Optimizing Pulse Shaping Filter for DOCSIS Systems

Tung T. Nguyen, Student Member, IEEE, Joseph Eric Salt, Member, IEEE, Ha H. Nguyen, Senior Member, IEEE, and Brian Berscheid Member, IEEE

Abstract—This paper proposes a cost efficient nearly linear phase approximation to an square root raised cosine (SRRC) pulse shaping filter that satisfies the out-of-band spectral constraints of DOCSIS down-stream channels. A nearly linear filter structure is converted to an SRRC filter using a weighted and sampled least-squares criterion to fit the magnitude and phase responses. To ensure stability, the search for optimum coefficients is constrained using the Steiglitz–McBride (SM) and Gauss–Newton (GN) methods, which unfortunately also eliminates sets of stable coefficients, one of which could be and probably is the optimum. To expand the sets of coefficients that produce stable filters in the SM and GN methods, the transfer function is parameterized in a special way. The effectiveness of the proposed filter is verified and compared with other approaches.

Index Terms—Modems, pulse shaping methods, IIR filters, matched filters, intersymbol interference.

I. INTRODUCTION

When a new communication standard is implemented, the inherited backward compatible functions can be redesigned to take advantage of the latest technology and algorithms. Of interest here is the pulse shaping filter, which is the Nyquist filter located in the transmitter, used in DOCSIS single carrier QAM. Improving the performance of Nyquist filters in terms of reducing ISI is still an active research area [1]–[4]. However, DOCSIS pulse shaping filters must not only have low ISI, they must also stringently suppress out-of-band emission to avoid adjacent channel interference. Since the DOCSIS standards dictate Square Root Raised Cosine (SRRC) Nyquist filters, of interest here are SRRC pulse shaping filters with low ISI and low out-of-band emissions. The ideal pulse shaping filter is an infinite-length Square-Root Raised Cosine (SRRC) filter with complex frequency-domain response described by: (1), as shown at the bottom of the next page, where \( \omega \) is frequency in radians/sample and \( \beta, L, \varphi \) are the filter’s roll-off factor, ratio of sampling rate to symbol rate and phase delay, respectively.

The SRRC filter can be approximated by either a finite impulse response (FIR) or an infinite impulse response (IIR) filter. While FIR approximations have been widely used, they require large hardware resources (in terms of arithmetic circuits) to meet the out-of-channel portion of the DOCSIS spectral mask requirements. The alternative is a recursive IIR structure.

The main thrust of the paper is to improve the implementation efficiency of the SRRC pulse shaping filter for legacy single carrier down-stream QAM channels in the implementation of DOCSIS 3.1 and beyond. The specific objective is to provide a trade-off between ISI reduction and filter cost under the constraint the filter meets the DOCSIS 3.1 out-of-channel spectral mask. The filter derived from the approach proposed in this paper will be discussed and compared with filters designed with other techniques.

II. REVIEW OF RELEVANT FILTER DESIGN TECHNIQUES

A. FIR Approach

1) Windowing Technique: The simplest method is to obtain the impulse response of the SRRC filter by taking the inverse discrete-time Fourier transform of (1). The impulse response has a “sinc-like” shape, therefore it clearly cannot be implemented in practice, unless the “sinc-like” function is truncated on both sides. As such, the coefficients for a causal FIR filter are obtained by taking the \( N_F \) values closest to \( n = 0 \). The result is the sample-space impulse response given by, (2) as shown at the bottom of the next page, where \( \hat{\varphi} = \frac{N_F - 1}{2} \) is delay in samples of the truncated FIR filter.

Fig. 1 presents the impulse response of a truncated SRRC filter as well as the magnitude responses of ideal and truncated SRRC filters. The filter parameters are \( \beta = 0.12 \) and \( L = 4 \), with the length of the truncated filter being \( N_F = 41 \).

As can be seen in Fig. 1, truncating the impulse response introduces ripple in both the pass band and stop band of the magnitude response. To suppress the stop band ripple, the truncated “sinc-like” function must be tapered to force its extremity to approach zero. This can be done by multiplying the truncated impulse response by a window function to get the windowed impulse response:

\[
h_w(n) = h(n)w(n), \quad n = 0, 1, \ldots, N_F - 1, \tag{3}
\]

where \( w(n) \) is the window function. The main source of signal distortion created by windowing is the widening of the transition band. The widening can be reduced by increasing the

Digital Object Identifier 10.1109/TBC.2015.2505412

Manuscript received August 7, 2015; revised October 26, 2015 and November 16, 2015; accepted November 17, 2015. Date of publication January 22, 2016; date of current version June 3, 2016. This work was supported by the Natural Sciences and Engineering Research Council of Canada under Grant IPS2 Scholarship.

T. T. Nguyen, J. E. Salt, and H. H. Nguyen are with the Department of Electrical and Computer Engineering, University of Saskatchewan, Saskatoon, SK S7N 5A9, Canada (e-mail: tung.nguyen@usask.ca).

B. Berscheid is with Vecima Network Inc., Saskatoon, SK S7K 1P4, Canada.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBC.2015.2505412

length of the filter or by pre-compensating by using a lower roll-off factor. There are many well-known window functions, such as Blackman, Hamming, Kaiser, Gaussian, Tukey, etc., each with its own stop band ripple rejection and transition band distortion characteristics. For an extensive comparison among various window functions, the interested reader is referred to [5]. An appealing window is the Kaiser window, whose stop band ripple suppression trade off. The Kaiser window method that minimizes the ISI for a filter of length $N_F$ that satisfies the out-of-band portion of the DOCSIS specified spectral mask works as follows:

1. Choose the filter length $N_F$.
2. Set the roll off factor $\beta$ to its nominal value as specified in the DOCSIS standard (0.18 for 64-QAM channels, or 0.12 for 256-QAM channels).
3. Decrease $\beta$ by a small amount.
4. Set the Kaiser window parameter, $\alpha$, to 0 and steadily increase it until the stop band response satisfies the out-of-channel portion of the spectral mask specified in DOCSIS.
5. Measure and record the values of $\alpha$, $\beta$ and ISI.
6. Reduce $\beta$ and repeat steps 3 to 5 to minimize the ISI.
7. Choose the set $\alpha$, $\beta$ that produces minimum ISI.

2) Fred Harris’ Technique: Harris approximates the SRRC filter with a linear phase 3-band minimax filter [6]. In addition to specifying the pass band and stop band, a very narrow third band is specified in transition region at the 3dB-down frequency. Harris designs the filter with the Parks McClellan algorithm using the following parameter vectors

$$\text{Frequency (radians/sample): } [0 \ o_{p}/L \ \pi/L \ \pi/L \ o_{s}/L \ \pi]$$
$$\text{Gain: } [1 \ 1 \ \sqrt{2}/2 \ \sqrt{2}/2 \ 0 \ 0]$$
$$\text{Weight: } [2.4535 \ 1 \ 1]$$

where $o_{p}$ and $o_{s}$ are the pass band and stop band corner frequencies, given by

$$o_{p} = \pi(1 - \beta), \quad \text{and} \quad o_{s} = \pi(1 + \beta). \quad (5)$$

The first coefficient in the weight vector is manually adjusted to minimize the ISI level when the matched filter is the ideal SRRC filter truncated to the same length.

$$D(\omega) = \begin{cases} 
e^{-j\omega o_p}, & \text{for } |\omega| \leq \frac{\pi(1-\beta)}{L} \\ e^{-j\omega o_p} \sqrt{\frac{1}{2} \left[ 1 + \cos\left(\frac{\pi}{4\beta} \left( \frac{|\omega|}{L} - 1 + \beta \right) \right) \right]}, & \text{for } \frac{\pi(1-\beta)}{L} < |\omega| \leq \frac{\pi(1+\beta)}{L} \\ 0, & \text{otherwise}, \end{cases}$$

$$h(n) = \begin{cases} \frac{1}{L} \left( 1 - \beta + \frac{4\beta}{\pi} \right), & \text{for } n = \frac{N_F - 1}{2} \\ \frac{\beta}{\sqrt{2L}} \left[ \left( 1 + \frac{2}{\pi} \right) \sin\left( \frac{\pi}{4\beta} \right) + \left( 1 - \frac{2}{\pi} \cos\left( \frac{\pi}{4\beta} \right) \right) \right], & \text{for } n = \frac{N_F - 1}{2} \pm \frac{L}{4\beta} \\ \sin\left( \frac{\pi(1-\beta)(n-\phi)}{L} \right) + \frac{4\beta(n-\phi)}{L} \cos\left( \frac{\pi(1+\beta)(n-\phi)}{L} \right), & 0 \leq n \leq N_F - 1, \quad n \neq \frac{N_F - 1}{2} \pm \frac{L}{4\beta} \\ \pi(n-\phi) \left[ 1 - \frac{4\beta(n-\phi)}{L} \right]^2, & \text{otherwise}, \end{cases}$$

where $I_0(\cdot)$ is the zeroth-order modified Bessel function of the first kind, and $\alpha$ is the parameter that determines the shape of the window and balances the transition band distortion and stop band ripple suppression trade off. The Kaiser window function is given by

$$w(n) = \frac{I_0\left( \frac{\pi\alpha}{I_0(\pi\alpha)} \sqrt{1 - \frac{2n}{N} \left( 1 - \frac{2n}{N} \right)} \right)}{I_0(\pi\alpha)}, \quad n = 0, 1, \ldots, N_F - 1,$$
The filter created in this way has an unwanted spectral bump at the edge of the pass band, which is believed to be a major contributor to the ISI. To suppress this bump, Harris suggests adjusting the coefficients in the frequency vector from \( \omega_1 \) to \( \epsilon_1 \omega_2 \), and also fine tuning the stop band edge from \( \omega_2 \) to \( \epsilon_2 \omega_2 \), where the values of \( \epsilon_1 \) and \( \epsilon_2 \) are slightly more than 1. The new target frequency vector is then given as

\[
\text{Frequency (radians/sample): } [0 \ \epsilon_1 \omega_2 / L \ \pi / L \ \pi / L \ \epsilon_2 \omega_2 / L \ \pi].
\]

The procedure is performed by first finding parameter \( \epsilon_1 \) that minimize the ISI and then proceeding to find \( \epsilon_2 \) with the same goal.

The problem with Harris’ approach is that it does not consider the out-of-channel emission and therefore Harris’ filter will not meet the out-of-channel portion of the spectral mask unless a relatively large filter order is chosen.

B. Nearly Linear-Phase IIR Filter Approximation

The design task is to find a stable IIR filter that has a system function of the form:

\[
H(z) = \frac{b_0 + b_1 z^{-1} + \ldots + b_m z^{-m}}{1 + a_1 z^{-1} + \ldots + a_n z^{-n}},
\]

where \( b_i \) and \( a_i \) are real coefficients and \( m \) is the order of the numerator polynomial and \( n \) is the order of the denominator polynomial. The corresponding filter’s frequency response is

\[
H(e^{j\omega}) = \frac{b(a)}{a(a)} = \frac{b^T e_{bm}(a)}{1 + a^T e_{an}(a)},
\]

where \( b(a) \) and \( a(a) \) are the frequency responses on the numerator and denominator polynomials in (6) and the superscript \( T \) indicates the transpose of a vector or matrix. The numerator and denominator coefficients are represented by vectors \( \mathbf{b} = [b_0, b_1, \ldots, b_m]^T \) and \( \mathbf{a} = [a_1, \ldots, a_n]^T \) respectively. In addition, \( [e_{bk}(a), \text{ for } l < k] \), is defined as

\[
e_{lk}(a) = [e^{-j\omega l}, e^{-j\omega(l+1)}, \ldots, e^{-j\omega k}]^T.
\]

It is convenient to define the frequency response \( H(e^{j\omega}) \) as a complex function of real parameters \( X = [\mathbf{a}^T, \mathbf{b}^T] \) and refer to the frequency response as \( H(e^{j\omega}; X) \). Furthermore, it can be seen that \( H(e^{j\omega}; X) \) is a non-linear function of \( \omega \) and has a non-linear dependency on parameters \( X = [a_1, \ldots, a_n, b_0, b_1, \ldots, b_m]^T \). When \( H(e^{j\omega}; X) \) is designed to approximate the ideal frequency response \( D(\omega) \) in (1), a weighted squared error is given by \( J(X) = \sum \frac{W_i |D(\omega_i) - H(e^{j\omega_i}; X)|^2}{|a(\omega_i; X_0)|^2} \), where \( W_i \) is a non-negative weighting coefficient associated with frequency \( \omega_i \in [0, \pi] \).

The goal is to find the parameter vector \( X \) that minimizes the sum of the weighted least square errors at frequencies \( \omega_1, \omega_2, \ldots, \omega_N \) under the constraint the parameter \( X \) yields a stable filter. Since the total squared error is given by

\[
J(X) = \sum_{i=1}^{M} J_i(X) = \sum_{i=1}^{M} W_i |D(\omega_i) - H(e^{j\omega_i}; X)|^2,
\]

finding the value \( X = X_{\text{min}} \) that minimizes \( J(X) \) is difficult due to the non-linear nature of \( J(X) \). However, the Steiglitz-McBride and Gauss-Newton non-linear iterative optimization approaches can be used to find \( X \).

1) Steiglitz–McBride (SM) Technique: Theoretically, it is possible for the SM method [7–9] to find the vector \( X \) that minimizes the objective function defined in (9). Practically, it finds a good estimate of that vector.

The SM method finds a sequence of vectors that minimize a carefully designed sequence of objective functions. The objective functions are denoted \( J_{SM,1}(X), J_{SM,2}(X), \ldots \) and \( X_{SM,\text{min},1}, X_{SM,\text{min},2}, \ldots \) are the vectors that minimize them. The SM sequence of objective functions are such that

\[
\lim_{k \to \infty} X_{SM,\text{min},k} = X_{\text{min}},
\]

where \( X_{\text{min}} \) is the vector that minimizes \( J(X) \).

The SM objective functions are obtained by reformating and rearranging the expression for \( J(X) \) given by (9). The reformating is done by substituting \( [a(\omega_i; X)] \) for \( H(e^{j\omega_i}; X) \) into (9). The rearranging is factoring \( a(\omega_i; X) \) outside the absolute value operator. The reformatted, rearranged expression for \( J(X) \) is:

\[
J(X) = \sum_{i=1}^{M} \frac{W_i}{|a(\omega_i; X_0)|^2} |D(\omega_i)a(\omega_i; X) - b(\omega_i; X)|^2.
\]

The sequence of SM objective functions are successively approximating \( |a(\omega_i; X)|^2 \) with a set of \( M+1 \) constants that converge to the true values. The constants used for the first objective function is obtained by guessing vector \( X_{\text{min}} \) to be \( X_0 \) and substituting it into \( |a(\omega_i; X)|^2 \). This produces

\[
J_{SM,1}(X) = \sum_{i=1}^{M} \frac{W_i}{|a(\omega_i; X_0)|^2} |D(\omega_i)a(\omega_i; X) - b(\omega_i; X)|^2,
\]

where \( |a(\omega_i; X_0)|^2, i = 0, 1, \ldots, M \) are the constants for iteration \( 1 \) and \( X_0 \) is the initial guess for \( X_{\text{min}} \). Unlike \( J(X) \), \( J_{SM,1}(X) \) is quadratic in \( X \), which means \( X_{SM,\text{min},1} \) can be found by a standard quadratic search algorithm.

In general, the \((k+1)\)\textsuperscript{th} SM objective function is the quadratic function of \( X \) given by

\[
J_{SM,k+1}(X) = \sum_{i=1}^{M} \frac{W_i}{|a(\omega_i; X_k)|^2} |D(\omega_i)a(\omega_i; X) - b(\omega_i; X)|^2,
\]

where \( X_k, k > 0 \) is the \( k \)\textsuperscript{th} estimate of \( X_{\text{min}} \), given by

\[
X_k = \mu X_{SM,\text{min},k} + (1 - \mu) X_{k-1},
\]

where \( \mu \) is a real number less than 1 used to damp the sequences \( X_k \).

The number of SM objective functions that need to be solved to get a good estimate of \( X_{\text{min}} \) depends on the order of the filter \( H(e^{j\omega_i}; X) \) and the step size \( \mu \). Making \( \mu \) small ensures convergence to \( X_{\text{min}} \), but increases the number of objective functions that must be solved. Typically \( \mu \) is 0.2 or less.
In most cases $X_{\text{min}}$ produces an unstable filter making the algorithm useless. To ensure a useful result from the SM algorithm the searches for $X_{\text{SM.min.k}}$, $k = 1, 2, \ldots$, must be constrained to the set $S_{ST}$ which contains all values of $X$ that produce a stable filter. Constraining the poles to be inside the unit circle, which is the ideal constraint, cannot be incorporated in the search algorithm. A constraint that can be incorporated into the search is Positive Realness (PR) [8], [10], which limits the magnitude response are obtained by modifying the set $X$ using the PR constraint may be sub-optimum.

Another implementable constraint that can be placed on the algorithm useless. To ensure a useful result from the search for the minimum is based on Rouche’s theorem [11]. This constraint limits the solution to set $S_{RT}$ which ensures a stable filter, but $S_{RT}$ is smaller in size than $S_{PR}$. Furthermore, it was shown in [8] that filters obtained with the PR constraint have less mean squared error than those obtained with Rouche’s method. Therefore the PR constraint is used with the SM method in this paper.

The issue with using the PR constraint is that solution can have sharp peaks in magnitude response creating pockets of large mean square error that is not accounted for in the sampling points used in $J(X)$. The optimization process places poles near the unit circle between two sampling points, e.g., $\omega_k$ and $\omega_{k+1}$. Solutions without these sharp peaks in the magnitude response are obtained by modifying the PR constraint to limit the largest radius of a pole to less than $1 - \eta$ in magnitude. This is done using the inequality $|\Re{[a(\omega; X)]} > \eta$, $\forall \omega \in [0, \pi]$, where a good value for $\eta$ is in the range of 0.01 to 0.05.

2) Gauss–Newton (GN) Method: The GN algorithm [8], [11], [12] is an iterative method based on the first-order Taylor approximation:

$$H(e^{j\omega}; X_k) \approx H(e^{j\omega}; X_{k-1}) + \nabla H(e^{j\omega}; X_{k-1}) \cdot \delta, \quad k = 1, 2, \ldots, (15)$$

where $X_0$ is the initial guess, $\delta = X_k - X_{k-1}$ and $\nabla H$ is the gradient vector of $H(e^{j\omega}; X)$ with respect to $X$. The vector $X_{k-1}$ contains the coefficients of a stable filter obtained from the $(k-1)$th iteration, where the $k$th iteration finds $\delta$ by solving

Minimize $J_{GN,k}(\delta) = \sum_{i=1}^{M} W_i |D(\omega) - H(e^{j\omega}; X_{k-1}) - \nabla H(e^{j\omega}; X_{k-1}) \cdot \delta|^2 \quad (16)$

Subject to $\delta + X_{k-1} \in S_{ST}$. (17)

The $k$th iteration coefficients are given by $X_k = X_{k-1} + \delta$.

Unfortunately, practical methods of constraining $X_{k-1} + \delta$ restricts the solution space to a subset of $S_{ST}$. Whether or not this subset contains the optimum solution or a near-optimum solution depends not only on the method of constraint but also on the way the filter’s system function is parameterized.

- **Parameterizing the filter’s transfer function in radii and angles form:**

  Similar to the approach introduced in [12], one can represent the transfer function in the form

  $$H(z) = \chi \prod_{i=1}^{p_r} \left(1 - r_i z^{-1}\right) \prod_{j=1}^{p_q} \left(1 - 2 r_j \cos(\phi_j) z^{-1} + r_j^2 z^{-2}\right) \prod_{j=1}^{p_p} \left(1 - 2 r_i \cos(\theta_i) z^{-1} + r_i^2 z^{-2}\right) \quad (18)$$

  where $\chi$ is the scaling factor. The scalars $R, Z, P, Q$ are the numbers of real zeros, complex zeros, real poles and complex poles, respectively. With the form of (18), a stable filter is ensured by constraining the pole radius, i.e., forcing $p_j, p_i \leq \rho$ where $\rho$ is a real positive number less than 1.

- **Parameterizing the filter’s transfer function in polynomial form:**

  $$H(z) = \frac{b_0 + b_1 z^{-1} + \ldots + b_m z^{-m}}{1 + a_1 z^{-1} + \ldots + a_n z^{-n}} \quad (19)$$

  With this representation, one can use the PR condition to force a stable solution. The PR constraint at the $k$th iteration limits $\delta$ to the set $\{\delta : |\Re{[a(\omega; X_k - 1 + \delta)]} > 0, \forall \omega \in [0, \pi]\}$. Another approach to constrain the pole radii is to exploit Rouche’s theorem [13], which states: if $f(z)$ and $g(z)$ are analytic inside and on a closed contour $C$, and $|g(z)| < |f(z)|$ on $C$, then $f(z) + g(z)$ has the same number of zeros inside $C$. Let the contour $C$ be a circle with radius $\rho$ centered at the origin of the complex plane. Then the denominator polynomial on the contour $C$ can be expressed as

  $$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_n z^{-n}$$

  $$+ a_n z^{-n} \bigg|_{z = \rho e^{j\omega}}, \quad \omega \in [0, \pi]. \quad (20)$$

  According to Rouche’s theorem, if the initial denominator polynomial $A_{k-1}(z) = 1 + a_{k-1,1} z^{-1} + a_{k-1,2} z^{-2} + \ldots + a_{k-1,n} z^{-n}$ has all of its zeros inside $C$, then the updated denominator polynomial, given as

  $$A_k(z) = A_{k-1}(z) + \mu \Delta(z) \quad (21)$$

  where $\Delta(z)$ is the function defined on domain $C$ as

  $$\Delta(z) = \delta_1 z^{-1} + \delta_2 z^{-2} + \ldots + \delta_n z^{-n} \bigg|_{z = \rho e^{j\omega}}, \quad \omega \in [0, \pi] \quad (22)$$

  with $\delta_i, i = 1, 2, \ldots, n$ are elements to update the denominator coefficients $a_i, i = 1, 2, \ldots, n$. The filter is stable if the update function (21) satisfies $|\Delta(z)| \leq |A_{k-1}(z)|, \forall \omega \in C$. Thus the optimization problem becomes:

  Minimize $J_{GN,k}(\delta) \quad (23)$

  Subject to $|\Delta(\rho e^{j\omega})| \leq |A_{k-1}(\rho e^{j\omega})|, \forall \omega \in [0, \pi] \quad (24)$

  which is solvable by the multiple exchange algorithm proposed in [11].

  While the PR and the RT conditions are more conservative than having a direct constraint on the poles’ radius, the polynomial representation of the transfer function $H(z)$ has a better convergence characteristic. In general, there is clearly
a tradeoff between parameterizations in radii-angles form and polynomial form.

III. NEW APPROACH FOR DESIGNING A STABLE SRRC FILTER

A. Hybrid Parametrization and the Stability Constraint

Conventionally the transfer functions of IIR filters are parameterized in either radii-and-angles form or polynomial form. In this section, the transfer function is expressed as a hybrid with the numerator in polynomial form and the denominator in radii-and-angles form. The hybrid form is given by

$$H(z) = \frac{\sum_{i=0}^{m} b_i z^{-i}}{\prod_{j=1}^{P} (1 - p_j z^{-1}) \prod_{i=P+1}^{P+Q} (1 - 2p_i \cos(\phi_i) z^{-1} + p_i^2 z^{-2})}.$$  (25)

The parameters of this transfer function are organized within the parameter vector $X$ as

$$X = [p_1, p_2, \ldots, p_{P+Q}, \phi_{P+1}, \phi_{P+2}, \ldots, \phi_{P+Q}, b_0, b_1, \ldots, b_m]^T.$$  (26)

The form in (25) allows the poles to be constrained which is a necessary and sufficient condition for the GN method to produce a stable filter. The $k^{th}$ iterative step still involves finding the $\delta$ that minimizes $J_{GN,k}(\delta)$, given by (16), but in this case subject to the constraint $p_i < \rho_i 1 \leq i \leq P + Q$.

B. Putting the Algorithm in Quadratic Form

Putting the algorithm in quadratic form begins by defining notation that simplifies the mathematical expressions involved. The gradient vector in the first-order Taylor approximation (15) is defined as

$$\nabla_e H(e^{jo}; X_{k-1}) = \begin{bmatrix} \frac{\partial H(e^{jo}; X_{k-1})}{\partial p_1}, & \ldots, & \frac{\partial H(e^{jo}; X_{k-1})}{\partial p_{P+Q}} \\ \frac{\partial H(e^{jo}; X_{k-1})}{\partial \phi_{P+1}}, & \ldots, & \frac{\partial H(e^{jo}; X_{k-1})}{\partial \phi_{P+Q}} \\ \frac{\partial H(e^{jo}; X_{k-1})}{\partial b_0}, & \ldots, & \frac{\partial H(e^{jo}; X_{k-1})}{\partial b_m} \end{bmatrix}. $$  (27)

The complex error term $D(o_i) - H(e^{jo}; X_{k-1})$ in (16) is defined by $E_{k-1,i}$. This allows $J_{GN,k}(\delta)$ to be expressed as

$$J_{GN,k}(\delta) = \sum_{i=1}^{M} W_i |E_{k-1,i} - \nabla_e H(e^{jo}; X_{k-1}) \cdot \delta|^2$$

$$= \sum_{i=1}^{M} W_i (E_{k-1,i} - \nabla_e H(e^{jo}; X_{k-1}) \cdot \delta)^* (E_{k-1,i} - \nabla_e H(e^{jo}; X_{k-1}) \cdot \delta)$$

$$= \sum_{i=1}^{M} W_i \left( E_{k-1,i}^* - \delta^T \cdot \nabla_e H(e^{jo}; X_{k-1})^* \right) \left( E_{k-1,i} - \nabla_e H(e^{jo}; X_{k-1}) \cdot \delta \right)$$

$$= \sum_{i=1}^{M} W_i \left( |E_{k-1,i}|^2 - E_{k-1,i}^* \cdot \delta^T \cdot \nabla_e H(e^{jo}; X_{k-1})^* \cdot \delta \right)$$

$$+ \delta^T \cdot \nabla_e H(e^{jo}; X_{k-1})^* \cdot \delta$$

$$+ \delta^T \cdot \nabla_e H(e^{jo}; X_{k-1}) \cdot \delta$$

$$= \sum_{i=1}^{M} W_i |E_{k-1,i}|^2$$

$$- \sum_{i=1}^{M} 2Re \{ W_i E_{k-1,i}^* \cdot \nabla_e H(e^{jo}; X_{k-1}) \} \cdot \delta$$

$$+ \delta^T \cdot \nabla_e H(e^{jo}; X_{k-1}) \cdot \delta.$$  (28)

where the superscript $\dagger$ denotes the complex transpose. To further simplify the expression, define $\Lambda_{k-1}$, $f_{k-1}$ and $c$ as follows

$$\Lambda_{k-1} = \sum_{i=1}^{M} \nabla_e H(e^{jo}; X_{k-1}) \cdot \nabla_e H(e^{jo}; X_{k-1})^* W_i \nabla_e H(e^{jo}; X_{k-1})$$

$$f_{k-1} = \sum_{i=1}^{M} 2Re \{ W_i E_{k-1,i}^* \cdot \nabla_e H(e^{jo}; X_{k-1}) \}$$

$$c = \sum_{i=1}^{M} W_i |E_{k-1,i}|^2.$$  (29)

It should be noted that $\Lambda_{k}$ is a Hermitian matrix, i.e., $\Lambda_{k-1} = \Lambda_{k-1}^*$, therefore $\delta^T \Lambda_{k-1} \delta = \delta^T A_{k-1} \delta$, where $A_{k-1} = 2Re\{ \Lambda_{k-1} \}$. Then substituting $A_{k-1}$, $f_{k-1}$ and $c$ into (28) produces

$$J_{GN,k}(\delta) = \delta^T A_{k-1} \delta + 2f_{k-1} \delta + c.$$  (30)

Finding the $\delta$ that minimizes $J_{GN,k}(\delta)$ is now a typical quadratic optimization problem that can be solved by a typical quadratic programming solver, like the solver in Matlab. Note that $c$ is a constant and does not affect the result.

C. Improving the Convergence Characteristic

Upon obtaining $\delta$, the new set of coefficients can be found as $X_k = X_{k-1} + \delta$. If $\delta$ is anything but infinitesimal, $X_k$ will likely be in error since a linearized model was used for the calculation. To ensure stability, the update equation is modified to be $X_k = X_{k-1} + \mu \delta$, where $\mu$ is a positive constant that is less than 1. Convergence is declared and the iteration process stopped as soon as $\delta$ is very small and there is no significant change from one iteration to the next.

However, since (30) needs to be optimized, it is convenient to directly limit the value of each element in $\delta$ at each step, i.e., incorporating a new constraint $|\delta_i| < \kappa_i$, $\forall i = 1, 2, \ldots, N$. Incorporating such a constraint eliminates overshoot if $\kappa < 1$ and also narrows the search domain of $\delta$, speeding up the search. Overall, the $k^{th}$ iteration of the optimization problem can be formulated as

$$\delta^T A_{k-1} \delta + 2f_{k-1} \delta$$  (31)
A. Comparison of FIR Designs

First an FIR square-root raised cosine filter that meets the out-of-band emission requirements for a DOCSIS CMTS transmitter is designed. The specific pulse shaping filter is for 256-QAM at a symbol rate of 5.36 Msym/sec, which requires a rolloff factor of 0.12. The most basic approach, which is a rectangular window applied to (2), requires a filter of order of 230 to meet the out-of-band emission requirement. Tapering the impulse response with a window function increases the stop band attenuation, which reduces the out-of-band emission, but increases the ISI. The ISI is compensated by windowing the ideal impulse response generated by (2), but with a reduced rolloff factor.

A design with a Kaiser window is examined by creating a filter with a rolloff factor equal to 0.05 and then applying a Kaiser window with shape factor \(\alpha = 3.5\) to increase the stop-band attenuation and satisfy the emission requirement. The best combination of rolloff factor and shape factor was found by experiment to be 0.05 and 3.5, respectively. This produces a filter of order 90 that has integrated powers of \(-58.122\) dB and \(-62.435\) dB in band OB1 and OB2, respectively.

A filter designed with Harris’ technique requires an order of 104 in order to meet the DOCSIS out-of-band emission requirements. Although this filter has a higher implementation cost than a Kaiser windowed filter, it provides significantly better ISI performance, as will be shown later.

Table I shows details of implementation cost among the three aforementioned windowed filters. The columns \(P_{OB1}\) and \(P_{OB2}\) present integrated powers in bands OB1 and OB2, respectively. Fig. 3 shows the pass-band and stop-band detail of the three filters. Thanks to the roll-off in its stop-band, the rectangular windowed filter has much less power in band OB2.

Table II in Section IV-B shows the ISI performance for many different combinations of pulse shaping and matched filters. In each case, the filter length used is the shortest which meets the DOCSIS spectral mask. Rectangular windowed, Kaiser windowed and Harris’ are among the filters compared. The convolution of a rectangular windowed impulse response with itself shows the least ISI, which is expected as its transition band is nearly ideal. In contrast, Kaiser’s and Harris’ filters sacrifice ISI in exchange for lower implementation costs. While Harris’ approach requires slightly more coefficients than the Kaiser window approach, the ISI power is significantly lower.

Measuring the ISI at the output of the matched filter is not necessarily a true indication of system performance since...
imperfect pulse shaping and matched filters [15], [16]. To provide some insight into the actual performance, the ISI power is also calculated at the output of a 24-tap symbol-spaced LMS equalizer that follows the matched filter. Such an equalizer may or may not be the type and size used in cable modems. The coefficients in \( g \) are chosen to minimize the mean squared error \( \| v \star g - \Delta_{K+8} \|^2 \), where \( \star \) denotes convolution operation and \( \Delta_{K+8} \) is a length \((N_v + 23)\) column vector where all elements are zeros except for the \((K + 8)\)th row, which has the value of 1. The optimum equalizer coefficients, denoted as \( g^{opt} \), therefore can be found by solving, (36) as shown at the bottom of this page, where \( V \) in (36) is a size \((N_v + 23) \times 24\) Toeplitz convolution matrix.

It is obvious that (36) is a conventional least squares problem that can be solved as

\[
g^{opt} = V^+ \Delta_{K+8}. \tag{37}
\]

where \( V^+ \) is Moore-Penrose pseudo-inverse matrix of \( V \) (which can be computed with Matlab function \( pinv() \)). It turns out the equalizer coefficients are the \((K+8)\)th row of \( V^+ \).

The ISI power after equalization can be approximated with high-precision by

\[
P_{ISI}^{(eq)} = \| Vg^{opt} - \Delta_{K+8} \|^2 \tag{38}
\]

The lower part of Table II shows the ratio of ISI power to signal power with the 24-tap symbol-spaced LMS equalizer in place for the three filters of discussion as well as others yet to be discussed.

Fig. 3. Pass-band and stop-band details of truncated, Kaiser windowed and Harris’ FIR filters.
TABLE II
ISI POWERS FOR DIFFERENT COMBINATIONS OF PULSE SHAPING FILTER AND MATCHED FILTER

<table>
<thead>
<tr>
<th>Window type (number of multiplies per sample)</th>
<th>Rectangular</th>
<th>Kaiser</th>
<th>harris</th>
<th>SMP</th>
<th>GNP</th>
<th>GNR</th>
<th>GNH</th>
<th>GNH2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISI before equalization ($P_{th}^{(dB)}$)</td>
<td>-55.9</td>
<td>-29.5</td>
<td>-53.3</td>
<td>-28.7</td>
<td>-24.3</td>
<td>-25.2</td>
<td>-35.4</td>
<td>-43.4</td>
</tr>
<tr>
<td>ISI after 24-taps equalization ($P_{th}^{(dB)}$)</td>
<td>-55.9</td>
<td>-29.5</td>
<td>-53.3</td>
<td>-28.7</td>
<td>-24.3</td>
<td>-25.2</td>
<td>-35.4</td>
<td>-43.4</td>
</tr>
</tbody>
</table>

TABLE III
IMPLEMENTATION COST, DESIGN PARAMETERS AND PERFORMANCE OF THE IIR DESIGNS

<table>
<thead>
<tr>
<th>IIR Design</th>
<th>Multiples</th>
<th>Design parameters</th>
<th>$P_{OB1}$</th>
<th>$P_{OB2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SM method with PR condition (SMP)</td>
<td>61</td>
<td>$\varphi = 34, m = 50, n = 10$, $\eta = 0.05, \mu = 0.2$, $W_{OB1} = 400, W_{OB2} = 800$</td>
<td>-58.6 dB</td>
<td>-61.4 dB</td>
</tr>
<tr>
<td>GN method with polynomial parameterization and the PR constraint (GNP)</td>
<td>51</td>
<td>$\varphi = 26, m = 40, n = 10$, $\eta = 0.05, \kappa = 1, \mu = 0.05$, $W_{OB1} = 500, W_{OB2} = 800$, $N = 200$</td>
<td>-58.1 dB</td>
<td>-62.7 dB</td>
</tr>
<tr>
<td>GN method with polynomial parameterization and the RT constraint (GNR)</td>
<td>41</td>
<td>$\varphi = 26, m = 40, n = 10$, $\rho = 0.985, \kappa = 1, \mu = 0.05$, $W_{OB1} = 200, W_{OB2} = 400$, $N = 1024$</td>
<td>-58.2 dB</td>
<td>-61.4 dB</td>
</tr>
<tr>
<td>Proposed GN method with hybrid parameterization (GNH)</td>
<td>34</td>
<td>$\varphi = 37, m = 16, P = 1, Q = 16$, $\rho = 0.985, \kappa = 0.2, \mu = 0.05$, $W_{OB1} = 150, W_{OB2} = 200$</td>
<td>-58.1 dB</td>
<td>-62.9 dB</td>
</tr>
<tr>
<td>2nd GNH design (GNH2)</td>
<td>41</td>
<td>$\varphi = 46, m = 24, P = 0, Q = 16$, $\rho = 0.985, \kappa = 0.2, \mu = 0.05$, $W_{P} = 6, W_{OB1} = 250, W_{OB2} = 400$</td>
<td>-58.4 dB</td>
<td>-60.9 dB</td>
</tr>
</tbody>
</table>

B. IIR Designs

First, a filter is obtained with the Steiglitz-McBride (SM) method, in which the objective function is minimized under the constraint of positive realness. This method can be applied to finding the coefficients for a filter that minimize the stop-band attenuation while maintaining the desired pass-band response. To find the lowest order filter requires a cut-and-try approach using different filter structures and parameters. The set of parameters that yielded a filter with the lowest multiplies to sample ratio is given in Table III, where the parameter $\varphi$ defines the phase delay of the target SRRC frequency response, $D(\omega)$, as in (1). The result was a filter with 50 zeros and 10 poles requiring 61 multiplies. The integrated powers in bands OB1 and OB2 are -58.6 dB and -61.4 dB, respectively. The frequency and phase characteristics of the filter is shown in Fig. 4.

Next, consider the GN method with radii-angle parameterization. This method does not converge well when the numerator’s order exceeds 12, and the method does not work at all with an order greater than 17. It is concluded that it is not possible to get a reasonable DOCSIS’s shaping filter with this method.

Then, the GN method with polynomial parameterization under the PR constraint is used to obtain a filter referred to as a GNP filter. The design parameters are specified in Table III. Fig. 5 provides results obtained from the proposed method. The algorithm takes 115 iterations to converge.

Finally, the GN method with polynomial parameterization and the RT constraint is used to obtain a filter referred to as a GNR filter. Its frequency domain characteristics are presented in Fig. 6. The constraint (24) is enforced with $N = 1024$ equally-spaced frequency points over $[0, \pi]$. Formulated with the multiple exchange algorithm, the GN method generates a very efficient filter requiring only 41 multiplies per sample. This is significantly less than any of the FIR filters. However, due to a conservative constraint, it does not yield a good result if the order of the denominator is increased further. The order is limited to $n = 10$ for both the GNP and GNR designs.

C. Improved Design With Hybrid Parameterization

The frequency response characteristics of a filter obtained using the proposed method are shown in Fig. 7. The numerator polynomial is order-16 and is initiated with the coefficients of
a truncated ideal FIR filter. The denominator contains 17 poles. Initially, all poles are set to have \( p_i = 0, \forall i \), whereas all angles, \( \phi_i, \forall i \), are equally distributed over the range \([-\pi, \pi]\). It appears that the algorithm cannot converge if all the angles are initially created equal. It is noticeable in Fig. 7 that the stop band of the proposed design has rolloff within it. This is a significant advantage for a shaping filter. Also, the least passband ripple is smallest among the IIR designs. The zeros and poles of the filter are given by

\[-9.7121, -0.9483, 0.3561 \pm 0.8812i, 0.5706 \pm 0.8010i, 0.6306 \pm 0.7722i, 0.9027 \pm 0.6585i, 1.0172 \pm 0.4873i, 1.1312 \pm 0.1005i, 1.0930 \pm 0.2989i\]

and

\[0.8879, 0.6404 \pm 0.7443i, 0.6495 \pm 0.6914i, 0.6885 \pm 0.6251i, 0.7326 \pm 0.5409i, 0.7829 \pm 0.4471i, 0.8272 \pm 0.3436i, 0.8611 \pm 0.2328i, 0.8817 \pm 0.1170i, \text{ respectively.}\]
Another example, which uses more multiplies/sample to achieve less ISI, is presented in Fig. 8. The filter is referred to as GNH2. Zeros of the filter are given by 
\[-0.9526, -0.8470 \pm 0.3714i, -0.5844 \pm 0.8261i, -0.0421 \pm 1.0101i, 0.3838 \pm 0.9284i, 0.5563 \pm 0.8282i, 0.6268 \pm 0.7774i, 0.8687 \pm 0.7000i, 0.9683 \pm 0.5800i, 1.0466 \pm 0.4452i, 1.1506 \pm 0.0000i, 1.1385 \pm 0.1517i, 1.1036 \pm 0.3013i \]
and the poles are \[0.6492 \pm 0.7233i, 0.6705 \pm 0.6564i, 0.7083 \pm 0.5834i, 0.7578 \pm 0.4918i, 0.8037 \pm 0.3925i, 0.8416 \pm 0.2859i, 0.8817 \pm 0.0581i, 0.8682 \pm 0.1737i.\]

Table II presents the ISI levels obtained when the pulse shaping and matched filters are any combination of
Fig. 8. 41-multiplies IIR filter designed by GN method with hybrid parameterization.

V. CONCLUSION

A properly designed IIR pulse shaping filter can satisfy the downstream DOCSIS 3.0’s out-of-band emission requirements, and yield sufficiently low ISI with better efficiency, i.e., fewer multiplies/sample, than linear-phase FIR filters, windowed or otherwise. It was demonstrated that IIR filters have a cost versus ISI trade off. The least expensive IIR filter is obtained by the proposed design algorithm, which considers a hybrid parameterization of the filter’s transfer function in combination with a constraint on maximum pole radius.

REFERENCES

Tung T. Nguyen received the B.Eng. degree from the Hanoi University of Technology, Hanoi, Vietnam, in 2008, and the M.Sc. degree from the University of Saskatchewan, Saskatoon, Canada, in 2013, where he is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering. His research interests involve digital signal processing, filters design, and channel estimation.

Joseph Eric Salt (M’88) received the B.Sc. degree from the University of Saskatchewan, Saskatoon, SK, Canada, in 1974, the M.Eng. degree from Carleton University, Ottawa, ON, Canada, in 1978, and the Ph.D. degree from the University of Saskatchewan, in 1987, all in electrical engineering. He joined the Department of Electrical Engineering and Computer Engineering, University of Saskatchewan, in 1985, where he currently holds the rank of Professor. His current research interests are in DSP algorithms and architectures for implementing CATV headend.

Ha H. Nguyen (M’01–SM’05) received the B.Eng. degree from the Hanoi University of Technology, Hanoi, Vietnam, in 1995, the M.Eng. degree from the Asian Institute of Technology, Bangkok, Thailand, in 1997, and the Ph.D. degree from the University of Manitoba, Winnipeg, MB, Canada, in 2001, all in electrical engineering. He joined the Department of Electrical and Computer Engineering, University of Saskatchewan, Saskatoon, SK, Canada, in 2001, and became a Full Professor in 2007. He has co-authored the textbook entitled *A First Course in Digital Communications* (Cambridge University Press). His research interests include diversity techniques in wireless communications, wireless relay networks, and error-control coding. He was an Associate Editor of the IEEE *Transactions on Wireless Communications* from 2007 to 2011. He currently serves as an Associate Editor for the IEEE *Transactions on Vehicular Technology* and the IEEE *Wireless Communications Letters*. He was a Technical Program Co-Chair of the Wireless Track, the IEEE Vehicular Technology Conference (Fall 2014, Vancouver, BC, Canada), the Multiple Antenna Systems and Space-Time Processing Track, the IEEE Vehicular Technology Conferences (Fall 2010, Ottawa, ON, Canada, and Fall 2012, Quebec, QC, Canada), and the 14th Canadian Workshop on Information Theory (2015, St. John’s, Canada). He is a Registered Member of the Association of Professional Engineers and Geoscientists of Saskatchewan.

Brian Berscheid received the B.Eng. and Ph.D. degrees from the University of Saskatchewan, Saskatoon, SK, Canada, in 2006 and 2011, respectively. Since 2006, he has been with Vecima Networks Inc., where he has been involved in the design of next-generation cable systems. His current research interests involve architectures for the efficient implementation of signal processing algorithms in FPGAs.