Moving-Horizon Optimal Quantizer for Audio Signals*

GRAHAM C. GOODWIN AND DANIEL E. QUEVEDO

School of Electrical Engineering and Computer Science, University of Newcastle, Callaghan, NSW 2308, Australia

AND

DAVID MCGRATH, AES Member

Lake Technology Ltd., Ultimo, NSW 2007, Australia

By analyzing the quantization of audio signals as a deterministic finite-set constrained quadratic optimization problem, a new scheme, called moving-horizon optimal quantizer (MHOQ), is developed. The MHOQ includes a model of the ear’s sensitivity to low-level noise power and minimizes directly the perceived error over a finite prediction horizon. Feedback is incorporated by means of the moving-horizon principle. With a prediction horizon equal to 1, the MHOQ reduces to the psychoacoustically optimal noise-shaping quantizer, widely used in practical applications. Larger prediction horizons outperform the noise shaper at the expense of only a small increase in computational complexity.

0 INTRODUCTION

In many applications continuous-valued signals have to be converted into digitized ones whose amplitude belongs to a finite set. This procedure is called quantization and forms an essential part of analog-to-digital converters.

The related issue of requantization corresponds to lowering the resolution of a finely quantized signal in order to provide a coarser quantized signal and can often be dealt with as if it were a quantization problem. (In the sequel we will use the term quantize to denote both situations.) This problem arises in the mastering process of compact discs, where each sample must be quantized to 16 bit, based on master recordings of higher resolution (such as 24 bit) or of analog nature.

Quantization unavoidably introduces loss of information. In the case of audio signals, the challenge is to preserve the perceived sound quality as much as possible. The final aim corresponds to making the difference between the original and the quantized signal, as far as possible, inaudible to the listener.

In view of the importance of the perceived sound quality, a significant research effort has been concentrated on the development of psychoacoustically optimal quantizers. The most popular scheme in this context is the noise-shaping quantizer [1]–[3].

In this engineering report we present a new converter, which can be used to both quantize and requantize audio signals. It is based on a deterministic optimization procedure which utilizes a model of the perceived quantization error power. It does not make use of any assumptions regarding the underlying audio signal, such as needing to be characterized by a certain probability density function or modeled by an underlying Markov chain.

The layout of the remainder of this study is as follows. Section 1 summarizes the state of the art in this area. The key contribution is then explained in Sections 2 and 3. These show how the quantization problem can be formulated as a moving-horizon optimization problem. The subsequent sections expand on the core idea. Specifically, Section 4 establishes the relationship to the noise-shaping quantizer, Section 5 shows how dithering can be incorporated, Section 6 studies some implementation issues, and Section 7 gives examples.

1 PROBLEM STATEMENT AND EXISTING SOLUTIONS

Suppose a sequence of sampled scalar audio data \{a(t)\}, \(t \in \mathbb{N}\), which are either analog or finely quantized. (For ease of notation, \(t\) has been normalized by the sampling period so that \(t \in \mathbb{N}\).) This signal is required to be quantized into a signal \{\(u(t)\)\}, \(t \in \mathbb{N}\), where each value \(u(t)\) is restricted to belong to the finite set \(U = \{s_1, \ldots, s_n\}\).
1.1 Direct Quantization

A direct approach to this problem corresponds to setting, at time \( t = k \),

\[
u(k) = q_u[a(k)]
\]

(1)

where \( q_u[\cdot] \) maps \( a(k) \) to the nearest available value belonging to the set \( U \). More formally, \( q_u[\cdot] \) can be defined (and generalized) as follows.

**Definition 1 (Nearest Neighbor Vector Quantizer):** Given a countable (not necessarily finite) set of nonequal vectors \( \mathcal{B} = \{b_1, b_2, \ldots \} \subset \mathbb{R}^n \), the nearest neighbor quantizer is defined as a mapping \( q_{\mathcal{B}}: \mathbb{R}^n \rightarrow \mathcal{B} \), which assigns to each vector \( c \in \mathbb{R}^n \) the closest element of \( \mathcal{B} \) (as measured by the Euclidean norm), that is, \( q_{\mathcal{B}}(c) = b \in \mathcal{B} \) if and only if \( c \) satisfies

\[
\|c - b\| \leq \|c - b_i\|, \quad \forall b_i \in \mathcal{B}.
\]

(2)

It should be emphasized that the nonlinearity of \( q_{\mathcal{B}}[\cdot] \) in Eq. (1) is a simple scalar quantizer where \( n_{\mathcal{B}} = 1 \) and certainly could be defined more easily.

The more general structure of Definition 1 accounts for vector inputs and outputs and is needed in Sections 2 and 3 in order to develop the MHOQ. Its output is, in general, different from that of componentwise scalar quantization. A thorough treatment of vector quantizers and their features can be found in [4].

1.2 Noise-Shaping Quantizer

A more sophisticated method compared with the one described in Section 1.1, and which has been implemented successfully in compact disc mastering applications [5], and in \( \Delta \Sigma \) converters [6], [7], [3], utilizes the noise-shaping quantizer depicted in Fig. 1. [Here and in the remainder of this study \( p \) denotes the forward shift operator, \( p \cdot(k) = \cdot(k+1) \), where \( (\cdot(k)) \) is any sequence.] This configuration includes a filter \( F(p) \) that feeds back the quantization error, defined as

\[
e(t) \triangleq u(t) - w(t).
\]

(3)

Since \( e(t) \) is not available until the quantization has taken place, the filter \( F(p) \) includes an (at least) unitary time delay.

In multibit applications the quantization error can often be well approximated by white noise, which is independent of \( a(t) \), leading to a linear quantization model (see, for example, [8]). In Fig. 1 the output of the system satisfies

\[
u(k) = a(k) + \left[ 1 - F(p) \right] e(k)
\]

so that one may treat \( 1 - F(p) \) as a filter that shapes the frequency content of the quantization noise present in \( a(t) \).

The purpose of quantizers such as that shown in Fig. 1 is to push the noise spectrum into less audible regions in order to improve the resulting sound quality as perceived by the human ear. Quantization is especially audible at low signal levels. Hence, in principle, it is useful to tune \( 1 - F(p) \) such that its frequency response approximates the threshold of detection curves like, for example, those included in [9]. Indeed, the whole design procedure can, in principle, be cast as a curve-fitting problem [1], [10], [3].

Note that the input to \( q_{\mathcal{B}}[\cdot] \) satisfies

\[
w(k) = a(k) - F(p) \left[ u(k) - w(k) \right]
\]

so that, solving for \( w(k) \), we see that

\[
w(k) = \frac{1}{1 - F(p)} a(k) - \frac{F(p)}{1 - F(p)} u(k).
\]

Hence the system of Fig. 1 can be redrawn as the two-degrees-of-freedom loop of Fig. 2. Since \( 1 - F(p) \) is commonly designed to be minimum phase [1], both transfer functions in this figure are stable.

It is worth nothing that the noise-shaping quantizer is a time-invariant structure and hence does not take into account aspects of audio perception which vary with time, such as, for example, frequency masking [11].

2 AN OPTIMALITY-BASED APPROACH

It is known that the noise-shaping quantizer of Fig. 1 provides a significant improvement with respect to the direct implementation of Eq. (1). However, it should be noted that there exist no reasons to believe that this architecture is the final word and that it cannot be further improved.

In this section we introduce a framework which allows one to develop an optimal scheme for quantization and, in addition, provides further insight into the noise-shaping quantizer described in Section 1.

2.1 Direct Incorporation of a Perception Filter

The final goal of the quantization process consists of reducing, as much as possible, the distortion introduced as perceived by a human listener. Hence a well-designed quantizing system should take into account the human auditory system, that is, it should incorporate psycho-
acoustical aspects. General aspects of psychoacoustics are outlined, for example, in [12] and [13, ch. 10]. Our specific interest will be in the use of time-invariant linear characteristics, although extensions to more general cases are undoubtedly feasible along similar lines.

In a simple psychoacoustic model the ear's sensitivity is characterized by means of a perception filter. In this case the quantization problem can be interpreted as follows (see Fig. 3).

Given a sequence \( \{a(t)\} \) and a perception filter, design an abstract quantizer which optimizes a function of \( \{e(t)\} \) by choosing a sequence \( \{a(t)\} \). Each of the \( u(t) \) values to be assigned is restricted to belong to the finite set \( U \).

The perception filter models the ear's sensitivity to low-level noise power, and can be obtained from existing equal-loudness contours (see, for example, [13]).

It will turn out (see Section 4) that the perception filter can be interpreted in terms of the standard noise-shaping filter. However, the key point that we will make is that the quantization problem can be cast in the more general setting of multihorizon optimization using the perception filter. This framework offers the potential for improved performance and reduces to the standard noise-shaping quantizer in a special case.

### 2.2 Choice of Design Criterion

We propose to measure performance by means of a quadratic cost over a finite horizon \( N \). At time \( t = k \) the cost function is given by

\[
V_N = \sum_{k=1}^{k+N-1} e^2(t) .
\]

(5)

This design criterion examines the perceived errors over a future horizon and therefore accounts for current and future errors. It is worth mentioning that a related decision criterion was proposed in [14] in a slightly different context.

If the perception filter is modeled as a stable linear time-invariant filter,

\[
H(p) = 1 + \sum_{i=1}^{\infty} h_i p^{-i}
\]

(6)

then the overall perceived error is given by

\[
e(t) = H(p) \left[ a(t) - u(t) \right].
\]

(7)

The cost function, Eq. (5), depends on \( N \) future values of \( a(t) \), which can be grouped to form the vector

\[
u(k) = \begin{bmatrix} a(k) & a(k+1) & \ldots & a(k+N-1) \end{bmatrix}^T.
\]

(8)

Then Eq. (5) can be rewritten as

\[
V_N(u(k)) = \sum_{k=1}^{k+N-1} \left[ H(p) \left[ a(t) - u(t) \right] \right]^2 .
\]

(9)

Minimization of the cost function, Eq. (9), yields the optimizing sequence

\[
u^*(k) = \arg \min_{u(k) \in \mathbb{U}^N} V_N(u(k))
\]

(10)

where the set \( U^N \subset \mathbb{R}^N \) is defined as \( U^N = U \times \ldots \times U \) and contains \( n^N \) elements.

Note that in this formulation we have implicitly assumed that future values of \( a(t) \) are known. This previewing of the audio signal introduces a delay of, at least, \( N - 1 \) sampling periods. This requires minimal storage and thus should not restrict the applicability of this setting significantly.

### 2.3 Closed-Form Solution

It is worth noting that the quantization problem, as stated here, can be regarded as a special control problem where the input to the system \( H(p) \) has to be chosen from a finite alphabet. In particular, the optimization problem, Eq. (10), corresponds to choosing \( u(k) \in U^N \) so that the signal \( H(p)u(t) \) tracks the reference \( r(t) = H(p)a(t) \) with performance measured in a mean-square sense. The quantization problem can thus be converted into a finite-alphabet optimal control problem.

A closed-form expression for \( u^*(k) \) in Eq. (10) can be obtained, as we will show in Theorem 1. In order to give a clear exposition, it is useful to describe the filter \( H(p) \) of Eq. (6) in state-space form as

\[
H(p) = 1 + C (p I - A)^{-1} B .
\]

(11)

The matrices \( A, B, \) and \( C \) are related to the impulse response description, Eq. (6), by

\[
h_i = C A^{i-1} B . \quad i = 1, 2, \ldots
\]

(12)

From Eq. (7) it follows that \( e(t) \) can alternatively be described as the output of the following state-space system:

\[
x(t+1) = A x(t) + B \left[ a(t) - u(t) \right] \in \mathbb{R}^n
\]

\[
e(t) = C x(t) + \left[ a(t) - u(t) \right]
\]

(13)

where \( x \in \mathbb{R}^n \) is the state vector.

The state-space approach is a powerful methodology, which allows for both time-invariant and also time-variant cases to be treated simultaneously. It is very commonly used in some areas, such as control system design, and is known to reduce to other methods, such as frequency-response approaches, as special cases. More details can be found in the textbook [15], for example.

The state-space structure of Eq. (11) includes both infinite impulse response and finite impulse response filters.
In the latter case, there exists a finite value \( m \in \mathbb{N} \) such that \( h_j = 0 \), for all \( j > m \), and \( A, B, \) and \( C \) can be chosen as

\[
A = \begin{bmatrix}
0 & 0 & \cdots & 0 \\
1 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & \cdots & \cdots & 0
\end{bmatrix}, \quad B = \begin{bmatrix}
0 \\
0 \\
\vdots \\
0
\end{bmatrix}, \quad C = \begin{bmatrix}
h_1 & h_2 & \cdots & h_m
\end{bmatrix}.
\]

(14)

With this as background, we have the following result.

**Theorem 1:** Suppose \( \mathcal{U}^N = \{ v_1, v_2, \ldots, v_r \} \), where \( r = n_0^N \), and \( H(p) \) has realization (11), then the optimizing sequence \( u^o(k) \) in Eq. (10) combined with Eq. (9) is given by

\[
u^o(k) = \Psi^{-1} q_{\mathcal{U}^N} \left[ \Psi a(k) + \Gamma x(k) \right]
\]

(15)

where

\[
a(k) = \begin{bmatrix}
a(k) \\
a(k+1) \\
\vdots \\
a(k+N-1)
\end{bmatrix}, \quad \Gamma = \begin{bmatrix}
C \\
CA \\
\vdots \\
CA^{N-1}
\end{bmatrix}, \quad \Psi = \begin{bmatrix}
h_0 & 0 & \cdots & 0 \\
h_1 & h_0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
h_{N-1} & \cdots & h_1 & h_0
\end{bmatrix}
\]

The diagonal entries of \( \Psi \) satisfy \( h_0 = 1 \) whereas the others obey Eq. (12). The nonlinearity \( q_{\mathcal{U}^N}[\cdot] \) is the nearest neighbor quantizer described in Definition 1. The image of this mapping is the set

\[
\tilde{\mathcal{U}}^N = \{ \tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_r \} \subseteq \mathbb{R}^N.
\]

(16)

**Proof:** For completeness, this result is established in Appendix A.

Starting from an initial condition \( x(0) = x_0 \), the state \( x(k) \) in Eq. (15) can be obtained by iterating the recursion (13). It should be emphasized that \( q_{\tilde{U}^N}[\cdot] \) in Eq. (15) is an \( N \)-dimensional vector quantizer and that the columns of the matrix \( \Psi \) correspond to truncated impulse responses of \( H(p) \).

**3 THE MOVING-HORIZON OPTIMAL QUANTIZER**

The solution \( u^o(k) \) provided by Theorem 1 allows for a batch implementation. If \( N \) is chosen as large as the complete audio signal, then this procedure will provide the lowest perceived mean-square distortion. However, this would not be very useful in practice due to prohibitive computational complexity. A more appropriate scheme is therefore described next.

We fix a relatively small value for the horizon length \( N \) and implement, at time \( t = k \), only the first element of the optimizer \( u^o(k) \), that is, we set

\[
u(k) = [1 \ 0 \ \cdots \ 0] \Psi^{-1} q_{\tilde{U}^N} \left[ \Psi a(k) + \Gamma x(k) \right].
\]

(17)

We then move the previous block of \( N \) samples forward and repeat the procedure. Thus at the next time instant, \( t = k + 1 \), we use the value

\[
u(k + 1) = [1 \ 0 \ \cdots \ 0] \Psi^{-1} q_{\tilde{U}^N} \left[ \Psi a(k + 1) + \Gamma x(k + 1) \right].
\]

Note that the same amount of future information, that is, the same number of future audio values, is used. The preview horizon is of fixed length \( N \) and moves (slides) forward as \( t \) increases, as shown in Fig. 4. We call this scheme the moving-horizon optimal quantizer (MHOQ).

A closed-loop implementation of the MHOQ is provided by the following corollary. Note that this result does not depend on the particular realization chosen in Eq. (11).

**Corollary 1:** The MHOQ (17) can be implemented as shown in Figs. 5 and 6, where \( \mathcal{N}(p) = I + \Psi^{-1} G(p) [1 \ 0 \ \cdots \ 0] \)

(18)

is a square matrix transfer function having \( N \) inputs and outputs. The matrix transfer function

\[
G(p) = \Gamma (pI - A)^{-1} B
\]

(19)

has 1 input and \( N \) outputs. It can be described based on the impulse response description, Eq. (6), according to

\[
G(p) = \left[ G_1(p) \ G_2(p) \ \cdots \ G_N(p) \right]^T
\]

\[
G_j(p) = p^{j-1} \sum_{n=j}^{\infty} h_n p^{-n}
\]

(20)

**Proof:** See Appendix B.

This corollary summarizes the main contribution of this engineering report. In particular, we have proposed a moving-horizon approach to quantization, which incorporates a perception filter and signal previewing. The following sections contain further embellishments of this idea.

**4 RELATIONSHIP TO NOISE-SHAPING QUANTIZER**

The MHOQ is psychoacoustically optimal in the sense that it minimizes a quadratic norm of perceived errors.
Although it is formulated in the time domain and does not directly use the linear model of quantization described in Section 1.2 or frequency weighting, a relationship to the noise shaper can be established as follows.

From Eq. (18) it follows that \( H(p) \) can alternatively be written as

\[
H(p) = \begin{bmatrix}
1 + H'_1(p) & 0 & \cdots & 0 \\
H'_2(p) & 1 & 0 & \cdots & 0 \\
H'_3(p) & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
H'_N(p) & 0 & \cdots & 1 \\
\end{bmatrix}
\]

\[
H'_i(p) = \left[ \Psi^{-1} \right]_{i*} \Gamma (pI - A)^{-1} B
\]

where \( [\Psi^{-1}]_{i*} \) denotes the \( i \)th row of \( \Psi^{-1} \).

Consider a unitary prediction horizon, namely, \( N = 1 \). With \( N = 1 \), \( H(p) \) reduces to its first element, \( 1 + H'_1(p) \), which according to the definitions of Theorem 1 and Eq. (11), satisfies

\[
1 + H'_1(p) = 1 + C (pI - A)^{-1} B = H(p)
\]

that is, it is the perception filter. Moreover, \( \Psi = 1 \in \mathbb{R} \) and \( \hat{U}^N = U \), so that the vector quantizer \( q_N[\cdot] \) reduces to the scalar quantizer \( q_1[\cdot] \). Hence the MHOQ of Fig. 5 reduces to the scheme depicted in Fig. 7.

Direct comparison with the two-degrees-of-freedom loop depicted in Fig. 2 shows that both schemes are equivalent if

\[
H(p) = \frac{1}{1 - F(p)}.
\]

\[\text{Fig. 4. Moving-horizon principle, } N = 5.\]

\[\text{Fig. 5. Implementation of MHOQ using } H(p).\]

\[\text{Fig. 6. Implementation of MHOQ using } G(p).\]
Notice that with this choice it follows that \( H(p) - 1 = F(p)[1 - F(p)] \) and that, due to the fact that \( F(p) \) includes a time delay and \( 1 - F(p) \) is minimum phase, the transfer function \( H(p) \) is stable and has unitary feedthrough in concordance with its description, Eq. (11).

Indeed, we see that the scheme of Fig. 2 is just a redrawing of Fig. 1. As a consequence, the noise-shaping quantizer corresponds to the MHOQ when the horizon is chosen as the special value of \( N = 1 \). It is then a symbol-by-symbol scheme, where each symbol \( u(k) \) is assigned based solely upon the present filtered error \( e(k) \).

It follows that the noise-shaping quantizer incorporates the perception filter implicitly and is, in the aforementioned sense, optimal. Note that instead of tuning by means of

\[
F(p) = \frac{H(p) - 1}{H(p)}
\]

the perception filter \( H(p) \) can be used directly. We thus see that the MHOQ scheme embeds the standard noise-shaping quantizer in a more general arrangement.

The main advantage of using a horizon larger than 1 in the MHOQ scheme derives from the fact that with \( N > 1 \), not only the present but also future values of the signal \( a(t) \) are taken into account in the determination of \( u(k) \). As illustrated by means of the example included in Section 7, this typically leads to improved performance.

5 INCLUSION OF DITHER

By extending the results documented in [16], [17], it can be shown that, depending on the signal \( a(t) \), the resulting sequence \( u(t) \) will be periodic and give rise to audible tones. As a consequence, although the cost, Eq. (5), is certainly intuitively appealing, its implementation may lead to unsatisfactory listening experiences.

This phenomenon is well known when dealing with structures such as the noise shaper of Fig. 1 (see, for example, [18]). Arguing from a probabilistic point of view, it is easy to see that the quantization noise \( e(t) \) defined in Eq. (3) and the audio signal \( a(t) \) are correlated. (They are actually related deterministically.) Hence the quantization noise is modulated by the music.

The inclusion of dither in quantization schemes has been recognized as a means to reduce signal-dependent distortion and noise modulation (see, for example, [19]–[21]. The main idea is to randomize the decisions made by the quantizer in order to break the correlation mentioned. This is usually accomplished by including a (pseudo)random signal at the input to the quantizer, which leads to the architecture of Fig. 8. This scheme may render any moment of the quantization noise independent of \( a(t) \). Although the overall noise level is higher, this signal independence improves the (subjective) sound quality and makes the linear model derived from Eq. (4) more accurate.

As shown in [22], given a quantizer \( q_{\nu}[-] \) whose output \( U \) consists of consecutive integers, the dither signal \( d(t) \), which renders the mean and variance of the quantization noise independent of \( a(t) \) and yields the lowest increase in variance, is

\[
d(t) = n_1(t) + n_2(t)
\]

where at time \( t = k \), \( n_1(k) \) and \( n_2(k) \) are independent and uniformly distributed over the interval \([-1/2, 1/2]\). Since the dependence on \( a(t) \) of higher moments of the quantization noise is believed to be inaudible, this choice has been adopted widely.

Dithering can be included in the framework of Section 2 by replacing the cost function, Eq. (9), with

\[
V_N'(u(k)) = \sum_{i=k}^{k+N-1} \left\{ H(p)[a(t) - u(t)] + d(t) \right\}^2
\]

where \( u(k) \) is defined as in Eq. (8) and \( d(t) \) is the exogenous dither signal, which is fixed and can be precomputed. Note

Fig. 7. MHOQ with horizon \( N = 1 \).

Fig. 8. Noise-shaping quantizer with dither.
that both \( a(t) \) and \( d(t) \) are previewed in this cost function.

Optimizing Eq. (24) will make the signal \( H(p)u(t) \) track the reference \( H(p)a(t) + d(t) \), or, equivalently, will make \( u(t) \) be close to the signal \( a(t) + [H(p)^{-1}] d(t) \). At first glance this sounds less appealing than the original design criterion, Eq. (5). However, as will be shown, the resulting dithered MHOQ generalizes the dithered noise shaper of Fig. 8 in the same way as the undithered MHOQ generalizes the undithered noise shaper depicted in Fig. 1. While Eq. (9) may lead to noise modulation, optimization of Eq. (24) might avoid it. The following results resemble those of Sections 2 and 3.

**Theorem 2:** Suppose \( \mathcal{U}^N = \{v_1, v_2, \ldots, v_r\} \), where \( r = n_o^N \) and \( H(p) \) has the realization (11), then the optimizing sequence

\[
u^*(k) = \arg \min_{u(k) \in \mathcal{U}^N} V_N^x(u(k))
\]

with \( V_N^x(u(k)) \) defined in Eq. (24), can be calculated as

\[
u^*(k) = \Psi^{-1} q_{\mathcal{U}^N} \left[ \Psi u(k) + \Gamma x(k) + d(k) \right]
\]

where

\[
d(k) = \begin{bmatrix} d(k) & d(k+1) & \cdots & d(k+N-1) \end{bmatrix}^T
\]

and \( a(k), \Gamma, \Psi, \) and \( q_{\mathcal{U}^N} \) are as in Theorem 1.

**Proof:** See Appendix C.

As in the proof of Corollary 1, it is also straightforward to show that the dithered MHOQ, which satisfies

\[
u(k) = \begin{bmatrix} 1 & 0 & \ldots & 0 \end{bmatrix} \Psi^{-1} q_{\mathcal{U}^N} \left[ \Psi u(k) + \Gamma x(k) + d(k) \right]
\]

can be implemented as shown in Figs. 9 and 10, where \( \mathcal{Y}(p) \) and \( G(p) \) are defined in Eqs. (18) and (19), respectively.

Not surprisingly, with \( N = 1 \) the dithered MHOQ provides the one-step optimal solution

\[
u(k) = q_{\mathcal{U}^N} \left[ H(p) a(k) - \left( H(p) - 1 \right) u(k) + d(k) \right]
\]

which describes the dithered noise shaper depicted in Fig. 8, provided Eq. (22) is satisfied.

The authors have found that the inclusion of dither appears to improve sound quality qualitatively. As illustrated in the example of Section 7, the sound quality is further enhanced as the optimization horizon increases. This is perhaps not surprising, given the well-known benefits of dithering in the standard noise-shaping quantizer, which is a special case of the methodology outlined here.

**6 IMPLEMENTATION ISSUES**

In some practical applications, such as the mastering of an audio CD where \( \mathcal{U} \) contains \( n_{\mathcal{U}} = 2^{16} \) elements, the computational complexity of implementing the MHOQ directly may be considerable. In this section two approaches that alleviate this burden are described.

**6.1 Partition of the State Space**

The optimizing sequence \( u^*(k) \) in Eq. (25) is obtained by means of the nearest neighbor quantizer of Definition 1. Direct manipulation yields that condition (2) is equivalent to

\[
q_{\mathcal{U}_v}(a) = b \iff 2 \left( b_v - b \right)^T a \leq b_v^T b - b^T b, \quad \forall b_v \in \mathcal{B}.
\]

Hence \( u^*(k) = v_i \in \mathcal{U}^N \) if and only if

\[
2 \left( v_i - v_i^+ \right)^T \Psi^T \left[ \Psi u(k) + \Gamma x(k) + d(k) \right] \leq v_i^T \Psi^T \Psi v_i - v_i^T \Psi^T v_i, \quad \forall v_i \in \mathcal{U}^N.
\]

Suppose now that \( N \leq n \) and that \( T \Gamma T^T \in \mathbb{R}^{N\times N} \) is invertible. (Note that with \( N = n, \Gamma \) is the observability matrix of the...
pair \((A, C)\) [15]. In this case it holds that

\[
\Psi a(k) + \Gamma x(k) + d(k) = \Gamma \left[ x(k) + \Gamma^T (\Gamma \Gamma^T)^{-1} \left( \Psi a(k) + d(k) \right) \right]
\]

so that expression (27) can be rewritten as

\[
2(\nu_i - \nu_j)^T \Psi^T \Gamma \tilde{x}(k) \leq \nu_i^T \Psi^T \Psi \nu_i - \nu_j^T \Psi^T \Psi \nu_j,
\]

\[\forall \nu_i \in \mathbb{U}^N \] (28)

where

\[
\tilde{x}(k) \equiv x(k) + \Gamma^T (\Gamma \Gamma^T)^{-1} \left( \Psi a(k) + d(k) \right) \in \mathbb{R}^n
\]
defines the shifted state.

As a consequence, the optimizing sequence is characterized by

\[
u^*(k) = \nu_i \in \mathbb{U}^N \Rightarrow \tilde{x}(k) \in \mathcal{R}_i
\]

where each region \(\mathcal{R}_i\) contains all elements \(\tilde{x} \in \mathbb{R}^n\) that satisfy Eq. (28). This condition specifies a set of linear inequalities, some of which may be redundant. In these cases the corresponding regions do not share a common border, that is, they are not adjacent.

All regions \(\mathcal{R}_i\) are polytopes. They can be written in compact form as

\[
\mathcal{R}_i = \{ \tilde{x} \in \mathbb{R}^n : D_i \tilde{x} \leq h_i \}
\] (29)

where the rows of \(D_i\) are equal to all terms \(2(\nu_i - \nu_j)^T \Psi^T \Gamma\) as required, whereas the vector \(h_i\) contains the scalars \(\nu_i^T \Psi^T \Psi \nu_i - \nu_j^T \Psi^T \Psi \nu_j\).

When a moving horizon is used, only the first element of \(u^*(k)\) is used. As a consequence, only \(n_U\) instead of \(n_U^N\) regions are needed to characterize the MHOQ. Each of these regions is given by the union of all regions \(\mathcal{R}_i\) in Eq. (29) corresponding to vectors \(\nu_i\) having the same first element.

Eq. (29) specifies a polyhedral partition of the shifted state space, which can be computed off-line. The on-line optimization reduces to the evaluation of a finite number of inequalities, which can be efficiently implemented using some off-line calculations.

Besides reducing the on-line computational burden, this characterization also enhances the verifiability of the complete range of dynamic behavior of the closed loop, including issues such as stability, without relying on extensive simulations. For more details see the finite-set constrained control case [23].

Unfortunately if \(N > n\), the matrix \(\Gamma \Gamma^T\) is not invertible, and this procedure is not possible. The optimizing sequence \(u^*(k)\) depend on \(a(k)\) and \(d(k)\), each containing \(N\) elements, and the partition induced by the quantizer has to be considered over the complete space \(\mathbb{R}^N\).

6.2 Suboptimal Schemes

If instead of searching through the whole set \(\mathbb{U}^N\), as prescribed in Eq. (15) or (25), the search is restricted to a smaller (possibly time-varying) subset \(\hat{\mathcal{J}}(k)\), the computational load needed can be reduced. In the dithered case, the sequence used is given by

\[
u^*(k) = \Psi^{-1}
\]

\[
q_{\hat{J}(k)} \left( \Psi a(k) + \Gamma x(k) + d(k) \right) ,
\]

\[
\hat{J}(k) \subset \hat{\mathcal{U}}^N
\] (30)

and gives rise to the suboptimal moving horizon quantizer

\[
u(k) = [1 \ 0 \ \cdots \ 0] u^*(k).
\]

[If no dither is included, set \(d(k) = 0 \in \mathbb{R}^N\) in Eq. (30).]

Although optimality is lost, an educated choice of \(\hat{J}(k)\) may give rise to performance which, in practical terms, is indistinguishable from that of the optimal MHOQ.

This can be accomplished if

\[
\tilde{J}(k) = \{ \tilde{\nu}_1, \tilde{\nu}_2, \ldots \tilde{\nu}_r \}
\] (31)

where \(\tilde{\nu}_i = \Psi \nu_i\), and the first element of every vector \(\nu_i\) is restricted to be either the one-step optimal value given in Eq. (26) or one of its (at most) two neighbors. Note that, by construction, this algorithm will always outperform the noise shaper, but is cheaper to implement than the \(N\)-step optimal MHOQ, since \(\tilde{J}(k)\) contains at most only \(3n_U^{-1}\) elements.

Another possibility for choosing \(\tilde{J}(k)\) in Eq. (31) derives from using, at time \(t = k\), the previous suboptimal sequence \(u^*(k-1)\) as a starting point. More precisely, \(\tilde{J}(k)\) can be obtained by setting the first \(N-1\) elements of \(\nu_i\) in Eq. (31) equal to the last \(N-1\) elements of \(u^*(k-1)\) or one of its (at most) \(2^{N-1}\) neighbors. The last element of \(\nu_i\) should be left free to be any element of \(\mathbb{U}\) in order to ensure a sufficiently large bandwidth of the resulting sequence \(\{u(t)\}\). In this case the set \(\tilde{J}(k)\) contains at most \(n_U^{(N-1)}\) elements. The scheme can be initialized, for example, with the optimizer \(u^*(0)\).

7 EXAMPLE

In order to illustrate the merits of the MHOQ, consider the following example, taken from [10]. Given a sampling frequency \(f_s = 44.1\) kHz, the filter \(F(p)\) is chosen as

\[
F(p) = p^{-1} \frac{2.245 - 0.664 p^{-1}}{1 + 0.91 p^{-1}}.
\]

This is equivalent to having

\[
H(p) = 1 + p^{-1} \frac{2.245 - 0.664 p^{-1}}{1 - 1.335 p^{-1} + 0.644 p^{-2}}.
\]

The frequency responses of these filters can be obtained by replacing \(p = e^{-2\pi f/2f_s}\), where the frequency \(f\) is restricted to belong to \([0, f_s/2]\). The two frequency responses are depicted in Fig. 11.

Simulations were carried out with different sets \(U\) and \(a(k), k = 1, 2, \ldots, T_i = 1000\), consisting of observations
of an independent random variable, uniformly distributed over an interval \( \mathcal{X} \) as detailed in Table 1.

Fig. 12 shows the performance achieved for different horizons \( N \) as quantified by means of the sample variance of the perceived error,

\[
\hat{V}^2 = \frac{1}{T_1} \sum_{i=1}^{T_1} \left( H(\rho) \left[ a(i) - u(i) \right] \right)^2.
\]

(A signal-to-noise measure that can be used to compare the performance with different word lengths can be obtained by normalizing with the signal amplitude.)

Fig. 13 shows the results from a simulation with the same specifications as before, but including triangular dither, as specified in Eq. (23). In both figures, \( N = 0 \) denotes direct quantization as in Eq. (1) and \( N = 1 \) corresponds to the standard noise-shaping quantizer.

It can be seen that the perceived error variance decreases monotonically with \( N \). The MHOQ outperforms the

\textbf{8 CONCLUSIONS}

The MHOQ was developed by using psychoacoustic principles expressed as a perception filter, which models the ear's sensitivity to low-level noise. The scheme is based

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|}
\hline
\textbf{U} & \textbf{F} & \textbf{f} \\
\hline
3 bit & [\(-4, -3, \ldots, 3\)] & [\(-4.5, 3.5\)] \\
4 bit & [\(-8, -7, \ldots, 7\)] & [\(-8.5, 7.5\)] \\
5 bit & [\(-16, -15, \ldots, 15\)] & [\(-16.5, 15.5\)] \\
\hline
\end{tabular}
\caption{Simulation study parameters.}
\end{table}

![Fig. 11. Frequency responses of \( H \) and \( F \).](image)

![Fig. 12. Effect of \( N \) on undithered MHOQ performance.](image)
on the minimization of predictions of perceived errors over a finite horizon. Choosing the horizon length equal to 1 reduces the scheme to the widely used noise shaper.

Advantages of using larger horizons arise from the fact that the future is not sacrificed by present decisions due to the ability to look ahead and that more information is taken into account in the bit allocation process. As a consequence, the MHOQ typically gives better performance than the noise shaper.

Further work could be carried out relating to the inclusion of more complex psychoacoustical phenomena, such as frequency masking [11].

9 REFERENCES


Fig. 13. Effect of N on dithered MHOQ performance.


APPENDIX A
PROOF OF THEOREM 1

Form the vector
\[ e(k) = \begin{bmatrix} e(k) & e(k + 1) & \cdots & e(k + N - 1) \end{bmatrix}^T \]
and iterate Eq. (13) in order to obtain
\[ e(k) = \Psi a(k) - \Psi u(k) + \Gamma x(k). \quad (32) \]

This allows one to write the cost function \( V_N \) in vector form as
\[ V_N(u) = \tilde{V}_N(x, a) + u^T \Psi^T \Psi u - 2u^T \Psi^T (\Psi a + \Gamma x) \quad (33) \]
where \( \tilde{V}_N(x, a) \) does not depend on \( u \). (Note that, for ease of notation, the dependence on \( k \) of all these vectors has not been explicitly included.)

We introduce the change of variables, \( \mu = \Psi u \). This transforms \( U^N \) into \( \tilde{U}^N \) defined in Eq. (16). Expression (33) then allows one to rewrite Eq. (10) as
\[ u^* = \Psi^{-1} \arg \min_{\mu \in \tilde{U}^N} J_N(\mu), \]
\[ J_N(\mu) = \mu^T \mu - 2\mu^T (\Psi a + \Gamma x). \]
The level sets of \( J_N \) are spheres in \( \mathbb{R}^N \), centered at \( \Psi a + \Gamma x \). Hence the constrained optimizer is given by
\[ \arg \min_{\mu \in \tilde{U}^N} J_N(\mu) = q_{\tilde{U}^N}(\Psi a + \Gamma x). \]
This establishes the result given in Eq. (15). (Note that this result is also related to the finite-alphabet optimal control problem addressed by the authors in [24].)

APPENDIX B
PROOF OF COROLLARY 1

From Eq. (13) one obtains that \( x(k) = (\rho l - A)^{-1} B(a(k) - u(k)) \) so that
\[ \Gamma x(k) = G(\rho)[a(k) - u(k)] = G(\rho)[1 \ 0 \ \cdots \ 0](a(k) - u^*(k)) \quad (34) \]
and
\[ \Psi a(k) + \Gamma x(k) = \Psi \left[ I + \Psi^{-1} G(\rho)[1 \ 0 \ \cdots \ 0] \right] a(k) - \Psi^{-1} G(\rho)[1 \ 0 \ \cdots \ 0] u^*(k) \right]. \]

Substituting this expression and Eq. (18) into Eq. (17) leads to
\[ u(k) = [1 \ 0 \ \cdots \ 0] \Psi^{-1} q_{\tilde{U}^N} \left[ \mathcal{H}(\rho) a(k) - [\mathcal{H}(\rho) - I] u^*(k) \right]. \quad (35) \]
This equation describes the scheme of Fig. 5.

In order to derive the scheme of Fig. 6, we first note that the first components of the vectors \( \tilde{v}_l \) defined in Eq. (16) are equal to the first components of the corresponding vectors \( v_l \), since the first row of \( \Psi \) (and of \( \Psi^{-1} \)) is equal to \( [1 \ 0 \ \cdots \ 0] \). As a consequence, \( [1 \ 0 \ \cdots \ 0] \Psi^{-1} = [1 \ 0 \ \cdots \ 0] \) and the MHOQ in Eq. (17) can be simplified to
\[ u(k) = [1 \ 0 \ \cdots \ 0] q_{\tilde{U}^N} \left[ \Psi a(k) + \Gamma x(k) \right]. \quad (36) \]
Fig. 6 implements Eq. (35).

Expression (20) can be proven as follows. According to the definition of \( \Gamma \) in Theorem 1, the entries of \( G(\rho) \) are
\[ G_j(\rho) = CA_j^{-1}(\rho I - A)^{-1} B \] and the Laurent series [25] around the origin of \( G_j(\rho) \) is
\[ G_j(\rho) = \sum_{l=1}^{\infty} \gamma_l \rho^{-l}, \quad \gamma_l = CA^{j+l-2} B \]
yielding
\[ G_j(\rho) = \rho^{j-1} \sum_{l=1}^{\infty} CA^{j+l-1} B \rho^{-(j+l-1)} \]
\[ = \rho^{j-1} \sum_{n=j}^{\infty} \sum_{n-j} A_j^{n-1} B \rho^{-n} \]
\[ = \rho^{j-1} \sum_{n-j} A_j^{n-1} B \rho^{-n} \]
\[ = \rho^{j-1} \left[ H(z) - \left( 1 + \sum_{n=1}^{j-1} h_n \rho^{-n} \right) \right] \]
after replacing Eq. (12).

APPENDIX C
PROOF OF THEOREM 2

Define $e(k)$ as in the proof of Theorem 1 and note that

$$V_N^d = [e(k) + d(k)]^T [e(k) + d(k)].$$

Eq. (32) allows one to write

$$V_N^d = V^d_N(x, a, d) + u^T \Psi^T \Psi u - 2u^T \Psi^T (\Psi a + \Gamma x + d)$$

where the term $V^d_N(x, a, d)$ does not depend on $u$. The rest of the proof parallels that of Theorem 1 and is therefore omitted in the interest of brevity.

THE AUTHORS

G. C. Goodwin

D. E. Quevedo

D. McGrath

Graham C. Goodwin received B.Sc. in physics, B.E. in electrical engineering, and Ph.D. degrees from the University of New South Wales, Australia.

From 1970 until 1974 he was a lecturer in the Department of Computing and Control, Imperial College, London. Since 1974 he has been with the Department of Electrical and Computer Engineering, The University of Newcastle, Australia. He is currently a professor of electrical engineering and an associate director of the Centre for Integrated Dynamics and Control at the University.


Dr. Goodwin is the recipient of several international prizes including the USA Control Systems Society 1999 Hendrik Bode Lecture Prize, Best Paper award by the IEEE Trans. Automatic Control, Best Paper award by the Asian Journal of Control, and Best Engineering Text Book award from the International Federation of Automatic Control. He is the recipient of the ARC Federation Fellowship. He is an honorary fellow of the Institute of Engineers, Australia; a fellow of the IEEE; the Australian Academy of Science, the Royal Society, London, and the Australian Academy of Technology, Science and Engineering; and a member of the International Statistical Institute.

D. E. Quevedo received Ingeniero Civil Electrónico and Magíster en Ingeniería Electrónica degrees from the Universidad Técnica Federico Santa María, Valparaíso, Chile, in 2000. During his time at the university he was supported by a full scholarship from the alumni association and upon graduating received several university-wide prizes.

Since 2001 he has been with the School of Electrical Engineering and Computer Science, The University of Newcastle, Australia, working toward a Ph.D. degree under the supervision of Professor Graham Goodwin.

Mr. Quevedo has lectured at both universities in several courses related to automatic control. He also has working experience at the VEW Energie AG, Dortmund, Germany, and at the Cerro Tololo Inter-American Observatory, Chile. He is a member of the IEEE Control Systems Society and was recently a finalist in the Student Paper Competition for the 41st IEEE Conference on Decision and Control. His main research interests cover several areas of automatic control, signal and audio processing, and communications.

D. McGrath received Bachelor of Science and Bachelor of Engineering degrees from the University of Sydney, Australia, in 1984 and 1986, respectively.

From 1986 to 1990 he worked as an integrated circuit designer at Austek Microsystems in Adelaide, Australia, where he was primarily involved in the design of digital signal processing circuits. In 1991 he worked with the CSIRO, an Australian government research group, on the development of a FIR processor chip for audio. In 1991 he co-founded Lake Technology in Sydney, Australia. His work at Lake has included the development of specialized hardware for acoustics research based on highly efficient convolution algorithms.

Mr. McGrath’s principal interests include psychoacoustics, particularly 3-D sound perception, and more recently noise-shaping techniques and methods for the processing of sigma–delta modulated 1-bit audio signals.