SUBBAND CODING FOR FAST CONTROLLER IMPLEMENTATION

*Soura Dasgupta**
Department of Electrical and Computer Engineering
University of Iowa
Iowa City, IA 52242, USA
dasgupta@eng.uiowa.edu

ABSTRACT

This papers concerns a new wavelet packet based design methodology for efficient implementation of digital controllers that operate at very high sampling rates, but for reasons of compactness are fabricated on VLSI chips with limited surface area. We argue that this brings about a fundamental tradeoff between sampling speed, accuracy due to signal quantization, and the chip area. Here we reconcile this tradeoff using multirate subband coding techniques, that exploit spectral disparities in the closed loop signals. A precise optimization problem is formulated, and its solution presented.

1. INTRODUCTION

We propose a new wavelet packet based design methodology for efficient implementation of digital controllers that operate at very high sampling rates, but for reasons of compactness are fabricated on VLSI chips with limited surface area. There is an implicit tradeoff between the chip surface area, A, the number of computations, N, that this chip can perform and the time, T, taken to effect these computations. Roughly speaking, [1] in a large variety of VLSI technologies, under suitable normalization

$$AT^2 = N. \tag{1}$$

In a control setting, under fixed sampling rate and a given controller this tradeoff limits the number of bits that can be processed over each sampling interval. Consequently, finite word length (FWL) effects come into play, and fast sampling impairs accuracy. We note that a number of emerging control applications require very fast controllers that must be fabricated on limited chip surfaces. See for example, [3, 4], which describes the stabilization of semiconductor laser arrays, with very high open loop bandwidth.

To overcome the limitations posed by (1), we propose to exploit spectral disparities among the different subbands of the input to the controller. The goal is to minimize FWL effects while maintaining the overall speed of control action without enlarging the controller chip surface area. While FWL effects can be sourced to both signal and coefficient quantization, it is the former that will be our subject. We assume floating point arithmetic and recognize that signals with larger energy need greater bit resolution to achieve a given level of accuracy. Traditional controller architectures uniformly apply the same resolution accross the signal spectrum, without heed to the fact that some spectral bands may be more energy rich than others. Improved accuracy can be expected by allocating bit resolutions in accordance with the energy content. This gain can be consolidated by using multirate methods.

* Supported by NSF grants ECS-9970105 and CCR-9973133.

As an example, suppose the input, $e(k)$ to the digital controller is dominantly low pass. Assume that the sampling rate is f_s Hz, and that (1) permits b bits to be processed over one sample interval. Now split $e(k)$ into two signals occupying half the bandwidth of $e(k)$. In view of their halved bandwidth they can be downsampled by a factor of two. Let $e_0(k)$ and $e_1(k)$ be these downsampled signals, $e_0(k)$ containing the low frequency components of $e(k)$ and $e_1(k)$ having the residual high frequency components. Both operate at the effective sampling rate of $f_s/2$. Process the $e_i(k)$ separately through two different units which operate at the lower sampling rate. Combine their outputs at the higher sampling rate of f_s Hz. If the upsampling scheme is chosen appropriately, and the original target controller $C(z)$ at the higher sampling rate is strictly proper, and is envisioned to have the same coefficient precision as the two blocks that replace it, then under infinite signal precision one can choose the two processing units in such a way as to ensure that the input output relation of this multirate controller is precisely $C(z)$. Now (1) permits the unit processing $e_i(k)$ to process b_i bits at the reduced sampling rate as long as

$$b = \frac{b_0 + b_1}{4}. \tag{2}$$

Since b_0, and even b_1 can be chosen to be greater than b, the accuracy of this processing improves and is further enhanced if b_0 and b_1 properly reflect the energy disparity between $e_0(k)$ and $e_1(k)$. The net effect is a preservation of the chip area while achieving a higher accuracy control action *at the original sampling rate of* f_s. In certain circumstances it may well be desirable to split the signal $e(k)$ into more than two subbands, of potentially unequal spectral support. In this case the sampling rate reduction prior to processing in these subbands must correspond to the size of their respective supports.

The key focus of this research is to formulate strategies for band splitting and bit allocation between the various subband signals to achieve optimal distortion for a given bit processing rate. To this end we are influenced by the subband coding literature[1] that effects optimal band splitting through sophisticated time-frequency analyses employing wavelet packet techniques, see [8, 9] for surveys directed at control audiences. Subband coding has had a substantial impact on data compression in signal processing. One example is in [11] which achieves a 0.48 bits/pixel representation of a 8 bits/pixel image. It is expected that the improvement in the control context may be even more dramatic. For, a major constraining factor on the bit rate in signal processing is bandwidth. This has an inverse as opposed to the squared inverse, relation with the bit rate. Thus halving the sampling ratio, would require the bit bud-

[1] Many thanks to Professor Karl Astrom, who suggested this application of subband coding in controller design, to the author.

get to be

$$b = \frac{b_0 + b_1}{2} \tag{3}$$

rather than (2). Since distortion decreases exponentially with the number of bits, this is a significant difference.

Section 2 reviews the subband coding literature. Section 3 makes concrete the basic approach. Section 4 extracts an optimization problem. Section 5 solves it.

2. A REVIEW OF THE SUBBAND CODING LITERATURE

Subband coding techniques in the multirate signal processing (MSP) literature use wavelet packet based time-frequency design to employ the type of exploitation of spectral discrepancies described briefly in the introduction. Most of these techniques can be captured within the unifying framework of multirate filter banks [7]. This review will thus take a filterbank point of view to subband coder design.

Figure 1 describes an M-channel Uniform Filter Bank (UFB) that acts as a subband coder. Here the down arrow and up arrow blocks respectively describe n-fold decimators and interpolators. In particular, with $u_i(k)$ the input to the i-th decimator,

$$v_i(k) = u_i(nk). \tag{4}$$

On the other hand the n-fold interpolator raises the sampling rate by a factor of n by padding the interval between the successive samples of $x(k)$ by $n-1$ zeros: i.e. with $\nu_i(k)$ the output of the i-th interpolator

$$\nu_i(k) = \begin{cases} w_i(k/n); & k \bmod n = 0 \\ 0; & else \end{cases} . \tag{5}$$

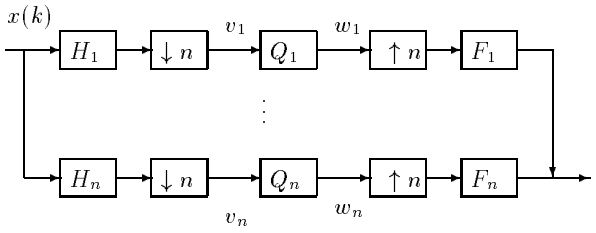

Figure 1. Uniform Filter Bank

In fig. 1, Q_i represents a b_i bit quantizer, $H_i(z)$ and $F_i(z)$ are respectively called the *analysis and synthesis* filters. The $H_i(z)$ split $x(k)$ the signal to be coded, into n subband signals $v_i(k)$, whose quantized versions are transmitted, and resynthesized at the receiver site by the $F_i(z)$. The arrangement to the left of the Q_i is called the *Analysis Bank (AB)*, and that to right of the Q_i, *Synthesis Bank (SB)*. In each channel, the subband signals $v_i(k)$, undergo a sampling rate reduction by the factor of n. The synthesis side recovers the original baud rate by raising the sampling rate by the same factor. The filter H_i extracts a suitable component of $x(k)$ occupying the bandwidth of $2\pi/n$ consistent with the rate reduction in the i-th channel. Frequently one enforces the *Perfect Reconstruction (PR)* proprty i.e. absent quantizers, the input output relation is z^{-n+1}. Most practical subband coders are also *Orthonormal*: i.e. input energy equals the collective energy in the $v_i(k)$.

Now turn to the subband coding problem. The average bit rate of transmission is kept constant, i.e. for a fixed b the b_i obey

$$b = \frac{1}{n}\sum_{i=1}^{n} b_i. \tag{6}$$

For $n = 2$, this reduces to (3). Under (6) and orthonormality, the goal is to select the b_i and the H_i and F_i to minimize the average quantization induced distortion at the output. See [10] for a solution.

To obtain some insights consider the case $n = 2$. As (3) holds, to improve fidelity one should assign more bits to the subband with higher energy and fewer bits to the other subband. This guides bit allocation. Taking this procedure further, one would then select the two analysis filters to ensure that one subband has as large a signal energy as possible and the other the smallest energy possible. This must be done within the constraint of orthonormality and band splitting of equal support. Thus, the optimal solution should have two ingredients: (i) *Energy Compaction* whereby energy is concentrated in one subband over the other, and (ii) *Optimum bit allocation* where, subject to the bit budget (3), more bits are assigned to the channel with larger energy concentration. In particular, energy compaction is effected by the selection of the H_i.

Typically the quantizer noise model for a b_i-bit quantizer Q_i is as follows. If the signal $v_i(k)$ is zero mean Wide Sense Stationary (WSS), its output obeys, [11]

$$w_i(k) = v_i(k) + q_i(k) \tag{7}$$

where $q_i(k)$ is *zero mean, WSS white, independent from* $v_i(k)$ and has variance

$$\sigma_{q_i}^2 = c2^{-2b_i}\sigma_{v_i}^2. \tag{8}$$

Here $\sigma_{v_i}^2$ is the variance of the subband signal v_i and c is a constant determined by the signal distribution. In fact one can further assume that $q_i(k)$ is also independent of $q_j(k)$, $i \neq j$. These assumptions generally hold at high bit rates, [11]. A key feature is that it captures our earlier observation that signals with large energy and consequently larger dynamic range, require larger number of bits to achieve a given level of distortion. These models have also been used in the controls literature under the name of Finite Signal to Noise models, [12].

The n-channel uniform filter bank is equivalent to its lifted version, [2], depicted in fig. 2. The Linear Time Invariant (LTI) $n \times n$ systems $E(z)$ and $R(z)$ have elements that bear a one to one correspondence with the $H_i(z)$ and $F_i(z)$, respectively. Specifically, if one expresses:

$$H_i(z) = \sum_{j=1}^{n} z^{-j+1} E_{ij}(z^n) \tag{9}$$

and

$$F_i(z) = \sum_{j=1}^{n} z^{-n+j} R_{ji}(z^n), \tag{10}$$

then the ij-th elements of $E(z)$ and $R(z)$ are E_{ij} and R_{ij}, respectively. Note that the elements of the *i-th row* of $E(z)$ provide $H_i(z)$, while the elements of the *i-th column* of $R(z)$ provide $F_i(z)$. In particular the $\mathcal{H}_2$ norm of $F_i(z)$ is given as below with e_i the vector with the i-th element 1, and the rest zero.

$$\|F_i(z)\|_2^2 = e_i' \int_0^{2\pi} R'(e^{-j\omega}) R(e^{j\omega}) e_i d\omega/2\pi, \tag{11}$$

This uniform filter bank is PR iff

$$R(z) = E^{-1}(z). \tag{12}$$

It is Orthonormal, if in addition to (12) $E(z)$ and hence also $R(z)$ is all pass, i.e.

$$E'(e^{-j\omega})E(e^{j\omega}) = R'(e^{-j\omega})R(e^{j\omega}) = I \quad \forall\omega. \tag{13}$$

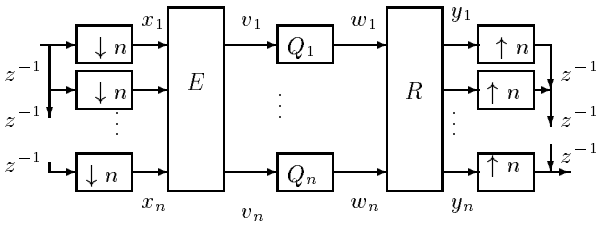

Figure 2. Lifted Description of Uniform Filter Bank

An important special case is when E and R are constant. Such a subband coder is often called a *Transform Coder* and will play an important role in this paper.

3. THE BASIC APPROACH

For a controller of relative degree $n-1$ or more, we split its input into n-bands of equal support, i.e. the equivalent of the n-channel uniform filter bank. Such a setting is depicted in fig. 3. Note that to avoid delay free loops one generally selects controllers to have relative degree of at least one. If the relative degree is one then a 2-channel arrangement must be employed. Higher relative degrees permit greater amount of band splitting.

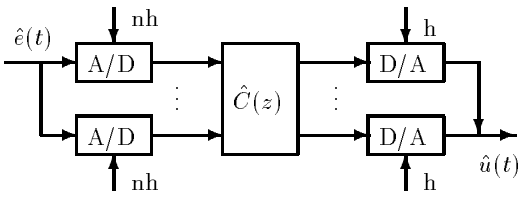

Figure 3. The Basic Setting

In fig. 3 the continuous time input $\hat{e}(t)$ is sampled via the A/D convertors at $1/n$-th the required sampling frequency $1/h$. Of course all A/D convertors require anti-aliasing *analog* filters. The A/D blocks subsume these filters. In this case, however, these filters also play the dual role of band splitting filters. The only thing being implemented in the digital domain is the $\hat{C}(z)$ the $n \times n$ implementation of target SISO controller $C(z)$. The relation between $C(z)$ and $\hat{C}(z)$ will be explained presently. For the moment, note that the processing of the bandsplit signals is performed in parallel, but separately by $\hat{C}(z)$ through its n-columns. Once n separate outputs are generated by the $\hat{C}(z)$ they are converted to the analog domain by the D/A convertors that must also bring the sampling rate back to $1/h$. Now, the upsampling process implicit in these D/A blocks generally creates redundant images in the spectral domain. For example the output of an n-fold interpolator spectrum is $2\pi/n$ rather than 2π periodic. In general these should be eliminated [7]. To this end one needs anti-imaging filters which are combined with the D/A blocks and are again implemented in the analog domain. It may appear at first sight that the need for these additional filters over and above the traditional anti-aliasing filters imposes an additional burden representing a disadvantage of this method. Yet, there is an advantage. Essentially, the presence of these filters can the relax the cutoff requirements on anti-aliasing filters. Finally, the D/A block outputs are combined to synthesize the continuous time control input $\hat{u}(t)$.

Consider now the discrete time equivalent closed loop depicted in fig. 4. Here $P(z)$ represents the discrete time plant

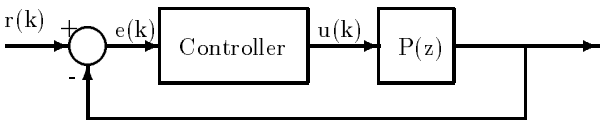

Figure 4. Discrete Time Closed Loop

model. The block labelled as "Controller" in that figure is as in fig. 5. In this figure the H_i and F_i are discrete time models of the analog antialiasing and anti-imaging filters respectively. It is also assumed that the upsampling processes implicit in the D/A blocks in fig. 3 are consistent with the upsampling definition given in Section 2. Since the antialiasing and anti-imaging filters are implemented in the analog domain, the only FWL issues are those associated with $\hat{C}(z)$. As a mathematical abstraction we will assume that the bit resolution associated with the processing of the two inputs to $\hat{C}(z)$ is lumped at the input points to $\hat{C}(z)$, specifically by the n-quantizers Q_i at the input of $\hat{C}(z)$. We assume that Q_i is a b_i-bit quantizer whose distortion is modelled by the quantizer noise $q_i(k)$, satisfying the assumptions given in Section 2.

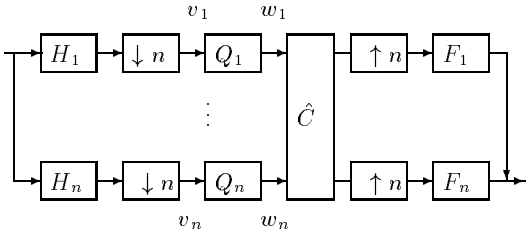

Figure 5. An Abstraction

We next turn to the relation between $\hat{C}(z)$ and the *desired* SISO controller $C(z)$. Write,

$$C(z) = z^{n-1} \sum_{i=0}^{n-1} z^{-1} C_i(z^n). \tag{14}$$

Define $\tilde{C}(z)$ as the matrix

$$\hat{C}(z) = \begin{bmatrix} C_0(z) & C_1(z) & \cdots & C_{n-1}(z) \\ z^{-1}C_1(z) & C_0(z) & \cdots & C_{n-2}(z) \\ \vdots & \vdots & \vdots & \vdots \\ z^{-1}C_{n-1}(z) & z^{-1}C_{n-2}(z) & \cdots & C_0(z) \end{bmatrix}. \tag{15}$$

Provided [7],

$$R(z)\hat{C}(z)E(z) = \tilde{C}(z), \tag{16}$$

the relationship from $e(k)$ to $u(k)$ is in fact $C(z)$. To preserve internal stability, it will be necessary for $E(z)$ and $R(z)$, to be minimum phase and stable. We comment more on the choice of $R(z)$, $\hat{C}(z)$ and $E(z)$ below.

Several technical differences from the subband coding case emerge. First the bit budget is n times more generous. However, though this has beneficial practical consequences in analysis terms it is inconsequential. More importantly the constraint (16) injects a shaping function $\hat{C}(z)$. Further, one cannot now assume $E(z)$ to be all pass, precluding orthonormality as an option. For, suppose $E(z)$ is all pass,

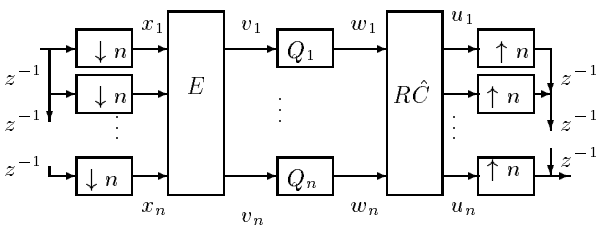

Figure 6. A Lifted Model

causal and stable. Then it must be nonminimum phase. Consequently its causal inverse is unstable, and the minimum phase stable requirement on $R(z)$, $E(z)$ is violated. Thus this additional constraint must be placed on $E(z)$. In general it is $E(z)$ that is determined first, and then (16) is invoked to obtain $R(z)$. In the next section we make precise an optimization problem, that subject to a fixed chip area, i.e. a bit budget, seeks to minimize the quantization induced mean square distortion at $u(k)$.

Note for a fixed area the bit budget constraint becomes:

$$b = \sum_{i=0}^{n-1} \frac{b_i}{n^2}, \tag{17}$$

the obvious extension of (6). The underlying optimization will be with respect to the selection of $E(z)$, and the *bit distribution* b_i. As will become evident in the sequel, the for a given plant and desired $C(z)$, the optimization is independent of the choice of $\hat{C}$, as long as (16) is satisfied. In fact the optimization yields either $E(z)$ or $R(z)$, depending on which way one phrases the problem, and the other quantities must simply obey (16). Obviously, the remaining quantities must be selected to make the implementation as efficient as possible. Our focus here is simply on the selection of either $E(z)$ or $R(z)$. Ways of searching for the other matrices will be the subject of a different article.

By selecting $E(z)$ and $R(z)$, one selects $H_i(z)$ and $F_i(z)$, which are in fact implemented in analog domain, and are part of the A/D and D/A convertors. Thus, there is also an implicit analog approximation problem, that is not the subject of this paper. Also observe, that the $H_i(z)$ are the discrete equivalent of analog filters that are part of anti-aliasing filters. Anti-aliasing filters are selected from a variety of considerations, [13] that range beyond the issues considered here. One way of reconciling the many requirements on anti-aliasing filter selection, is to design, a single, anti-aliasing filter in a standard way, [13], (call this the standard part), and treat the analog, counterparts of the bandsplitting filters yielded by the design methodology given here, as adjuncts of this standard anti-aliasing filter, e.g. through the arrangement in fig. 7, where the $\hat{H}_i(s)$ are the analog equivalents of $H_i(z)$. In this approach the band splitting optimization will be conducted with the standard part incorporated in the continuous time plant model.

4. AN OPTIMIZATION PROBLEM

In this Section we extract a core optimization problem that guides the selection of the b_i, $R(z)$, $\hat{C}(z)$ and $E(z)$. In the sequel assume that $\tilde{P}(z)$ is the n-fold lifted version of the discrete time plant. Our goal is to minimize the mean square distortion induced by the quantizer at the output of fig. 6. We assume that the quantizer noise generated by the i-th quantizer obeys the assumptions in Section 2. Specifically, they obey (7) and (8), together with whiteness and independence. In general in a closed loop setting, the the

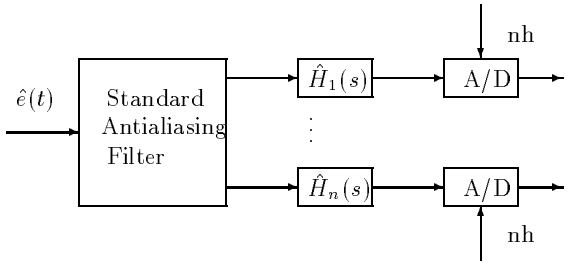

Figure 7. Anti-Aliasing Implementation

feedback effect of $\sigma_{q_i}^2$ on itself creates a nonlinear system. However one can formally show that for large b in (17), because of (7) and (8), the following constitutes a good first approximation: Compute $\sigma_{v_i}^2$, from the statistics of $r(k)$, assuming no quantization. Compute $\sigma_{q_i}^2$ using this value of $\sigma_{v_i}^2$. The resulting relative error, i.e. the ratio of the error in computing the distortion using this approximation, and the actual error is only, $O(2^{-2b})$, and for large b, does not impact the optimization procedure.

In the sequel assume that $\tilde{P}(z)$ is the n-fold lifted version of the discrete time plant $P(z)$ in fig. 4 and $\tilde{r}(k) = [r(nk), r(nk-1), \cdots, r(nk-n+1)]'$. In fig. 6, define $\tilde{u}(k) = [u_1(k), \cdots, u_n(k)]'$, $v(k) = [v_1(k), \cdots, v_n(k)]'$ and $w(k) = [w_1(k), \cdots, w_n(k)]'$. The elements of $\tilde{u}(k)$ are certain samples of the output in fig. 6.

Because of (16) that the transfer function relating $\tilde{r}(k)$ to $v(k)$, absent quanitizers, is

$$T_{rv}(z) = E(z)(I + \tilde{P}(z)\tilde{C}(z))^{-1}. \tag{18}$$

Call $q(k) = [q_1(n), \cdots, q_n(k)]'$. Further, one has:

Lemma 1 *With the quantizers absent, the closed loop transfer function relating $q(k)$ to $\tilde{u}(k)$ is given by*

$$T_{qu}(z) = (I + \tilde{C}(z)\tilde{P}(z))^{-1}\tilde{C}(z)E^{-1}(z). \tag{19}$$

Proof: One has from (16) and figs 4 and 6 that

$$\begin{aligned} T_{qu}(z) &= R(z)\hat{C}(z)(I + E(z)\tilde{P}(z)\hat{C}(z)R(z))^{-1} \\ &= ((\hat{C}^{-1}(z)R^{-1}(z) + E(z)\tilde{P}(z))^{-1} \\ &= ((E(z)\tilde{C}^{-1}(z) + E(z)\tilde{P}(z))^{-1} \\ &= (I + \tilde{C}(z)\tilde{P}(z))^{-1}\tilde{C}(z)E^{-1}(z). \end{aligned}$$

■

We assume that $r(k)$ is a *zero mean WSS, Gaussian process*, and that the Power Spectral Density (PSD) matrix of $\tilde{r}(k)$ is $\tilde{S}_r(\omega)$. Note the output of fig. 6 is no longer WSS but has n-periodic statistics. It thus makes sense to minimize the *average* variance of the effect of the $q_i(k)$. Define

$$\tilde{F}_i(z) = \sum_{j=1}^{n} z^{-n+j} \left[T_{qu}(z^n)\right]_{ji}. \tag{20}$$

Because of (8), and the whiteness assumption on the q_i, one must thus minimize

$$\frac{1}{n}\sum_{i=1}^{n} 2^{-2b_i}\sigma_{v_i}^2 \|\tilde{F}_i\|_2^2. \tag{21}$$

Invoke the well known AM-GM inequality that states that the arithmetic mean of a set of numbers is bounded from below by their geometric mean, with the bound met iff all the numbers are the same. One obtains that under (17),

$$\frac{1}{n}\sum_{i=0}^{n-1} 2^{-2b_i}\sigma_{v_i}^2\|\tilde{F}_i\|_2^2 \geq c2^{-2nb}\left(\prod_{i=0}^{n-1}\sigma_{v_i}^2\|\tilde{F}_i\|_2^2\right)^{1/n},$$

with equality iff for all i, j,

$$2^{-2b_i}\sigma_{v_i}^2\|\tilde{F}_i\|_2^2 = 2^{-2b_j}\sigma_{v_j}^2\|\tilde{F}_j\|_2^2. \tag{22}$$

This constitutes **optimum bit allocation.** It is consistent with the intuition of assigining fewer bits to lower energy signals. Consequently, subject to the orthonormality condition, the filter design reduces to the minimization of

$$J = \prod_{i=0}^{n-1}\sigma_{v_i}^2\|\tilde{F}_i\|_2^2. \tag{23}$$

Observe also, from (20) that

$$\|\tilde{F}_i\|_2^2 = e_i'\int_0^{2\pi} T_{qu}'(e^{-j\omega})T_{qu}(e^{j\omega})e_i d\omega/2\pi, \tag{24}$$

and from fig. 6

$$\sigma_{v_i}^2 = e_i'\int_0^{2\pi} T_{rv}(e^{j\omega})\tilde{S}_r(\omega)T_{rv}'(e^{-j\omega})e_i d\omega/2\pi. \tag{25}$$

Now note, under internal stability of the unquantized closed loop, and provided $E(z)$ is stable minimum phase, both $T_{qu}(z)$ and $T_{rv}(z)$ are stable. Thus, defining

$$\begin{aligned} S_1(\omega) &= (I+\tilde{P}(e^{j\omega})\tilde{C}(e^{j\omega}))^{-1}\tilde{S}_r(\omega) \\ &\quad \left[(I+\tilde{P}(e^{-j\omega})\tilde{C}(e^{-j\omega}))^{-1}\right]' \end{aligned} \tag{26}$$

and

$$\begin{aligned} S_2(\omega) &= \left[(I+\tilde{C}(e^{-j\omega})\tilde{P}(e^{-j\omega}))^{-1}\tilde{C}(e^{-j\omega})\right]' \\ &\quad (I+\tilde{C}(e^{j\omega})\tilde{P}(e^{j\omega}))^{-1}\tilde{C}(e^{j\omega}). \end{aligned} \tag{27}$$

Then,

$$\|\tilde{F}_i\|_2^2 = e_i'\int_0^{2\pi}\left[E^{-1}(e^{-j\omega})\right]' S_2(\omega)E^{-1}(e^{j\omega})d\omega/2\pi e_i, \tag{28}$$

and

$$\sigma_{v_i}^2 = e_i'\int_0^{2\pi} E(e^{j\omega})S_1(\omega)E'(e^{-j\omega})e_i d\omega/2\pi. \tag{29}$$

Thus subject to optimum bit allocation, one must find $E(z)$, stable and minimum phase, for which (23) is minimized under (28) and (29). *Note, neither $S_1(\omega)$, nor $S_2(\omega)$ depend on either $R(z)$ or $\hat{C}(z)$.* Thus, one optimizes for $E(z)$ and then solves (16) to find $\hat{C}(z)$ and $R(z)$. A reinterpretation of (16) is as follows. Since $E(z)$ is stable, minimum phase $\tilde{C}(z)E^{-1}(z)$ can be seen as being in the left coprime factor form. Find a right coprime factorization, choose $R^{-1}(z)$ and $\hat{C}(z)$ as the "numerator and denominator" of this right factorization.

In the sequel we will restrict attention to the analog of the *Transform Coding* problem, in that in the optimization above, our search space will be restricted to *constant E*. Note that unlike Transform coding, we do not assume a unitary E. Nor is R a constant. Under these conditions, the following problem must be solved.

Problem 1 Define $\Sigma_i = \int_0^{2\pi} S_i(\omega)d\omega/2\pi = \Sigma_i' > 0$. Find a *constant E* that minimizes:

$$J_T = \prod_{i=1}^{n}\left[\left(e_i'E\Sigma_1E'e_i\right)\left(e_i'(E^{-1})'\Sigma_1E^{-1}e_i\right)\right]. \tag{30}$$

5. SOLUTION

The transform coding problem, Problem 1, is solved through the two Theorems given below.

Theorem 1 *For a given pair of real matrices $\Sigma_i = \Sigma_i' > 0$, suppose $E = \hat{E}$ optimizes Problem 1. Then for at least one such $\hat{E}$,*

$$\hat{E}\Sigma_1\hat{E}' = (\hat{E}^{-1})'\Sigma_2\hat{E}^{-1} = \Lambda, \tag{31}$$

where Λ is a real diagonal matrix.

Proof: Call

$$\Sigma_1^* = \hat{E}\Sigma_1\hat{E}', \text{ and } \Sigma_2^* = (\hat{E}^{-1})'\Sigma_2\hat{E}^{-1}.$$

Now consider,

$$J_T(E) = \prod_{i=1}^{n}\left[\left(e_i'E\Sigma_1^*E'e_i\right)\left(e_i'(E^{-1})'\Sigma_1^*E^{-1}e_i\right)\right].$$

Then for all i, j,

$$\left.\frac{\partial J_T(E)}{\partial E_{ij}}\right|_{E=I} = 0.$$

Call σ_{li}^* the ij-th element of Σ_l^*. For $i \neq j$, direct evaluation reveals

$$\frac{\sigma_{1ij}^*}{\sigma_{1ii}^*} = \frac{\sigma_{2ji}^*}{\sigma_{2ii}^*}.$$

Observe, that premultiplying of E by a diagonal matrix, does not change J_T. Thus, without loss of generality, one can assume that

$$\sigma_{1ii}^* = \sigma_{2ii}^*.$$

Thus, as the Σ_i^* are symmetric, the first equality in (31) holds, i.e. $\Sigma_1^* = \Sigma_2^* = \Sigma^*$. Now it is well known, [14], that the product of the diagonal elements of a positive definite symmetric matrix is lower bounded by the product of its eigenvalues. Choose Λ to be the diagonal matrix of the eigenvalues of Σ^*. There is a unitary matrix Ω such that $\Omega\Sigma^*\Omega' = \Lambda$. Then the result follows by noting that with $T = \Omega\hat{E}$, $T\Sigma_1T' = (T^{-1})'\Sigma_2T^{-1} = \Lambda$. ∎

We now show that such an E always exists, and that the target Λ is in fact *unique* to within a permutation.

Theorem 2 *Consider a pair of positive definite real symmetric matrices A and B. Then there exists a T and a diagonal, positive definite Λ for which*

$$TAT' = (T^{-1})'BT^{-1} = \Lambda. \tag{32}$$

Further Λ is unique to within a permutation.

Proof: There exists positive definite diagonal matrix Λ_A and an unitary matrix U_A, for which

$$A = U_A \Lambda_A U_A'. \tag{33}$$

Define the unitary matrix W and the real positive definite diagonal matrix Λ for which

$$\Lambda_A^{1/2} U_A' B U_A \Lambda_A^{1/2} = W \Lambda^2 W'.$$

Choose

$$T = \Lambda^{1/2} W' \Lambda_A^{-1/2} U_A'.$$

Then,

$$\begin{aligned} TAT' &= \Lambda^{1/2} W' \Lambda_A^{-1/2} U_A' \left[U_A \Lambda_A U_A' \right] U_A \Lambda_A^{-1/2} W \Lambda^{1/2} \\ &= \Lambda. \end{aligned}$$

Further,

$$\begin{aligned} (T^{-1})' B T^{-1} &= \Lambda^{-1/2} W' \left[\Lambda_A^{1/2} U_A' B U_A \Lambda_A^{1/2} \right] W \Lambda^{-1/2} \\ &= \Lambda^{-1/2} W' \left[W \Lambda^2 W' \right] W \Lambda^{-1/2} \\ &= \Lambda. \end{aligned}$$

Thus the required T exists. We now prove its uniqueness to within a permutation. Suppose, there are two real diagonal $\Lambda_i > 0$ and T_i, $i = 1, 2$ for which

$$T_i A T_i' = \Lambda_i = (T_i^{-1})' B T_i^{-1}. \tag{34}$$

Observe, for any Y and positive definite symmetric X,

$$YY' = X$$

implies that

$$Y = X^{1/2} U$$

where, [14] $X^{1/2}$ is the unique positive definite, symmetric square root of X and U is any unitary matrix. Thus every T_i, Λ_i that satisfy the first equality in (34), obey for arbitrary unitary matrices U_i,

$$T_i = \Lambda_i^{1/2} U_i A^{-1/2}.$$

Likewise every T_i, Λ_i that satisfy the first equality in (34), obey for arbitrary unitary matrices V_i,

$$(T_i^{-1})' = \Lambda_i^{1/2} V_i B^{-1/2}.$$

Thus,

$$T_i = \Lambda_i^{1/2} U_i A^{-1/2} = \Lambda_i^{-1/2} V_i B^{1/2}.$$

Define $G = B^{1/2} A^{1/2}$. Then

$$G = V_1' \Lambda_1 U_1 = V_2' \Lambda_2 U_2.$$

Thus, both Λ_i are matrices whose diagonal elements are the singular values of G. Consequently they must be the same to within a permutation. ■

This Theorem, also gives a method for constructing T. Together, the two theorems show how, the optimizing solution can be obtained. Notice in particular this ensures that the subband signals v_i are uncorrelated. This is tantamount to a energy compaction condition.

6. CONCLUSION

We have proposed a new approach to designing fast controllers that have to be fabricated on chips of limited area. Our approach involves multirate implementations of the target controller, and employs subband coding techniques. Specifically, the target SISO controller is implemented as a multi-input multi-output system operating at a reduced sampling rate, and whose inputs are signals occupying different spectral bands of the the target controller input. Our focus is on signal quantization rather than coefficient quantization. The optimizing framework considered here involves a variation of transform coding, and seeks to optimally allocate bits and select band splitting analog filters to minimize the quantizer induced distortion. Future directions of research include: (i) relaxing the transform coding requirement of a constant E; (ii) detailed consideration of synthesizing the analog filters from the H_i, F_i; (iii) the use of nonuniform filter banks; and (iv) direct anti-aliasing and anti-imaging device design, using sampled data techniques. We are currently pursuing these extensions, together with experimental validation of these results using laser control.

REFERENCES

[1] J. D. Ullman, *Computational Aspects of VLSI*, Computer Science Press, 1984.

[2] B. Friedland, "Sampled data control systems containing periodically time varying members", *Proc. Ist IFAC World Congress.*

[3] S. Dasgupta and D.R. Andersen, "Feedback Stabilization of semiconductor laser arrays", *Journal of Optical Society of America B*, vol. 11 , No. 2, pp 290-296, 1994.

[4] D.E. Hill, S. Dasgupta, K.M. Nagpal and D.R. Andersen, "Feedback Stabilization of semiconductor laser arrays with complex feedback coefficients", *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 1, pp 150-164, June, 1995.

[5] D. Williamson and K. Kadiman, "Optimal finite wordlength linear quadratic regulation", *IEEE Transactions on Automatic Control*, vol. AC-34, pp 1218-1288, December 1989.

[6] G. Li and M. Gevers, "Optimal finite precision implementation of a state-estimate feedback controller", *IEEE Transactions on Circuits and Systems*, pp 1487-14987, December 1990.

[7] P.P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice Hall, 1992.

[8] S. Dasgupta, "A Glimpse of Multirate Signal Processing", *Proc. of the European Control Conference: Compilation of Plenary and Semiplenary talks*, 1997.

[9] B. Francis and S. Dasgupta, "Signal compression by subband coding", *Invited Paper*. Special issue on control methods in communications. *Automatica*, December 1999.

[10] P.P. Vaidyanathan, "Theory of optimal orthonormal subband coders", *IEEE Transactions on Signal Processing*, pp 1528-1543, June, 1998.

[11] N.S. Jayant and P. Noll, *Digital Coding of Waveforms*, Prentice Hall, 1984.

[12] R. E. Skelton and J. Lu, "Iterative identification and control of finite signal to noise models", *Mathematical Modeling of Systems*, Vol. 2, 1996.

[13] K. Astrom and B. Wittenmark, *Computer Controlled Systems: Theory and Design*, 2nd Ed. Prentice Hall, 1990.

[14] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, Academic Press, 1985.