The 33rd Annual Conference of the IEEE Industrial Electronics Society (IECON)
Nov. 5-8, 2007, Taipei, Taiwan

System Analysis of an Optimal Noise Shaped Quantizer for Audio-band Digital Amplifier

Jwu-Sheng Hu, Member, IEEE, and Keng-Yuan Chen
Department of Electrical and Control Engineering,
National Chiao Tung University, Hsinchu 300, Taiwan, ROC
jshu@cn.nctu.edu.tw bettery.ece94g@nctu.edu.tw

Abstract- **This paper reports a novel application of receding finite horizon constrained optimization techniques to design an audio-band full digital amplifier. The digital amplifier using power MOSFET requires a modulator to modulate the reference signal into binary sequences. The modulator can be viewed as an over-sampled quantizer with noise-shaping requirement. One of the main issues in designing the modulator is to maintain stability while maximizing the range of the reference input signal. The quantization scheme resulting from the optimization is derived in detail to arrive at the stability condition under zero initial conditions. It shows the stability condition becomes more stringent when the relative degree of the shaping filter is high. Simulations results are given to illustrate the effectiveness of the design methodology.**

I. INTRODUCTION

Class D amplifier, a more efficient way for audio power amplification than Class A/B amp, have drawn a lot of attention in recent years [7]. To drive the power stage to produce high fidelity sound, it is necessary to convert the audio PCM data into the driving signal (the binary sequence). The binary sequence is usually referred as the quantized 1-bit signal and has a higher sampling rate (called over-sampling) than the original source signal. The binary sequence can be generated using the digital PWM technique. To reduce the total harmonic distortion and enhance the signal/noise ratio, research efforts were reported by using sigma-delta modulation [13-14][18], DSP techniques [2] [5], feedback control [6], and interpolation methods [4] [11].

Recently, feedback quantization using the technique of receding horizon linear quadratic control with finite input constraint was reported [16-17]. It was also applied to design analog-to-digital conversion [15]. The technique offers a systematic way of designing the quantizer (or modulator) with various performance measures. However, to apply the method for generating the modulation sequence for a class-D amplifier, it is necessary to maximize the input range while keeping the system to be stable. This paper reports a detail analysis of the stability of the nonlinear feedback under zero initial condition. The analysis can be applied to the cases of horizon one and two in the optimization scheme. Further, since the computing resource for real-time application is usually limited, operations such as nearest neighborhood quantization are carefully designed to fit into the platform. Simulation results show that the method is quite effective in generating the control sequences.

II. BASIC CONCEPT

Fig. 1 shows a typical power stage (H-bridge) in a class-D amplifier. The control signal u controls the current direction of the load. Suppose that u is updated every T_1 second and the value of u also contains the net duration of the current. The simplest case is $u = 1$ or -1 which means either positive or negative current direction is allowed within T_1. If a higher clock rate is available, say T_1 / n_H where n_H is an integer, there could be $2n_H + 1$ cases of the duration of the current. For example, 5 cases for $n_H = 2$ are shown in Fig. 2. It is easy to see that the value of u can be represented by a 2.5-bit variable. The explanation allows us to apply the constrained optimization control [15] to design a multi-bit quantizer using over-sampling technique. This section presents a concise introduction of the work in [15].

A. Problem Statement

Suppose u is characterized as it belongs to a finite set of scalars U:

$$U = \{s_1, \cdots, s_{n_U}\}$$

where n_U denotes the cardinality of U. For example, in a typical full-bridge topology, U might be {1, -1} to represent the current in both directions. For a digital audio input signal r, the purpose is to obtain the signal u that can represent the signal r in the chosen bandwidth. Under over-sampling, we can only take into account the distortion in input band (here for audio signal is 0-22.05kHz) instead of overall frequency band. Therefore, the distortion can be weighted via a stable, linear, time-invariant filter W, which can be written as

$$W(z) = D + C(zI - A)^{-1}B \tag{1}$$

The modulator then quantizes the signal r in a way that the amplitude of the corresponding filtered quantization noise is minimized.

It is then straightforward to define the system filtered noise power V as cost function [15]:

$$V \overset{\Delta}{=} \frac{1}{2\pi}\int_0^{2\pi} \left| W(e^{jw})(R(e^{jw}) - U(e^{jw})) \right|^2 dw \tag{2}$$

where $W(e^{jw})$ denotes the frequency response of the filter W, while $R(e^{jw})$ and $U(e^{jw})$ are the discrete Fourier transforms of the signal r and u, respectively.

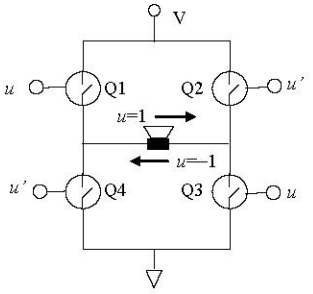

Fig. 1 The full bridge topology and associated coding of the control signal relative to the current direction

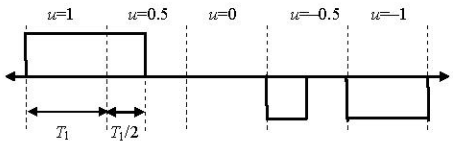

Fig.2 The switching waveform of a 2.5 bit control signal

The constraint output u is chosen to minimize the cost function V. If the value of V is minimized, the minimization will be more emphasized in the frequency band where W has larger magnitude, thus resulting in a U better approximating R in that band. In audio applications it makes sense to choose W as a low pass filter with cutoff frequency higher than 22.05kHz.

The expression in (2) can be translated into time domain by using Parseval's Theorem [15]:

$$V = \sum_{l=0}^{\infty} (e(l))^2 \tag{3}$$

where $e(l)$ are samples of filtered distortion as,

$$e \triangleq W(r-u) \tag{4}$$

The state-space equation of the filter W is:

$$\begin{aligned} x(m+1) &= Ax(m) + B(r(m) - u(m)) \\ e(m) &= Cx(m) + D(r(m) - u(m)) \end{aligned} \tag{5}$$

where $x \in R^n$ is the state vector of dimension n, i.e. the order of the filter W. Equation (3) is further simplified for practical use with finite decision number:

$$V_N \triangleq \sum_{m=k}^{k+N-1} (e(m))^2 \tag{6}$$

That means the cost function V_N only takes into account N number of constrained values u, which can be grouped into the vector

$$\vec{u}(k) \triangleq [u(k)\ u(k+1)\ \cdots u(k+N-1)] \tag{7}$$

B. Solution and Moving Horizon

The optimal constrained solution corresponds to finding $\vec{u}(k) \in U^N$, such that V_N is minimized. The following is the definition of the vector quantizer.

Definition 1 (Nearest Neighbor Vector Quantizer) [16]

Given a countable (not necessarily finite) set of non-equal vectors $B = \{b_1, b_2 \cdots\} \subset R^{n_B}$, the nearest neighbor quantizer is defined as a mapping $q_B : R^{n_B} \to B$ which assigns to each vector $c \in R^{n_B}$ the closest element of B(as measured by the Euclidean norm), i.e., $q_B(c) = b \in B$ if and only if c satisfies:

$$\|c-b\| \le \|c-b_i\|, \quad \forall b_i \in B \tag{8}$$

The optimal solution to the cost function V_N can be derived as stated in [15].

Theorem 1. Suppose $U^N = \{v_1, v_2 \cdots, v_{m_1}\}$, where $m_1 = (n_U)^N$, then the sequence $\vec{u}^*(k)$ of (7) which optimizes (6) under (5) is given by:

$$\vec{u}^*(k) = \Psi^{-1} q_{\tilde{U}^N}(\Psi \vec{r}(k) + \Gamma x(k)) \tag{9}$$

where:

$$\Psi = \begin{bmatrix} D & 0 & \cdots & 0 \\ h_1 & D & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ h_{N-1} & \cdots & h_1 & D \end{bmatrix}, \Gamma = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{N-1} \end{bmatrix}, \quad \vec{r}(k) = \begin{bmatrix} r(k) \\ r(k+1) \\ \vdots \\ r(k+N-1) \end{bmatrix}$$

and $h_i = CA^{i-1}B,\ i = 1,2,\cdots,N-1$

The nonlinearity $q_{\tilde{U}^N}(.)$ is the nearest neighbor quantizer described in Definition1. The image of this mapping is the set:

$$\tilde{U}^N = \{\tilde{v}_1, \tilde{v}_2, \cdots, \tilde{v}_{m_1}\} \subset R^N \ \text{ with } \tilde{v}_i = \Psi v_i,\ v_i \in U^N$$

Although, in principle, one might think of choosing N as large as the length of the complete audio data stream, the implementation complexity increases with the horizon length N. Therefore, the work in [15] proposed to fix the horizon N to a small value and use a moving horizon form. That means at time t=k, only the first element of the optimizing sequence is used as the output of converter, i.e.

$$u(k) = u^*(k) = [1 \ \ 0 \ \ \cdots \ \ 0]\Psi^{-1} q_{\tilde{U}^N}(\Psi r(k) + \Gamma x(k)) \tag{10}$$

Also, $u^*(k)$updates the states according to (5), and the new states are used to minimize the costV_N, yielding$u(k+1)$. The horizon moves (slides) forward as time increases.

To illustrate the effect of noise shaping, we write the z-transform of the optimized output $u^*(k)$ by (5), i.e.

$$U^* = R - W^{-1}E \tag{11}$$

where$R$and E are the z-transforms of r and$e$, respectively. Clearly, the output contains two parts of signal; the first term on the right hand side of (11) is the input signal, and the second term is the quantization error, which is the residual error E filtered by the inverse of the system filter, W^{-1}. Consequently, W^{-1} is responsible for spectrally shaping the noise.

III. Bounding the Error Signal and System States

To determine the design parameters, it is necessary to analyze the stability of (5) under the control law of (10). A general setting of stability was investigated in [15]. However, the results cannot be applied to the present case. In particular, it is usually required to maximize the allowable range of the input signal r. The technique presented in [13] and [14] for Σ–Δ modulator is adopted to access the stability. Notice that the major difference between Σ–Δ modulator and the optimal quantizer is the quantization scheme for the latter case is more complicated (see (10)). This requires further analysis into the quantization scheme to derive the stability boundary. This paper considers the case of horizon length one and two (i.e. N=1, 2 of (6)) and further assume that the system satisfies $0 < CB < D$, i.e. the relative degree is zero. In the following, we divide the problem into two parts. First, we assume that the error signal *e(k)* is bounded and develop a condition to bound the system states. In the second part, we explain how to bound *e(k)* when initial values of the system are all zero by limited input amplitude.

A. The condition to bound the system states

The constrained signal *u(k)* can be expressed as (12). By substituting (12) into the first equation of (5), we can write the system as (13).

$$u(k) = r(k) - D^{-1}e(k) + D^{-1}Cr(k) \tag{12}$$

$$x(k+1) = (A - BD^{-1}C)x(k) + BD^{-1}e(k) \tag{13}$$

If *e(k)* is bounded, the states are bounded if the system (13) is stable. It means that the eigenvalues of the matrix $(A - BD^{-1}C)$ must lie inside the unit circle. If the shaping

filter $W(z)$ is designed without pole zero cancellation, the eigenvalues of $(A - BD^{-1}C)$ are exactly the zeros of $W(z)$.

B. Bounding error signal for horizon length one

For N=1, the optimized sequence are

$$u^*(k) = q_{\tilde{U}^1}(Cx(k) + Dr(k)),$$

where $q_{\tilde{U}^1}(.)$ is a scalar quantizer. If the quantizer has the resolution of b-bit, there are K^1 levels in the constrained set and can be expressed as,

$$\tilde{U}^1 = \{D, D-\Delta D, D-2\Delta D \cdots, -D+\Delta D, D\} \quad (14)$$

where $\Delta = 2/(2^b - 1)$, $K^1 = 2^b$ if $b \in Z$, and $\Delta = 2/2^{b_1}$, $K^1 = 2^{b_1} + 1$ if $b = b_1 + 0.5$ and $b_1 \in Z$.

Therefore, the quantizer divides a line into K^1 segments considering the nearest distance between the quantizer input to the elements of constrained set. Fig.3 plots a general partitions of the scalar quantizer where d_1 is the input of it. We denote $\tilde{u}_j^1$ as the j-th element of $\tilde{U}^1$ and the output of quantizer in region D_i is $\tilde{u}_i^1$. The error signal $e(k)$ can be written as,

$$\begin{aligned} e(k) &= Cx(k) + D(r(k) - u(k)) \\ &= \{Cx(k) + Dr(k)\} - q_{\tilde{U}^1}(Cx(k) + Dr(k)) \\ &\triangleq d_1(k) - q_{\tilde{U}^1}(d_1(k)) \end{aligned} \quad (15)$$

According to (15), the bound of $e(k)$ is limited to $D \times \Delta/2$ under the condition:

$$|Cx(k) + Dr(k)| = |d_1(k)| \le D(1 + \Delta/2) \quad (16)$$

To satisfy the inequality of (16), we consider the condition of zero initial states. If the initial values of the system are all zero, then the error signal is bounded initially. As a result, we can limit the amplitude of input to prevent the violation of (16) [14]:

$$|r|_\infty \le 1 + \Delta/2 - \|P_1(z)\|_\infty \Delta/2 \quad (17)$$

where $P_1(z) = C(zI - A + BD^{-1}C)^{-1}BD^{-1}$. And from (13), it can be shown that,

$$Z\{Cx\} = P_1(z)E(z) \quad (18)$$

where $Z\{y\}$ denotes the z-transform of y.

C. Bounding error signal for horizon length two

For horizon length two, the optimized sequence are:

$$u^*(k) = [1 \quad 0]\Psi^{-1} q_{\tilde{U}^2}(\Gamma x(k) + \Psi \bar{r}(k))$$

where $\Psi = \begin{bmatrix} D & 0 \\ CB & D \end{bmatrix}$, $\Gamma = \begin{bmatrix} C \\ CA \end{bmatrix}$, and $q_{\tilde{U}^2}(.)$ is a vector quantizer. If the quantizer has the resolution of b-bit, there are $(K^1)^2$ conditions in the constrained set $\tilde{U}^2$. Therefore, the vector quantizer partitions the x-y plane into $(K^1)^2$ portions basing on the nearest neighbor vector quantizer. If we consider the first element of constrained set $\tilde{U}^2$ only, $(K^1)^2$ partitions can be grouped into K^1 regions and the output $\bar{u}^*$ is one of the levels in (19).

$$\tilde{U}_1^2 \triangleq [1 \quad 0]\tilde{U}^2 = \{D, D-\Delta D, D-2\Delta D \cdots, -D+\Delta D, -D\} \quad (19)$$

where $\Delta = 2/(2^b - 1)$ if $b \in Z$, and $\Delta = 2/2^{b_1}$ if $b = b_1 + 0.5$ and $b_1 \in Z$.

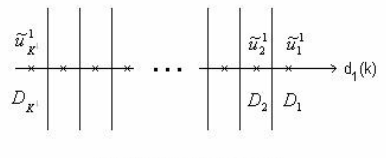

Fig.3 scalar quantizer

A general quantizer with resolution of b-bit for horizon length two is shown in Fig.4. The dots are constrained set $\tilde{U}^2$ which partitions the plane into K^1 regions. Obviously, the partition lines are saw-toothed that are different from the straight lines in horizon length one. It is caused by taking signal d_2 into account. Therefore, the k-th output of quantizer is influenced not only by $u(k)$ but also $u(k+1)$. A simulation example in next section shows that this kind of quantizer can further minimize the error signal and have better performance in SNR.

If we denote $\tilde{u}_{ij}^2$ as the i-th element of j-th column of $\tilde{U}^2$, where $\tilde{U}^2$ is shown in (20), the coordinates of the point A_{ij} can be expressed as the $(K^1(i-1)+j)$-th column of $\tilde{U}^2$.

$$\tilde{U}^2 = \Psi \begin{bmatrix} \tilde{u}_1^1 & \tilde{u}_2^1 & \cdots & \tilde{u}_{(2^{b_1}+1)}^1 & \tilde{u}_1^1 \cdots & \tilde{u}_{2^{b_1}}^1 & \tilde{u}_{(2^{b_1}+1)}^1 \\ \tilde{u}_1^1 & \tilde{u}_1^1 & \cdots & \tilde{u}_1^1 & \tilde{u}_2^1 \cdots & \tilde{u}_{(2^{b_1}+1)}^1 & \tilde{u}_{(2^{b_1}+1)}^1 \end{bmatrix} D^{-1} \quad (20)$$

$$\begin{aligned} e(k) &= Cx(k) + D(r(k) - u(k)) \\ &= \{Cx(k) + Dr(k)\} - [1 \quad 0]\Psi^{-1} q_{\tilde{U}^1}(\Gamma x(k) + \Psi \bar{r}(k)) \end{aligned} \quad (21)$$

The quantizer output is the first element of constrained set to which the quantizer input belongs. From this general horizon length two quantizer, we can calculate the error bound when its output is $\tilde{u}_i^1, 2 \le i \le K^1 - 1$. For example, if $i = 2$, the amplitude of error (see (21)) is dominated by the line (a) and (b) in Fig.2 if $|d_2| \ge y_0$,

$$y_0 > 0 \in R, and \left| \frac{-CB}{D}\left(-y_0 - \frac{\tilde{p}_{21} + \tilde{p}_{22}}{2}\right) + \frac{\tilde{p}_{11} + \tilde{p}_{12}}{2} \right| \ge x_b \quad (22)$$

where point (x_b, y_b) is the intersection of line (c) and (d) in Fig.4.

Now suppose $y_0 \le |d_2| \le y_1$, then the error bound for quantizer output level equal to $\tilde{u}_2^1$ is,

$$|err|_2 \le \max(e_1, e_2),$$

where $e_1 = |\tilde{u}_2^1 - x_1|$, $e_2 = |\tilde{u}_2^1 - x_2|$

and $x_1 = -\frac{CB}{D}(y_1 - \left[\frac{\tilde{v}_{22} + \tilde{v}_{23}}{2}\right]) + \frac{\tilde{v}_{12} + \tilde{v}_{13}}{2}$

$$x_2 = -\frac{CB}{D}(-y_1 - \left[\frac{\tilde{p}_{21} + \tilde{p}_{22}}{2}\right]) + \frac{\tilde{p}_{11} + \tilde{p}_{12}}{2}$$

$\tilde{v}_{ij}$ ($\tilde{p}_{ij}$) is the i-th element of j-th column of $\tilde{V}$ ($\tilde{P}$) defined below.

$$\tilde{V} = \Psi \begin{bmatrix} \tilde{u}_1^1 & \tilde{u}_2^1 & \cdots & \tilde{u}_{K^1}^1 \\ \tilde{u}_1^1 & \tilde{u}_1^1 & \cdots & \tilde{u}_1^1 \end{bmatrix} D^{-1}, \tilde{P} = \Psi \begin{bmatrix} \tilde{u}_1^1 & \tilde{u}_2^1 & \cdots & \tilde{u}_{K^1}^1 \\ \tilde{u}_{K^1}^1 & \tilde{u}_{K^1}^1 & \cdots & \tilde{u}_{K^1}^1 \end{bmatrix} D^{-1} \quad (23)$$

Note that x_1 (x_2) is the left (right) boundary of d_1 limited by $|d_2| \le y_1$ in region D_2, i.e. they are calculated by the equation of line (a) (line (b)) and the maximum error must occur at one of these points. As a result, the maximum error for output $\tilde{u}_i^1, 2 \le i \le K^1 - 1$ can be calculated by,

$|err|_{\max} = \max(err_2, err_3 \cdots, err_{K^1-1})$, where (24)

$err_i = \max(e_{i1}, e_{i2})$ and

$e_{i1} = |\tilde{u}_i^1 - x_{i1}|$, $e_{i2} = |\tilde{u}_i^1 - x_{i2}|$ and

$$x_{i1} = -\frac{CB}{D}(y_1 - \left[\frac{\tilde{v}_{2i} + \tilde{v}_{2(i+1)}}{2}\right]) + \frac{\tilde{v}_{1i} + \tilde{v}_{1(i+1)}}{2}$$

$$x_{i2} = -\frac{CB}{D}(-y_1 - \left[\frac{\tilde{p}_{2(i-1)} + \tilde{p}_{2i}}{2}\right]) + \frac{\tilde{p}_{1(i-1)} + \tilde{p}_{1i}}{2}$$

Also, $x_{i1}(x_{i2})$ is the left (right) boundary of d_1 limited by $|d_2| \le y_1$ in region D_i. If we further limit the maximum amplitude of d_1 (see (25)), the error signal is totally bounded no matter what the quantizer output level is.

$$|d_1| \le \tilde{u}_1^1 + |e|_\infty \overset{\Delta}{=} |d_1|_{\max} \quad (25)$$

where $|e|_\infty \overset{\Delta}{=} \max(|err|_{\max}, err_1, err_{K^1})$, $|err|_{\max}$ is depicted in (24), and ,

$err_1 = |\tilde{u}_1^1 - x_{11}|$, $err_{K^1} = |\tilde{u}_{K^1}^1 - x_{K^1 1}|$

$$x_{11} = -\frac{CB}{D}(y_1 - \left[\frac{\tilde{v}_{21} + \tilde{v}_{22}}{2}\right]) + \frac{\tilde{v}_{11} + \tilde{v}_{12}}{2}$$

$$x_{K^1 1} = -\frac{CB}{D}(-y_1 - \left[\frac{\tilde{p}_{2(K^1-1)} + \tilde{p}_{2K^1}}{2}\right]) + \frac{\tilde{p}_{1(K^1-1)} + \tilde{p}_{1K^1}}{2}$$

To ensure that the inequality constraints of d_1 and d_2 ((25) and $|d_2| \le y_1$) are always satisfied when initial conditions of the system are all zero, we have to limit the maximum input amplitude (see (26)). Because the error is bounded initially, the error will be bounded all the time if input satisfies (26) [14].

$$|r|_\infty \le \min(r_1, r_2) \quad (26)$$

, where $r_1 = (|d_1|_{\max} - \|P_1(z)\|_\infty |e|_\infty) D^{-1} > 0$

$r_2 = (y_1 - \|P_2(z)\|_\infty |e|_\infty) \times (CB + D)^{-1} > 0$

$P_1(z) = C(zI - A + BD^{-1}C)^{-1} BD^{-1}$

$P_2(z) = CA(zI - A + BD^{-1}C)^{-1} BD^{-1}$

And from (13), it can be shown that,

$Z\{Cx\} = P_1(z)E(z)$, $Z\{CAx\} = P_2(z)E(z)$.

Note that the error is proportional to y_1. Therefore, the increase in y_1 will result in decrease of r_1 but also increase of r_2.

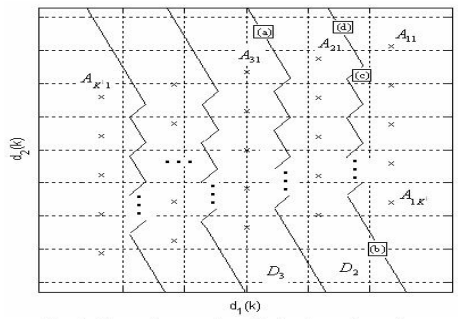

Fig.4 General quantizer for horizon length two

As a result, there exists an optimal value of y_1 that achieves maximum value of $|r|_\infty$. To find out the optimal value of y_1, we suppose the maximum error in (24) is dominated by the line,

$$x_{e2} = \frac{-CB}{D}\left(-y_1 - \frac{\tilde{p}_{2(e-1)} + \tilde{p}_{2e}}{2}\right) + \frac{\tilde{p}_{1(e-1)} + \tilde{p}_{1e}}{2} \quad (27)$$

where $e \in Z, 2 \le e \le K^1 - 1$

The error then can be expressed as $(x_e - \tilde{u}_e^1)$, and from (25) and (26), we get,

$$r_1 = (\tilde{u}_1^1 + (1 - \|P_1(z)\|_\infty)(x_{e2} - \tilde{u}_e^1))D^{-1} \quad (28)$$

$$r_2 = (y_1 - \|P_2(z)\|_\infty (x_{e2} - \tilde{u}_e^1)) \times (CB + D)^{-1} \quad (29)$$

By substituting (27) into (28) and (29) and setting $r_1 = r_2$, the optimal value of y_1 is obtained as,

$y_1 = y_{num} / y_{den}$, and (30)

$$y_{num} = D^{-1}\tilde{u}_1^1 + \left(D^{-1}(1 - \|P_1\|_\infty) + (CB + D)^{-1}\|P_2\|_\infty\right) \times \left(\frac{CB}{D}\frac{\tilde{p}_{2(e-1)} + \tilde{p}_{2e}}{2} + \frac{\tilde{p}_{1(e-1)} + \tilde{p}_{1e}}{2} - \tilde{u}_e^1\right)$$

$$y_{den} = \left((CB + D)^{-1}\left(1 - \|P_2\|_\infty \frac{CB}{D}\right) - (1 - \|P_1\|_\infty)\frac{CB}{D^2}\right)$$

Although the general partitions in Fig.4 is calculated for the system satisfying $0 < CB < D$, we can always derive the error bound for different relative degree and any order of the system in the same procedure proposed here.

IV. A DESIGN EXAMPLE

An example for audio-band full digital amplifier is recommended here. The reference signal bandwidth is 48 KHz and the oversampling ratio is 64. Shaping filter is designed with cutoff frequency 65 KHz and the resolution of quantizer is 2.5-bit. Converting the 2.5-bit result into the switching control signal is introduced in Fig.2. In the following, we first design the loop filter which has stable zeros. Second, a 2.5-bit quantizer with step size one and two are shown along with the block diagrams for quantizer implementation.

A. Weighting Filter Design

We use a third order lowpass filter with relative degree zero as our system:

$$W(z) = \frac{1.24 - 1.95z^{-1} + 0.8z^{-2}}{1 - 2z^{-1} + z^{-2}} \quad (31)$$

Therefore, the noise transfer function is fixed and its frequency response in magnitude is shown in Fig.5. Also, the state space matrix of the system is obtained:

$$A = \begin{bmatrix} 0 & -1 \\ 1 & 2 \end{bmatrix},\ B = \begin{bmatrix} -0.44 \\ 0.53 \end{bmatrix},\ C = [0 \quad 1],\ D = 1.24 \quad (32)$$

Since the zeros are (0.786±0.176j), the system states are stable if the error signal is bounded.

B. Quantizer Design

With the use of 2.5-bit resolution quantizer, the output constrained set is U={-1 -0.5 0 0.5 1}. Because the oversampling ratio is 64, the system operates at a clock rate of 3.072 MHz. To produce the switching signal as in Fig.2, the actual clock rate is 6.144 MHz. The zero level output in 2.5-bit switching signal can be regarded as the zero voltage

difference between loudspeakers. We shall discuss it as follows,

1) For Horizon one (N=1), $\Psi = D = 1.24$, the constrained set is mapped to the same space, hence, $\tilde{U} = \Psi U = \{1.24, 0.62, 0, -0.62, -1.24\}$. Fig.6 shows the relation of quantizer input $d_1(k) = Cx(k) + Dr(k)$ and output $u^*(k)$ which can be written as,

$$u^*(k) = \frac{1}{4}\left(q_U\{d_1(k)+0.31\} + q_U\{d_1(k)-0.31\} + q_U\{d_1(k)+0.93\} + q_U\{d_1(k)-0.93\}\right) \quad (33)$$

The diagram of implementing the scalar quantization is presented in Fig.7. Note that the operation of $q_U(.)$ is defined as:

$$q_U(a) = \begin{cases} 1 & if\ a > 0 \\ -1 & else \end{cases}$$

From (17), the input level is limited to:

$$\|r\|_\infty \le 1 + 2^{-2} - 1.36 \times 2^{-2} = 0.9$$

2) For horizon two (N=2), $\Psi = \begin{bmatrix} D & 0 \\ CB & D \end{bmatrix} = \begin{bmatrix} 1.24 & 0 \\ 0.53 & 1.24 \end{bmatrix}$, and U={-1 -0.5 0 0.5 1}, the input of the quantizer is a vector. There are 25 conditions in the constrained set U^2. The vector quantizer $q_{\tilde{U}^2(.)}$ partitions its input space R^2 into twenty-five regions according to the nearest neighborhood rule. Since we are only interested in the first element of $u^*(k)$, only five regions are of significance. These are shown in Fig.8. The optimized output is then given by

$$u^*(k) = \begin{cases} 1, & if\ d(k) = [d_1(k)\ \ d_2(k)] \in D_1 \\ 1/2, & if\ d(k) = [d_1(k)\ \ d_2(k)] \in D_2 \\ 0, & if\ d(k) = [d_1(k)\ \ d_2(k)] \in D_3 \\ -1/2, & if\ d(k) = [d_1(k)\ \ d_2(k)] \in D_4 \\ -1, & if\ d(k) = [d_1(k)\ \ d_2(k)] \in D_5 \end{cases} \quad (34)$$

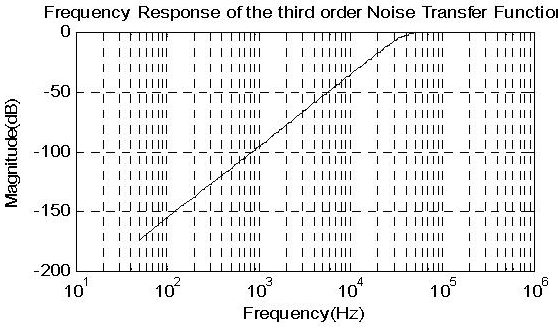

Fig.5 Frequency Response of the noise transfer function in design

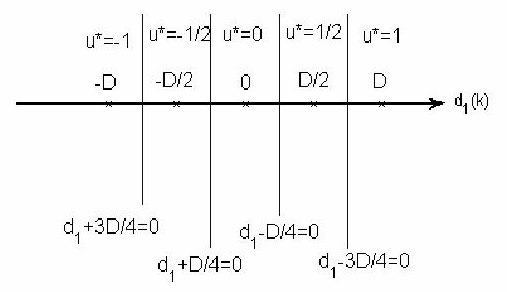

Fig.6 scalar quantizer

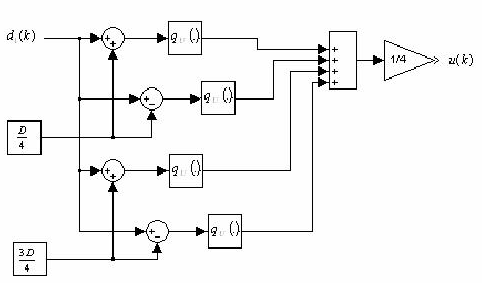

Fig.7 block diagram of scalar quantizer

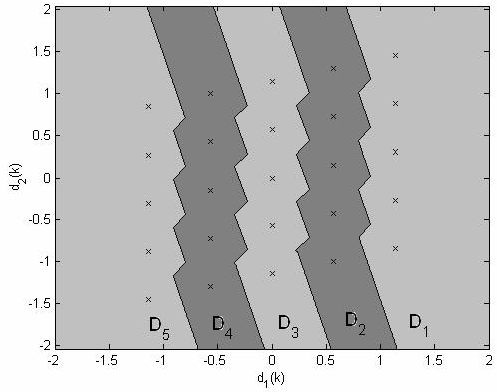

Fig.8 Partition induced by the quantizer

From (30), optimized bound of d_2 is 3.07 and error is dominated by the equation,

$$x_{e2} = \frac{-CB}{D}\left(-y_1 - \frac{\tilde{p}_{2(e-1)} + \tilde{p}_{2e}}{2}\right) + \frac{\tilde{p}_{1(e-1)} + \tilde{p}_{1e}}{2},\ where\ e = 1$$

Its bound is 1.26 if $|d_1|_{\max} < 2.51$ (see (24) and (25)). The maximum input range is 0.65 for horizon length two. Geometrical arguments allow us to describe the partition of Fig.8 by means of the following relation,

$$u^*(k) = \frac{1}{4}\sum_{i=1}^{4} q_U\left\{\sum_{j=1}^{5} f_{ij} + \sum_{j=1}^{4} g_{ij}\right\}$$

, where $f_{ij} = q_U\left\{d_1(k) + \frac{CB}{D}\left(d_2(k) - y_{m_{ij}}\right) - x_{m_{ij}}\right\}$

$$y_{m_{ij}} = \left(\tilde{u}^2_{2(K^1(j-1)+i)} + \tilde{u}^2_{2(K^1(j-1)+i+1)}\right)/2$$

$$x_{m_{ij}} = \left(\tilde{u}^2_{1(K^1(j-1)+i)} + \tilde{u}^2_{1(K^1(j-1)+i+1)}\right)/2$$

, and $g_{ij} = q_U\left\{d_1(k) - \frac{(CB-D)}{D}\left(d_2(k) - y_{m_{ij}}\right) - x_{m_{ij}}\right\}$

$$y_{m_{ij}} = \left(\tilde{u}^2_{2(K^1(j-1)+i+1)} + \tilde{u}^2_{2(K^1 j+i)}\right)/2$$

$$x_{m_{ij}} = \left(\tilde{u}^2_{1(K^1(j-1)+i+1)} + \tilde{u}^2_{1(K^1 j+i)}\right)/2$$

As a consequence, the vector quantizer $q_{\tilde{U}^2}(.)$ can be realized with thirteen standard scalar quantizers $q_U(.)$, which operate on scalar signals.

C. Simulation Results

A 1kHz sine wave with normalized amplitude 0.65 sampled at 48kHz is applied to the system as an input. Fig.10、Fig.11 plots the output spectrum obtained by taking DFT (Discrete Fourier Transform) of the binary sequence produced with horizon two and horizon one. Simulation time is 100ms and the resulting SNDR is 89.5dB、87.6dB.

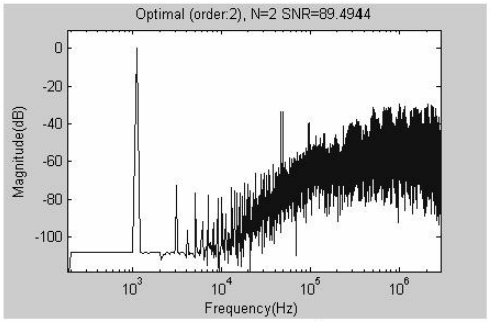

Fig.10 Spectrum of output u^* (horizon 2)

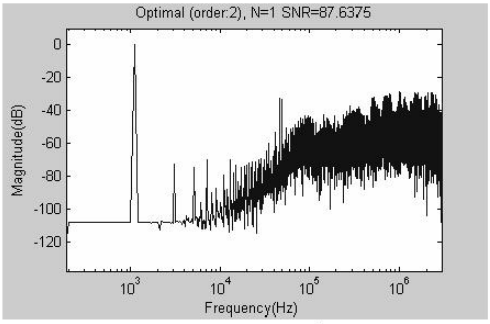

Fig.11 Spectrum of output u^* (horizon 1)

V. CONCLUSION

A novel application of finite horizon constrained optimization techniques to design an audio-band full digital amplifier is presented in this paper. Stability analysis to arrive at the bound of the input reference signal is given by a detailed derivation of the optimal quantizer. The paper only reports the case of moving horizon of length 1 and 2. The general case of an arbitrary horizon length will be developed in the future work.

ACKNOWLEDGMENT

This work was supported by National Science Council of the R.O.C. under grant no. NSC94－2218－E009064 and MOE ATU Program under the account number 95W803E.

REFERENCES

[1] Engelen, J. van and Plassche, R. van de, "New stability criteria for the design of lowpass sigma-delta modulators," *In Proc. of Int. Symp. Low Power Electronics and Design*, 1997,pp. 114-118.

[2] Hiorns, R.E. and Sandler, M.B., "Power digital to analogue conversion using pulse width modulation and digital signal processing," *In IEE Circuits, Devices and Systems, Proceedings G*, **140**, Issue: 5, Oct., 1993, 329 - 338

[3] Hu, J and Yu S-H, "Analysis and design of 1-bit noise-shaping quantizer using variable structure control approach," *In 2004 American Control Conference, Boston, USA, 2004.*

[4] Li H.; Gwee, B.H., and Chang, J.S., "A digital Class D amplifier design embodying a novel sampling process and pulse generator," *In 2001 IEEE International Symposium on Circuits and Systems (ISCAS'01),2001,* **4** , 826 – 829.

[5] Logan, S. and Hawksford, M.O.J., "Linearization of class D output stages for high-performance audio power amplifiers," *In Second International Conference on Advanced A-D and D-A Conversion Techniques and their Applications*, Jul. 1994, 136 - 141

[6] Park J.; Kim, C.G., Jeong J.; Cho, B.H.A, "A novel controller for switching audio power amplifier with digital input," *In IEEE 33rd Annual Power Electronics Specialists Conference*, June 2002, **1**, 39 – 44.

[7] Putzeys, B., "Digital audio's final frontier," *In IEEE Spectrum*, 2003, **40** , Issue 3 , March, 34-41

[8] Schreier, R., "On the use of chaos to reduce idle-channel tones in delta-sigma modulators," *In IEEE Transactions on, Circuits and Systems I: Fundamental Theory and Applications*, **41**, Issue: 8, Aug., 1994, 539 – 547.

[9] Schreier R. and Yang Y., "Stability tests for single-bit sigma-delta modulators with second-order FIR noise transfer functions," *In Proc.* IEEE Int. Symp. Circuits and Systems, 1992, 1316-1319.

[10] Silva J. F., "PWM audio power amplifiers: sigma delta versus sliding mode control," In IEEE Int. Conference on Electronics, Circuits and Systems, 1998, 1, 359-362.

[11] Smedley, K.M., "Digital-PWM audio power amplifiers with noise and ripple shaping," In 25th Annual IEEE Power Electronics Specialists Conference, June 1994, 1, 566 – 570.

[12] Wang H., "A geometric view of Modulations," In IEEE Trans. Circuits and Systems-II, 1992, 39, no. 2, 402-405.

[13] Yu S-H and Hu, J, "Sigma-delta modulators operated in optimization mode," In 2004 IEEE International Symposium on Circuits and Systems (ISCAS'04), May, British Columbia, Canada.

[14] Shiang-Hwua Yu, "Noise-Shaping Coding Through Bounding the Frequency-Weighted Reconstruction Error," IEEE Transactions On Circuit And Systems I: Express Briefs, Vol. 53, No. 1, January 2006

[15] Daniel E. Quevedo and Graham C. Goodwin, "Multi-Step optimal analog-to-digital conversion, " IEEE Transactions On Circuits and Systems I: Regular Papers, Volume: 52, pp503-505, March 2005

[16] Daniel E. Quevedo, José A. De Doná, and Graham C. Goodwin, "Receding horizon linear quadratic control with finite input constraint sets," IFAC 15th Triennial World Congress, Barcelona, Spain, 2002.

[17] Daniel E. Quevedo and Graham C. Goodwin,"Audio Quantization from a Receding Horizon Control Perspective," Proceedings of the American Control Conference, Denver, Colorado June 4-6,2003

[18] Jwu-Sheng Hu and Keng-Yuan Chen, "Implementation of a Full Digital Amplifier Using Feedback Quantization," 2006 American Control Conference, Minneapolis, Minnesota USA. June 14-16, 2006.

[19] Shiang-Hwua Yu, "Analysis and Design of Single-Bit Sigma-Delta Modulators Using the Theory of Sliding Modes," IEEE Transactions On Control Systems Technology, Vol. 14, No. 2, March 2006