.

The sequence $c_\ell^*[\ell]$ of step 4 forms the output of the moving horizon sampler quantizer. The algorithm reduces filtered aliasing and quantization noise. For that purpose, it does simultaneous adaptive filtering on the input signal and adaptive noise shaping of the quantization noise, thus respecting the interaction between both phenomena.

It is interesting to note that, as the horizon is made larger, the output of the sampler-quantizer defined above approaches the optimal feasible output sequence possible in (4)

4. EXPERIMENTAL RESULTS

The moving horizon sampler-quantizer defined in 3 was simulated, utilizing a Riemann approximation of the integrals, for three audio signals 22.6 [ms] long. The input sequences had zero DC value and were normalized such that $\max|s(t)| = 1$. Although the signals were bandlimited to around 20 [kHz], the effects of aliasing were induced by choosing sampling frequencies significantly below the Nyquist rate.

In all cases the forward horizon N was fixed to 4 samples. This horizon length provides near- optimal performance for a reasonably low computational cost.

For each signal, the simulation is performed for 3 to 8 quantization levels, 2 different bit-rates and two different weighting filters. The first type of weighting filter is $\widehat{w}(\Omega) = 1$. The second type of weighting filter, denoted by w_{psycho}, is an approximation of the psycho-acoustic response of the human ear [19]. Figure 2 a plot of the frequency response of $\widehat{w}_{psycho}(\Omega)$

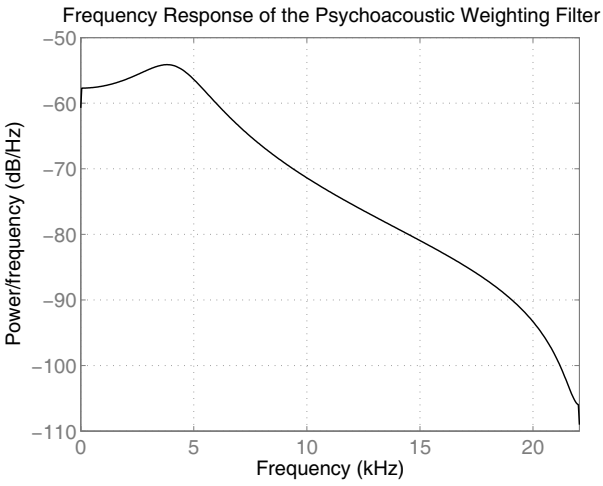

Fig. 2: Frequency Response of the Psychoacoustic Weighting Filter

For every sampling rate, the chosen reconstruction filter is formed by a linear interpolation filter followed by a high order Butterworth low-pass filter, with cut-off frequency fixed at 22.05 [kHz]. The chosen interpolation filter has impulse response given by a symmetrical B-spline of order 1, expanded according to the sampling rate:

$$\varphi(t) = \beta^1(\tfrac{t}{T}) \tag{18}$$

where

$$\beta^1(t) = (1+t)\mu(t+1) - 2t\mu(t) + t\mu(t-1) \tag{19}$$

and $\mu(t)$ is the unit step function

$$\mu(t) = \begin{cases} 0 & t < 0 \\ 1 & t \geq 0 \end{cases} \tag{20}$$

The expansion by $1/T$ adapts the frequency response of the interpolation filter to eliminate the sampling frequency from the output. At the same time, the interpolation filter attenuates the images in the output spectrum below the 22.05 [kHz].

Figure 3 shows that, as expected, decreasing the sampling rate while keeping the same number of quantization levels produces higher output distortion. However, the rate of increase in the normalized error is relatively slow, which suggests that the optimization algorithm has the potential to give reductions in normalized distortion by reducing the sampling frequency while keeping the bit-rate constant.

The latter claim is verified in table 1, which shows the normalized distortion, $\|\epsilon\|^2_{\mathbf{L_2}}/\|w * g\|^2_{\mathbf{L_2}}$ (with $\|\epsilon\|^2_{\mathbf{L_2}}$ as defined in (1)), at the output of the system for one of the input signals, using w_{psycho} as error weighting filter. For each of the two bit-rates, several combinations of sampling frequency (f_s) and number of quantization levels were simulated.

As expected, for a given number of quantization levels, a lower sampling rate yields higher distortion. However, if the bit-rate is

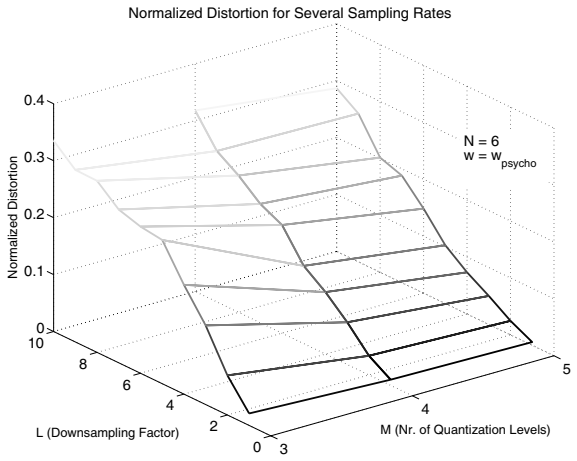

Fig. 3: Normalized Distortion as a function of sampling rate and number of quantization levels. $L \triangleq 44.1[kHz]/f_s$ is the downsampling factor.

Table 1: Normalized Distortion as a function of Bit-Rate and Number of Quantization Levels, $w = w_{psycho}$. f_s is the sampling rate

Norm. Distortion	Number of Quantization Levels					
Bit Rate	3	4	5	6	7	8
Bit-rate = 22 Kbps	.22	.12	.11	.12	.13	.15
f_s [kHz]	13	11	9.6	8.5	7.8	7
Bit-rate = 14 Kbps	.25	.20	.19	.19	.22	.24
f_s [kHz]	9.2	7	6.3	5.6	5.2	4.7

kept constant, reducing the sampling frequency can yield lower distortion, due to the reduction in the quantization step. Interestingly, as seen in table 1, the lowest distortion at a bit-rate of 22 Kbps is obtained for a sampling rate of 9.6 kHz, whereas the optimal at a bit-rate of 14 Kbps is obtained at 5.6 kHz, sampling frequencies well below the Nyquist rate in both cases.

5. CONCLUSIONS

This paper has presented a sampled data quantized moving horizon approach to the joint design of a quantization and sampler for not necessarily band limited signals. This allowed us to analyse the trade-off between quantization errors and aliasing due to sampling, obtaining the joint optimal sampler and quantizer. We showed that, for band-limited signals, particularly at low bit-rates, the optimal sampling rate can be lower than the Nyquist rate. Experimental results confirm the validity of this approach.

6. REFERENCES

[1] Michael Unser, "Sampling – 50 years after Shannon," in *Proceedings of the IEEE*, April 2000, vol. 88.

[2] P.P. Vaidyanathan, "Generalizations of the sampling theorem: Seven decades after Nyquist," *IEEE transactions on circuits and systems - I: Fundamental Theory and Applications.*, vol. 48, no. 9, pp. 1094–1109, September 2001.

[3] A. Aldroubi and M. Unser, "Sampling procedures in function spaces and asymptotic equivalence with Shannon's sampling theory," *Numer. Funct. Anal. Opt.*, vol. 15,, no. 1-2, pp. 1–21, Feb 1994.

[4] Gilbert G. Walter, "A sampling theorem for wavelet subspaces," *IEEE Transactions on information theory*, vol. 38, pp. 881–884, March 1992.

[5] Igor Djokovic and P. P. Vaidyanathan, "Generalized sampling theorems in multiresolution subspaces," *IEEE Transactions on Signal Processing,*, vol. 45, pp. 583–599, March 1997.

[6] Michael Unser and Zerubia Josiane, "A generalized sampling theory without band-limiting constraints," *IEEE Tansactions on Circuits and Systems-II: Analog and Digital Signal Processing*, vol. 45, no. 8, pp. 959–969, August 1998.

[7] Hidemitsu Ogawa, "A unified approach to generalized sampling theorems," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 11, pp. 1657–1660, April 1986.

[8] S. R. Norsworthy, R. Schreier, and G. C. Temes, Eds., *Delta–Sigma Data Converters: Theory, Design and Simulation*, IEEE Press, Piscataway, N.J., 1997.

[9] C.G. Goodwin, D.E. Quevedo, and D. McGrath, "Moving horizon optimal quantizer for audio signals," *Journal of the Audio Engineering Society*, vol. 51, March 2003.

[10] Daniel E. Quevedo and Graham C. Goodwin, "Multistep optimal analog-to-digital conversion," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 52, Issue 3, pp. 503–515, March 2005.

[11] Daniel E.. Quevedo, Graham .C. Goodwin, and Helmut Bölcskei, "Multi-step optimal quantization in oversampled filter banks," *CDC. 43rd IEEE Conference on Decision and Control, 2004.*, vol. 2, pp. 1442–1447, December 2004.

[12] T. Chen and B. A. Francis, *Optimal Sampled-Data Control Systems*, Springer-Verlag, London, 1995.

[13] A. Feuer and G. C. Goodwin, *Sampling in digital signal processing and control*, Birkäusser Boston, Cambridge, Mass., 1996.

[14] H. Bölcskei and F. Hlawatsch, "Noise reduction in oversampled filter banks using predictive quantization," *IEEE Trans. Inform. Theory*, vol. 47, no. 1, pp. 155–172, Jan. 2001.

[15] H. Bölcskei, F. Hlawatsch, and H. G. Feichtinger, "Frame-theoretic analysis of oversampled filter banks," *IEEE Trans. Signal Processing*, vol. 46, no. 12, pp. 3256–3268, Dec. 1998.

[16] Zoran Cvetković, Ingrid Daubechies, and Benjamin Logan, "Interpolation of bandlimited functions from quantized irregular samples," in *Proceedings of the data compression conference (DCC'02)*, April 2002, pp. 412–421.

[17] Terry Blu and Michael Unser, "Quantitative Fourier analysis of approximation techniques: Part I– interpolators and projectors," *IEEE Transactions on signal processing*, vol. 47, pp. 2783–2795, Oct. 1999.

[18] G. C. Goodwin, M. M. Serón, and J. A. De Doná, *Constrained Control & Estimation – An Optimization Perspective*, Springer-Verlag, London, 2005.

[19] R. A. Wannamaker, "Psycho-acoustically optimal noise shaping," *J. Audio Eng. Soc.*, vol. 40, no. 7/8, pp. 611–620, July/Aug. 1992.