PROCEEDINGS OF THE
2001 FINNISH SIGNAL PROCESSING SYMPOSIUM

FINSIG'01

Helsinki University of Technology
Department of Electrical and Communications Engineering
Signal Processing Laboratory, and
Laboratory of Acoustics and Audio Signal Processing
Espoo, Finland
June 5, 2001

Edited by Jarno Tanskanen and Jarno Martikainen

ISBN: 951-22-5497-2

Report - Helsinki University of Technology.
Laboratory of Signal Processing and Computer Technology
Report 33

Distributor:
Helsinki University of Technology
Department of Electrical and Communications Engineering
Signal Processing Laboratory
P.O.Box 3000
FIN-02015 HUT
FINLAND
http://wooster.hut.fi/

ISBN: 951-22-5497-2 (CD-ROM Edition)
ISBN: 951-22-5496-4 (Electronic Edition, available from http://wooster.hut.fi/publications/finsig01/)
ISSN: 1456-6907

Chairmen's Message

FINSIG (The Finnish Signal Processing Symposium) is a one-day symposium on signal processing organized biannually. The meeting is well suited for post-graduate students working in different areas of DSP. FINSIG´01 is organized by the Signal Processing Laboratory and the Laboratory of Acoustics and Audio Signal Processing at the Helsinki University of Technology.

We received this year 18 papers from which 16 were accepted. The papers were divided into five oral sessions the topics of which are

- communications,
- signal processing,
- medical signal processing,
- digital filters and systems,
- DSP architectures and multimedia systems, and
- audio and image processing.

We would like to acknowledge all the individuals, who have contributed in organizing FINSIG'01. Thanks especially to all the speakers and authors, who have created the real contents of the symposium. The sponsors City of Espoo and IEEE Finland Section provided financial support that made the social program possible. Thanks also to our secretaries, Mrs. Anne Jääskeläinen and Mrs. Marja Leppäharju, for their invaluable assistance.

Iiro Hartimo
General Chair

Paavo Alku
Technical Program Chair

Jarno Tanskanen
Organization Chair

FINSIG'01 Organization

General Chair

Iiro Hartimo

Technical Program Chair

Paavo Alku

Organizing Chair

Jarno Tanskanen

Technical Program Committee

Jaakko Astola, Ramin Baghaie, Petri Haavisto,
Hannu Hakalahti, Iiro Hartimo, Jari Iinatti,
Markku Juntti, Visa Koivunen, Timo Laakso,
Jukka Lahti, Jorma Lilleberg, Aarne Mämmelä,
Markku Renfors, Juha Röning, Jukka Saarinen,
Harri Saarnisaari, Tapio Saramäki, Seppo Seikkala,
Olli Silven, Olli Simula, Vesa Välimäki

Organizing Committee

Paavo Alku, Jarno Tanskanen, Anne Jääskeläinen, Marja Leppäharju, Jarno Martikainen

The symposium is co-sponsored by

Signal Processing Laboratory
Helsinki University of Technology

Laboratory of Acoustics and Audio Signal Processing
Helsinki University of Technology

IEEE Finland Section

The City of Espoo

Table of Contents

AUDIO AND IMAGE PROCESSING

Author Index

Decimation by Non-Integer Factor using CIC Filter and Linear Interpolation

Djordje Babic[1†], *Jussi Vesma*[2], *and Markku Renfors*[1]

[1]Tampere University of Technology
Telecommunications Laboratory
P.O. Box 553, FIN-33101 Tampere
FINLAND
[†]Tel. +358-3-365 3910, Fax: +358-3-365 3808
[†]E-mail: babic@cs.tut.fi

[2]NOKIA GROUP
Nokia Research Center
P.O. Box 407 FIN-00045
FINLAND

ABSTRACT

Recently we have developed an efficient flexible multirate signal processing structure with high oversampling ratio and adjustable fractional or irrational sampling rate conversion factor. One application area is a multistandard communication receiver which should be adjustable for different symbol rates utilised in different systems. The proposed decimation filter consists of parallel CIC (cascaded integrator-comb) filters followed by a linear interpolation filter. The idea in this paper is to use two parallel CIC filters to calculate the two needed sample values for linear interpolation. In this paper we give a modification of the proposed structure and its control logic that enables better image and aliasing attenuation. The modification is based on the observation of the dependence of behaviour of the control logic on the fractional part of the sampling rate conversion factor.

1. INTRODUCTION

In multistandard receivers, the hardware should be configurable or programmable for the reception of different types of signals having different symbol rates. After the AD conversion, utilizing commonly the delta-sigma AD-conversion principle and high oversampling ratio, the sampling rate is reduced to be a low integer multiple of the symbol rate. In this decimation, the desired channel is preserved and other channels and noise are attenuated. The problem is that the needed decimation factor can be a difficult fractional number or even an irrational number and, for instance, FIR filters used for integer or fractional decimation cannot be efficiently utilized. Another problem is that there can be disturbing channels that are much stronger (e.g. 80-100 dB) than the desired channel. Therefore, the frequency bands which cause aliasing in decimation should have good attenuation. In addition to these requirements, the overall implementation should be simple because this decimation filter is used in the digital front-end of mobile receivers where the sampling rate is high [1], [2]. Based on these requirements (low complexity and possible irrational decimation factor), in [2] we have introduced a decimation filter structure which consists of two parallel CIC (cascaded integrator-comb) filters followed by linear interpolation. As it was shown this structure is easy to implement because the CIC filter does not need any multiplications and the linear interpolation requires only one multiplication. This structure has good anti-aliasing and anti-imaging properties.

In the general case, the decimation factor is a very difficult non-integer, thus the overall decimation factor is expressed as

$$R = \frac{F_{in}}{F_{out}} = R_{int} + \varepsilon, \qquad (1)$$

where $F_{in} = 1/T_{in}$ and $F_{out} = 1/T_{out}$ are the input and output sampling frequencies, whereas R_{int} is the integer part and ε is the decimal part of the overall decimation ratio. In [2] we have restricted discussion only for $\varepsilon \in [0,1)$. However, it was shown that sometimes it is better to use negative ε in order to increase aliasing band attenuation level. Therefore, in this paper we introduce modifications of the structure and control logic proposed in [2], in order to use the system for $\varepsilon \in (-1,0]$ as well. In that way characteristics of the proposed structure are improved, especially the worst case aliasing attenuation level.

2. BUILDING UNITS

Cascaded integrator-comb (CIC) filters are commonly used for decimation and interpolation by integer ratio providing efficient anti-image and anti-alias filtering [3]. These filters have a simple regular structure without multipliers. CIC decimation filter (see [3]) consists of N cascaded digital integrator stages operating at high input data rate F_{in}, followed by N cascaded comb or differentiator stages operating at low sampling rate F_{in}/R. Its frequency response is given by

$$H_{CIC}\left(e^{j\omega}\right) = e^{-j\omega N(R-1)/2}\left(\frac{\sin(\omega R/2)}{R\sin(\omega/2)}\right)^N, \qquad (2)$$

where $\omega = 2\pi f/F_{in}$ is the normalized input frequency.

When the decimation factor is an irrational number, the filters intended for integer or fractional decimation can not be directly used. One solution is to use polynomial-based interpolation filters. Among them, linear interpolation filter has a simple implementation structure, only one multiplication is needed [4]. Because interpolation is basically a reconstruction problem, polynomial-based interpolation can be analysed using the hybrid analog/digital model shown in Fig. 1, [4]. In this model, the interpolated output samples $y(l)$ are obtained by sampling the reconstructed signal $y_a(t)$ at the time instants $t = (n_l + \mu_l)\, T_{in}$. Here n_l is any integer, $\mu_l \in [0,1)$ is the adjustable fractional interval, and T_{in} is the sampling interval of the input signal $x(n)$.

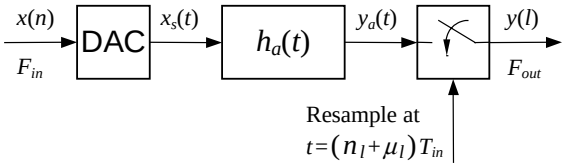


Fig. 1. The hybrid analog/digital model for the linear interpolation filter.

For linear interpolation, the impulse response of the reconstruction filter $h_a(t)$ is a triangular function, and thus, its frequency response is given by

$$H_a(f) = \left(\frac{\sin(\pi f / F_{in})}{\pi f / F_{in}} \right)^2 . \qquad (3)$$

The digital implementation of the linear interpolation, which needs only one multiplication, can be based on the following equation:

$$y(l) = x(n_l) + [x(n_l + 1) - x(n_l)]\mu_l . \qquad (4)$$

3. PROPOSED STRUCTURE FOR NON-INTEGER DECIMATION IN THE CASE OF $\varepsilon \in (-1,0]$

Figure 2 illustrates the proposed structure for the decimation filter. The input signal $x(n)$ is divided into polyphase components $x_k(m)$ for $k = 0, 1, \cdots, R_{int} - 1$ by using delay line and parallel CIC filters. Therefore, the sampling rate at the output of the CIC filters is F_{in}/R_{int}. The final decimation by $1+\varepsilon/R_{int}$ is done using linear interpolation between some of the two signal pairs $x_k(m)$ and $x_{k\oplus 1}(m)$, where $\oplus$ denotes the modulo R_{int} summation. The linear interpolation block in Fig. 2 is shifted by one branch according to some condition (to be discussed later on). Because of the modulus R_{int} summation mentioned above, the next signal pair for linear interpolation after $x_0(m)$ and $x_1(m)$ is $x_{R_{int}-1}(m)$ and $x_0(m)$. The fractional interval μ_l is recalculated for each output sample $y(l)$ for $l = 0, 1, 2, \cdots$. The time interval between samples $x_k(m)$ and $x_{k\oplus 1}(m)$ equals to T_{in} and, thus, the linear interpolation is done at the high input sampling frequency F_{in}. This means better image attenuation. The CIC filters attenuate the disturbing channels and noise which would cause

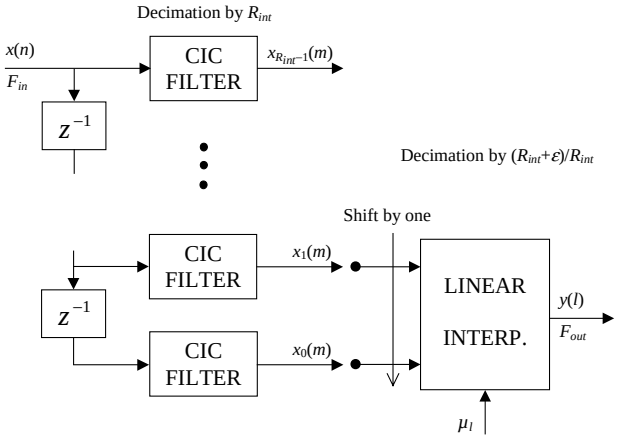


Fig. 2. Model of proposed decimation filter.

aliasing in linear interpolation. In other words, the CIC filters and linear interpolation take care of anti-aliasing and anti-imaging property, respectively. It should be pointed out that the filter structure of Fig. 2 is not the final implementation form. All the CIC filter branches are not needed and some of the blocks can be shared to make the final implementation easier, as will be discussed in Section 3.

As an example, Fig. 3 shows the input and output signals as well as some of the polyphase signals of the decimation filters for the decimation factor of $R = 3.9$. These polyphase signals $x_0(m)$ and $x_1(m)$ shown in Fig. 3(b) are obtained from $x(n)$ using a delay and two parallel CIC filters as shown in Fig. 2. Linear interpolation is then applied between these two signals to obtain the output samples $y(l) = y(lT_{out})$ for l–1, l, l+1 and l+2. After sample $y(l+1)$, the next output sample $y(l+2)$ falls outside the interval $x_0(m)$ and $x_1(m)$. When this occurs, the linear interpolation is shifted by one interval (as indicated by an arrow in Fig. 2) and the interpolation is done between signals $x_{R_{int}-1}(m)$ and $x_0(m)$.

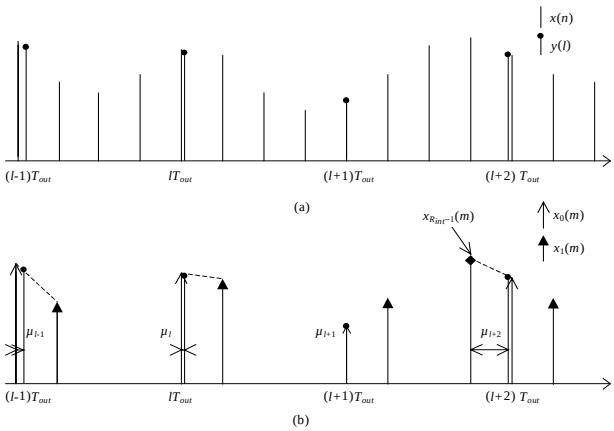


Fig. 3. (a) The input and output samples of the proposed decimation filter for $R = 3.9$. (b) The output samples of the two parallel CIC filter branches $x_0(m)$ and $x_1(m)$.

3.1. The frequency response of the overall system

The overall frequency response of the decimation filter

structure in Fig. 2 is a product of the frequency responses of the CIC filter and linear interpolation filter. Note that the former response is periodical whereas the latter is not. The frequency response of the parallel CIC filter stage is simply the same as the response of one CIC filter given by Eq. (2), where, however, R has to be replaced by R_{int}. Since the linear interpolation is done at the higher input rate F_{in}, its frequency response is given by Eq. (3). Consequently, the overall zero-phase frequency response of the proposed decimation filter, relative to the input sampling frequency, is given by

$$H_T(\omega) = H_{CIC}(\omega) H_a\left(\omega F_{in} / 2\pi\right) = \left(\frac{\sin\left(\frac{\omega R_{int}}{2}\right)}{R_{int}\sin\left(\frac{\omega}{2}\right)}\right)^N \left(\frac{\sin\left(\frac{\omega}{2}\right)}{\frac{\omega}{2}}\right)^2. \quad (5)$$

where $\omega = 2\pi f / F_{in} = 2\pi f / (RF_{out})$.

4. IMPLEMENTATIONS

The implementation structure in the case of negative ε, given in Fig. 4, is exactly the same as in the case of positive ε that is explained in [2], only the control logic is changed. However, here we shortly describe the implementation structure for the completeness of the paper. In the general case the number of the parallel CIC filters B, that is a number of comb filter branches, is given with $B=2+N$, where N is the order of the CIC filter. Two branches are used for calculating the output samples and the remaining N branches are used for initializing the state-variables of the branches needed later. However, the number of required comb branches can be reduced to the minimum. It is possible to use only $B=3$ branches in the comb section if following condition holds

$$|\varepsilon| \le 1/N. \quad (6)$$

The integrator stage is shared among the branches. The commutators COM1 and COM2 are used to select the correct input branch for the B comb sections and for linear interpolation, respectively.

As it was mentioned the control logic algorithm is different in the case of negative ε. Using analysis in time as in Fig. 3, one can notice that operations for $\varepsilon'>0$ and $\varepsilon<0$ are complementary, where $\varepsilon'=1+\varepsilon$. That means, there is shifting performed for the case of $\varepsilon<0$ whenever there is no shifting in the case of $\varepsilon'>0$. Using this observation the structure of the algorithm remains the same as in [2], however some changes are required. In Fig. 5(a) the control logic in the case of negative ε is given. The first step in this algorithm is the initial set up of the index value l as well as the fractional interval $\mu_0 = 0$. The next step is the interpolation which is expressed by

$$y(l) = I(\mu_l, u_0(m), u_1(m)), \quad (7)$$

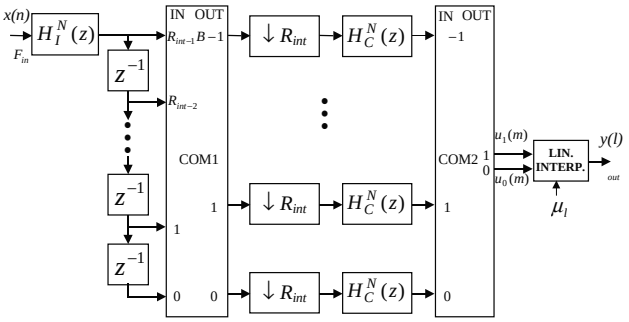


Fig. 4. Implementation structure for the proposed decimation filter.

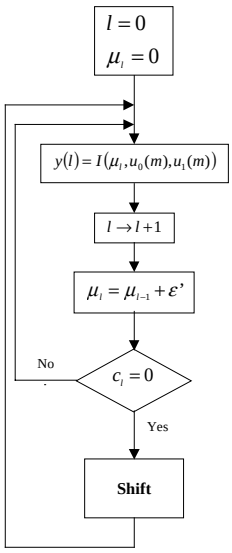
where $I(\ )$ denotes the linear interpolation between the samples $u_0(m)$ and $u_1(m)$ with the fractional interval of μ_l. After interpolation, l is incremented by one and the fractional interval can be computed by

$$\mu_l = \mu_{l-1} \oplus \varepsilon', \quad (8)$$

with the initial condition $\mu_0 = 0$, note that here we use complementary value ε' instead of ε and this is a main difference in the algorithm. In Eq. (8) the modulo summation indicates that only the decimal part of the result is used. According to Eq. (8), the calculation of μ_l can be implemented by using an adder with fixed point arithmetic. The shifting in the interpolation has to be performed whenever there is no overflow while calculating μ_l. Therefore, the overflow bit c_l of the adder can be used as a shifting condition. The shift block in Fig. 5(a) means that the interpolation is shifted by one branch (see Fig. 2). This shifting operation is implemented using the commutators COM1 and COM2 as it is shown in Fig. 5(b). The commutator COM1 has R_{int} inputs and B outputs. The commutator COM2 has B inputs and two outputs. In order to describe the function of the commutators we use variables for the outputs of the commutators. There are B variables for the outputs of COM1 denoted by OUT_i^1 for $i=0,1\ldots B-1$ and two variables for COM2 denoted by OUT_i^2 for $i=1$ and 2. The values of these variables determine what input sample is connected to the i^{th} output. The switching algorithm for COM1 and COM2 is given in Fig. 5(b). When shifting occurs, only one output of COM1, numbered by p, should be switched to the another input. Hence, only the value of the variable OUT_p^1 is changed. In COM2, when shifting occurs, both output branches should be switched to the another input. This is done because the order of the interpolator inputs must be preserved.

5. EXAMPLES

The bands that cause aliasing to the desired band are positioned around frequencies that are multiples of F_{out}. However the zeros of CIC filters are at the points which are multiples of RF_{out}/R_{int}. The minimum aliasing attenuation occurs at the edge of the first aliased band. Figure 6 shows the minimum attenuation of the aliasing



(*initial set up*)

$OUT_0^1 = 0$, $OUT_1^1 = 1$;

for i=2 to B-1

$OUT_i^1 = R_{int} - B + i$;

end

$OUT_0^2 = 0$, $OUT_1^2 = 1$;

p=1;

(*shifting*)

if $c_l = 0$ then

$OUT_p^1 = OUT_p^1 \oplus_{R_{int}} (-B)$;

$p = p \oplus_B (-1)$;

$OUT_0^2 = OUT_0^2 \oplus_B (-1)$;

$OUT_1^2 = OUT_1^2 \oplus_B (-1)$;

end.

(a) (b)

Fig. 5. (a) The state flow diagram of the control logic. (b) Algorithm for switching of COM1 and COM2. Here $\oplus_i$ denotes modulo i summation.

bands in a range of the fractional decimation factor ranging from 32 to 34 in the case of the third order CIC filter. As it can be seen the minimum attenuation of the aliasing bands depends on ε. As ε increases the minimum attenuation reduces, but this is avoided if we use positive ε algorithm when ε<0.5 and negative ε algorithm when ε>0.5, as explained above.

6. CONCLUSIONS

Since the whole structure requires only one multiplier and since it offers good anti-aliasing and anti-imaging properties, the proposed decimation filter is considered as power-efficient, relatively simple, and flexible solution for non-integer factor decimation in the multistandard radio receivers. We have shown that the negative ε algorithm can be implemented to the proposed structure. It was also shown that the negative ε algorithm allows us to use reduced number of comb branches in the actual implementation, and that means reduction in power consumption of the overall structure. Further, better aliasing attenuation is achieved using proposed negative ε algorithm for certain range of ε and positive ε algorithm for other values of ε.

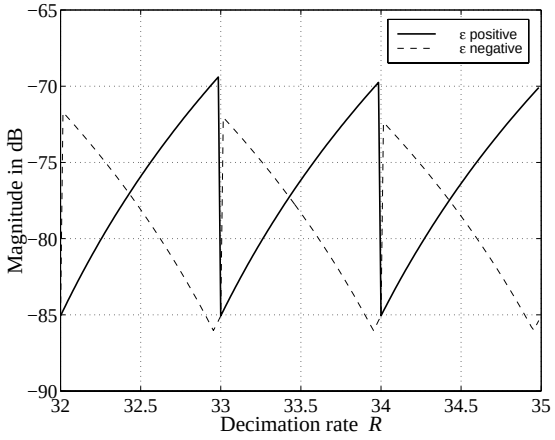


Fig. 6. The maximum value of the aliasing bands for positive and negative ε and the third order CIC filter.

ACKNOWLEDGMENT

This work has been supported by Graduate School TISE.

REFERENCES

[1] T. Hentschel and G. Fettweis, "Software radio receivers," Chapter 10 in *CDMA Techniques for Third Generation Mobile Systems*, edited by F. Swarts, P. van Rooyan, I. Oppermann, and M. P. Lötter, Kluwer Academic Publishers, 1999, pp. 255-283.

[2] D. Babic, J. Vesma, M. Renfors, "Decimation by irrational factor using CIC filter and linear interpolation," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing, ICASSP2001*, Salt Lake City, USA, 2001, in press.

[3] E. B. Hogenauer, "An economical class of digital filters for decimation and interpolation," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-29, pp. 155-162, April 1981.

[4] J. Vesma, *Optimization and Applications of Polynomial-Based Interpolation Filters*; Doctoral Thesis, Tampere University of Technology, Publications 254, 1999.

A COMPLEX ADAPTIVE NOTCH FILTER BASED ON THE STEIGLITZ-MCBRIDE METHOD

Yaohui Liu and Timo I. Laakso

Helsinki University of Technology
Signal Processing Laboratory
P.O.Box 3000, Fin-02150, Finland
email:{yaohui.liu,timo.laakso}@hut.Þ

Paulo S. R. Diniz

COPPE
Universidade Federal do Rio de Janeiro
Caixa Postal 68504 RJ, Brazil, 21945-970
email: diniz@lps.ufrj.br

ABSTRACT

This paper propose a new complex adaptive notch Þlter (ANF) structure based on the Steiglitz-McBride (SM) method. Recursive least square (RLS) algorithm is applied to the proposed ANF with optimized stepsize. Simulations show that RLS-SM ANF converges fast and requires less computational complexity than the conventional ANF using recursive prediction error (RPE) algorithm.

1. INTRODUCTION

Adaptive notch Þlters (ANF) are widely used in many signal processing applications to extract, eliminate or trace narrowband or sinusoidal signals embedded in broadband noise [2]. If such signal consists of in-phase and quadrature components, a complex coefÞcient ANF must be implemented. Most of such applications are in radar and communication systems. An early contribution to ANF algorithms by Nehorai [3] imposed constraints on a notch transfer function, which leads to simple relations between poles and zeros in adaptive Þlter design. Nehorai also derived the Gauss-Newton type recursive prediction error (RPE) algorithm [3] whose structure is shown in Fig. 1. The algorithm adjusts the Þlter coefÞcients to minimize the cost function $E\{|e(k)^2|\}$ by calculating the gradient recursively. Based on the same objective function, Pei [4] extended the RPE algorithm to complex coefÞcient ANF, which converges to a small biased solution. Cheng [1] derived a new real-valued ANF algorithm using the well-known SM method. ChengÕs idea comes from the system identiÞcation application by using delayed signal as the reference signal [7]. In this paper, we extend the result of [1] and derive complex coefÞcient adaptive notch Þlter algorithm using SM method. Furthermore, optimized stepsize is employed in our algorithm. Simulation results show that the complex SM method converge faster than RPE algorithms in [4].

The paper is organized as follows. In Section 2, Complex ANF algorithms using the SM method are derived. Section 3 analyzes the ANF convergence. In Section 4, simulation results show the improved performance of ANF using SM method. Finally, Section 6 concludes the paper.

2. SYSTEM MODEL

Consider a measured stationary data $y(k)$ which comprises a known number of complex sinusoids and a white noise $\epsilon(k)$,

$$y(k) = \sum_{i=1}^{M} R_i exp(j\omega_i k + \phi) + \epsilon(k) \qquad (1)$$

where the amplitudes $\{R_i\}$, phases $\{\phi_i\}$ and the frequencies $\{\omega_(k)\}$ are unknown constants. $\epsilon(k)$ is a sequence of i.i.d. complex random variable with zero mean and variance denoted by σ^2. It is known that (1) can be represented by an ARMA model [6],

$$A(q^{-1})y(k) = A(\rho q^{-1})\epsilon(k) \qquad (2)$$

where $A(q^{-1})$ is a monic polynomial of order M and its roots are on the unit circle with arguments equal to $\{\omega_i\}$. The parameter $\rho \in (0, 1)$ is a pole radius which keeps the Þlter $A(q^{-1})/A(\rho q^{-1})$ stable. Such Þlter is also known as constrained form notch Þlter.

3. COMPLEX SM ANF ALGORITHMS

The idea of ANF algorithm using the SM method comes from the system identiÞcation application by using delayed signal as the reference signal [7]. The resulting block diagram is depicted in Fig. 2. The function of the delay factor Δ in the Þgure is to decorrelate the the preÞlter outputs $g(k)$ and $h(k)$ in the upper and lower paths. If the noise is white,$\Delta = 1$ is enough to decorrelate the signals. By letting the structure to approximate a notch Þlter, the structure shown in Fig. 3 is obtained.

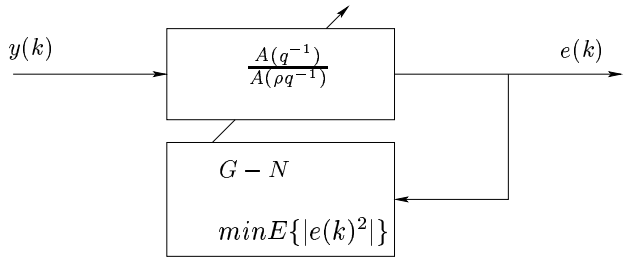


Figure 1: Adaptive notch Þlter with recursive prediction error algorithm for coefÞcient adjustment

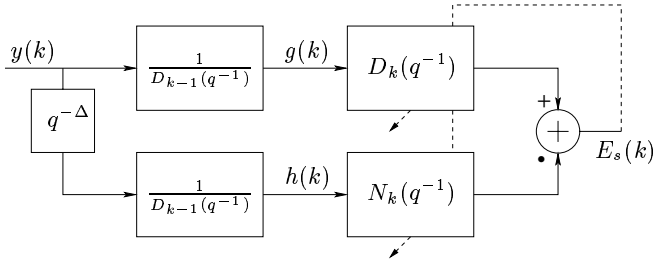


Figure 2: Adaptive notch Þlter based on SM method

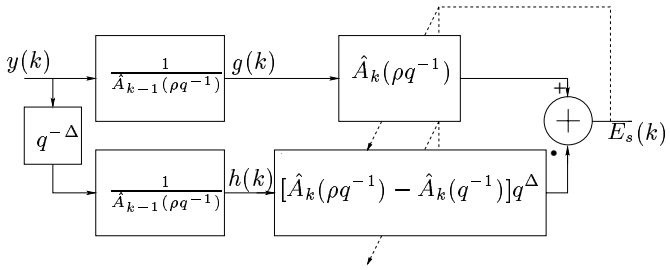


Figure 3: Block diagram of ANF using SM method in adaptive line enhancer structure

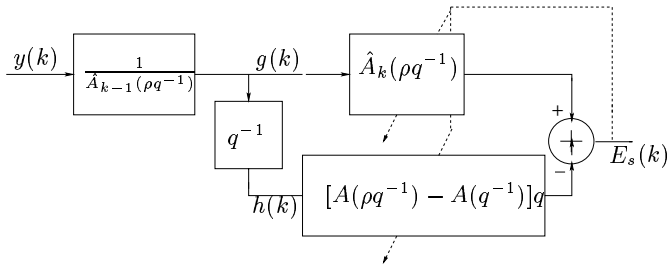


Figure 4: Block diagram of modiÞed ANF using SM method

3.1. Simplified ANF structure

Assuming the broadband signal is stationary, we can move the delay operation at the lower branch after the preÞlter shown in Fig. 2, such arrangement can save one preÞlter block. Larger Δ can be chosen in other applications when the noise is colored. Since the resulting transfer function after the convergence is desired to be a notch Þlter, the following equation should be satisÞed:

$$\lim_{k\to\infty}\left(1-q^{-1}\frac{N_k(q^{-1})}{D_k(q^{-1})}\right)=\frac{A(q^{-1})}{A(\rho q^{-1})} \tag{3}$$

This yields the block diagram shown in Fig 4. Therefore the polynomials $D_k(\rho q^{-1})$ and $N_k(q^{-1})$ in modiÞed Fig. 2 can be deÞned as

$$\begin{aligned} D_k(q^{-1}) &= A(\rho q^{-1}) \\ N_k(q^{-1}) &= [A(\rho q^{-1})-A(q^{-1})]q \end{aligned} \tag{4}$$

3.2. Algorithm derivation

The adaptive algorithm can be derived directly from Fig 4. Let the estimated coefÞcient vector $\mathbf{\Theta}_{k-1} = [a_1, a_2, \cdots, a_M]^T_{k-1}$ where the superscript T denotes the transpose operation. Using the recursive least square (RLS) procedure, we derive the detailed algorithm as below.

Step1: PreÞlter

$$g(k)=\frac{1}{\hat{A}_{k-1}(\rho q^{-1})}y(k) \tag{5}$$

where $\hat{A}_{k-1}(\rho q^{-1}) = 1+a^*_{1,k-1}\rho q^{-1}+a^*_{2,k-1}\rho^2 q^{-2}+\cdots+a^*_{M,k-1}\rho^M q^{-M}$. Rearranging the input-output, we obtain,

$$g(k)=y(k)-\mathbf{\Theta}^H_{k-1}\mathbf{G}_k \tag{6}$$

where the superscript H denotes conjugate transpose, and $\mathbf{G}_k = [\rho g(k-1), \rho^2 g(k-2), \cdots, \rho^M g(k-M)]^T$. Since the preÞlter output for lower branch $h(k) = g(k-1)$ is a delayed version of $g(k)$, one preÞlter can be saved.

Step 2: Output expression
The output can also be arranged in vector form,

$$\begin{aligned} \epsilon(k) &= g(k)\hat{A}_k(q^{-1}) - h(k+1)[\hat{A}_k(\rho q^{-1}) - \hat{A}_k(q^{-1})] \\ &= g(k) - \mathbf{\Theta}^H_{k-1}\mathbf{\Phi}(k) \end{aligned} \tag{7}$$

where $\mathbf{\Phi}(k) = [\phi_1(k), \phi_2(k), \cdots, \phi_M(k)]^T$ and $\phi_i(k) = -\rho^i g(k-i) + (\rho^i - 1)h(k-i+1)$.
Step 3: Covariance matrix update

$$\mathbf{P}(k+1)=\frac{1}{\lambda(k)}\left[\mathbf{P}(k)-\frac{\mathbf{P}(k)\mathbf{\Phi}(k)\mathbf{\Phi}^H(k)\mathbf{P}(k)}{\frac{\lambda(k)}{\alpha(k)}+\mathbf{\Phi}^H(k)\mathbf{P}(k)\mathbf{\Phi}(k)}\right] \tag{8}$$

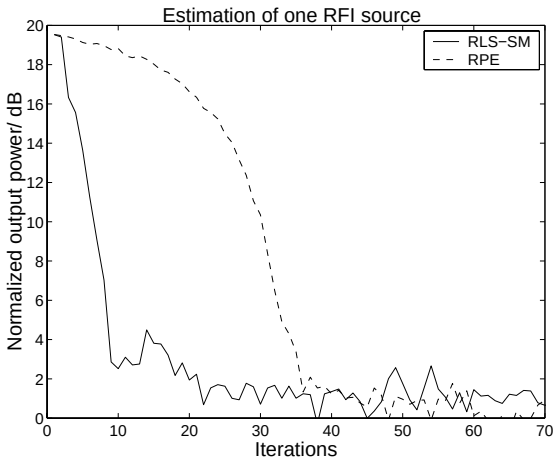


Figure 5: The output MSE when estimating 1 sinusoid

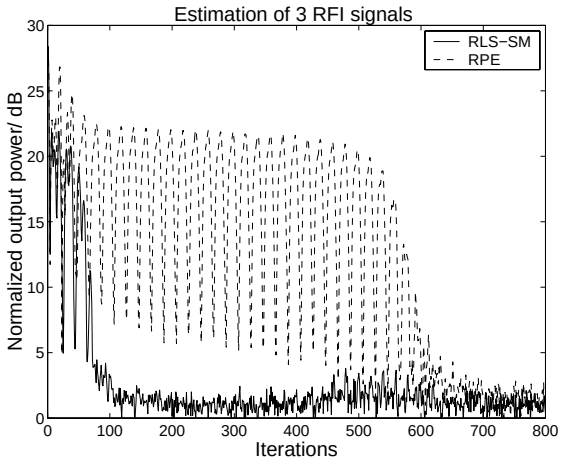


Figure 6: the output MSE when estimating 3 sinusoids(Þxed α and ρ for RPE)

where $\lambda(k) = 1 - \alpha(k)$ is the forgetting factor in the RLS algorithm.

Step 4: Estimation parameter update

$$\mathbf{\Theta}(k+1) = \mathbf{\Theta}(k) + \alpha(k)\mathbf{P}_{k+1}\mathbf{\Phi}(k)\epsilon(k)^* \qquad (9)$$

The algorithm is in the RLS form. The differences between the RPE algorithm and our RLS algorithm using the SM methods are on the choice of regression vector $\mathbf{\Phi}(k)$ and error $\epsilon(k)$. A better choice of these parameters can lead to faster convergence and less excess mean square error (MSE) at the output.

4. CONVERGENCE CONSIDERATIONS

In both RPE and RLS-SM algorithms, the convergence speed and the excess MSE depends on two parameters: the pole radius ρ and the stepsize α.

4.1. Time-varying ρ

In most of the ANF algorithms the pole radius is a time-varying function [4, 1]. The reason is that ρ determines the bandwidth of the notches. Practically, if no *a priori* information is available on the input sinusoid, when the notches are too narrow, the algorithm may not converge. On the other hand, a larger pole radius ($\rho \to 1$) will lead to less excess MSE after convergence. Therefore an exponential function is often used for $\rho \to 1$ by letting ρ grow from an initial value $\rho(1)$ to the desired value $\rho(\infty)$ according to

$$\rho(k+1) = \rho_0\rho(k) + (1-\rho_0)\rho(\infty) \qquad (10)$$

where ρ_0 determines the rate of change in $\rho(k)$.

4.2. Optimal α for SM method

In the algorithms derived by Pei [4] and Cheng [1], the stepsize is treated in the same way as ρ, which approaches exponentially the predeÞned value. Without*a priori* knowledge of the input, the choice of α is usually a difÞcult task. In this paper, we apply the optimal stepsize derived in [5] for IIR Þlters using the SM method. An optimal stepsize puts a proper weight on the new incoming data at each updating step, which will lead to the maximal reduction of MSE, thus speeding up the convergence. The optimal α has the form

$$\alpha(k) = \frac{\kappa}{1+\tau(k)} \qquad (11)$$

where $0 < \kappa < 1$ is a reduction factor which is related to the Þlter order and $\tau(k) = \mathbf{\Phi}^H(k)\mathbf{P}(k)\mathbf{\Phi}(k)$. Note that $\tau(k)$ is an intermediate result of (8), so that Þnding the optimal convergence factor does not increase the complexity of the algorithm.

5. SIMULATIONS

We apply the proposed complex RLS-SM ANF and RPE algorithm to suppress sinusoid in white noise. In all the following experiments, The pole radius is time-varying according to (10), where $\rho_0 = 0.99$, $\rho(1) = 0.7$ and $\rho(\infty) = 0.995$. We use optimal stepsize derived in (11) for RLS-SM algorithms, whereas the optimal stepsize is used in RPE algorithm. The input signal is modelled as in (1). The sinusoid and white noise are chosen such that the noise is 20 dB below the sinusoidal level. The output signal is normalized with respect to the white noise power, in other words, 0 dB output is the

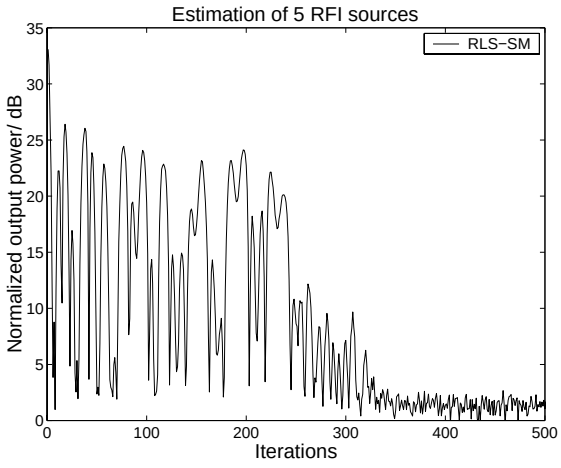


Figure 7: The output MSE when estimating 5 sinusoids using RLS-SM algorithm

best suppression result that we can achieve. The simulation results are averaged over 100 independent runs.

5.1. The first-order notch filter

There is one complex sinusoid signals embedded in white noise, whose frequency is $\omega_1 = 0.015$. The Þlter output MSE is shown in Fig. 5. It can be seen that using RLS-SM algorithm leads to the optimal solutions in ca.15 iterations, whereas the RPE algorithm requires ca.40 iterations to converge.

5.2. The third-order notch filter

This experiment is for the case of three sinusoids where their frequencies are $\omega_1 = 0.1$, $\omega_2 = 0.2$ and $\omega_3 = 0.4$. In this case, RPE algorithm with time-varying pole radius can not converge, therefore ρ is set Þxed at 0.8. As can be seen in Fig. 6, the RLS-SM algorithm works very well, whereas the RPE algorithm converges much more slowly and generates higher excess MSE.

5.3. Estimation of 5 sinusoids using RLS-SM algorithm

This is an extreme case that there are 5 sinusoids exist. where their frequencies are $\omega_1 = 0.1$, $\omega_2 = 0.2$, $\omega_3 = 0.25$, $\omega_4 = 0.3$ and $\omega_5 = 0.4$ respectively. Since this is a difÞcult situation for ANF to converge, we loose the criteria on extra MSE and let $\rho(\infty) = 0.95$. As can be seen in Fig. 7, the algorithm converges in ca. 400 iterations, whereas RPE algorithm fails to converge within 1024 iterations, even when the optimal stepsize is utilized.

6. CONCLUSION

In this paper, the ANF using SM method proposed by Cheng [1] is extended to the complex-coefÞcient case. We propose a simpliÞed structure by relocating the delay elements on one branch of the Þlter assuming the broadband process is white and stationary. Simulations show that the RLS-SM algorithm converges faster than RPE algorithm when suppressing sinusoid embedded in white Gaussian noises. Our algorithm is more robust compared with the RPE algorithm since it can deal with up to 5 sinusoids. Furthermore, optimized stepsize developed in [5] for RLS-SM algorithms is also employed to speed up the convergence.

7. REFERENCES

[1] M. H. Cheng and J. L. Tsai, ÒA new adaptive notch Þlter with constrained poles and zeros using steiglitz-mcbride method,Ó in*Proceedings of ICASSP98*, Seattle, Washington, USA, May 12-15 1998, pp. 1469Ð1472.

[2] L. Ljung and T. Sderstrm, *Theory and Practice of Recursive Identification*, MIT Press, Cambridge, 1983.

[3] A. Nehorai, ÒA minimal parameter adaptive notch Þlter with constrained poles and zeros,Ó*IEEE Trans.on ASSP*, vol. ASSP-33, no. 4, pp. 983Ð996, 1985.

[4] S. C. Pei and C. Tseng, ÒComplex adaptive iir notch Þlter algorithm and its applications,Ó*IEEE Trans. on Circuits and Systems-II: Analog and Digital Signal Processing*, vol. 41, no. 2, pp. 158Ð163, 1994.

[5] J. E. Cousseau and P. S. R. Diniz, ÒOn optimal convergence factor for iir adaptive Þlters,Ó in*Proceedings of ISCAS '94*, Tokyo, Japan, October 1994, pp. 137Ð140.

[6] J. A. C. Bingham, *ADSL, VDSL, and Multicarrier Modulation*, Wiley, 2000.

[7] P. S. R. Diniz, *Adaptive filtering: algorithm and practical implementation*, Kluwer, 1997.

Flat Systems in Discrete Signal Processing

Markku T. Nihtilä

University of Kuopio
Department of Computer Science and Applied Mathematics
P.O.Box 1627, FIN-70211 Kuopio
FINLAND

Tel. +358-17-162571, Fax: +358-17-162595

E-mail: Markku.Nihtila@uku.fi

ABSTRACT

The concept of flatness is here introduced for dynamic discrete-time systems analogously to the flatness of continuous-time systems. This concept gives a way for open-loop as well as closed-loop control design for dynamic systems when the goal is to drive the system from one steady-state to another. The successive derivatives of the so-called flat output and the control of a continuous-time system are substituted by their backward shifts in discrete approach. Some flatness based properties are preliminarily studied via a linear example. Relations to dead-beat control are also pointed out

1. INTRODUCTION

In many dynamical control and signal processing systems an intermediate goal is to drive the output of a system from one steady-state to another as quickly as possible. The recently coined and studied concept of flatness of nonlinear as well as linear differential equations and systems points out a way for straightforward open-loop control design. Differential flatness has been developed in the works of Michel Fliess and his co-workers in France, see, *e.g.*, [7], [8], [10], [12]. It has its origin in the beginning of 1900s in the studies of Elie Cartan on underdetemined differential systems, i.e. on the sets of differential equations (without control) having a lesser number of equations than variables. Flatness issues have also been studied from another viewpoint by utilizing differential forms and exterior algebras, *c.f.* Murray and co-workers in [20], and [21], again originating in the work of Cartan.

If the differential system is flat, its input and state can be expressed as functions of another variable, called a flat output, and of a finite number of its time derivatives, and *vice versa* this another variable can be respresented as a function of the state and control and of devivatives of the control. Then, for control design, starting from a desired flat output, one can construct the actual state (and output) and the open-loop control, which produces the output. This can be done easily just by differentiating sufficiently many times the flat output and by using the known, system dependent functions, see some design examples in [11].

Definition (Differential flatness) [10]. Consider a nonlinear ordinary differential system

$$\frac{dx}{dt} = f(x,u); \quad x(t) \in \mathbb{R}^n, \quad u(t) \in \mathbb{R}^m \tag{1}$$

where x and u denote the state and the control of the system, respectively. If there exist algebraic functions $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ and finite integers α, β, and γ such that for any (sufficiently differentiable) pair (x,u) satisfying the dynamics (1) there exists a vector-valued sufficiently differentiable function z ($z(t) \in \mathbb{R}^m$) satisfying

$$\begin{aligned} x &= \mathcal{A}(z, \dot{z}, \ldots, z^{(\alpha)}) \\ u &= \mathcal{B}(z, \dot{z}, \ldots, z^{(\beta)}) \\ z &= \mathcal{C}(x, u, \dot{u} \ldots, u^{(\gamma)}) \end{aligned} \tag{2}$$

then the system (1) is called *differentially flat* and the variable z is called a *flat, or linearizing output.*

Remark. It has to be noted that the flatness concept actually does not include the output, say y, at all. Then, in fact, the inclusion of the output equation for the considerations is, in principle, unnecessary. However, in any practical control problem there is an output to be controlled.

In this paper the discrete-time flatness is introduced according to Fliess & Marquez [13], which is based on the original study of Fliess [6]. Here it is demonstrated that minimal linear state-variable representations describe flat systems. A scalar example is given to illustrate open-loop control design based on flatness. Corresponding feedback control and nonlinear problems are discussed, too.

2. DIFFERENCE FLATNESS

There are two possibilities to extend the concept of flatness to discrete-time systems by using a definition analogous to that of differential flatness. The derivatives can be substituted by forward shifts or backward

shifts, *i.e.* the term, say, $z(t+i) = q^i z(t)$ substitutes the derivative $\frac{d^i z(t)}{dt^i}$ or $z(t-i) = q^{-i} z(t)$ substitutes $\frac{d^i z(t)}{dt^i}$. Forward shifts were discussed, however, without further explicit constructions in Aranda-Bricaire *et al.* [1], p. 2016. Here we apply the following backward shift definition.

Definition (Difference flatness). Consider a nonlinear ordinary difference system

$$x(t+1) = f(x(t), u(t)); \quad x(t) \in \mathbb{R}^n, \quad u(t) \in \mathbb{R}^m \qquad (3)$$

If there exist algebraic functions $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ and finite integers α, β, and γ such that for any pair (x, u) satisfying the dynamics (3) there exists a vector-valued function z ($z(t) \in \mathbb{R}^m$) satisfying

$$\begin{aligned} x &= \mathcal{A}(z, q^{-1}z \ldots, q^{-\alpha}z) \qquad (4)\\ u &= \mathcal{B}(z, q^{-1}z, \ldots, q^{-\beta}z)\\ z &= \mathcal{C}(x, u, q^{-1}u, \ldots, q^{-\gamma}u) \end{aligned}$$

then the system (3) is called *differencely flat* and the variable z is called a *flat, or linearizing output.*

2.1. Flatness in Linear SISO-Systems

Here we study a class of linear single-input - single-output (SISO) systems, which are controllable and observable, and which have the polynomial representation of the form ($a_n \neq 0$)

$$\begin{aligned} A(q^{-1})y(t) &= B(q^{-1})u(t) \qquad (5)\\ A(q^{-1}) &= 1 + a_1 q^{-1} + a_2 q^{-2} + \ldots + a_n q^{-n}\\ B(q^{-1}) &= b_1 q^{-1} + b_2 q^{-2} + \ldots + b_n q^{-n} \end{aligned}$$

Theorem 1. Linear difference systems of the form (5), where $A(q^{-1})$ and $B(q^{-1})$ are coprime, are differencely flat when represented in the controllable and observable form

$$\begin{aligned} x(t+1) &= Fx(t) + Gu(t) \qquad (6)\\ y(t) &= Cx(t). \qquad (7) \end{aligned}$$

A flat output is defined, see [13], by

$$z(t) = S(q^{-1})y(t) + R(q^{-1})u(t). \qquad (8)$$

where S and R satisfy, due to coprimeness of A and B, Bezout's equation

$$R(q^{-1})A(q^{-1}) + S(q^{-1})B(q^{-1}) = 1. \qquad (9)$$

Furthermore, the flat output gives

$$\begin{aligned} u(t) &= A(q^{-1})z(t) \qquad (10)\\ y(t) &= B(q^{-1})z(t) \qquad (11) \end{aligned}$$

The proof is omitted.

Remark. For a practical trajectory design the input-output description with its flatness defining equations (8)-(11) are a feasible way to proceed instead of using the state variable representation.

3. FLATNESS IN LINEAR MULTIVARIABLE SYSTEMS

Controllable and observable linear multivariable systems has two equivalent polynomial matrix fraction representations, see [5], p. 599. The left coprime fraction representation resembles the representation of a SISO system. The input-output system having the control $u(t) \in \mathbb{R}^m$, and the output $y(t) \in \mathbb{R}^k$ has a representation of the form

$$A(q^{-1})y(t) = B(q^{-1})u(t)$$

For the present author this representation did not open the way for flatness.

The *right coprime fraction representation* of the system transfer matrix is of the form ($D_n \neq 0$)

$$\begin{aligned} T(q^{-1}) &= N(q^{-1})[D(q^{-1})]^{-1} \qquad (12)\\ D(q^{-1}) &= I + D_1 q^{-1} + \ldots + D_n q^{-n} \qquad (13)\\ N(q^{-1}) &= N_1 q^{-1} + N_2 q^{-2} + \ldots + N_n q^{-n} \qquad (14) \end{aligned}$$

where the matrix coefficients are $D_i \in \mathbb{R}^{m\times m}$ and $N_i \in \mathbb{R}^{k\times m}$. In other words the shift polynomial matrices are $D(q^{-1}) \in \mathbb{R}(q^{-1})^{m\times m}$ and $N(q^{-1}) \in \mathbb{R}(q^{-1})^{k\times m}$.

This representation gives an obvious way to define a candidate for a flat output ($z(t) \in \mathbb{R}^m$).

$$z(t) = S(q^{-1})y(t) + R(q^{-1})u(t). \qquad (15)$$

where S and R satisfy, due to coprimeness of A and B, Bezout's matrix equation

$$\underbrace{R(q^{-1})}_{m\times m}\underbrace{D(q^{-1})}_{m\times m} + \underbrace{S(q^{-1})}_{m\times k}\underbrace{N(q^{-1})}_{k\times m} = I_{m\times m}, \qquad (16)$$

where the new shift polynomial matrices are $R(q^{-1}) \in \mathbb{R}(q^{-1})^{m\times m}$ and $S(q^{-1}) \in \mathbb{R}(q^{-1})^{m\times k}$. Without going into details it can be shown, by keeping in mind the matrix nature of the polynomial matrices, that

$$\begin{aligned} u(t) &= D(q^{-1})z(t) \qquad (17)\\ y(t) &= N(q^{-1})z(t). \qquad (18) \end{aligned}$$

Consequently, the equations (15), (17), and (18) can be used for the trajectory design for the output y by starting from a feasibly chosen flat output z.

It seems to be the case that a proper rational transfer matrix having an irreducible right coprime fraction representation (12) is differencely flat. So, we present the following result in the form of a conjecture, because all the details have not yet been verified.

Conjecture. The multivariable input-output system presented in an irreducible right coprime fraction form

$$y(t) = N(q^{-1})D(q^{-1})^{-1}u(t) \tag{19}$$

is differencely flat when represented in a nonreducible (minimal) observable and controllable state variable form

$$\begin{aligned} x(t+1) &= Fx(t) + Gu(t) && (20)\\ y(t) &= Cx(t) && (21) \end{aligned}$$

Outline of Proof. The equation (17) is clear. From the equation (15) the output y and its delayed values can be eliminated by using the time-reversed state equation. This results in the desired form

$$z(t) = Vx(t) + W(q^{-1})u(t), \tag{22}$$

where $V \in \mathbb{R}^{m\times n}$, $x(t) \in \mathbb{R}^n$, and $W(q^{-1}) \in \mathbb{R}(q^{-1})^{m\times m}$. It seems that the idea of constructing for each row $T_i(q^{-1})$, corresponding to each output y_i, of the transfer matrix

$$T(q^{-1}) = \begin{bmatrix} T_1(q^{-1}) \\ T_2(q^{-1}) \\ \vdots \\ T_k(q^{-1}) \end{bmatrix}$$

a state variable representation, with appropriate subsystem matrices F_i, G_i, and C_i, of the form

$$\begin{aligned} x^i(t+1) &= F_i x^i(t) + G_i u(t) \\ y_i(t) &= C_i x^i(t) \end{aligned}$$

according to Chen [5], p. 265, is fruitful. Then along the same lines as in the SISO case one could construct for the state components $x^i_j(t)$ representations of the form

$$x^i_j(t) = \Lambda^i_j(q^{-1})z(t).$$

which then form a long vector related to the flat output z by

$$x(t) = \Lambda(q^{-1})z(t).$$

4. NONLINEAR SYSTEMS

Local equivalence of a nonlinear system with a linear one has in continuous-time framework studied by using a so-called Brunowský canonical form. It is simply a set of subsequent integrators driven by a single input. The number of the integrator sets is equal to the dimension of the control. The numbers of the integrators in each chain are so-called controllability indices, *c.f.* Chen [5], p. 190. In discrete-time systems the canonical form corresponding to the Brunowský one is called a *prime system*. It is a set of forward shift lines, each of which is driven by a separate input, see Aranda-Bricaire *et al.* [2], [3], & Marino *et al.* [17]. Consequently, the numbers of the forward shifts are the controllability indices, too.

Discrete-time systems are obtained via sampling of a correspondind continuous-time ones with the goal of obtaining a discrete model amenable for numerical calculations. Then the zeros of the sampled system may cause problems in control design. Inclusion of sampled systems to a general framework of discrete-time systems has in this respect studied in Monaco & Normand-Cyrot [18], [19]. Feedback linearisation was studied in [14]. On the other hand, open-loop control design based on flatness avoids these nonminimum phase problems. A comprehensive framework for studying nonlinear discrete-time systems was developed by Grizzle [15].

Theorems including linearizability via static diffeomorphisms or via state feedback are all valid for demonstrating the flatness of the original nonlinear system. Quite generally, without presenting detailed conditions or giving a proof, it can be stated the following.

Theorem. If the nonlinear discrete-time system

$$x(t+1) = f(x(t), u(t)) \tag{23}$$

is linearizable to a controllable linear system

$$\bar{x}(t+1) = F\bar{x}(t) + Gv(t)$$

via (sufficiently differentiable) transformations (*c.f.* Jakubczyk [16])

$$\bar{x} = \Phi(x); \quad u = k(x, v)$$

i.e. via the diffeomorphism Φ and the state feedback k then the system (23) is differencely flat.

5. EXAMPLE

A scalar discrete-time second order system

$$\begin{aligned} &y(t) + a_1 y(t-1) + a_2 y(t-2) = \\ &b_1 u(t-1) + b_2 u(t-2) \end{aligned} \tag{24}$$

is studied. Application of the polynomials $A(q^{-1}) = 1 + a_1 q^{-1} + a_2 q^{-2}$ and $B(q^{-1}) = b_1 q^{-1} + b_2 q^{-2}$ in the Bezouz's identity gives the first order polynomials

$$R(q^{-1}) = r_0 + r_1 q^{-1};\ S(q^{-1}) = s_0 + s_1 q^{-1}$$

The explicit flatness equations between the variables are then

$$\begin{aligned} z(t) &= s_0 y(t) + s_1 y(t-1) + r_0 u(t) + r_1 u(t-1) \\ u(t) &= z(t) + a_1 z(t-1) + a_2 z(t-2) \\ y(t) &= b_1 z(t-1) + b_2 z(t-2) \end{aligned} \tag{25}$$

If the goal now is to drive the output y from 0 to $\bar{y}$ ($\neq 0$) as quickly as possible by suitably manipulating the control variable u, we have a dead-beat control problem. It can be shown that the minimal number

of time steps required $N_{min} = \mathtt{deg}B^*(q) + 1$, where deg denotes the degree of the reciprocal polynomial $B^*(q) = q^{\mathtt{deg}A}B(q^{-1})$. In the model (24) $N_{min} = 2$. The design starts by choosing a step change for the flat output z:

$$z(t) = \begin{cases} 0, & t \leq 0 \\ \bar{z}, & 1 \leq t \end{cases}$$

The the input obtained form (25)

$$\begin{cases} u(0) &= 0 \\ u(1) &= \bar{z} \\ u(2) &= (1 + a_1)\bar{z} \\ u(t) &= (1 + a_1 + a_2)\bar{z}; \quad t = 3, 4, \ldots \end{cases}$$

produces the dead-beat output, *i.e.*

$$\begin{cases} y(1) &= 0 \\ y(2) &= b_1\bar{z} \\ y(t) &= \bar{y} = (b_1 + b_2)\bar{z}; \quad t = 3, 4, \ldots \end{cases}$$

If we want a smoother transfer of the output y from 0 to $\bar{y}$ then a gradual change of the flat output z form 0 to $\bar{z}$ can be applied via some intermediate values $0, \bar{z}_1, \bar{z}_2, \ldots, \bar{z}_n$ giving the corresponding smoother control and output.

6. CONCLUDING REMARKS

The concept of flatness in discrete-time systems facilitates control design for dynamic systems via so-called flat output variables. A typical problem encountered in traditional control design is the non-minimum phase problem. Then controlled systems may become unstable. Design via flatness is independent of this property. On the other hand direct design gives the control only in open-loop mode. A conversion to practical closed-loop mode can be carried out via the flatness relations (8)-(11), *c.f.* [13]. Then the flatness-based control can be applied also under model uncertainties. Multivariable extensions work analogously to the example above. A corresponding nonlinear scalar study was reported in [4].

REFERENCES

[1] Aranda-Bricaire, E., Kotta, Ü., and Moog, C.H., Linearization of discrete-time systems. *SIAM J. Control & Optimiz.* **34**(1996)6, 1999-2023.

[2] Aranda-Bricaire, E. and Kotta, Ü., Equivalence of discrete-time systems to prime systems. *J. Math. Systems, Est. & Control* **8**(1998)4, 471-474 (summary). Full paper 12 pp.

[3] Aranda-Bricaire, E. and Hirschorn, R.M., Equivalence of nonlinear systems to prime systems under generalized output transformations. *SIAM J. Control & Optimiz.* **37**(1999)1, 118-130.

[4] Bastin, G., Jarachi, F., and Mareels, I.M.Y., Output deadbeat control of nonlinear discrete-time systems with one-dimensional zero dynamics: Global stability conditions. *IEEE Trans. Autom. Control* **44**(1999)6, 1262-1266.

[5] Chen, C.-T., *Linear System Theory and Design.* Holt, Rinehart and Winston, New York, 1984.

[6] Fliess, M., Automatique en temps discret et algèbre aux corps différences. *Forum Math.* **2**(1990), 213-232.

[7] Fliess, M., Lévine, J., Martin, P., and Rouchon, P., Sur les systèmes nonlinéaires différentiellement plats. *C.R. Acad. Sci. Paris* **I-315**(1992), 619-624.

[8] Fliess, M., Lévine, J., Martin, P., and Rouchon, P., On differentially flat nonlinear systems. *Proc. IFAC Nonlinear Control Systems Design Symposium, NOLCOS*, 24-26 June 1992, Bordeaux, France, M. Fliess, ed., pp. 408-412.

[9] Fliess, M., Reversible linear and nonlinear discrete-time dynamics. *IEEE Trans. Autom. Control* **37**(1992)8, 1144-1153.

[10] Fliess, M., Lévine, J., Martin, P., and Rouchon, P., Flatness and defect of nonlinear systems: Introductory theory and examples. *International Journal of Control* **61**(1995)6, 1327-1361.

[11] Fliess, M., H. Sira-Ramírez, and R. Marquez., Regulation of non-minimum phase outputs: a flatness based approach. *Perspectives in Control - Theory & Applications*, Colloq. in honor of I.D. Landau, Paris, France, June 1998, Springer-Verlag

[12] Fliess, M., Lévine, J., Martin, P., and Rouchon, P., A Lie-Bäcklund Approach to Equivalence and Flatness of Nonlinear Systems, *IEEE Transactions on Automatic Control* **AC-44**(1999)5, 922-937.

[13] Fliess, M. and Marques, R., Towards a module-theoretic approach to discrete-time linear predictive control, *Proc. Mathematical Theory of Networks and Systems, MTNS'2000*, Perpignan, France, June 2000, 7 pp.

[14] Grizzle, J.W. and Kokotovic, P.V., Feedback linearization of sampled-data systems. *IEEE Trans. Autom. Control* **33**(1988)9, 857-859.

[15] Grizzle, J.W., A linear algebraic frawork for the analysis of discrete-time nonlinear systems, *SIAM J. Control & Optimiz.* **31**(1993)4, 1026-1044.

[16] Jakubczyk, B., Feedback linearization of discrete-time systems. *Systems & Control Lett.* **9**(1987), 411-416.

[17] Marino, R., Respondek, W., and van der Schaft, A.J., Equivalence of nonlinear systems to input-output prime forms. *SIAM J. Control & Optimiz.* **32**(1994)2, 387-407.

[18] Monaco, S., and Normand-Cyrot, D., Zero dynamics of sampled nonlinear systems. *Systems & Control Lett.* **11**(1988), 229-234.

[19] Monaco, S., and Normand-Cyrot, D., A unifying representation for nonlinear discrete-time and sampled dynamics. *J. Math. Systems, Est. & Control* **5**(1995)1, 103-105.

[20] Murray, R.M., Rathinam, M., and Sluis, W., Differential flatness of mechanical control systems: A catalog of prototype systems, *1995 ASME International Mechanical Engineering Congress*, 12-17 November, San Francisco, CA, U.S.A.

[21] Rathinam, M, *Differentially Flat Nonlinear Control Systems*, Technical Report CDS 96-008, California Institute of Technolgy, Control and Dynamical Systems Option, Pasadena, CA, U.S.A , 96 pp. (Doctoral dissertation)

Polynomial-Predictive FIR Design – A Review

Jarno M. A. Tanskanen

Helsinki University of Technology
Institute of Intelligent Power Electronics
P.O. Box 3000, FIN-02015 HUT, FINLAND
E-mail: jarno.tanskanen@hut.fi

ABSTRACT

In this paper, polynomial-predictive FIR (PPF) design is reviewed. Step-by-step instructions are given starting from the signal model, up to designing ideal fixed-point PPFs. This paper is a one-stop starting point for immediate application of PPFs. Also, a literature review is given, including examples of analogously designable filter types.

1. INTRODUCTION

Most real world signals exhibit smooth behavior, if adequately sampled, and the achieved noise level is sufficiently low. With smooth signals, piecewise polynomial signal model can be employed. Polynomial signals offer themselves for efficient prediction with polynomial-predictive FIRs (PPFs) [1]. PPFs, and their more sophisticated augmented versions, find applications in control field, for example, where they can be advantageously applied in fighting control loop delay.

Derivation of PPFs is reviewed in Section 2, and designing for exact polynomial prediction in fixed-point environments in Section 3. Magnitude response shaping feedback design is reviewed in Section 4. A literature review is given in Section 5, and Section 6 concludes the paper.

2. POLYNOMIAL-PREDICTIVE FIR DESIGN

The goal of PPF design [1] is to design such FIR coefficients $h(k)$, $k = 1, 2, ..., m$, where m is FIR length, that a piecewise polynomial input signal is exactly predicted. Thereafter, noise gain (1) of the FIR is minimized.

$$NG = \sum_{k=1}^{m} h(k)^2 \tag{1}$$

2.1. Derivation of Constraints for FIR Coefficients

Here, constraints on the filters coefficients are derived for (p+1)-steps-ahead PPFs (e.g. $p = 1$ yields two-steps-ahead prediction). With the latest input sample taken at time n–1, prediction of an input signal $x(n)$ is generally given by

$$x(n+p) = \sum_{k=1}^{m} h(k)x(n-k). \tag{2}$$

Polynomial signal model yields constraints on the PPF coefficients for each polynomial degree i up to the maximum input polynomial degree I.

0th degree constraint: Prediction of a constant signal $x(n) = c$ yields the constant c itself, and constraint g_0 as

$$\sum_{k=1}^{N} h(k)c = c \Leftrightarrow \sum_{k=1}^{m} h(k) = 1, \tag{3}$$

$$g_0 = \sum_{k=1}^{m} h(k) - 1 = 0. \tag{4}$$

1st degree constraint: prediction of a ramp signal $x(n) = an$ with a slope a, is given by

$$\sum_{k=1}^{m} h(k)a(n-k) = a(n+p) \Leftrightarrow \tag{5}$$

$$\sum_{k=1}^{m} h(k)(n-k) = n+p \Leftrightarrow \sum_{k=1}^{m} h(k)n - \sum_{k=1}^{m} h(k)k = n+p \tag{6}$$

$$\Leftrightarrow \sum_{k=1}^{m} h(k)n - \sum_{k=1}^{m} h(k)k = n+p \tag{7}$$

$$\Leftrightarrow \sum_{k=1}^{m} (h(k)-1)n = \sum_{k=1}^{m} h(k)k + p, \tag{8}$$

which, with (4), yields the 1st degree constraint g_1 as

$$g_1 = \sum_{k=1}^{m} h(k)k + p = 0. \tag{9}$$

2nd degree constraint: prediction of a parabola $x(n) = an^2$ is given by

$$\sum_{k=1}^{m} h(k)a(n-k)^2 = a(n+p)^2 \tag{10}$$

$$\Leftrightarrow \sum_{k=1}^{m} h(k)(n^2 - 2nk + k^2) = n^2 + 2np + p^2 \tag{11}$$

$$\Leftrightarrow \sum_{k=1}^{m} h(k)n^2 - 2\sum_{k=1}^{m} h(k)nk + \sum_{k=1}^{m} h(k)k^2 = n^2 + 2np + p^2 \tag{12}$$

$$\Leftrightarrow \sum_{k=1}^{m} (h(k)-1)n^2 - 2n\sum_{k=1}^{m} h(k)k + \sum_{k=1}^{m} h(k)k^2 = p^2. \tag{13}$$

With (4) and (9), (13) yields the 2nd degree constraint g_2.

$$g_2 = \sum_{k=1}^{m} h(k)k^2 - p^2 = 0 \qquad (14)$$

In [1], the constraint for one-step-ahead prediction $p = 0$ of an Ith degree polynomial input signal is given as

$$g_I = \sum_{k=1}^{m} h(k)k^I = 0\,. \qquad (15)$$

2.2. Derivation of FIR Coefficients from Their Constraints

To obtain closed form presentations for the filter coefficients from the constraints, method of Lagrange multipliers [2][1] may be applied. The objective is to minimize a function of the filter coefficients and Lagrange multipliers λ_i, $i = 1, 2, ..., I$, given by

$$\begin{aligned} &L(h(1),h(2),\ldots,h(m),\lambda_0,\lambda_1,\ldots,\lambda_I) = \\ &NG + \lambda_0 g_0 + \lambda_1 g_1 + \cdots + \lambda_I g_I\,. \end{aligned} \qquad (16)$$

Function (16) includes the noise gain (1), and the constraints g_i, $i = 1, 2, ..., I$. (16) is minimized when its partial derivatives with respect to the filter coefficients and Lagrange multipliers are zero. For example, for the second degree PPFs ($I = 2$),

$$\begin{aligned} &L = NG + \lambda_0 g_0 + \lambda_1 g_1 + \lambda_2 g_2 = \\ &\sum_{k=1}^{m} h(k)^2 + \lambda_0\left(\sum_{k=1}^{m} h(k) - 1\right) + \\ &\lambda_1\left(\sum_{k=1}^{m} h(k)k + p\right) + \lambda_2\left(\sum_{k=1}^{m} h(k)k^2 + p^2\right). \end{aligned} \qquad (17)$$

Next, the partial derivatives are calculated and set to zero:

$$\frac{\partial L}{\partial h(k)} = 2h(k) + \lambda_0 + \lambda_1 k + \lambda_2 k^2 = 0\,, \qquad (18)$$

$$\frac{\partial L}{\partial \lambda_0} = \sum_{k=1}^{m} h(k) - 1 = 0\,, \qquad (19)$$

$$\frac{\partial L}{\partial \lambda_1} = \sum_{k=1}^{m} h(k)k + p = 0\,, \qquad (20)$$

$$\frac{\partial L}{\partial \lambda_2} = \sum_{k=1}^{m} h(k)k^2 - p^2 = 0\,. \qquad (21)$$

The system (18)–(21) is solved by first solving $h(k)$ from (18), inserting $h(k)$ into (19)–(21) and solving for λ_0, λ_1, and λ_2, which are substituted back into (18), yielding the filter coefficients $h(k)$. Mathematica code for solving this system of equations is given in Fig. 1, along with the resulting closed form expression for $I = 2$ PPF coefficients for any p, and the coefficient values for $p = 1$, $m = 10$. Three lowest degree $p = 0$ PPFs are given below [1].

$$I = 0:\ h(k) = 1/m \qquad (22)$$

```
Solve[{Sum[(-(λ0 + λ1 k + λ2 k^2) / 2) - 1 == 0, {k, 1, m}],
  Sum[(-k (λ0 + λ1 k + λ2 k^2) / 2) + p == 0, {k, 1, m}],
  Sum[(-k^2 (λ0 + λ1 k + λ2 k^2) / 2) - p^2 == 0, {k, 1, m}], hk == -(λ0 + k λ1 + λ2 k^2) / 2}, {hk, λ0, λ1, λ2}]

{{hk -> (30 k^2 (2 + 3 m + m^2 + 6 p + 6 m p + 6 p^2)) / (m (-4 + m^2) (-1 + m^2)) +
    (3 (2 + 3 m + 3 m^2 + 6 p + 12 m p + 10 p^2)) / ((-2 + m) (-1 + m) m) -
    (6 k (6 + 21 m + 21 m^2 + 6 m^3 + 22 p + 60 m p + 32 m^2 p + 30 p^2 + 30 m p^2)) /
     ((-1 + m) m (1 + m) (-4 + m^2)), λ0 -> -((6 (2 + 3 m + 3 m^2 + 6 p + 12 m p + 10 p^2)) / ((-2 + m) (-1 + m) m)),
  λ1 -> (12 (6 + 21 m + 21 m^2 + 6 m^3 + 22 p + 60 m p + 32 m^2 p + 30 p^2 + 30 m p^2)) / ((-1 + m) m (1 + m) (-4 + m^2)),
  λ2 -> -((60 (2 + 3 m + m^2 + 6 p + 6 m p + 6 p^2)) / (m (-4 + m^2) (-1 + m^2)))}}

% /. {m -> 10, p -> 1, k -> {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}}

{{h{1,2,3,4,5,6,7,8,9,10} -> {27/22, 19/30, 37/220, -37/220, -62/165, -5/11, -89/220, -149/660, 9/110, 57/110},
  λ0 -> -39/10, λ1 -> 1039/660, λ2 -> -17/132}}
```

Fig. 1. Solving the system of equations (18)–(21) with Mathematica. Also calculated are the values of the $p = 1$, $m = 10$, PPF coefficients. Notation: $h_k = h(k)$, $\lambda_i = \lambda(i)$.

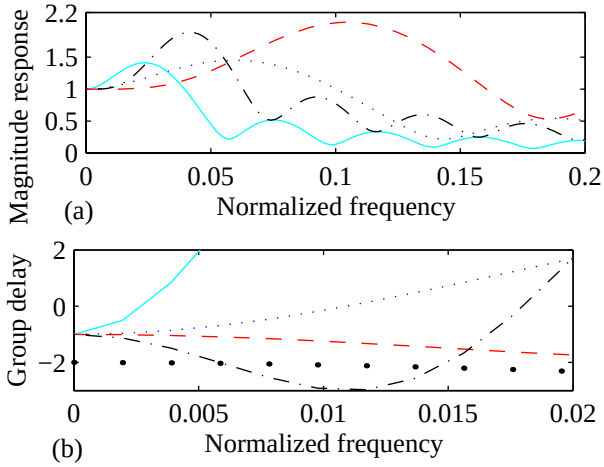


Fig. 2. Frequency responses (a) and group delays (b) of exemplary PPFs: $I = 1$, $p = 0$: $m = 20$ (dotted) and $m = 50$ (solid); $I = 2$, $p = 0$: $m = 20$ (dashed) and $m = 50$ (dash-dot). In (b), also the group delay of the PPF calculated in Fig. 1 (dark dotted).

$$I = 1:\ h(k) = \big(4m - 6k + 2\big)/\big[m(m-1)\big] \qquad (23)$$

$$I = 2:\ h(k) = \frac{9m^2 + (9 - 36k)m + 30k^2 - 18k + 6}{m^3 - 3m^2 + 2m} \qquad (24)$$

In (22)–(24), $k = 1, 2, ..., m$. In Fig. 2, exemplary PPF magnitude responses and group delays are shown. Also generally, passband peak is suppressed and the passband width gets narrower as PPF length increases or polynomial degree decreases. Prediction band, i.e., frequency range with group delay sufficiently close (depends on the application) to $-p-1$, gets narrower with increased PPF length or polynomial degree.

3. FIXED-POINT PPF DESIGN

Requirement for exact polynomial prediction in fixed-point environments is that the fixed-point PPF (FPPPF) coefficients $h_q(k)$ must exactly satisfy the constraints (4), (9), (14), and up to the constraint for degree I [3]. At simplest, FPPPF design can be an exhaustive search [4] over a limited region of a quantized coefficient space H, to find quantized coefficients $h_q(k) \in$ H, $k = 1, 2, ..., m$, which

fulfill the constraints. An example of a search region spawning two quantization levels above and below the exact coefficients $h(k)$, given by (24), is seen in Fig. 3.

After finding all the sets of $h_q(k) \in$ H, $k = 1, 2, ..., m$, which exactly satisfy our constraints within the search region, the set that minimizes the noise gain (1) is selected as the ideally quantized coefficient PPF. One such set of coefficients is seen in Fig. 3. Though the ideally quantized coefficient PPFs make no assumptions of global noise gain minimization, their noise gain losses are negligible when compared with the noise gains of the corresponding non-quantized coefficient PPFs [3].

4. MAGNITUDE RESPONSE SHAPING

Magnitude responses of PPFs exhibit high inherent passband peaks, which is a drawback in most applications. By augmenting PPFs with appropriate feedbacks [5], PPF magnitude responses can be shaped without affecting the predictive properties. An augmented PPF of length $m = 2$ is illustrated in Fig. 4. This feedback has smoothing function, since the leftmost summation point yields a weighted sum of an input sample and its prediction, like also the second summation point within the delayline, as seen from Fig. 4. Thus, the feedback does not affect the predictive properties of the PPF. Conditions for the feedback to exactly preserve the predictive properties of the basis PPF are only that $\{b(k), 1 - b(k)\} \in$ H, $k = 1, 2, ..., m$, i.e., the feedback coefficients belong to the same quantized coefficient space H as the PPF coefficients $h(k)$ [3].

In Fig. 5, magnitude response and group delay of the shortest $I = 1$, $p = 0$, PPF (23), $m = 2$, is shown along with two augmented PPFs with the same basis PPF [3]. Using feedbacks, passband peak is clearly reduced. All filters in Fig. 5 are coefficient quantization error free with eight bit coefficients, i.e., they fulfill all the required constraints.

5. LITERATURE REVIEW

PPFs were derived by Heinonen and Neuvo in [1], where the case $p = 0$ is presented. Derivation of PPFs with any p is given in the appendix of [6], and in [7], least squares formulation of PPF design for any p and I is given in a matrix form suitable for numerical calculations. In [7], asymptotic noise gain of PPFs is also derived. In [8], a computationally efficient structure for implementing PPFs is given; complexity does not depend on the PPF length but only on I. In [9], PPFs [1] were formulated for complex-valued signals. Derivation and an example of an interpolated FIR approach to PPFs, resulting in lower complexity, is presented in [10]. To suppress passband peaks of PPFs, a prefiltering approach was proposed in [11]. Thereafter, *PPF feedback augmentation design* is given in [5]. In [5], examples are given on a modified first-degree PPF filter with a notch for power line frequency suppression, and on predictor-estimator cascade design providing for better stopband attenuation. Feedback augmentation is also presented in [12]. Finite coefficient word length is

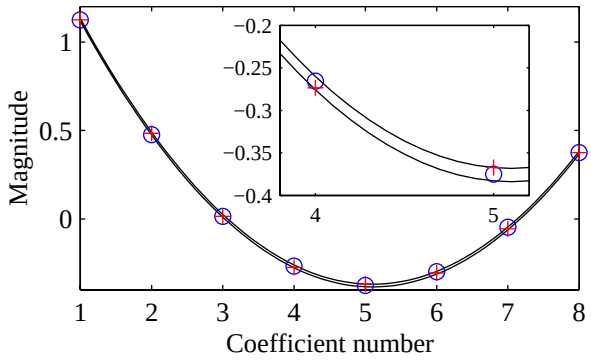


Fig. 3. A search band (between solid lines) for quantized coefficients of the $I = 2$, $p = 0$, $m = 8$, PPF with 8-bit coefficient precision. Circles 'o' denote the magnitude truncated, and plusses '+' the ideally quantized coefficients.

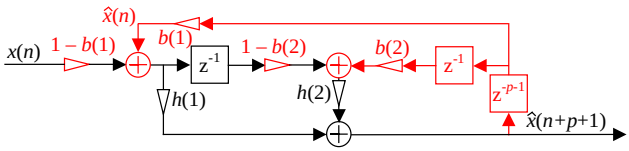


Fig. 4. Feedback augmented PPF of length $m = 2$. Hat denotes predictive estimate.

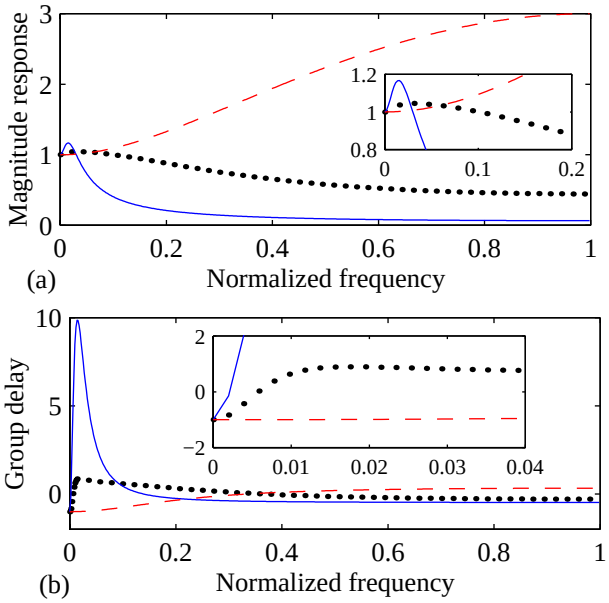


Fig. 5. Magnitude responses (a) and group delays (b) of augmented coefficient quantization error free $I = 1$, $p = 0$, $m = 2$ PPFs with the feedback coefficients $\{b(1), b(2)\} = \{0.6875, -0.9375\}$ (dark dotted) and $\{b(1), b(2)\} = \{0.9375, -0.9375\}$ (solid), along with their basis PPF $\{h(1), h(2)\} = \{2 \ -1\}$ (dashed).

the topic of [13][14]. PPF feedback coefficients are optimized using a genetic algorithm in [15], and rigorously in [16]. *Design of exact fixed-point PPF implementations* is presented in [4] (also for predictive polynomial differentiator FIRs), along with integer programming interpretation of the design method. *Conditions for exact fixed-point implementations of predictive polynomial differentiator FIRs, and of their feedback augmentations*, are given in [3], where also a quantization error feedback approach for

roundoff noise alleviation is proposed. PPFs are also the topic of [17]. Polynomial predictive filtering in control instrumentation is reviewed in [18] with several predictor structures. Matlab packages for PPF and augmented PPF design are available at [19] (note on the nomenclature: in [19] PPFs are referred to as Heinonen-Neuvo (H-N) filters, and augmentation can be done with the "FIR2IIR" designer).

Predictive polynomial differentiator FIRs (PPDFs) are derived analogously to PPFs; PPDFs for $I = 1$ are derived in [20], and for $I = 2$ in [21]; in them, also efficient recursive implementations are given. PPDFs get their feedback augmentations in [22]. PPDFs are also discussed in [23], and their coefficient quantization sensitivity is illustrated in [24]. PPDFs for angular acceleration measurement are reviewed in [25].

Sinusoidal predictors (SPs), also derived in time domain, are derived in [26] for application in power line frequency zero crossing detection. Evolved version of the system in [26] is presented in [27], where SPs are derived with constraints for suppression of DC and the first odd harmonic of the nominal SP frequency. IIR based computationally efficient implementation of SPs is given in [10]. A comprehensive presentation of SPs is given in [28], where also feedback augmentation is added to SPs.

6. CONCLUSIONS

Polynomial-predictive filtering, derived in time domain, is generally not well-known, and usually not mentioned in standard text books. Still, it is efficient and beneficial in processing many real-world signals. For most applications, it is sufficient to apply low degree polynomial predictors, i.e., of the maximum degree of $I = 1$, 2, or 3, which is also recommendable from the predictor magnitude response characteristics point of view.

In this paper, step-by-step instructions for designing polynomial-predictive filters are given, along with a fixed-point design method. A literature review on polynomial-predictive FIRs and associated time domain filters is given.

REFERENCES

[1] P. Heinonen and Y. Neuvo, "FIR-median hybrid filters with predictive FIR substructures," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 36, pp. 892–899, June 1988.

[2] D. Bertsekas, *Constrained Optimization and Lagrange Multipliers Methods*. New York, NY, USA: Academic Press, 1982.

[3] J. M. A. Tanskanen, "Coefficient quantization error free fixed-point IIR polynomial predictor design," in *Proc. 2000 IEEE Nordic Signal Processing Symp.*, Kolmården, Sweden, June 2000, pp. 219–222.

[4] J. M. A. Tanskanen and V. S. Dimitrov, "Round-off Error Free Fixed-Point Design of Polynomial FIR Predictors and Predictive FIR Differentiators," *Digital Signal Processing, A Review Journal*, accepted for publication. Also: Helsinki University of Technology Institute of Intelligent Power Electronics Publications, publication 4, Espoo, Finland, Aug. 2000 [online]. Available http://wooster.hut.fi/publications/tanskanen/round_off_error_free_predictors.pdf.

[5] S. J. Ovaska, O. Vainio, and T. I. Laakso, "Design of predictive IIR filters via feedback extension of FIR forward predictors," *IEEE Trans. Instrumentation and Measurement*, vol. 46, pp. 1196–1201, Oct. 1997.

[6] T. G. Campbell, "Design and implementation of image filters," Doctoral dissertation, Tampere University of Technology Publications 97, Tampere, Finland, 1992.

[7] P. Händel and P. Tichavský, "Asymptotic noise gain of polynomial predictors," *Signal Processing*, vol. 62, pp. 247–250, 1997.

[8] T. G. Campbell and Y. Neuvo, "Predictive FIR filters with low computational complexity," *IEEE Trans. Circuits and Systems*, vol. 38, pp. 1067–1071, Sept. 1991.

[9] T. Harju and T. I. Laakso, "Polynomial predictors for complex-valued vector signals," *Electronics Letters*, vol. 31, pp. 1650–1652, Sept. 1995.

[10] O. Vainio, "Design and efficient implementations of predictive FIR filters," *IEEE Trans. Instrumentation and Measurement*, vol. 44, pp. 864–868, Aug. 1995.

[11] T. I. Laakso and S. J. Ovaska, "Prefiltering approach for optimal polynomial prediction," *IEEE Trans. Signal Processing*, vol. 44, pp. 701–705, Mar. 1996.

[12] S. J. Ovaska and O. Vainio, "Predictive compensation of time-varying computing delay on real-time control systems," *IEEE Trans. Control Systems Tech.*, vol. 5, pp. 523–526, Sept. 1997.

[13] P. T. Harju, "Finite wordlength implementation of IIR polynomial predictive filters," in *Proc. 1997 IEEE Instrumentation and Measurement Tech. Conf.*, Ottawa, Canada, May 1997, pp. 60–65.

[14] P. T. Harju, "Roundoff noise properties of IIR polynomial predictive filters," in *Proc. 1997 IEEE Instrumentation and Measurement Tech. Conf.*, Ottawa, Canada, May 1997, pp. 66–71.

[15] P. T. Harju and S. J. Ovaska, "Optimization of polynomial predictive IIR filters using genetic algorithms," in *Proc. 3rd International Conf. on Signal Processing*, Beijing, China, Oct. 1996, pp. 68–71.

[16] P. T. Harju and S. J. Ovaska, "Optimization of IIR polynomial predictive filter magnitude response," *Signal Processing*, vol. 56, pp. 219–232, Feb. 1997.

[17] J. M. A. Tanskanen, "Polynomial predictive filters: implementation and applications," Doctoral dissertation, Helsinki University of Technology Institute of Intelligent Power Electronics Publications, Publication 5, Espoo, Finland, Nov. 2000.

[18] S. Väliviita, S. J. Ovaska, and O. Vainio, "Polynomial predictive filtering in control instrumentation: a review," *IEEE Trans. Industrial Electronics*, vol. 46, pp. 876–888, Oct. 1999.

[19] J. Martikainen, "Soft Filtering, an Emerging DSP Technique," [WWW Site]. Espoo, Finland: Helsinki University of Technology, Institute of Intelligent Power Electronics, Apr. 2000 [cited 9 Apr. 2001]. Available http://www.hut.fi/Units/PowerElectronics/soft/.

[20] O. Vainio, M. Renfors, and T. Saramäki, "Recursive implementation of FIR differentiators with optimum noise attenuation," *IEEE Trans. Instrumentation and Measurement*, vol. 46, pp. 1202–1207, Oct. 1997.

[21] S. Väliviita and O. Vainio, "Delayless differentiation algorithm and its efficient implementation for motion control applications," *IEEE Trans. Instrumentation and Measurement*, vol. 48, pp. 967–971, Oct. 1999.

[22] S. Väliviita and S. J. Ovaska, "Delayless recursive differentiator with efficient noise attenuation for control instrumentation," *Signal Processing*, vol. 69, pp. 267–280, Sept. 1998.

[23] S. Väliviita, "Predictive filtering methods for motor drive instrumentation," Doctoral dissertation, Helsinki University of Technology Institute of Intelligent Power Electronics Publications, publication 1, Espoo, Finland, Aug. 1998.

[24] J. M. A. Tanskanen and S. J. Ovaska, "Coefficient sensitivity of polynomial-predictive FIR differentiators: Analysis," in *Proc. 42nd IEEE Midwest Symp. on Circuits and Systems*, Las Cruces, NM, USA, Aug. 1999, pp. 405–408.

[25] S. J. Ovaska and S. Väliviita, "Angular acceleration measurement: a review," *IEEE Trans. Instrumentation and Measurement*, vol. 47, pp. 1211–1217, Oct. 1998.

[26] O. Vainio, "Noise reduction in zero crossing detection by predictive digital filtering," *IEEE Trans. Industrial Electronics*, vol. 42, pp. 58–62, Feb. 1995.

[27] O. Vainio and S. J. Ovaska, "Digital filtering for robust 50/60 Hz zero-crossing detectors," *IEEE Trans. Instrumentation and Measurement*, vol. 45, pp. 426–430, Apr. 1996.

[28] P. Händel, "Predictive digital filtering of sinusoidal signals," *IEEE Trans. Signal Processing*, vol. 46, pp. 364–374, Feb. 1998.

Adaptive Channel Equalizer for WCDMA Downlink

Kari Hooli, Matti Latva-aho, and Markku Juntti

University of Oulu, Centre for Wireless Communications
P.O. Box 4500 FIN-90014 University of Oulu
FINLAND
email: kari.hooli@ee.oulu.fi

ABSTRACT

The main 3rd generation cellular communications standard is based on wideband code-division multiple-access (WCDMA). For WCDMA downlink, receivers based on channel equalization at chip level have been proposed to ensure adequate performance even with a high number of active users. These receivers equalize the channel prior to the despreading, thus restoring the orthogonality of users and resulting multiple access interference (MAI) suppression. In this paper, ideas of generalized side-lobe canceler (GSC) and minimum output energy (MOE) are applied to the adaptation of downlink channel equalizer. The performance of the adaptation scheme is compared to the performance of conventional Rake receiver as well as to the performance of equalizer adapted with Griffiths' algorithm. Numerical results show significant performance gain over Rake receiver and some performance gain over the Griffiths' algorithm in a fading channel.

1. INTRODUCTION

Wideband code-division multiple-access is the main air interface of the 3rd generation cellular mobile communications standards. The downlink capacity is expected to be more crucial than the capacity of the uplink due to the asymmetric capacity requirements, i.e., the downlink direction should offer higher capacity than the uplink [1]. Therefore the employment of efficient downlink receivers is important. In order to avoid performance degradation near-far resistant (or multiuser) receivers can be used. Several suboptimal receivers feasible for practical implementations have been proposed, including linear minimum mean squared error (LMMSE) receivers [2]. The adaptive versions of the symbol level LMMSE receivers rely on cyclostationary of multiple access interference (MAI), and thus require periodic spreading sequences with a very short period. Hence they can not be applied on the WCDMA downlink, which uses spreading sequences with 1 radio frame (10 ms) period.

In a synchronously transmitted downlink employing orthogonal spreading codes MAI is mainly caused by multipath propagation. Due to the non-zero cross-correlations between the spreading sequences with arbitrary time shifts, there is interference between propagation paths (or Rake fingers) after the despreading causing multiple access interference. If the received chip waveform, distorted by the multipath channel, is equalized prior to the correlation by the spreading code or matched filtering, there is only a single path in the despreading. With orthogonal spreading sequences the equalization effectively retains, to some extent, the orthogonality of users lost due to the multipath propagation, thus suppressing MAI. Since the signal is equalized on the chip level, not on the symbol level, they can also be applied in systems using long spreading sequences. Such a receiver, discussed e.g. in [3]-[6], consists of a linear equalizer followed by a single correlator and a decision device, as depicted in Fig. 1.

Several adaptive versions of chip-level channel equalizers have been presented, e.g., in references in [6]-[7]. In this paper the ideas of generalized side-lobe canceler and minimum output energy are applied, resulting a novel adaptation scheme, channel-response MOE (CR-MOE). The bit error rate (BER) in a Rayleigh fading multipath channel was numerically evaluated for CR-MOE equalizer and compared to the performance of conventional Rake receiver and equalizer adapted with Griffiths' algorithm, suggested in [8]. Comparison between CR-MOE and Griffiths' algorithm [9] is quite natural due to the similarities of the schemes.

2. SYSTEM MODEL

Since the downlink is considered, synchronous transmission of all signals through the same multipath channel is assumed. The discrete-time received signal at user terminal can be written as

$$\boldsymbol{r} = \sum_{k=1}^{K} \boldsymbol{D}\boldsymbol{C}\boldsymbol{A}_k\boldsymbol{S}_k\boldsymbol{b}_k + \boldsymbol{n}, \qquad (1)$$

where K is the number of users, $\boldsymbol{D}$ is a path delay and chip waveform matrix whose columns contain samples from appropriately delayed chip waveforms, and $\boldsymbol{C}$ is a block diagonal channel matrix containing channel coefficients for L propagation paths. Diagonal matrix $\boldsymbol{A}_k$ contains the average received amplitudes and $\boldsymbol{S}_k$ is a block diagonal matrix

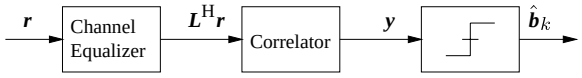


Fig. 1. Conceptual structure of chip-level channel equalizer.

The research described in the paper has been supported by Nokia and Texas Instruments. The contents of the paper have been presented in COST 262 workshop in Ulm, Germany, on 17.-18.1. 2001.

containing the spreading sequence for the kth user. The cell specific scrambling sequence is included to the spreading sequence, and the sequences are normalized so that $\boldsymbol{S}_k^{\mathrm{H}}\boldsymbol{S}_k = \boldsymbol{I}$. Vector $\boldsymbol{b}_k$ contains the transmitted symbols of the kth user, and vector $\boldsymbol{n}$ contains samples from white complex Gaussian noise process with variance σ_n^2. A more detailed description of the system model is given e.g. in [6].

3. RECEIVERS

In this section, the conventional Rake receiver as well as the adaptive equalizers are discussed. First, the Rake receiver, the linear minimum mean square error (LMMSE) chip-level channel equalizer, and Griffiths' algorithm are shortly defined, followed by discussion of the channel-response MOE equalizer.

In the Rake receiver, the received signal is filtered by the chip waveform, appropriately time-aligned and despread by correlation with the spreading sequence in the each of the Rake fingers. To obtain the decision variable, the Rake fingers are weighted by the channel coefficient estimates and combined in the maximal ratio combining (MRC).

The chip-level LMMSE equalizer is obtained by minimizing the mean square error between the equalizer output and the total transmitted signal, i.e., by solving

$$\boldsymbol{w}_L = \arg\min_{\boldsymbol{w}} \mathrm{E}\left[\left|\boldsymbol{w}^{\mathrm{H}}\boldsymbol{r} - \sum_{k=1}^{K}\boldsymbol{A}_k\boldsymbol{S}_k\boldsymbol{b}_k\right|^2\right], \qquad (2)$$

where minimization is carried out elementwise. The optimization problem in (2) can be easily solved. However, the exact LMMSE solution depends on the spreading sequences of all users following from the dependency between consecutive chips to be estimated. In [6] it was shown that the chip dependency has only minor effect on the performance, and the LMMSE solution can be approximated as

$$\boldsymbol{w}_L \approx \left(s^2\sum_{k=1}^{K}A_k^2\boldsymbol{D}\boldsymbol{C}\boldsymbol{C}^{\mathrm{H}}\boldsymbol{D}^{\mathrm{H}} + \sigma_n^2\boldsymbol{I}\right)^{-1}\boldsymbol{D}\boldsymbol{C}. \qquad (3)$$

The decision variable of the chip-level LMMSE equalizer after the correlation with a spreading sequence is

$$\begin{aligned}\boldsymbol{y} &= \boldsymbol{S}_1^{\mathrm{H}}\boldsymbol{w}_L^{\mathrm{H}}\boldsymbol{r}\\ &= \boldsymbol{S}_1^{\mathrm{H}}\boldsymbol{C}^{\mathrm{H}}\boldsymbol{D}^{\mathrm{H}}\left(s^2\sum_{k=1}^{K}A_k^2\boldsymbol{D}\boldsymbol{C}\boldsymbol{C}^{\mathrm{H}}\boldsymbol{D}^{\mathrm{H}} + \sigma_n^2\boldsymbol{I}\right)^{-1}\boldsymbol{r}. \qquad (4)\end{aligned}$$

In the adaptive chip-level channel equalizers, the received signal is filtered by chip waveform, equalized and correlated with the spreading sequence. The decision variable for arbitrary selected user 1 after the correlation with spreading sequence is given by $\boldsymbol{y} = \boldsymbol{S}_1^{\mathrm{H}}\boldsymbol{z}$, where vector $\boldsymbol{z}$ contains equalizer outputs for corresponding chip intervals. The nth element of $\boldsymbol{z}$ is defined by $\boldsymbol{w}(n)^{\mathrm{H}}\bar{\boldsymbol{r}}(n)$, where $\boldsymbol{w}(n) \in \mathbb{C}^{(2D+1)N_s}$ contains the equalizer taps and $\bar{\boldsymbol{r}}(n) = [r((n-D)N_s)\ldots r(nN_s),\ldots r((n+D+1)N_s-1)]^{\mathrm{T}}$ is a vector of output samples from the chip waveform matched filter within equalizer at nth chip interval. N_s is the number of samples per chip.

Several adaptation algorithms are obtained through different approximations of gradient vector

$$\frac{\nabla J}{\nabla \boldsymbol{w}} = -2\mathrm{E}[d^*\bar{\boldsymbol{r}}] + 2\mathrm{E}[\bar{\boldsymbol{r}}\bar{\boldsymbol{r}}^{\mathrm{H}}]\boldsymbol{w}, \qquad (5)$$

where $J = \mathrm{E}[|d - \boldsymbol{w}^{\mathrm{H}}\bar{\boldsymbol{r}}|^2]$ is the mean square error and d^* is the desired equalizer output's complex conjugate [10]. For example, the standard LMS algorithm is obtained by replacing expectations with instantaneous estimates, i.e., signal vectors $\bar{\boldsymbol{r}}(n)$. In [8], the Griffiths' algorithm is used for the adaptation of chip-level channel equalizer. The algorithm is obtained from (5) by replacing $\mathrm{E}[d^*\bar{\boldsymbol{r}}]$ with $\tilde{\boldsymbol{p}}$, the channel response matched filter. The resulting adaptation becomes

$$\boldsymbol{w}(n+1) = \boldsymbol{w}(n) - \mu(z^*(n)\bar{\boldsymbol{r}}(n) - \tilde{\boldsymbol{p}}), \qquad (6)$$

where μ is the adaptation step size and $z(n)$ is the equalizer output at nth chip interval.

In the channel-response constrained minimum-output-energy (CR-MOE) equalizer, the equalizer is decomposed into a constraint (or non-adaptive) component and to an adaptive component. This is the well known idea of generalized side-lobe canceler, described, e.g., in [10, chap. 5]. The same approach has been applied in blind MOE multiuser receivers, in which the spreading sequence of a desired user is used as the constraint [11]–[12]. As mentioned, the equalizer is decomposed into two parts, i.e., $\boldsymbol{w} = \tilde{\boldsymbol{p}} + \boldsymbol{x}$. The channel response matched filter $\tilde{\boldsymbol{p}}$ is used as the non-adaptive part, and the adaptive part $\boldsymbol{x}$ is constrained onto subspace orthogonal to $\tilde{\boldsymbol{p}}$ to avoid suppression of the desired signal. Now the mean square error J can be written as

$$J = \mathrm{E}[d^2] - 2\tilde{\boldsymbol{p}}^{\mathrm{H}}\tilde{\boldsymbol{p}} + (\tilde{\boldsymbol{p}} + \boldsymbol{x})^{\mathrm{H}}\mathrm{E}[\bar{\boldsymbol{r}}\bar{\boldsymbol{r}}^{\mathrm{H}}](\tilde{\boldsymbol{p}} + \boldsymbol{x}). \qquad (7)$$

Clearly the mean square error for given $\tilde{\boldsymbol{p}}$ is optimized by minimizing the last term of J, i.e., equalizer output energy. To obtain adaptive algorithm for $\boldsymbol{x}$, stochastic approximation is applied to the gradient of output energy $\boldsymbol{w}^{\mathrm{H}}\mathrm{E}[\bar{\boldsymbol{r}}\bar{\boldsymbol{r}}^{\mathrm{H}}]\boldsymbol{w}$. The orthogonality condition is maintained at each iteration by projecting the gradient onto the subspace orthogonal to $\tilde{\boldsymbol{p}}$. The orthogonal component of gradient is given by

$$\nabla\tilde{J}_{\perp\tilde{\boldsymbol{p}}} = \left(\bar{\boldsymbol{r}} - \tilde{\boldsymbol{p}}\frac{\tilde{\boldsymbol{p}}^{\mathrm{H}}\bar{\boldsymbol{r}}}{\tilde{\boldsymbol{p}}^{\mathrm{H}}\tilde{\boldsymbol{p}}}\right)\bar{\boldsymbol{r}}^{\mathrm{H}}\boldsymbol{w}. \qquad (8)$$

The resulting adaptation algorithm is given by

$$\boldsymbol{x}(n+1) = \boldsymbol{x}(n) - \mu z^*(n)(\bar{\boldsymbol{r}}(n) - z_p(n)\tilde{\boldsymbol{p}}), \qquad (9)$$

where $z_p(n) = \tilde{\boldsymbol{p}}^{\mathrm{H}}\bar{\boldsymbol{r}}(n)/(\tilde{\boldsymbol{p}}^{\mathrm{H}}\tilde{\boldsymbol{p}})$ is the output of channel response matched filter normalized with the energy of channel response. The CR-MOE equalizer is depicted in Fig. 2.

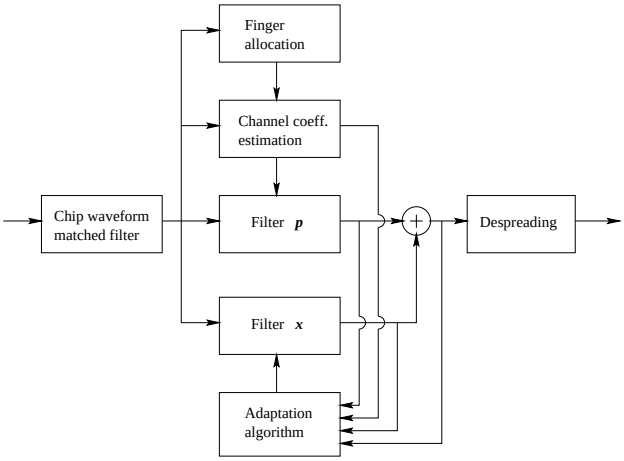


Fig. 2. Structure of CR-MOE equalizer.

The CR-MOE has the typical weaknesses of MOE adaptation [12]. The orthogonality between $\boldsymbol{x}$ and $\tilde{\boldsymbol{p}}$ is lost when the channel response estimate is updated. Thus periodical re-orthogonalization of $\boldsymbol{x}$ is required, given by

$$\boldsymbol{x}_{\perp\tilde{\boldsymbol{p}}} = \boldsymbol{x} - \frac{\tilde{\boldsymbol{p}}^{\mathrm{H}}\boldsymbol{x}}{\tilde{\boldsymbol{p}}^{\mathrm{H}}\tilde{\boldsymbol{p}}}\tilde{\boldsymbol{p}}. \tag{10}$$

The second problem of the MOE adaptation is the unavoidable estimation error in $\tilde{\boldsymbol{p}}$. Due to the estimation error, $\boldsymbol{x}$ has small projection on true $\boldsymbol{p}$ while maintaining orthogonality with $\tilde{\boldsymbol{p}}$. Since $\boldsymbol{x}$ is adapted to minimize output energy, the projection on $\boldsymbol{p}$ translates to partial suppression of the desired signal component. Since the channel estimation error is usually relatively small, suppression of the desired signal means large $||\boldsymbol{x}||^2$ values[1] and significant noise enhancement. Therefore, in noisy environments the suppression remains at acceptable levels. However, to avoid the desired signal suppression at high SNR, $||\boldsymbol{x}||^2$ values must be restricted. One solution is to introduce tap leakage [12]

$$\boldsymbol{x}(n+1) = (1-\mu\alpha)\boldsymbol{x}(n) - \mu z^*(n)(\bar{\boldsymbol{r}}(n) - z_p(n)\tilde{\boldsymbol{p}}), \tag{11}$$

where α, a small positive constant, controls the tap leakage. On the other hand, too low $||\boldsymbol{x}||^2$ values prevent efficient channel equalization. Therefore α must be adjusted to changing conditions. This can be achieved by periodically observing $||\boldsymbol{x}||^2/||\tilde{\boldsymbol{p}}||^2$ ratio and adjusting α if necessary.

It can be easily noted that the considered equalizers have distinctively similar properties. For example the part of adaptation step orthogonal to $\tilde{\boldsymbol{p}}$ in (6) is equal to the adaptation step in (9), assuming the same equalizer taps $\boldsymbol{w}(n)$. However, the estimated channel response is directly inserted to the equalizer in CR-MOE, whereas in Griffiths' algorithm it is gradually introduced through the adaptation. It is clear that both adaptive algorithms rely on the channel response estimate, obtained, e.g., with the help of common or dedicated pilot channel. Also the whole transmitted signal from the desired base-station is utilized in the adaptation instead of, e.g., using only the signal of desired user, thus significantly enhancing the available SNR in the equalizer adaptation. Finally, both equalizers have relatively low complexity with linear dependence on the channel delay spread.

[1] $||\boldsymbol{x}||^2 = \boldsymbol{x}^{\mathrm{H}}\boldsymbol{x}$

4. NUMERICAL RESULTS

To obtain a good understanding and comparison of the presented receivers' performance, BER's were evaluated in a Rayleigh fading channel. The channel had three propagation paths with delays of 0 ns, 521 ns and 1042 ns, and the relative average powers of the paths were 0 dB, -3 dB and -6 dB. QPSK modulation was used employing root raised cosine pulses with roll-off factor of 0.22. Random cell specific scrambling code and Walsh channelization codes were used, and the chip rate was set to 3.84 Mchip/s corresponding to 260 ns chip interval.

The BER's were evaluated for the receivers in a Rayleigh fading channel with 4 users employing spreading factor 8 and common pilot channel (CPICH) using spreading factor 64. The transmission power of pilot channel was scaled to be 11% from the total transmitted power. The terminal velocity was assumed to be 60 km/h. The fingers in the Rake receiver were allocated at correct path delays. For the equalizers, two samples per chip were taken from the output of chip waveform matched filter. The channel response matched filter $\tilde{\boldsymbol{p}}$ had non-zero values at correct path delays as well as on the adjacent samples, due to oversampling of chip waveform. The channel coefficients were estimated with common pilot channel and a moving average filtering.

The BER's are presented in Fig. 3, with the performance of ideal LMMSE equalizer and theoretical single-user bound. From the results it can be seen that both equalizers provide significant performance gain over the Rake receiver. It can be also noted that CR-MOE equalizer provides performance improvement over the equalizer adapted with Griffiths' algorithm in a fading channel.

5. CONCLUSIONS

One approach to improve the performance of WCDMA downlink receivers was studied in this paper, namely channel equalization prior to despreading. The presented receivers restore to some extent the orthogonality of users, and thus suppress the multiple access interference when orthogonal spreading sequences are employed. The filter decomposition idea of generalized side-lobe canceler is applied to the chip-level channel equalization, resulting a novel adaptive equalizer, channel-response MOE equalizer. CR-MOE equalizer consists of two parallel filters. Channel response matched filter is used as the non-adaptive filter, and the other filter minimizes the equalizer output energy.

The performance was numerically evaluated for CR-MOE equalizer and compared to the performance of conventional

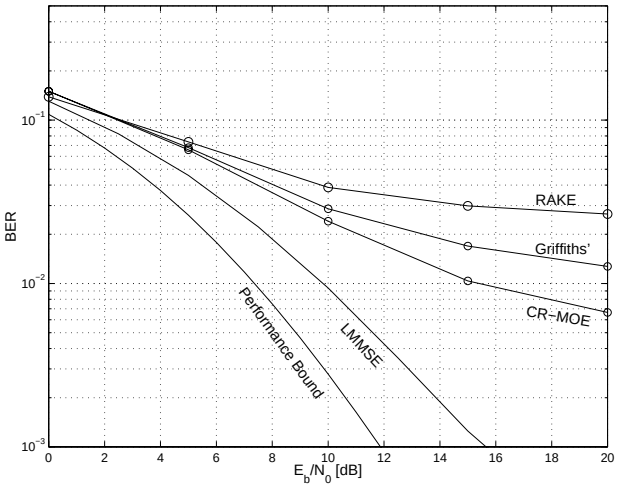


Fig. 3. BER vs. E_b/N_0 in a Rayleigh fading channel for 4 users with spreading factor 8 and CPICH with spreading factor 64.

Rake receiver and an equalizer adapted with Griffiths' algorithm. Results show significant performance gain over the Rake receiver. In a fading channel, CR-MOE equalizer provides performance improvement also over the equalizer adapted with Griffiths' algorithm.

ACKNOWLEDGMENT

Mr E. Corrales is acknowledged for the help with Rake simulations.

REFERENCES

[1] H. Holma and A. Toskala, Eds., *Wideband CDMA for UMTS*, John Wiley and Sons, New York, 2000.

[2] P. B. Rapajic and B. S. Vucetic, "Adaptive receiver structures for asynchronous CDMA systems," *IEEE J. Select. Areas Commun.*, vol. 12, no. 4, pp. 685–697, May 1994.

[3] A. Klein, "Data detection algorithms specially designed for the downlink of CDMA mobile radio systems," in *Proc. IEEE Vehic. Tech. Conf.*, Phoenix, USA, May 4–7 1997, vol. 1, pp. 203–207.

[4] C. D. Frank and E. Visotsky, "Adaptive interference suppression for direct-sequence CDMA systems with long spreading codes," in *Proc. Annual Allerton Conf. Communication Control and Computing*, Allerton House, Monticello, USA, Sept. 23-25 1998.

[5] I. Ghauri and D. T. M. Slock, "Linear receivers for the DS-CDMA downlink exploiting orthogonality of spreading sequences," in *Proc. 32th Asilomar Conf. on Signals, Systems and Comp.*, Asilomar, CA, Nov. 1–4 1998, vol. 1, pp. 650–654.

[6] K. Hooli, M. Latva-aho, and M. Juntti, "Multiple access interference suppression with linear chip equalizers in WCDMA downlink receivers," in *Proc. IEEE Glob. Telecommun. Conf.*, Rio de Janeiro, Brazil, Dec. 5–9 1999, vol. 1, pp. 467–471.

[7] K. Hooli, M. Latva-aho, and M. Juntti, "Performance evaluation of adaptive chip-level channel equalizers in WCDMA downlink," in *Proc. IEEE Int. Conf. Commun., to appear*, Helsinki, Finland, June 11–15 2001.

[8] M. Heikkilä, P. Komulainen, and J. Lilleberg, "Interference suppression in CDMA downlink through adaptive channel equalization," in *Proc. IEEE Vehic. Tech. Conf.*, Amsterdam, The Netherlands, Sept. 19–22 1999, vol. 2, pp. 978–982.

[9] J. R. Treichler, C. R. Johnson, and M. G. Larimore, *Theory and Design of Adaptive Filters*, John Wiley and Sons, New York, USA, 1987.

[10] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, Upper Saddle River, NJ, USA, 3rd edition, 1996.

[11] M. Honig, U. Madhow, and S. Verdú, "Blind adaptive multiuser detection," *IEEE Trans. Inform. Theory*, vol. 41, no. 3, pp. 944–960, July 1995.

[12] S. Verdu, *Multiuser Detection*, Cambridge University Press, Cambridge, UK, 1998.

Space-Frequency Turbo Coded OFDM for Future High Data Rate Wideband Radio Systems

Djordje Tujkovic, Markku Juntti, and Matti Latva-aho

Centre for Wireless Communications (CWC)
University of Oulu, P.O.Box 4500, Tutkijantie 2E, FIN-90401 Oulu, Finland
Email: djordje.tujkovic@ee.oulu.fi
fax: +358 8 553 2845, tel: +358 8 553 2887

ABSTRACT

We propose a new bandwidth and power efficient signaling scheme for achieving high data rates over wide-band radio channels exploiting bandwidth efficient OFDM modulation, multiple transmit and receive antennas and large frequency selectivity offered in typical low mobility indoor environments. Due to its maximum achievable transmit diversity gain and large coding gain, space-frequency turbo coded modulation strongly outperforms other space-frequency coding schemes recently proposed in literature. We also propose a simple way of combining space-frequency coding with OFDM delay diversity as a cost-effective method for further bandwidth efficiency increase by exploiting more than two antennas at the transmitter.

1. INTRODUCTION

Due to its high bandwidth efficiency and suitability for high data rate applications, OFDM was chosen as a modulation scheme for a physical layer in the several new wireless standards, i.e. digital audio and video broadcasting (DAB, DVB) [1,2] in Europe and the three broadband wireless local area networks (WLAN) [3], European HIPERLAN/2, American IEEE 802.11a and Japanese MMAC.

Recent results in literature [4-7] demonstrate that multiple-input multiple-output (MIMO) wireless channels, apart from spatial diversity against detrimental effects of fading, enable increased information theoretic capacity as compared to single-input single-output (SISO) channels. A number of transmit diversity schemes for multi-antenna OFDM systems has been proposed recently that exploit a form of simple spatial processing at transmitter to overcome link budget limitations, moving a complexity burden from mobile terminals to access points. As seen by single antenna error control codes employed therein, the given diversity scheme over MIMO channel creates an equivalent SISO channel with characteristics, either desirably close to Gaussian [8] or with artificially increased frequency selectivity [9,10]. Therefore, potentially increased capacity of MIMO channels is not exploited in a proper way.

For the perfectly known channel state information (CSI) at both ends of a wireless link, optimal and capacity approaching signaling strategy impose the initial singular value decomposition (SVD) of MIMO channel into a number of parallel, SISO sub-channels. Single antenna error control codes, with optimal power and bit allocation, are employed then on each of parallel SISO sub-channels [7]. Sub-carrier based spatial sub-channel adaptive coding/modulation suggested in [11] results in large complexity even for a limited number of transmit antennas. Also it is not directly applicable for broadcast channels, i.e. in DAB and DVB.

When the channel state information (CSI) is not available at the transmitter, space-time coding (STC) is an optimal signaling strategy, designed to achieve potentially high capacity of MIMO Rayleigh fading channels by jointly exploiting the benefits of spatial and temporal diversity. Application of STC to space-frequency domain is however not always straightforward. For layered STC architectures [12], complexity reduction due to the applied single antenna channel codes is difficult to justify in a situation where large frequency selectivity may result in complex sub-carrier based spatial filtering at a receiver. Also the required number of receive antennas should be greater than or equal to the number of transmit antennas. Therefore, maximum likelihood detection (MLD) based STC [13,14] becomes again a cost-effective way of exploiting the frequency selectivity in the channel.

In [15], STC's from [13] were applied as space-frequency codes. Large bandwidth and power efficiency gains were reported as compared to single antenna channel codes employed with OFDM transmit diversity [10]. The concept of recursive space-time trellis codes (Rec-STTrC) for parallel-concatenated space-time turbo coded modulation (STTuCM) was introduced in [16] and further generalized in [17]. The proposed parallel concatenated scheme was designed to preserve the maximum transmit diversity gain but simultaneously enhance the coding gain as compared to STC's in [13]. In this paper, we advocate application of STTuCM on space-frequency domain and demonstrate significant performance improvements when compared to some other space-time (turbo) coding schemes applied to multi-antenna OFDM systems under somewhat realistic ITU and ETSI BRAN channel models and physical layer parameters. We also propose a simple way of combining space-frequency coding with OFDM delay diversity for cost effective exploitation of more than two transmit antennas.

2. SYSTEM MODEL

We consider system employing a $R=(L_a+1)N$ transmit and M receive antennas depicted in Fig. 1. Applied STC is

The research has been supported by Elektrobit, Nokia, Finnish Air Force, the National Technology Agency of Finland (Tekes), Academy of Finland and Graduate School of Electronics, Telecommunications and Automation (GETA).

designed for N transmit antennas and L_a is the order of artificial multi-path introduced by additional OFDM delay diversity. Delays T_{la}, l_a=1..L_a are chosen in an ascending order $T_1<T_{la}<..<T_{La}$ as multiples of the OFDM sampling period T_s. Therefore the equivalent sampling rate discrete-time channel from any of the first N transmit to any of M receive antennas can be represented with an equivalent $L=\lceil (T_{La}+T_{DS})f_s \rceil+1$ order FIR filter with filter taps $\boldsymbol{h}_{n,m}^{\ k}=[h_{n,m,0}^{\ k}.. \ h_{n,m,L}^{\ k}]$ where T_{DS} denotes the maximum delay spread in the channel.

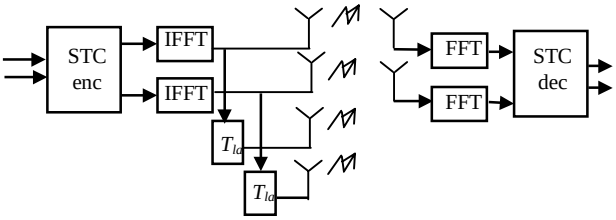


Fig. 1. Block diagram of space-frequency coded OFDM system

At each discrete time instant k, k=1..B, the input sequence of CZ bits $\boldsymbol{b}^k=[b_1^{\ k} \ b_2^{\ k}.. \ b_{CZ}^{\ k}]$ enters STC encoder where C is a number of sub-carriers in the OFDM symbol. Corresponding output of the STC encoder is a tall $C\times N$ matrix $\boldsymbol{S}^k=[\boldsymbol{S}_1^{\ k} \ \boldsymbol{S}_2^{\ k} .. \ \boldsymbol{S}_N^{\ k}]$ of coded complex symbols such that $\boldsymbol{S}_n^{\ k}=[S_{1,n}^{\ k}.. \ S_{C,n}^{\ k}]^T$ with $S_{c,n}^{\ k}$ denoting a point in complex constellation of 2^Z symbols. As in [8] let $\boldsymbol{F}=[\boldsymbol{F}_1 \ \boldsymbol{F}_2 .. \ \boldsymbol{F}_C]$, $\boldsymbol{T}_{cp}=[\mathbf{I}_{LxC}^{\ T} \ \mathbf{I}_{CxC}^{\ T}]^T$ and $\boldsymbol{R}_{cp}=[\mathbf{0}_{CxL} \ \mathbf{I}_{CxC}]$ denote the $C\times C$ fast Fourier transform (FFT) matrix, $(L+C)\times C$ cyclic prefix insertion matrix and $C\times(L+C)$ cyclic prefix removal matrix respectively. After OFDM demodulation at the receiver, complex base-band $C\times 1$ signal vector at receive antenna m can be expressed

$$\boldsymbol{r}_m^{\ k}=\sum_{n=1}^{N}\boldsymbol{D}_{n,m}^{\ k}\boldsymbol{S}_n^{\ k}+\boldsymbol{F}\boldsymbol{R}_{cp}\boldsymbol{\eta}_m^{\ k}, \quad m=1.. \ M \tag{1}$$

where $\boldsymbol{\eta}_m^{\ k}$ denotes $(C+L)\times 1$ vector of noise samples, mutually independent zero mean complex Gaussian random variables with variance σ^2 per complex dimension. Diagonal matrix $\boldsymbol{D}_{n,m}^{\ k}$ is given as $\boldsymbol{D}_{n,m}^{\ k}=\boldsymbol{F}\boldsymbol{R}_{cp}\boldsymbol{H}_{n,m}^{\ k}\boldsymbol{T}_{cp}\boldsymbol{F}^H=diag[\alpha_{n,m,1}^{\ k},.. \ \alpha_{n,m,C}^{\ k}]$ with $\alpha_{n,m,c}^{\ k}=[\boldsymbol{h}_{n,m}^{\ k} \ \mathbf{0}_{1x(C-L)}]\boldsymbol{F}_c$ and where $\boldsymbol{H}_{n,m}^{\ k}$ denotes $(C+L)\times(C+L)$ Toeplitz matrix with its (x,y) entry $h_{n,m,(x-y)}^{\ k}$. We assume in general that input information frame $\boldsymbol{b}=[\boldsymbol{b}^1.. \ \boldsymbol{b}^k.. \ \boldsymbol{b}^B]$ consists of $V=BCZ$ bits, so that one coded information frame covers multiple of B successive OFDM symbols. For the perfect knowledge of channel state information (CSI) at the receiver, maximum likelihood detection (MLD) metric for Viterbi and maximum *a posteriori* (MAP) probability decoder is given by

$$\hat{\boldsymbol{S}}=[\ \hat{\boldsymbol{S}}_1^{\ 1}..\hat{\boldsymbol{S}}_N^{\ 1}..\hat{\boldsymbol{S}}_1^{\ k}..\hat{\boldsymbol{S}}_N^{\ k}..\hat{\boldsymbol{S}}_1^{\ B}..\hat{\boldsymbol{S}}_N^{\ B}\ ] \tag{2}$$
$$=\underset{Q_1^1..Q_N^1..Q_1^k..Q_N^k..Q_1^B..Q_N^B}{\arg\min}\sum_{k=1}^{B}\sum_{m=1}^{M}\left\|\boldsymbol{r}_m^{\ k}-\sum_{n=1}^{N}\boldsymbol{D}_{n,m}^{\ k}\boldsymbol{Q}_n^{\ k}\right\|$$

where the minimization is done over all possible code-words of the space-time code used for transmission.

3. SPACE-FREQUENCY CODING

3.1 Space-Frequency Trellis Coded OFDM

Based on the large effective code length, *Lu et al.* proposed a new family of space-time trellis codes for multi-antenna OFDM systems in [18]. Codes were designed upon already existed trellis coded modulation schemes optimized for frequency flat fading channels. A class of rate 2/3 8PSK TCM for single antenna transmission was transformed into rate 2/4 QPSK code for two transmit antennas by splitting the original 8PSK mapper into two QPSK mappers, one for each transmit antenna. We refer to this space-frequency trellis code approach as SFTrC-L to distinguish between *Tarokh et al.* codes in [15] which we denote as SFTrC-T. In both cases, Viterbi decoder is used for STC decoding.

3.2 Space-Frequency Turbo Coded OFDM

In case system applies STTuCM, STC encoder and decoder in Fig. 1 are depicted in Figs. 2 and 3 respectively. We refer to [17] for detailed description of encoding and decoding operations. We only outline that component STC in Fig. 2 are recursive non-systematic space-time trellis codes (Rec-STTrC) introduced in [16]. Also interleaving actually consists of two half-length bit-wise pseudorandom interleavers. One interleaving is scrambling the input bits on the odd input symbol positions, another is independently from the first one, scrambling the input bits on the even input symbol positions. This will assure that due to puncturing each input information bit contributes once and only once to the output STTuCM code-word.

In order to enable pseudo-random bit-wise interleaving at encoder, additional symbol-to-bit reliability transformation is performed at the output of component symbol-by-symbol MAP decoders. This result in log-likelihood ratio for each information bit b_i

$$L(b_i)=Log\frac{\sum_{d_t|b_{i=1}}\Pr\{d_t=\left[b_{(t-1)Z+1}..b_{tZ}\right]\ |\boldsymbol{r}\}}{\sum_{d_t|b_{i=0}}\Pr\{d_t=\left[b_{(t-1)Z+1}..b_{tZ}\right]\ |\boldsymbol{r}\}} \tag{3}$$

for $\forall i\in\{(t\text{-}1)Z+1..tZ\}$, t=1..CB, $\boldsymbol{r}=[\boldsymbol{r}_1^{\ 1}.. \ \boldsymbol{r}_1^{\ B}.. \ \boldsymbol{r}_M^{\ 1}.. \ \boldsymbol{r}_M^{\ B}]$ being the total observation of the channel output and d_t taking values in $\{(0)_2, \ (1)_2,..(2^Z\text{-}1)_2\}$, where subscript 2 functions the Z-bits long binary representation of the value in brackets. Bit-level *extrinsic* information is now extracted

$$L_{ext}(b_i)=L(b_i)-L_{apri}(b_i) \tag{4}$$

with $L_{apri}(b_i)$ being *a priori probability* of the information bit b_i. After being bit-wise interleaved it becomes $\tilde{L}_{ext}$ and is passed through bit-to-symbol reliability transformation to compute *a priori probability* for another symbol-by-symbol MAP decoder

$$\Pr\{d_t = [b_{(t-1)Z+1}..b_{tZ}]\} = \prod_{j=1}^{Z} \frac{\exp(b_{(t-1)Z+j} \cdot \tilde{L}_{ext}(b_{(t-1)Z+j}))}{1+\exp(\tilde{L}_{ext}(b_{(t-1)Z+j}))}. \quad (5)$$

Consequently, the resultant encoder and iterative (turbo) decoder operate on bit level. We refer to above bit-wise interleaved Space-Frequency Turbo Coded Modulation scheme as SFTuCM-Dbit.

Independently, *Cui et al.* proposed somewhat similar space-time turbo coded modulation scheme in [19]. Implemented codes were recursive systematic space-time trellis codes (RecSys-STTrC), somewhat different from Rec-STTrC. It is worth nothing that the component codes are systematic so the scheme in [19] can be depicted with the same block diagrams of encoder and decoder as those in Figs. 2 and 3. The major difference between two schemes lays in the structure of interleaving. In [19] *Z*-wise or symbol level pseudo-random interleaving between two constituent codes is applied. As a consequence resultant encoder and iterative (turbo) decoder operate on symbol level. Symbol-to-bit and bit-to-symbol reliability transformations in Fig. 3 are avoided and the exchange of log-likelihood information between the two component symbol-by-symbol MAP decoders is done directly on the symbol level. Therefore the extraction of *extrinsic* information is done in the following manner

$$L_{ext}(d_t) = L(d_t) - L_{apri}(d_t) \quad (6)$$

with $L_{apri}(d_t)$ being the *a priori probability* of the information symbol d_t. We refer to the above symbol-wise interleaved Space-Frequency Turbo Coded Modulation scheme as SFTuCM-Csymb. We also consider the parallel concatenation of two 8-state RecSys-STTrC's with bit-wise interleaving and symbol-to-bit and bit-to-symbol reliability transformations in decoder. We refer to this scheme as SFTuCM-Cbit.

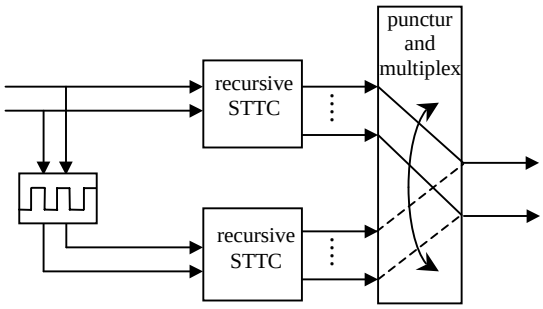


Fig. 2. Block diagram of STTuCM encoder

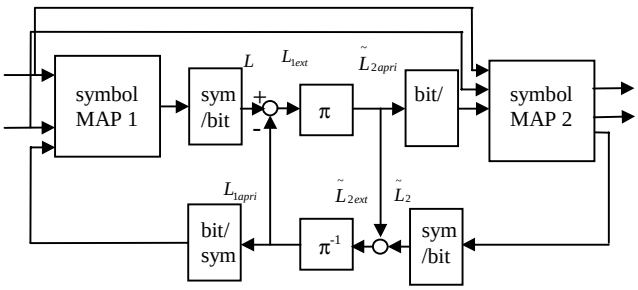


Fig. 3. Block diagram of STTuCM decoder

Finally, we applied two 8-state maximum effective code length *Lu et al.* space-time trellis codes from [18] with generator polynomials [11; 02; 04] in octal form as constituent codes in bit-wise interleaved space-time turbo coded modulation. We refer to this scheme as SFTuCM-Lbit.

4. SIMULATION RESULTS

All implemented STC's were designed to achieve spectral efficiency of 2 bits/sec/Hz using QPSK modulation and two transmit antennas (N=2). Penalty in band-with efficiency due to trellis termination is however lower with SFTuCM than with SFTrC's, because the number of tail bits is proportional to number of trellis states. We assume perfect frame and sample clock synchronization between the transmitter and the receiver. Prior to OFDM modulation at transmitter, complex code-word symbols were interleaved with length BC channel interleaving.

We adopted HIPERLAN/2 physical layer parameters [21] (the same as those for IEEE 802.11a) and evaluated performance under some specific ITU and ETSI BRAN, mainly indoor, low mobility channel models. Available bandwidth was 20 MHz with 64 sub-carriers in OFDM symbol corresponding to sub-channel separation of 3.125 kHz and OFDM frame duration of 3.2 μs. To each frame a guard period of 0.8 μs was added and a total of 48 sub-carriers were used for data transmission. Additional 4 sub-carriers were assigned for pilots though CSI was assumed to be perfectly estimated at receiver. A R=2 transmit and M=1 receive antennas were employed without optional delay diversity. Coded frame was spread across five consecutive OFDM symbols (B=5) during which fading is assumed to be quasi-static. The performance comparison between the considered schemes is depicted in Figs. 4 and 5.

In Fig. 4 the performance was evaluated on ITU-B [22], six path indoor, non-line of sight (NLOS) office channel model. The best performance is achieved with SFTuCM-Dbit which outperforms SFTuCM-Csymb and 256-state SFTrC-L by more than 2 dB and 32-state SFTrC-T by more than 4 dB at frame error rate (FER) of 10^{-2}. The performance of rather complex 256-state SFTrC-L can be achieved with lower complexity and more bandwidth efficient SFTuCM-Lbit, bit-wise interleaved parallel concatenation of two 8-state encoders of the same family. More than 2.5 dB performance loss as compared to SFTuCM-Dbit results from the fact that the large effective code length design criteria developed in [18] represent rather brutal force method not taking into account transmit diversity properties. Employing RecSys-STTrC from [19] in bit-wise interleaved manner, i.e. SFTuCM-Cbit, improves performance over symbol-wise interleaved version SFTuCM-Csymb by almost 1 dB but still suffers from more than 1 dB performance loss as compared to SFTuCM-Dbit as constituent Rec-STTrC's are better optimized for parallel concatenation than RecSys-STTrC's.

In Fig. 5 the NLOS large open space office environment ETSI BRAN-B [23] channel model with the total of 18 paths and 100nS *rms* delay spread was considered. SFTuCM-Dbit outperforms SFTuCM-Csymb and 256-state SFTrC-L by more than 2.5 dB and 32-state SFTrC-T by more than 5 dB, at frame error rate (FER) of 10^{-2}. The

performance gain of SFTuCM-Dbit over SFTuCM-Cbit is further increased to 1.5 dB.

5. CONCLUSIONS

In this paper, we proposed a bandwidth and power efficient signaling scheme for achieving high data rates over wide-band radio channels exploiting bandwidth efficient OFDM modulation, multiple transmit and receive antennas and large frequency selectivity offered in typical low mobility indoor office environments, e.g. ITU and ETSI BRAN channel models. Due to its maximum achievable transmit diversity gain and large coding gain, space-frequency turbo coded modulation strongly outperforms other space-frequency coding schemes recently proposed in literature. We have demonstrated that space-frequency turbo coded modulation owes its good performance to mainly two important features. Relatively simple 8-state recursive space-time trellis codes are optimized for both, multi-antenna transmission and parallel concatenation. Another distinctive feature is the bit-wise interleaving between two constituent codes. We also proposed a simple way of combining space-frequency coding with OFDM delay diversity for cost effective exploitation of more than two transmit antennas.

ACKNOWLEDGMENT

The authors gratefully acknowledge Esa Kunnari, Torsti Poutanen, Reijo Savola, Pavel Loskot and Ulrico Celentano for their helpful comments and discussions.

REFERENCES

[1] ETSI ETS 300 401, “Radio broadcasting systems; digital audio broadcasting (DAB) to mobile portable and fixed receivers”, Feb.1995.

[2] ETSI ETS 300 744, “Digital video brodcasting (DVB); frame structure, channel coding and modulation for digital terrestrial television (DVB-T)”, Mar.1997.

[3] R. van Nee, G. Awater, M. Morikura, H. Takanashi, M. Webster, K. W. Halfold, “New high rate wireless LAN standards”, *IEEE Comm. Mag*., pp. 82-88, Dec.1999.

[4] G. J. Foschini Jr, M. J. Gans, “On limits on wireless communication in a fading environment when using multiple antennas”, *Wireless Personal Communication*, March 1998.

[5] E. Telatar, “Capacity of multi-antenna Gaussian channels”, *Euro. Trans. Comm.* , vol.10, No. 6, November-December 1999.

[6] G. Raleigh, J. M. Cioffi, “Spatio-Temporal Coding for Wireless Communication”, *IEEE Trans. Comm*. Vol 46, No 3, March 1998.

[7] G. Caire, G. Taricco, E. Biglieri, “Capacity of multi-antenna block fading channels”, in *Proc. ICC 1999*, Vancuver, Canada, June 11-15, 1999.

[8] Z. Liu, G. B. Giannakis, S. Zhou, B. Muquet, “Space-time coding for broadband wireless communications”, *Wirel. Commun. Mob. Comput*., vol 1, No. 1, Jan-Mar 2001.

[9] Y. Li, J. C. Chuang, N. R. Sollenberger., “Transmitter diversity for OFDM systems and its impact on high-rate data wireless networks”, *JSAC*, Volume: 17 7 , July 1999, Page(s): 1233 –1243.

[10] L. J. Cimini, D. Babak, N. R. Sollenberger, “Clustered OFDM with transmitter diversity and coding”, In *Proc GLOBECOM'96*, London, UK, November 18 – 22, 1996.

[11] K. Wong; S. Lai; R.S. Cheng; K.B. Letaief; R.D. Murch, “Adaptive spatial subcarrier trellis coded MQAM and power optimization for OFDM transmission”, in *Proc. VTC 2000-Spring*, Tokyo, Japan, 2000.

[12] D. Shiu, J. M. Kahn, “Scalable layered space-time codes for wireless communications: Performance analysis and design criteria”, In *Proc WCNC 1999*, New Orleans, USA, September 22-24, 1999.

[13] V. Tarokh, N. Seshadri, A. R. Calderbank, ”Space-time codes for high data rate wireless communication: performance criterion and code construction”, *IEEE Trans. Inf. Th.*, vol. 44, no. 2, March 1998.

[14] J. Grimm, M. P. Fitz, J. V. Krogmeier, “Further results on space-time coding for Reyleigh fading”, presented at *36th Annual Allerton conf. on Commun., Control and Computing*, Monticello, USA, September 23 – 25, 1998.

[15] D. Agrawal, V. Tarokh, A. Naguib, N. Seshadri, “Space-time coded OFDM for high data-rate wireless communication over wide-band channels”, in *Proc. VTC'98*, Ottawa, Canada, May 18 – 22, 1998.

[16] Dj. Tujkovic, “Recursive space-time trellis codes for turbo coded modulation”, in *Proc. IEEE GLOBECOM'00*, San Francisco, USA, November 27 – December 1, 2000.

[17] Dj.Tujkovic, "High Bandwidth Efficiency Space-Time Turbo Coded Modulation", accepted for publication at *ICC 2001*, Helsinki, Finland, June 11-15, 2001.

[18] B. Lu, X. Wang, “Space-time coding design in OFDM systems”, in Proc. *IEEE GLOBECOM'00*, San Francisco, USA, November 27 – December 1, 2000.

[19] D. Cui, A. M. Haimovich, “Design and performance of turbo space-time coded modulation”, in Proc. *IEEE GLOBECOM'00*, San Francisco, USA, November 27 – December 1, 2000.

[20] Dj. Tujkovic, “Performance analysys and code optimization for space-time turbo coded modulation“, in preparation.

[21] ETSI TS 101 475 V1.1.1 (2000-04), “Broadband Radio Access Networks (BRAN), HIPERLAN Type 2, Physical (PHY) layer“.

[22] ITU Proposal for WCDMA, channel models.

[23] ETSI EP BRAN 30701F, “Criteria for Comparison“, R. Kopmeiners, P. Wijk, May 1998.

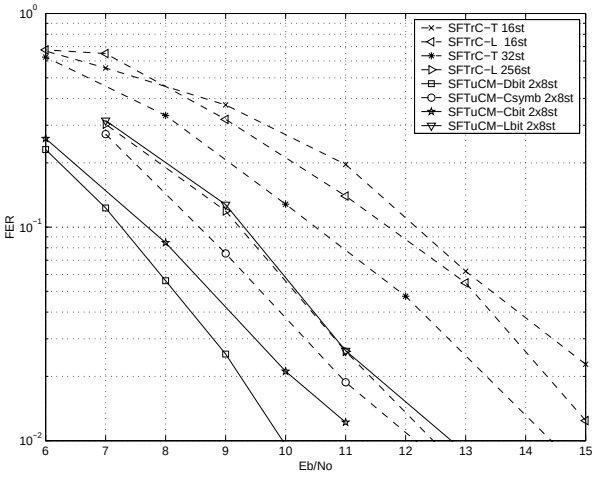


Fig. 4. ITU-B, six path indoor office, NLOS

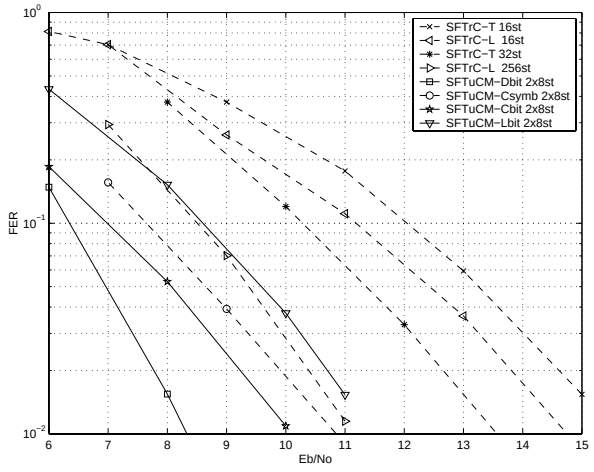


Fig. 5. ETSI BRAN-B, 18 path typical large open space office environment, NLOS

Action Potential Analysis by Real Time DSP Hardware and Software for Odour Exposure Responses

Matti Huotari, Vilho Lantto,
University of Oulu,
Microelectronics and Materials Physics Laboratories
P.O.Box 45000, FIN-90014 OULU UNIVERSITY
FINLAND
Tel. +358-8-5532728, Fax +358-8-2720

ABSTRACT

Tools for assorting and classifying neuronal action potentials can be either on-line or off-line. Here we examine on-line tools. We concentrate on an analysis of insect action potentials by using a DSP-based multi spike detector (MSD© by Alpha Omega Engineering). The MSD software is a modular program which is user-friendly and has optimal speed. In the action potential analysis neural firings are decomposed into waveforms, which are specific for each cell. Based on this comparative analysis and assorting of action potentials, many series of action potentials from the olfactory receptor neurones of small insects were analysed. The assorting algorithm compares the neural signals to three templates, which can be adaptively changed and fixed by the user. The action potentials are produced either spontaneously or at an odour exposure. In the response analysis they are further plotted as a function of the odour exposure sequence.

1. INTRODUCTION

Detection of action potentials (APs, spikes), which are large in relation to the electrode or membrane and other biological noise is unproblematic in sensory neurophysiology. There are two common methods for APs detection. However, until recently DSP-based devices have been the option of choice for AP assorting and further analysis. In the development of task-specific devices, it seems feasible to utilize an efficient on-line AP detector based on DSP hardware and its compatible software.

In the analysis of biological functions, it is important to distinguish exactly APs generated by other cells in the close vicinity of the cell from the primary olfactory receptor neurons (ORN) of the cell. There is free software available, which can separate action potentials from three different sources off-line and classify an action potential rate into three classes depending on the shape of each action potential. However, a separation into two classes was successful enough in practice. The recorded action potentials can be stored for further analysis and printing.

In the MSD, detection and assorting are simultaneous in each channel, which makes it possible to detect the interspike interval periods or action potential firing rates for each sensory neuron. A match between the AP and its template is found, when the local minimum of the sum of squared deviations between the template and AP is within the given criteria limit, and this means a detection, as reported in Fig.1 for three cells. The left window shows that the all AP matched, in the second window shows by the green color that there are double or triple match happened and in the third window three double match.

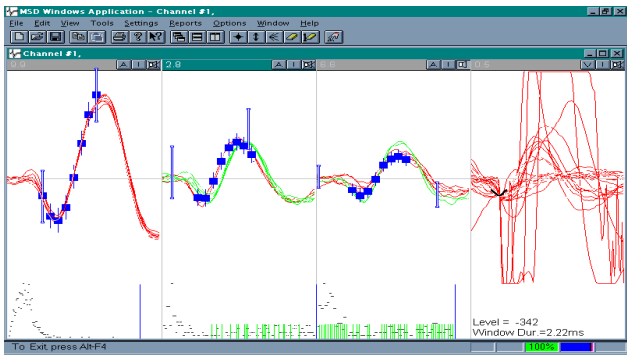


Figure 1. Three action potential shapes in the first three windows together with the laboratory noise in the fourth window. The match histogram shown by green sticks develops on the window bottom.

Fig. 2 shows action potential rates in pps (pulses per second) from two ORNs. A neuron is firing action potentials up to ca. 75 s around 45 - 75 pps and sometimes over, while the other neuron is firing about 5 pps continuously active without any break.

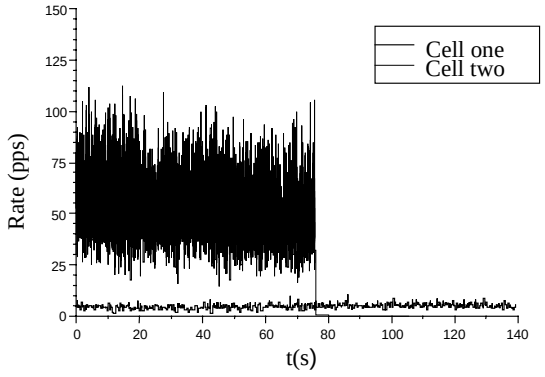


Figure 2. A high-rate action potential series together with a slow-rate at the bottom.

Before a response analysis of ORNs, it is important to

monitor the spontaneous activity of an ORN. The monitoring yields important information about the ORNs' activity. Therefore, our recordings last longer (up to 2 min) before and after the exposure to an odour.

2. MSD SOFTWARE OPERATIONS

Based on the differences in AP shapes obtained in single cell recordings from olfactory receptor neurons the APs can be assorted by means of the MSD hardware and software. It assorts up to 3 AP shapes per electrode signal. In a cascade it is possible to assort up to 3 multiplied channel numbers.

2.1. General Operations

The MSD environment is an advanced system designed to detect action potentials such as signals and assort them according to their shape. After detecting and sorting the AP shapes, the results are reported, allowing the user to measure the action potential rate, firing times, delays or construct histograms.
In the MSD software is based on modular sections. Its parameters, such as sampling rate, histogram bin, adaptive factor, double match, and polarity flag, can be changed through the dialog box in each module. Unsorted APs are displayed in two modes: All APs (like memory scope) or the last N APs (N=50 … 100) are displayed. A detection is indicated when the template and the corresponding AP match. APs are assorted simultaneously and saved on the basis of their match to each template. A typical AP pattern or template is defined to generate an optimal template for detecting and assorting the APs.

2.2. MSD Operations in AP analysis

The MSD software includes step by step instructions for the user to define clusters of APs, how to detect and isolate APs, the allowed noise level, and how to monitor the quality of multi-signal detection and isolation. The MSD hardware can filter the AP signal internally by a single-pole high-pass filter having a -3dB point at 200 Hz, and by a double-pole low-pass filter having a -3 point at 10 kHz. Because of this filtering the recorded signal bandwidth is also limited for the APs in the same or narrower band. However, this band contains all the information that is characteristic enough in order to assort AP signals.
The template definition program of the MSD software has three steps. In the first step, the input signal is displayed on the PC screen with the adjusted trigger level. In the second step, three groups of the APs are defined by the user. In the third step, the program computes and displays templates for all the selected AP groups. After these steps, the user switches to AP detection mode.
In the MSD software two separate programs cooperate together. One part operates the DSP, while the other operates the PC. The PC downloads a portion of the software from the logic board to the program-space. This program waits until a new value of the sampled input data is available. The new sample is transferred from the analog board into a buffer holding the last 100 samples. The first eight samples are compared to the first template and the sum of squared differences is computed. This sum is compared to the previous one. The minimum is reported by raising a flag in the program response, if this sum is a minimum and below the defined threshold.
In the PC the program examines if any data is available in the buffer. If the data is an AP, it displays it on the PC screen with appropriate options to define a new template if necessary.
In practice the MSD software functions as follows. Ongoing detected AP activity is displayed in three frames along with their eight template points and a line cursor. An additional graph develops at the bottom of each frame (Fig. 1). This graph contains an updated histogram, which displays the distribution of the sum of squared differences between the ongoing AP activity and the corresponding template. This sum of squared differences is called *spike distance*. The left point of the histogram is the zero distance or perfect match and it is used as a detection threshold. When the AP has its spike distance smaller than the detection threshold in two or three templates, a double or triple match is in process. These APs are shown in different colors, and the histogram is updated to show assessment of the situation.
In MSD operations it is recommended to assort as many AP shapes as possible in order to get used to the program's advantages and limits. In addition, one can develop his/her own strategy of satisfactory operations. The templates can be modified online without interrupting the sorting. Action potentials, which are detected and assorted the authentic spikes as they come and produces a 100 ms pulse following every real AP occurrence.
Action potential acquisition continues uninterrupted. The MSD software allows handling of AP intervals up to a memory size makes it possible.

3. DATA COLLECTION

We have reported elsewhere on several aspects of the AP responses to the odour exposures of individual olfactory receptor neurons [2]. Shortly, AP data were obtained from blowflies (*Calliphora vicina*), mosquitoes (*Culex pipiens*) and fruitflies (*Drosphila virilis*). The MSD collects the data on line. The analyzed data can be exported to other programs such as Excel™, Origin™ or MatLab™.

3. RESULTS

In the case of the blowfly ORNs the following histograms originate from two different olfactory receptor neurons, Figure 3. The histograms show that these ORNs have different distributions of AP intervals. One has longer intervals, while the other one has a normal distribution. The distribution function was fitted in the Origin™

software at the normal (red line) and Gaussian distribution (black line). The normal curve faithfully follows the measured values.

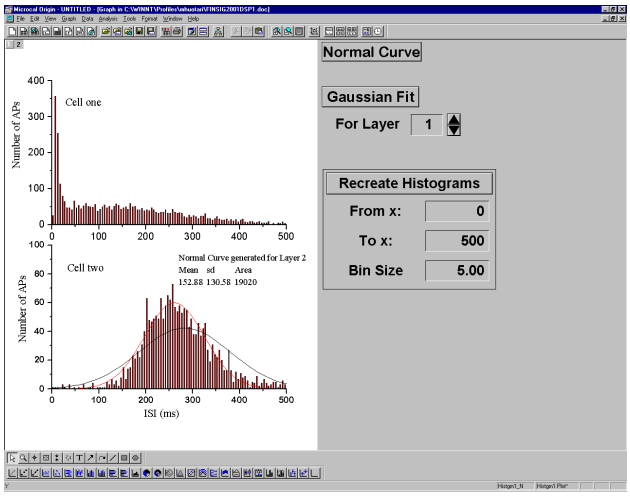


Figure 3. High-rate and a low-rate AP histograms of two ORNs.

5. CONCLUSIONS

Many studies of AP firing require the isolation of the AP generated by an ORN. The on-line peak detection method described in this paper facilitates the characterization of firing patterns during the recording of AP activity, for instance in ORNs. It makes it possible to extract information about the occurrence of APs from multiple sources. Unlike other much used techniques, the MSD method is user-friendly, economic, fast and its parameters are quantitatively defined in real time. In addition, this software and hardware do not require the use of a particularly powerful computer. This system provides the user with a whole set of graphical and statistical analyses of AP interval data, such as distribution, correlation and other functions.

Because AP trains can theoretically carry information in several ways (firing rate, interval distributions), it is important to record and save the actual AP interval history. These constitute the first step in any dynamic analysis, such as phase portraits, joint-interval distribution histograms and phase space of divided differences. The dynamic analysis makes it possible to make further evaluations on topological dimensions. On the basis of the histogram analysis ORNs were assigned to ten histogram classes. These AP interval histograms of each ORN class exhibit an obvious mathematical distribution function that is unambiguous.

Because the AP analysis is done partly by the researcher and partly by the computer, it is impossible to define any exact time or frequency characteristics for analysis purposes but we use common sense and utilize what we have learned.

ACKNOWLEDGMENT

We acknowledge the Tauno Tönning Foundation for the scholarship to MH.

REFERENCES

[1] MSD Multi-Spike Detector User's Manual Version 3.20, Alpha Omega Engineering, Biomedical Division, Nazareth, Israel.

[2] Matti j. Huotari, Biosensing by insect olfactory receptor neurons Sensors and Actuators B 71, 212-222, 2000.

Time-Varying ARMA modelling of Nonstationary EEG using Kalman Smoother Algorithm

Mika P. Tarvainen[1,2]†, Perttu O. Ranta-aho[1,2], and Pasi A. Karjalainen[1]

[1]University of Kuopio
Department of Applied Physics
P.O.Box 1627, FIN-70211 Kuopio
FINLAND
† Tel. +358-17-162584, Fax: +358-17-162585
† E-mail: Mika.Tarvainen@uku.fi

[2]Kuopio University Hospital
Department of Clinical Neurophysiology
P.O.Box 1777, FIN-70211 Kuopio
FINLAND

ABSTRACT

An adaptive autoregressive moving average (ARMA) modelling of nonstationary EEG by means of Kalman smoother is presented. The main advantage of the Kalman smoother approach compared to other adaptive algorithms such as LMS or RLS is that the tracking lag can be avoided. This advantage is clearly presented with simulations. Kalman smoother is also applied to tracking of alpha band characteristics of real EEG during an eyes open/closed test. The observed tracking ability of Kalman smoother, compared to other methods considered, seemed to be better.

1 INTRODUCTION

In the analysis of nonstationary EEG the interest is often to estimate the time-varying spectral properties of the signal. A traditional approach to this is the spectrogram method, which is based on Fourier transformation. Disadvantages of this method are the implicit assumption of stationarity within each segment and the rather poor time/frequency resolution. A better approach is to use parametric spectral analysis methods based on e.g. time-varying autoregressive moving average (ARMA) modelling. The time-varying parameter estimation problem can be solved with adaptive algorithms such as least mean square (LMS) or recursive least squares (RLS). These algorithms can be derived from the Kalman filter equations [1], [2].

In this paper we use the Kalman smoother algorithm in tracking of nonstationary properties of EEG. Kalman smoother is compared to LMS and RLS algorithms in tracking of alpha band characteristics of EEG measured during an eyes open/closed test. The Kalman smoother approach is also applied to the detection of alpha waves of EEG. The main advantage of the Kalman smoother algorithm compared to other adaptive algorithms is the fact that the tracking lag can be avoided. This is demonstrated with simulations. Kalman filter has been previously used in EEG analysis in e.g. [3], [4], [5].

2 METHODS

If the signal to be modelled is nonstationary it cannot be modelled as an output of a time-invariant system. It is natural in this case to assume that the system has time-varying parameters.

2.1 Time-varying linear regression

Here we use the time-varying autoregressive moving average ARMA(p,q) model for the signal

$$z(t) = -\sum_{j=1}^{p} a_j(t) z(t-j) + \sum_{k=1}^{q} b_k(t) e(t-k) + e(t) \quad (1)$$

where $a_j(t)$ and $b_k(t)$ are the time-varying ARMA parameters and $e(t)$ is the driving white noise process. By denoting

$$\theta_t = (-a_1(t), \ldots, -a_p(t), b_1(t), \ldots, b_q(t))^T \quad (2)$$

$$\varphi_t = (z(t-1), \ldots, z(t-p), e(t-1), \ldots, e(t-q))^T \quad (3)$$

the model can be written in the form

$$z_t = \varphi_t^T \theta_t + e_t \quad (4)$$

where $z_t = z(t)$ and $e_t = e(t)$. This is clearly a linear observation model, with φ_t^T being the observation matrix and e_t being the observation error. A typical description for the parameter variation when no *a priori* information is available, is the random walk model [6]. Thus for the parameters θ_t we write a state equation of the form

$$\theta_{t+1} = \theta_t + w_t \quad (5)$$

where w_t is a noise process. Equations (4) and (5) form a specific form of the general state space equations, with the input process w_t. Now the problem is to estimate the time-varying parameters θ_t, according to the state space model.

2.2 Kalman filter

The Kalman filtering problem is to find the minimum mean square estimator $\hat{\theta}_t$ for state θ_t given the observations $z_1, \ldots, z_t$. This has been shown to be equal to

the conditional expectation value

$$\hat{\theta}_t = E\{\theta_t | z_1, \ldots, z_t\} \quad (6)$$

We assume here the state and measurement noises w_t and e_t to be uncorrelated, zero mean, random processes with covariance matrices $C_{w_t} = \sigma_w^2 I$ and $C_{e_t} = \sigma_e^2 I$, so that the individual parameter evolutions are assumed to be independent. The initial state θ_0 is assumed to be uncorrelated with e_t and w_t with finite variance. The Kalman filter equations can be written in the form

$$\hat{\theta}_{t|t-1} = \hat{\theta}_{t-1} \quad (7)$$

$$C_{\tilde{\theta}_{t|t-1}} = C_{\tilde{\theta}_{t-1}} + C_{w_{t-1}} \quad (8)$$

$$K_t = C_{\tilde{\theta}_{t|t-1}} \varphi_t \left(\varphi_t^T C_{\tilde{\theta}_{t|t-1}} \varphi_t + C_{e_t} \right)^{-1} \quad (9)$$

$$C_{\tilde{\theta}_t} = \left(I - K_t \varphi_t^T \right) C_{\tilde{\theta}_{t|t-1}} \quad (10)$$

$$\epsilon_t = z_t - \varphi_t^T \hat{\theta}_{t|t-1} \quad (11)$$

$$\hat{\theta}_t = \hat{\theta}_{t|t-1} + K_t \epsilon_t \quad (12)$$

where $\hat{\theta}_{t|t-1}$ is the mean square estimator for state θ_t given the observations $z_1, \ldots, z_{t-1}$, $\tilde{\theta}_t$ is the state estimation error $\tilde{\theta}_t = \theta_t - \hat{\theta}_t$ and K_t is the Kalman gain matrix. The adaptation of the filter is primarily affected by C_{w_t}.

2.3 Fixed-interval smoother

The fixed-interval smoothing problem is to determine estimates

$$\hat{\theta}_{t|T} = E\{\theta_t | z_1, \ldots, z_T\} \quad (13)$$

for fixed T and for all t in the interval $1 \leq t \leq T$. The solution for this can be written in the form [7]

$$\hat{\theta}_{t-1|T} = \hat{\theta}_{t-1} + A_{t-1} \left(\hat{\theta}_{t|T} - \hat{\theta}_{t|t-1} \right) \quad (14)$$

$$A_{t-1} = C_{\tilde{\theta}_{t-1}} C_{\tilde{\theta}_{t|t-1}}^{-1} \quad (15)$$

where A_{t-1} includes the error covariances stored in the forward run of Kalman filter. Also the state estimates $\hat{\theta}_t$ and $\hat{\theta}_{t|t-1}$ need to be stored. The smoothed estimates $\hat{\theta}_{t-1|T}$ are then obtained by running the stored estimates backwards in time by taking $t = T, T-1, \ldots, 2$. The initialization is evidently with the filtered estimate.

2.4 Spectral estimation

Once the time-varying coefficients of the ARMA(p,q) model (1) are solved the time-varying power spectral density (PSD) estimation can be obtained in the terms of the estimated coefficients

$$P_t(\omega) = \sigma_e^2(t) \frac{|1 + \sum_{k=1}^{q} b_k(t) e^{-i\omega k}|^2}{|1 + \sum_{j=1}^{p} a_j(t) e^{-i\omega j}|^2} \quad (16)$$

where $\sigma_e^2(t)$ is the prediction error variance. After the adaptive algorithm, used to estimate the time-varying ARMA parameters, converges power spectrum can be calculated for each time instant.

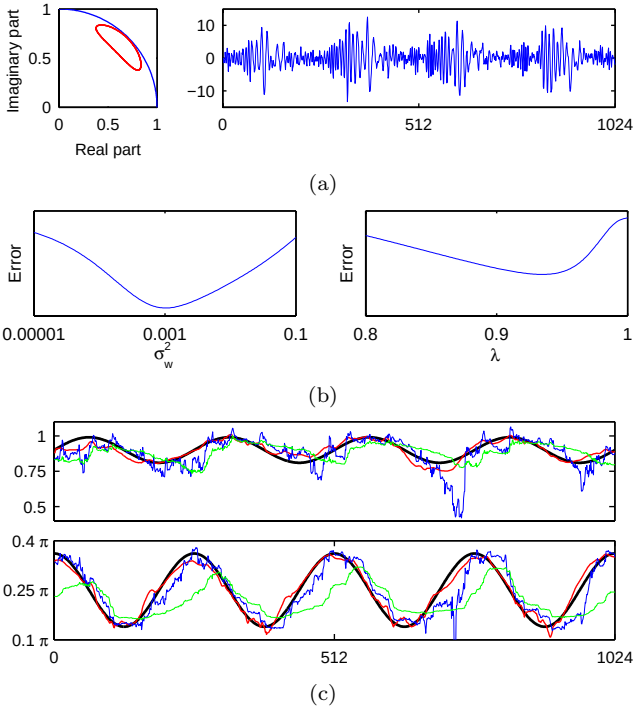


Fig. 1. AR(2) process estimation with RLS and Kalman smoother algorithms. The root evolution and the realization are presented in block (a). Both algorithms were optimized (b). Optimal value for the state noise covariance coefficient of the Kalman smoother was $\sigma_w^2 = 0.001$ and the forgetting factor of RLS was $\lambda = 0.935$. The estimates of the modulus and phase angle of the root are shown in block (c). The true values (black), Kalman smoother estimates (red) and optimal RLS estimates (blue). The smoother RLS estimates (green) were calculated by using $\lambda = 0.98$.

3 RESULTS

In order to evaluate the performance of the Kalman smoother algorithm we conduct two simulations, where Kalman smoother is compared to the popular forgetting factor RLS algorithm. Finally the Kalman smoother is applied to time-varying spectrum estimation of real EEG and for alpha wave detection.

3.1 Simulations

In the first simulation a time-varying signal was generated as an AR(2) process. The root evolution and a typical realization are presented in Fig. 1 (a). The modulus and phase angle of the root were estimated with Kalman smoother and RLS algorithms. Parameters controlling the adaptation were optimized in both algorithms to obtain the minimum error in AR coefficient estimation. The estimation errors as a function of adaptation parameters for both algorithms are presented in Fig. 1 (b). The estimates are shown in Fig. 1 (c).

RLS estimates with the optimal value for the forgetting factor have only a small tracking lag but the estimates are far more unstable compared to the Kalman smoother estimates. By increasing λ RLS estimates become more stable but the tracking lag increases at

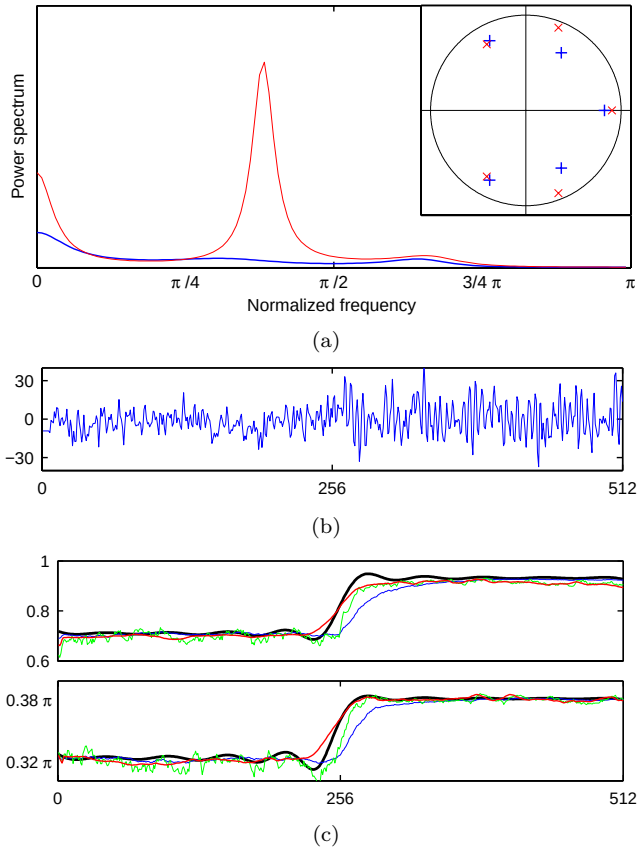


Fig. 2. A realistic simulation of EEG transition as an AR(5) process. (a) The roots and the corresponding spectra before (blue) and after (red) the transition. (b) A typical realization of the process. Averaged estimates over 100 realizations of the modulus and phase angle of the root corresponding to alpha activity are presented in (c), where true values (black), Kalman smoother estimates (red) and RLS estimates (blue/green) are shown. The state noise covariance coefficient of the Kalman smoother was $\sigma_w^2 = 8 \cdot 10^{-5}$ and the forgetting factors of RLS were $\lambda_1 = 0.98$ (blue) and $\lambda_2 = 0.9$ (green).

the same time. This simulation shows clearly the advantages of the Kalman smoother compared to the RLS algorithm. However not much can be said about the performance of the Kalman smoother in tracking of nonstationary EEG based on this simple simulation. Hence we aim to a more realistic simulation of EEG.

In many cases we are interested in tracking of narrow band characteristics of the EEG signal. One such case is the event related desynchronization/synchronization (ERD/ERS) of alpha waves. The occipital EEG recorded while patient having eyes closed shows high intensity in the alpha band (7-13 Hz). With the opening of the eyes this intensity decreases or even vanishes. It can be assumed that EEG exhibits a transition from a stationary state to another. Such a transition was here simulated as an AR(5) process. The roots of the system for both stationary states (obtained from real EEG measurements) and the corresponding power spectrums are presented in Fig. 2 (a).

In order to make the simulation more realistic abrupt transitions of AR coefficients were smoothed as described in [8]. A typical realization of the simulated

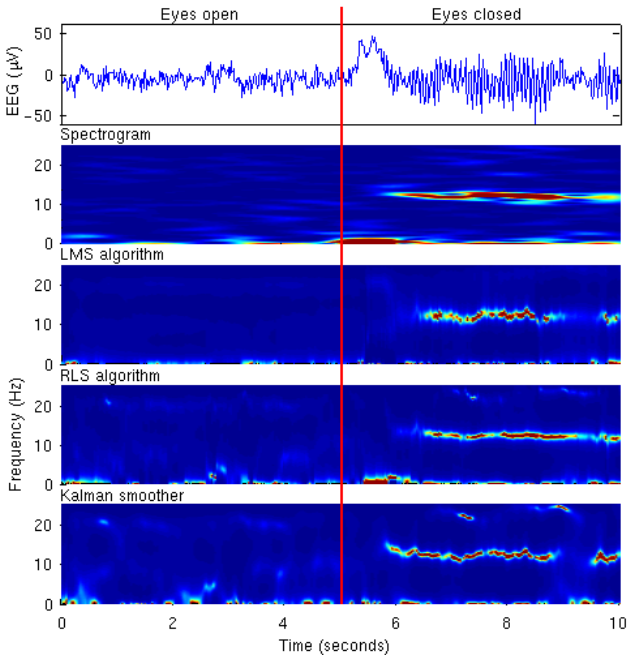


Fig. 3. Time-varying spectral analysis of ERD/ERS test of alpha waves of EEG. The measured EEG from channel O2 is shown on the topmost axis. The time window used in the spectrogram was 2 seconds. The step size of LMS algorithm was $\mu = 0.0002$ while the forgetting factor of RLS was $\lambda = 0.95$. The state noise covariance coefficient of the Kalman filter was $\sigma_w^2 = 0.0003$.

AR(5) process is presented in Fig. 2 (b). Results of tracking the alpha band characteristics are presented in Fig. 2 (c), where averaged estimates over 100 realizations of the phase angle and the magnitude of the root corresponding to alpha activity are presented. In order to obtain as smooth estimates with RLS as is obtained with Kalman smoother the forgetting factor λ must be quite small. However this leads to substantial tracking lag. With larger values of λ the tracking lag can be attenuated, but estimates become now more unstable.

3.2 ERD/ERS of alpha waves of EEG

The eyes open/closed test is a typical application of testing the desynchronization/synchronization of alpha waves of EEG. One such transition from desynchronized state to synchronized state is presented in Fig. 3. The performance of the Kalman smoother in tracking of alpha band characteristics is compared to most commonly used adaptive algorithms RLS and LMS and also to the traditional spectrogram method. An ARMA(6,2) model was used in all adaptive algorithms. The length of the time-window used in spectrogram was 2 seconds, which is long enough when considering the frequencies of the alpha band (7–13 Hz). The step size of the LMS algorithm was $\mu = 0.0002$ and the forgetting factor of RLS was chosen to be $\lambda = 0.95$ resulting in quite stable estimates and still rather fast adaptivity. The state noise covariance coefficient of the Kalman smoother was $\sigma_w^2 = 0.0003$.

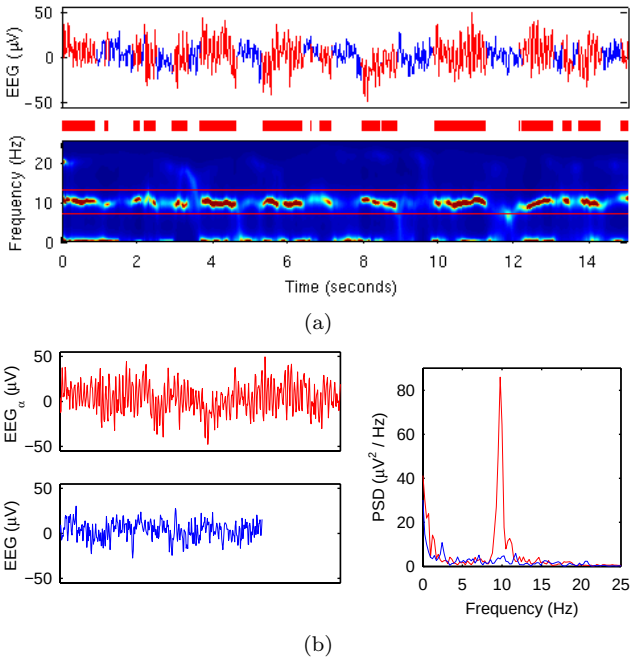


Fig. 4. Kalman smoother applied to alpha rhythm detection. (a) An EEG sample of 15 seconds from channel O2 measured from subject having eyes closed and the corresponding time-varying PSD. Detection is based on thresholding the power integral over the alpha band (7–13 Hz) with a threshold of 10 μV^2/Hz. Block (b) presents PSD estimates (calculated with traditional FFT based method) for the signals obtained by concatenating the EEG epochs where alpha activity was detected (red) or not detected (blue).

The tracking speed of the Kalman smoother seems to be fastest and an interesting gap in alpha rhythm is observed after 9 seconds. The contents of this kind of gaps is considered more closely in Fig. 4.

3.3 Detection of alpha rhythm of EEG

The aim of automatic EEG analysis is often the detection of certain waveforms. Hence the performance of the Kalman smoother on detection of alpha waves of EEG is considered here. Fig. 4 (a) presents a time-varying spectrum for an EEG sample of 15 seconds measured from healthy subject having eyes closed. Alpha wave detection was obtained by thresholding the power integral over the alpha band (7–13 Hz). The threshold was set to 10 μV^2/Hz. The performance of the alpha detector was verified by concatenating the EEG epochs where alpha waves were detected and those were no detection was made. The PSD estimates, calculated with a traditional FFT based periodogram method, for these concatenated signals are presented in Fig. 4 (c) verifying the absence of alpha rhythm in the lower concatenated signal.

4 DISCUSSION

The Kalman smoother algorithm was applied to tracking of nonstationary EEG. The performance of Kalman smoother in tracking of alpha band characteristics seemed to be most reliable compared to LMS and RLS algorithms. Kalman smoother was also applied to the detection of alpha waves of EEG with success. Also two simulations were conducted showing clearly the main advantages (smooth estimates without tracking lag) of Kalman smoother compared to other adaptive algorithms. The implementation and usability of the Kalman smoother approach are straightforward. The adaptation rate is adjusted simply by setting the state covariance coefficient σ_w^2.

One problem in modelling the data with adaptive algorithms is the selection of the model order. For time-invariant systems there exist various criteria for selecting the model order [9]. All these criteria are based on the compromise between model fit and model complexity. Also in the time-varying case there exist some criteria for selecting the model order. For example in [10] the use of Akaike's information criterion (AIC) was justified in the time-varying case under certain conditions. However in the case of tracking alpha rhythm of EEG the ARMA model of order $p = 6$ and $q = 2$ seems to be suitable. The same model order was also used in [11], [12].

REFERENCES

[1] L. Ljung, "General structure of adaptive algorithms: adaptation and tracking," Tech. Rep. LiTH-ISY-I-1294, Linköping University, December 1991.

[2] P. Karjalainen, "Estimation theoretical background of root tracking algorithms with application to EEG." Phil. Lic. thesis, Univ. of Kuopio, Dept. of Applied Physics, 1996.

[3] T. Bohlin, "Analysis of EEG signals with changing spectra using a short-word Kalman estimator," *Math Biosci*, vol. 35, pp. 221–259, 1977.

[4] A. Isaksson and A. Wennberg, "Spectral properties of non-stationary EEG signals, evaluated by means of Kalman filtering: Application examples from a vigilance test," in *Quantitative Analytic Studies in Epilepsy* (P. Kellaway and I. Petersen, eds.), pp. 389–402, Raven Press, 1976.

[5] A. Isaksson, A. Wennberg, and L. Zetterberg, "Computer analysis of EEG signals with parametric models," *Proc IEEE*, vol. 69, pp. 451–461, 1981.

[6] S. Haykin, *Adaptive Filter Theory*. Englewood Cliff, NJ: Prentice-Hall Inc., 1986. haussa.

[7] B. Anderson and J. Moore, *Optimal Filtering*. Prentice Hall, 1979.

[8] J. Kaipio and P. Karjalainen, "Estimation of event related synchronization changes by a new TVAR method," *IEEE Trans Biomed Eng*, vol. 44, no. 8, pp. 649–656, 1997.

[9] F. Gustafsson and H. Hjalmarsson, "Twenty-one ML estimators for model selection," *Automatica*, vol. 31, pp. 1377–1392, 1995.

[10] F. Kozin and F. Nakajima, "The order determination problem for linear time-varying AR models," *IEEE Trans Automat Contr*, vol. AC-25, pp. 250–257, 1980.

[11] L. Patomäki, J. Kaipio, P. Karjalainen, and M. Juntunen, "Tracking of nonstationary EEG with the polynomial root perturbation," in *EMBS'96*, 1996.

[12] J. Kaipio, P. Karjalainen, and M. Juntunen, "Perturbation expansions in polynomial root tracking," *Signal Processing*, vol. 80, pp. 515–523, 2000.

Pipeline Architecture for 8×8 IDCT with Fixed-Point Error Analysis

Jari A. Nikara, Rami J. Rosendahl, and Jarmo H. Takala

Tampere University of Technology
Digital and Computer Systems Laboratory,
P.O.B. 553, FIN-33101 TAMPERE, FINLAND
E-mail: jari.nikara@tut.fi, rami.rosendahl@tut.fi, jarmo.takala@tut.fi

ABSTRACT

In this paper, a sequential architecture for 8×8 inverse discrete cosine transform (IDCT) based on row-column decomposition is described. The sequential one-dimensional IDCT kernel is derived by utilizing vertical projection to fast IDCT algorithm. The matrix transposition network is realized with a register-based sequential permutation network and the resulting modular two-dimensional architecture can be freely pipelined. Moreover, the accuracy of the proposed architecture is analyzed in order to fulfil the IEEE standard for 8×8 *IDCT.*

1. INTRODUCTION

Discrete cosine transform (DCT) and its inverse (IDCT) are widely used tools in digital signal processing. Several architectures for DCT and/or IDCT implementations have been proposed for multimedia purposes. Typically high speed operation is achieved with the aid of parallelism. In principle, parallel architectures can be developed by exploiting inherent spatial and/or temporal parallelism in fast algorithms for DCT and IDCT. However, such algorithms are often irregular, which may limit the exploitation level of the parallelism. In addition to high data rates, the accuracy of the implementation is important; e.g., IEEE Standard 1180-1990 [1] defines accuracy requirements for two-dimensional 8×8 IDCT implementations.

Direct mapping of algorithm will result in architecture with both spatial and temporal parallelism. In general, the cost of the implementation should be low, i.e., the resources, especially number of arithmetic units, in the architecture should be minimized. Exploitation of spatial parallelism results in column architectures where operands are fed into the architecture in parallel. The arithmetic units are recursively used to compute the entire transform. Exploitation of temporal parallelism, in turn, results in pipeline architectures (or systolic array) where data is fed into the architecture sequentially. For this purpose the linear array processor approach described in [2] can be used.

In this paper, a sequential two-dimensional IDCT architecture is presented utilizing the principles used in architectural derivation of fast Fourier transform [2]. Vertical projection is applied to signal flow graph of IDCT, which results in cascaded one-dimensional IDCT architecture. The row-column decomposition is used for constructing two-dimensional transform. The required matrix transposition network is sequential and register-based with optimal number of registers. Due to the loop free structure, the architecture can be freely pipelined for improving throughput. Furthermore, the internal word width requirements are determined and analyzed for reaching the IEEE standard [1].

2. ARCHITECTURE

Architectural derivation is based on rescheduled constant geometry DCT algorithm of type II presented earlier in [3]. Since the DCT is orthogonal transform, the corresponding signal flow graph of IDCT in Fig. 1 is achieved by transposing the signal flow graph of the DCT. In addition, the signal flow graph is flipped in order to have the operands for each operation available when the result is needed. Such an arrangement offers an advantage in serial realization; every sample is not delayed in implementation thus decreasing the latency. The coefficients d_i can be generated recursively as

$$d_1 = \sqrt{\frac{1}{2}}, \quad d_{2i} = \sqrt{\frac{(1+d_i)}{2}},$$
$$d_{2i+1} = \sqrt{\frac{(1-d_i)}{2}} \qquad (1)$$

In order to reduce the dimensionality of the signal flow graph, the vertical projection [4] is applied to the operational stages in Fig. 1; the stages are collapsed into a one dimension resulting in basic sequential blocks. In order to

The first author acknowledges Nokia Foundation for financial support.

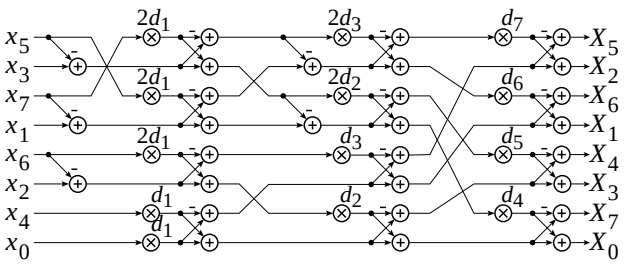


Fig. 1. Signal flow graph of fast algorithm for inverse 8-point DCT-II.

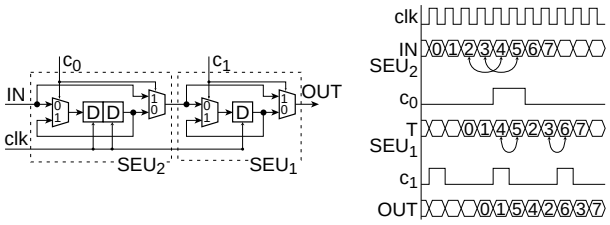


Fig. 3. Block diagram and timing diagram of sequential permutation network of 8-point perfect shuffle permutation. D: Delay register. c_k: control signal.

guarantee correct operation, the causality, i.e., the order of computation should be concerned. Furthermore, the resulting processing element realizing only one operation simultaneously introduces the requirement of unambiguity.

The operational stages of IDCT algorithm in Fig. 1 are actually similar to stages in DCT algorithm in [3]. Thus, the sequential basic blocks introduced in [3] can be utilized for realizing the sequential IDCT kernel. The basic data processing blocks needed in addition to multiplier are butterfly unit and local subtraction unit, which is capable of performing the first operational stage of IDCT. The block diagrams of the blocks are depicted in Fig. 2 (a) and (b).

The functionality of processing blocks in Fig. 2 can be explained as follows. In order to compute both operations of butterfly, subtraction and addition, each sample is stored for two sample periods introducing two storage elements into butterfly unit. The computation of operations requires one arithmetic unit that can be controlled to perform either subtraction or addition. The local subtraction unit in Fig. 2 (b) passes samples through but when subtraction is needed, it is computed between incoming and delayed value.

All the needed data reorderings in signal flow graph of IDCT in Fig. 1 can be performed with a sequential permutation network constructed of shift-exchange units (SEU) as proposed in [5]. A shift-exchange unit of size K (SEU_K) depicted in Fig. 2 (c) is capable of exchanging data elements K samples apart in sequential data stream. In general, perfect shuffle reorders elements of a sequence in such a way that the elements of the first half of a sequence are

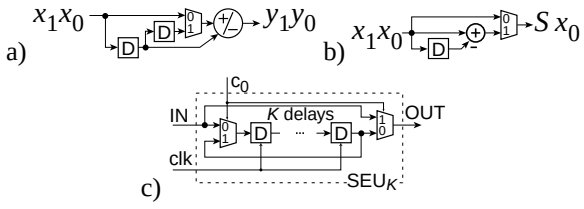


Fig. 2. Block diagrams of (a) butterfly unit, (b) local subtraction unit, and (c) shift-exchange unit of size K (SEU_K). D: Delay register

interlaced with the elements of the second half of the sequence. In other words, perfect shuffle permutation of a vector $\mathbf{x} = (x_0, x_1, \ldots, x_{K-1})^T$ results in a vector $\mathbf{y} = (x_0, x_{K/2}, x_1, x_{K/2+1}, x_2, \ldots, x_{K-1})^T$. Now, a 4-point perfect shuffle permutation can be realized with a single SEU_1 unit and an 8-point perfect shuffle with cascade of SEU_2 and SEU_1 units as illustrated with a timing diagram in Fig. 3. Apart from the global reorderings between the operational stage, there is also a single local reordering, i.e., exchanging of data elements two samples apart, before first multiplications in the signal flow graph in Fig. 1. Such a sample exchange can be realized with a single SEU_2 unit.

By cascading the basic data processing and data reordering blocks described previously, the sequential 8-point IDCT kernel can be constructed as illustrated in Fig. 4. Each unit in the 1-D IDCT architecture corresponds to a specific operational stage in Fig. 1. The loopfree structure enables the efficient pipelining. It should be noted that the pipeline registers are not included in Fig. 4. However, the degree of pipelining is a compromise between latency and throughput. Assuming that each arithmetic unit is followed by pipeline register, the latency of one-dimensional IDCT kernel equals to 17 cycles.

In two-dimensional IDCT architectures, which are based on row-column method, silicon area may be consumed into realization of the intermediate matrix transposition. The implementation efficiency is mainly dependent on interpretation of matrix transposition. The most straightforward way to realize the matrix transposition is its direct interpretation, i.e., rows in, columns out. However, such an approach will introduce double buffering with large silicon area and increased latency, since every sample is stored before reading [6].

The other difference is the way of storing the samples, i.e., the realization may be either memory-based or register-based. Here, the matrix transposition is realized with the register-based sequential permutation network presented in [7]. The corresponding structure and principal operation is illustrated in Fig. 5. It should be noted that the network is optimal from latency point of view since the maximum distance of the element to be moved in sequence equals to latency, which is 49 cycles.

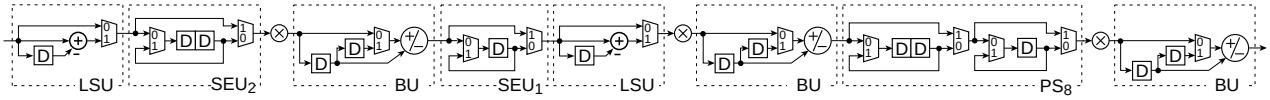


Fig. 4. Block diagram of sequential IDCT architecture. LSU: Local subtraction unit. D: Delay register. SEU$_k$: Shift-exchange unit of size k. BU: Butterfly unit. PS$_8$: 8-point perfect shuffle permutation. Clock and control signals are omitted for clarity.

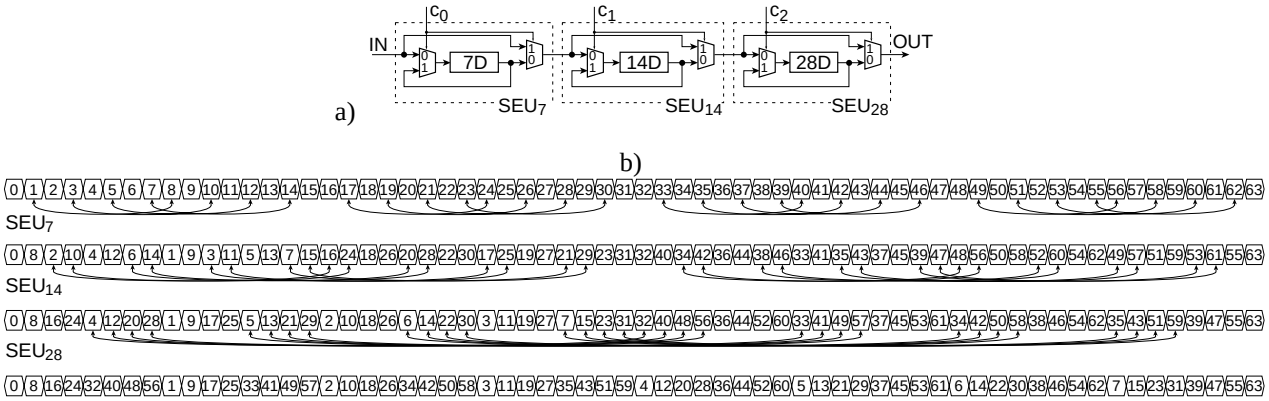


Fig. 5. Sequential 8×8 matrix transposition: (a) structure and (b) principal operation [7]. SEU$_k$: Shift-exchange unit of size k. KD: Shift register of size K.

3. ACCURACY ANALYSIS

The hardware implementations are often based on fixed-point, i.e., fractional, number representation due to the area friendly realization. This requires the scaling of intermediate signal levels for avoiding overflow during the computations. Typically scaling without additional hardware costs is done by rewiring, i.e., scaling factors are powers of two. Due to the fact, that all the intermediate data vectors are passed through multipliers in the proposed architecture, the signal levels can be adjusted at multipliers. This, on the contrary, allows scaling factors to be selected with finer resolution without additional hardware costs.

In the realizations of fixed-point number representation, the main error is caused by the finite word width in the intermediate arithmetic. This error is also known as quantization error. A test suite for the accuracy analysis of the proposed IDCT architecture is made according to the IEEE Standard 1180-1990 [1].

The performance of the pipeline architecture based on the IDCT algorithm shown in Fig. 1 is analyzed with simulations. First, six random test data sets are generated as specified in IEEE standard. Next, the proposed architecture is simulated with different word widths for estimating the error behaviour. Furthermore, two different quantization methods, rounding to the nearest integer and truncation of two's complement ("rounding towards minus infinity") are utilized. The coefficients d_i are rounded to the same word width as the internal data.

The obtained error values, mean error and mean square error per pixel and overall mean error and mean square error, are presented in Fig. 6 with different word widths. Overall mean square error reveals to be the limiting factor in simulations and, thus, 17 bits are required to fulfil the specification if rounding is used. It should be noted, however, that this method is more expensive from implementation area point of view.

If the hardware optimal quantization method, i.e., the truncation of two's complement is utilized, 22 bits are required for internal arithmetic. The quantized values are biased always towards minus infinity and, therefore, the sign of the error is negative at each pixel location. This removes the variance present in rounding method and makes the mean error value almost the same as the mean square error. Word width can be reduced if the error with opposite sign can be generated, i.e., introduce some variance to the error.

4. CONCLUSION

In this paper, a pipeline architecture for 8×8 IDCT is proposed. The architecture is based on row-column decomposition where the IDCT kernel is obtained by projecting the signal flow graph of fast IDCT algorithm vertically. The matrix transposition is realized with register-based sequential permutation network with optimal number of registers. The architecture can be freely pipelined for increasing throughput. The internal word width requirements in case of fixed-point realization was analysed with the aid of

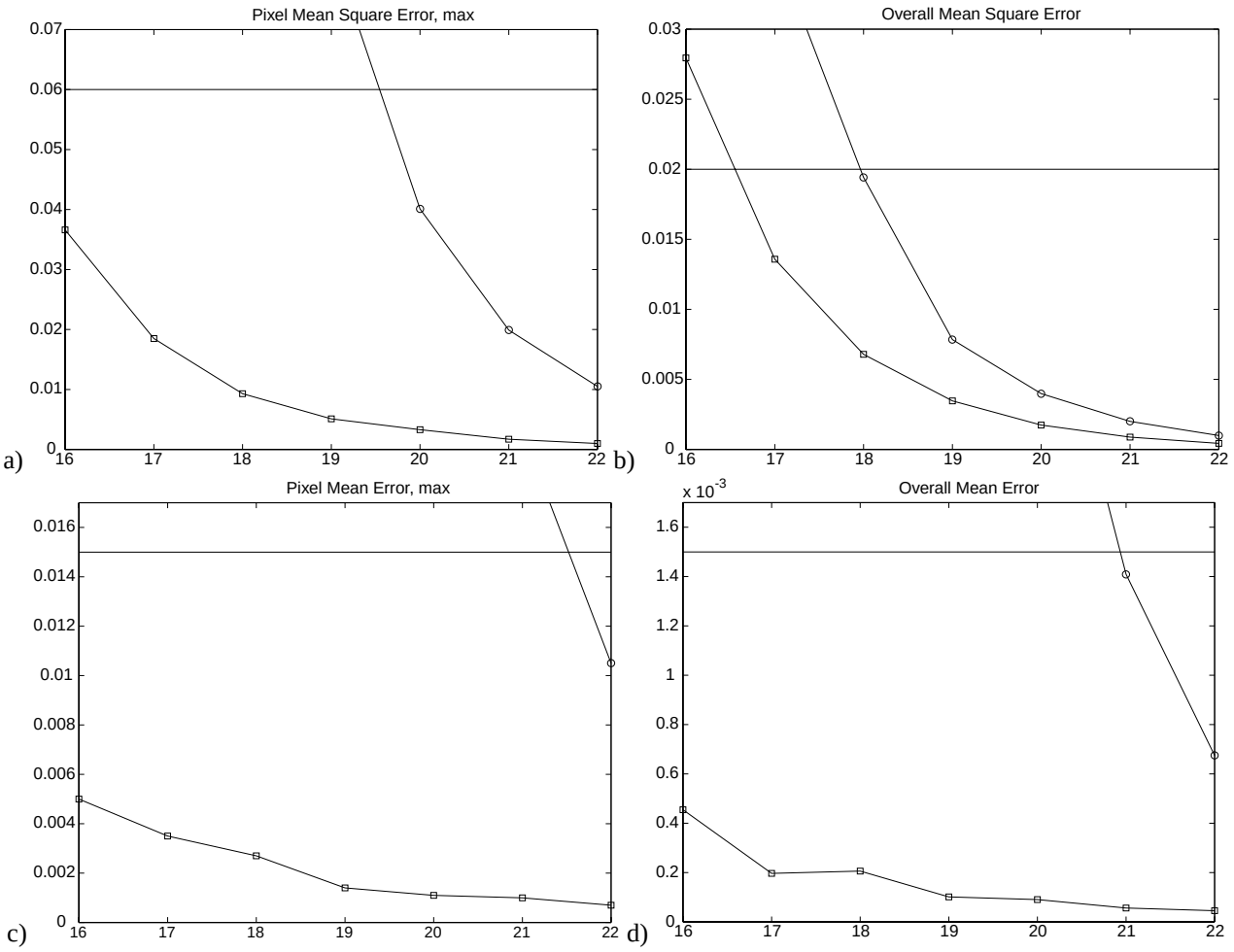


Fig. 6. Error behaviour of the proposed architecture as a function of internal word width: a) pixel mean square error, b) overall mean square error, c) pixel mean error, and d) overall mean error. Line with squares: rounding, line with circles: two's complement, and the solid line: requirement of the IEEE Standard.

simulations. The architecture requires internal word width of 17 bits with rounding and 22 bits with truncation of two's complement to satisfy IEEE Standard 1180-1990. The two-dimensional IDCT architecture yields arithmetic complexity of 6 multipliers, 6 adder/subtractors, and 4 adders. The overall latency with pipeline stages of single arithmetic unit is 83 system cycles.

REFERENCES

[1] IEEE Std 1180-1990, "IEEE standard specification for the implementations of 8x8 inverse discrete cosine transform," International Standard, Institute of Electrical and Electronics Engineers, New York, USA, Dec. 1990.

[2] H. L. Groginsky and G. A. Works, "A pipeline fast Fourier transform," *IEEE Trans. Comput.*, vol. 19, no. 11, pp. 1015–1019, Nov. 1970.

[3] J. Nikara, J. Takala, D. Akopian, J. Astola, and J. Saarinen, "Sequential architecture for discrete cosine transform," in *Proc. 18th NORCHIP Conference*, Turku, Finland, Nov. 6–7 2000, pp. 279–282.

[4] P. Pirsch, *Architectures for Digital Signal Processing*, John Wiley & Sons, Ltd., Chichester, United Kingdom, 1998.

[5] C. B. Shung, H.-D. Lin, R. Cybher, P. H. Siegel, and H. K. Thapar, "Area-efficient architectures for Viterbi algorithm I. Theory," *IEEE Trans. Commun.*, vol. 41, no. 4, pp. 636–644, Apr. 1993.

[6] J. C. Carlach, P. Penard, and J. L. Sicre, "TCAD: a 27 MHz 8x8 discrete cosine transform chip," in *Proc. IEEE Int. Conference on Acoustics, Speech, and Signal Processing*, Glasgow, UK, May 23–26 1989, pp. 2429–2432.

[7] J. Takala, J. Nikara, D. Akopian, J. Astola, and J. Saarinen, "Pipeline architecture for 8×8 discrete cosine transform," in *Proc. IEEE Int. Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, June 5–9 2000, pp. 3303–3306.

EFFICIENT IMPLEMENTATION OF MULTIMEDIA ALGORITHMS ON STANDARD PROCESSORS

Prakash Sastry and Irek Defée
Digital Media Institute,
Tampere University of Technology,
P.O. Box 553, FIN-33101 Tampere, FINLAND.
Tel: +358 40 073 6612; fax: +358 3 365 3857
e-mail: prakash,defee@cs.tut.fi

ABSTRACT

Multimedia algorithms are computationally intensive as they require a huge processing power. With the advent of multimedia processors, fast graphics adapters and the evolution of the operating systems, most of these algorithms can be realised in software. As an example for such algorithms, we consider decoding Standard digital TV signals on low end multimedia capable processor such as a Celeron 366Mhz and also High Definition digital TV signals on a fast and powerful processor such as a Pentium IV. The comparision between the Pentium-II,Pentium-III and Pentium IV architecture for multimedia processing is given. The figure 1 shows the simple decoder architecture which is implemented.

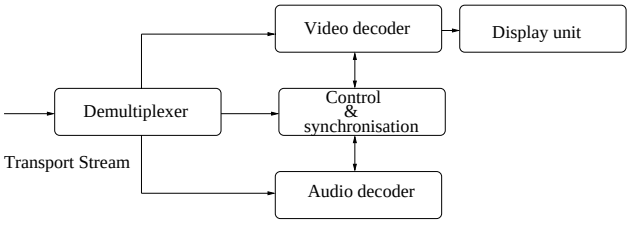


Figure 1. Software decoder architecture

1 Standard definition and High definition digital TV

The digital TV in the either format conforms to MPEG-2 MP@ML or MP@HL (Main Profile at Main Level, Main Profile at High Level) compression scheme respectively. This is based on traditional block based approach employing DCT(Discrete Cosine Transfrom) and motion estimation to remove the redundancy within a frame and also across the frames respectively. The standard definition tv signals are restricted to a maximum frame size of 720x576 for PAL and 720x480 for NTSC with an additional restriction on bitrate going as high as 15 mbps. The High Definition TV signals have a maximum frame size of 1920x1088 and the bitrate starts from 20 mbps. Just comparing the frame sizes for the two standards, we can say that the computation complexity for High Definition TV increases by a factor of five when compared to that of the Standard Definition TV. The internal memory requirements jump from 4MB to 12MB when decoding High Definition TV data. This large requirement on memory and on computation power put some stringent requirements on the processor as well as on the graphics adapter. During our experiments with the High Definition TV data, to achive a real time 30 fps, we would require a transfer of 180 MB/s of raw RGB data from the system memory to the graphics memory. We found that not all the graphics card could handel such large amount of raw data for real time display. Tests were conducted on Linux running latest freely availabel X-Server, using available programs such as DGA(Direct Graphics Access). The graphics card used in the tests were ATI-Radeon and Creative Gforce-2. These requirements will certainly saturate any low end processor.

Bit Rate in Mbps	No of Intra Blocks	No of Non-Intra Blocks
4	279462	1685217
6	308094	1847514
20	1929300	9411275

Comparision of No. of Intra and Non-Intra Blocks

Figure 2. Comparision of No of Blocks

2 Implementation HDTV Decoder

In our implementation of the decoder, the entire process is completely implemented in software on operating systems such as Linux and Windows. The performance of the decoder on these two operating systems for Standard Definition TV data is on par with each other. The transport stream demultiplexer is responsible for demultiplexing the audio and video data from the multiplexed stream. The audio standard used in High Definition TV is Dolby AC3 whose performance and resource requirements are quite small when compared to that of the video and hence it is not reported in this paper. The combined resource requirements for the demultiplexer and the audio decoder is comparitively small to that imposed by the video decoder and the display units combined. This implementation for High Definition TV is on Pentium-IV processor running at 1.4Ghz. The running system is profiled on both Windows and Linux using V-Tune and gprof utilities respectively. The following sections we discuss the various factor affecting the performance of the High Definition TV decoder.

3 Factors affecting the performance

There are many factors which affect the performance of the software only decoder when trying to decode High Definition TV data. They can be broadly classified into processor architecture, clock speed, internal bus architecture, capabilities of the graphics adapters to display High Definition TV and operating system issues. These issues affect the overall performance of the system, especially true in the case of High Definition TV data. The processor architecture, clock speed and the internal bus architecture determine the amount of data that can be pushed to the graphics adapter and the graphics adapter capabilities determine the final throughput of the system.

3.1 Comparison between the processors

Apart form the increase in clock speed from P-III and P-IV there are many architectural differences between the two processors with respect to processing the multimedia data. The additional 144 multimedia instructions available for integer arithmetic on P-IV help in motion compensation routine as we can operate on 128 bits of data at a time. Coupled with 400 Mhz system bus provides a 3.2 GB/s transfer speed between the Pentium-IV processor and the memory controller. With an increased level 2 cache also provides a considerable thourghput.

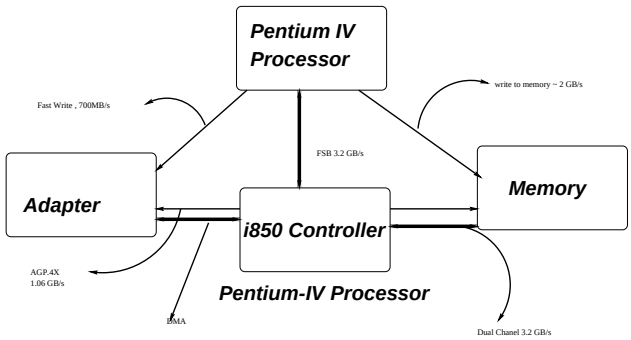


Figure 3. Pentium 4 bus architecture

3.2 Operating Systems Issues

The development tools availabel for implementing the new set of multimedia instructions on Pentium-IV are quite well developed under Windows platform while the same are not yet stabel under Linux. Since the latest multimedia instructions requires the support from the operating system, We have used, not yet stabel compilers and assemblers on the Linux platform to implement the critical loops such as motion compensation and inverse discrete cosine transform(IDCT). Considerabel throughput is achieved in the above critical loops as we can manipulate 128 bits at a time reducing the loops by another factor of two.

4 Performance

The figure show the performance of the software based decoder working at various bitstream speeds. The performance were measured on Standard Definition TV running on Celeron and a Pentium-III processor. The High Definition decoder works at 16.02 fps on a Pentium-IV processor at 1.4Ghz.

5 Future

The theoritical requirement for a full software only HDTV decoding would require a 2GHz processor. Since the processors, especially Pentium-IV are getting faster and faster, we should be abel to decode a fully complaint HDTV streams by the end of this year.

6 References

1. "Digital Video: An Introduction To MPEG-2" Barry G. Haskell, Atul Puri, Arun N. Netravali, Chapman and Hall 1997
2. "Intel Architecture Software Developer's Manual: Instruction Set Reference", http://developer.intel.com
3. "A SoftWare-Based Real-Time MPEG-2 Video Encoder", IEEE Transactions on Ciruits and Systems for Video Technology, Vol 10, No. 7, October 2000

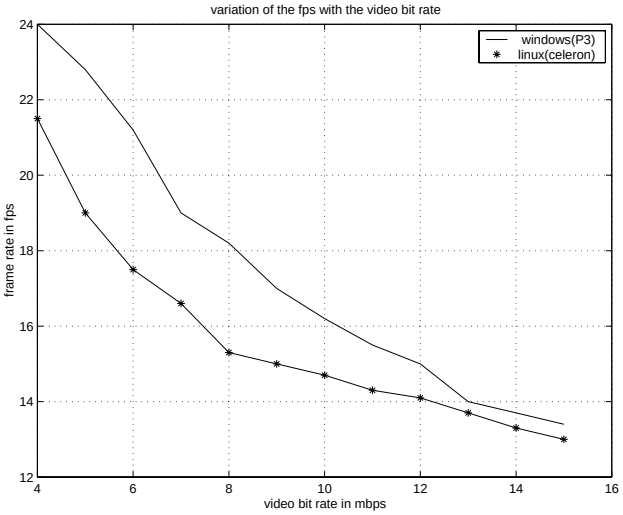


Figure 4. Comparison between Celeron/Pentium-III

The FutureTV Project: MHP Compliant Software Development

Artur Lugmayr, Seppo Kalli, Teemu Lukkarinen, Arttu Heinonen, Perttu Rautavirta, Mikko Oksanen, Florina Tico, Jens Spieker, and Mathew Anurag

Digital Media Institute
Technical University of Tampere
P.O. Box 553, Hermiankatu 3A, 33101 Tampere
FINLAND
Tel. +358-40-821 0558, Fax: +358-3-365 3966
E-mail: {*}@tut.fi

ABSTRACT

In Europe Digital, Interactive Television (digiTV), based on the Multimedia Home Platform (MHP) [1] as specified by the Digital Video Broadcast (DVB) group, is going "on-air" in the next few months and seems to be a rich platform for the next generation of television. The MHP is an excellent starting point in the development of services and covers a complete technical solution of technologies involved in digi-TV. Service and content development will be the major issues in enhancing this new type of multimedia. This paper describes our research work and results done within the FutureTV project, to enrich this platform with value added, content synchronized applications and their development.

1. INTRODUCTION

Fig. 1 shows an overview of the sub-parts of the FutureTV project done at the Digital Media Institute, Tampere. It can be divided in three major groups: Transmission Media and Transport Protocols, Application Development Environments, and Value Added, Content and Content Synchronized Applications.

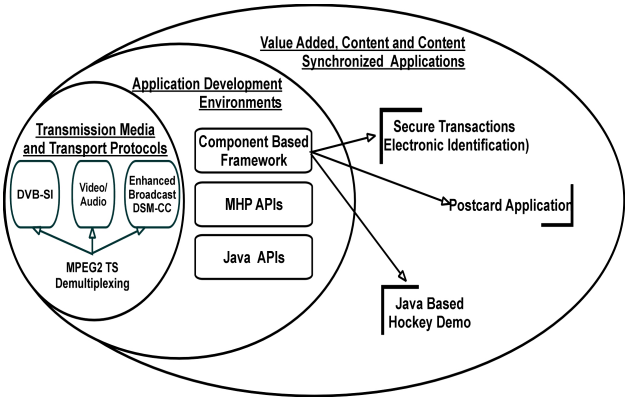


Fig. 1. FutureTV's Sub-Projects.

The following sections will describe them more detailed and allow a quick overview of our experiences and results done to converge digiTV to a complete integrated multimedia solution.

2. TRANSMISSION MEDIA AND TRANSPORT PROTOCOLS

A complete digiTV environment utilizes two different network channels. The high-bit-rate broadcast channel carries a MPEG2 *Transport Stream (TS)*, which consists of multiplexed audio, video, and data. Data can be arbitrary, such as objects, applications, content information, service content, applications, etc. On the end-user side the first essential task is to retrieve information about the stream structure, to be able to access its content.

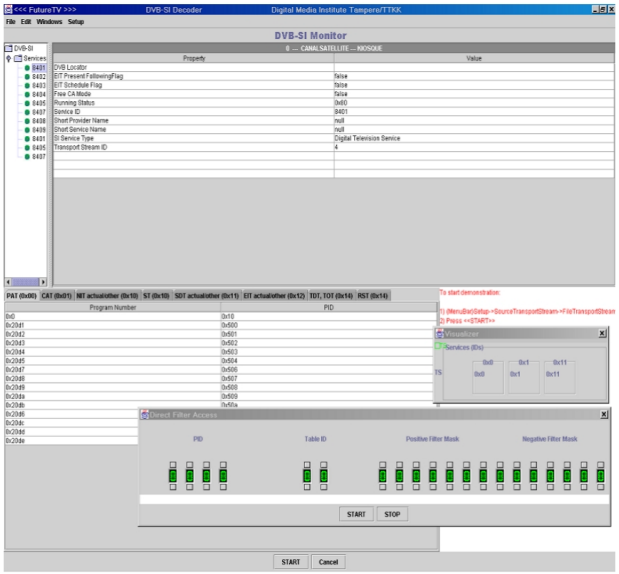


Fig. 2. SI Decoder.

The information needed for this process is called *Program Service Information (PSI)*. Additional information about stream content, pure data transmission and its description, and content can be found in *DVB-Service Information (DVB-SI)*. This part of the project focused on the signaling of DVB transmissions, live stream access from a satellite receiver card, and transmitting of signaling information over the Internet. The data is organized and described in tables, which are split into a stream of sections encapsulated within TS packets. On the client

side, the whole SI has to be decoded to demultiplex the content of the TS, and to be stored in a database from where it can be accessed by applications that rely on SI.

The current SI decoder can be configured to retrieve SI from different sources: satellite receiver card, data storage, or the Internet by utilizing RTP or UDP as transmission protocol. The sources deliver a MPEG2 TS or a section stream that has to be filtered to obtain the different tables and their content, encoded as descriptors.

The Digital Audio-Video Council (DAVIC) standard specifies this major integral task and defines how interfaces between SI sources and higher application layers have to look like. The current software system implements a subset of DAVIC's reference APIs. Where DAVIC standards mainly cover access to sources and filtering aspects, MHP defines how SI tables and their content can be re-assembled and stored in a database. The current implementation is capable of this task and provides a simple database for accessing SI asynchronously. The SI framework has different use scenarios: firstly a common used scenario is where SI feeds the navigator for visualizing services deployed within the TS multiplex; secondly as centralized server home solution for distributing SI in a LAN environment; and last as multicast source for TV-on-demand solutions by utilizing real-time streaming protocol solutions.

3. VALUE ADDED, CONTENT AND CONTENT SYNCHRONIZED APPLICATIONS AND THEIR DEVELOPMENT

Xlets, the lightweight version to Applets in web-based applications, represent the basic structure for digi-TV applications. Their life cycle is well defined and they provide comprehensive solutions for distributing applications to MHP compliant devices over the interaction- or broadcast-channel. Based on this technical solution sample content synchronized value added services have been developed and tested on currently available MHP compliant devices: Firstly an "Electronic Postcard Application" enables end-users to send and retrieve the electronic version of postcards over this platform; Secondly a hockey demonstration shows how content data can be synchronized with video and audio content of a hockey game; third - based on the experiences of the previous projects - how MHP compliant application development and the component based paradigm could enhance the process of compliant software development.

3.1. Hockey Demonstration

The Hockey-Demonstration illustrates how information about players, game scores, etc. can be retrieved and synchronized with a video/audio broadcast of a currently running game. Future trends in user-interface design, content synchronization, information access has been shown. The implementation allows stopping, pausing, fast forward playback, and browsing through the information during a digiTV broadcast. The whole framework is based on pure Java and utilizes the Java Media Framework API and MHP compliant APIs.

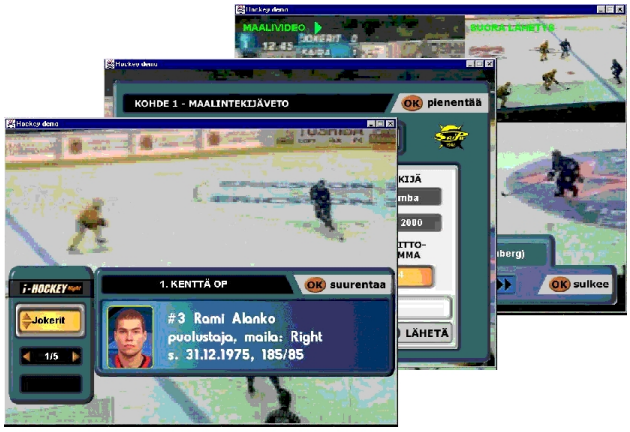


Fig. 3. Hockey Demonstration.

3.2. Electronic Postcard Exchange Tool

A more functionality-oriented development of an Electronic Postcard Exchange Tool has been established. The basic idea is to be able to exchange simple greeting cards over a MHP compliant device. It allows the end-user to load motives from multiple sources, manage the cards via a card management system, various transmission protocols (e.g. SMTP, POP3, etc.), and editing greeting texts.

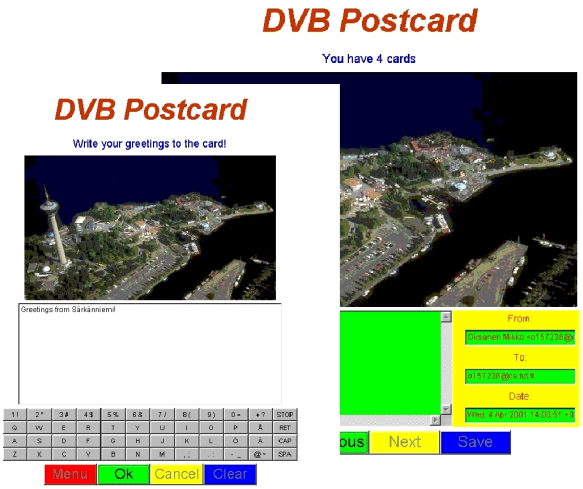


Fig. 4. Electronic Postcard Exchange Tool

The motives can be accessed either from the broadcast or interaction channel, hard disk, or from a linked camera.

3.3. Converging digiTV with the Component-Based Software Development Paradigm

Based upon the previous applications the goal was to present a complete solution for combining different

multimedia entities, such as graphics, text, video, audio, data, and interaction possibilities within a component based development software editor. Application development enhancements considers the following question: "How can services for MHP compliant platforms developed more rapidly and provide an integral solution for content, rather than application development?", was the question considered by this group of the project. Involving component-based paradigms into the life cycle of the software-development of digi-TV services should provide a generic, reusable framework for rapid content production and service deployment.

Fife major groups have been identified, that are involved in the process of developing MHP compliant software: Creative *Content Contributors* or authors focus on the development of content and not on underlying software and hardware technology. *Content Management* deals with the production and re-production of content by utilizing a content repository. It compromises content acquisition, authoring, service management, multi-platform support, the production of multiple services, managing the interaction between end-user and content, providing a set of rules and processes for content navigation, and life-cycle issues. The maintenance and realization of components is obeyed by the *Application Programmer*. His responsibility is to deploy a well tested, domain dependent, and utilizable set of components that are utilized by the content contributors or application engineers. The later group - the *Application Engineers* - interconnects and parameterizes single entities taken out of a component repository by utilizing an appropriate component-authoring tool. The development of the whole component based framework is the main responsibility of the *Architecture Group*, which designs, defines, and specifies the completely component-based repository.

A component-based application development approach seems to provide several advantages in comparison to current solutions: it gains productivity, minimizes required coding, allows visualized programming, reduces time-to-market, involves software behavior models, re-use of software sub-parts, and the development of a concurrent component repository. The aim of this research group was to point out how such a framework might look like, by utilizing the *Unified Modeling Language (UML)*, the JavaBeans concept, and exploring visualized programming environments.

4. ELECTRONIC IDENTIFICATION AND SECURE APPLICATIONS

The basic considered question was how it is possible to guarantee a secure environment for the end-user. Two aspects where covered by our research work: electronic identification and secure applications.

Electronic Identification (EID) based on a smart card solution allows identity proving of the end-user and secure transmissions between him and a third party. This provides comprehensive solutions for secure transaction management. To prevent applications causing damages to the digi-TV system a mechanism have to be introduced, validating the authentication and rights of one application. Therefore, our research work tried to evolve strategies for chain-of-trust based, certified applications and introduced security mechanisms to restrict resource access rights.

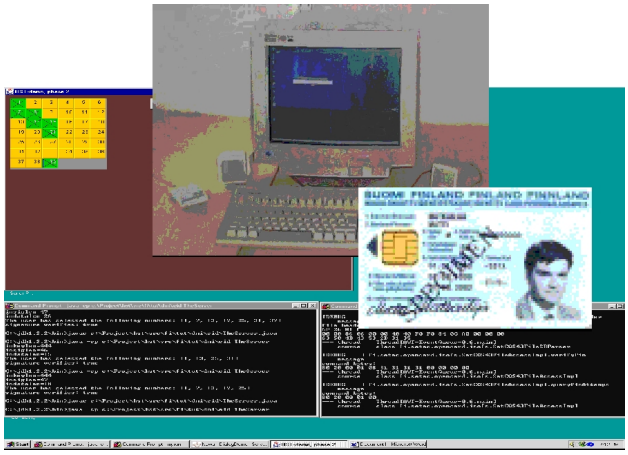


Fig. 5. Electronic Identification

To converge Electronic Identification (EID) with MHP based digiTV environments allows identification of the end-user based on a identification card. EID compromises solutions for electronic identification and digital signatures. The current implementation was tested and deployed firstly for a PC environment, where reusability and full MHP compliance were major integral requirements.

The development was divided in three steps: The first was to test and explore the possibilities of the EID card; by utilizing, the Java based OpenCard Framwork (OCF) as interface between card reader and the application. The second step extended this solution to a fully network solution of a lotto application, that should ensure secure transfer and identification of the end-user. The final step was the deployment of the application onto a fully MHP compliant device, therefore was migrated to an Xlet and embedded in a DVB-J application.

4.1 Secure Applications

With the introduction of Java JDK 1.2 a more flexible possibility was given for tailoring the security manger to fit own needs and to emphasis secure application development. The MHP defines both, keystore creation, and certificate factory therefore provides a unified solution for the realization of certified Xlets, remote resource access within certain access rights, and involving trusted java source code.

The strategies, how rights are assigned to applications is defined firstly by MHP, OS level pre-defined

configuration, and run-time assignments by authorized authorities. They rely on three different types of messages: Cryptographic hash codes, Signatures delivered by a master hash code, and certifications to enable the *Chain of Trust*.

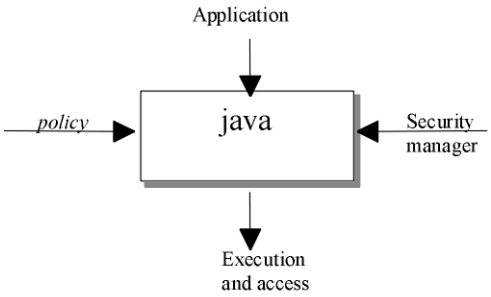


Fig. 6. Secure Applications

Once an application requests some additional rights, the procedure goes as shown in Figure5. The security manager will read all specified policy files and evaluates for this application whether rights can be granted; if yes, the code execution continues as desired, in case of an authentication failure any continuation is not permitted.

A MHP compliant demo application is currently under development and will be used to illustrate secure application development and show, how different security settings enhance security policies for this platform.

5. CONCLUSIONS AND FUTURE WORK

We presented our research work done within the FutureTV project and pointed out firstly, which enhancements in MHP still have to be made and how common applications could look like. MHP is going to be realized, but it will highly depend on the customer which services he accepts and whether television stays a pure broadcast medium or shifts to a full interactive multimedia home platform. Fore more detailed information about our research work, we would like to reference to your publications [2, 3, 4, 5, 6, 7, 8].

Our future research work focuses on the enhancement of application development, emphasizes security and identification issues, and will try to find possibilities and limitations given by MHP to develop content and services.

ACKNOWLEDGEMENTS

The authors would like to thank all their research colleagues at the DMI/TTKK for their help and discussions.

REFERENCES

[1] TS 101812. DVB: Multimedia Home Platform Specification. European Broadcast Union, 2000. Vers. 1.1.1. http://www.etsi.org.

[2] C. Peng, A. Lugmayr, and P. Vuorimaa. A Digital Television Navigator. Journal of Multimedia Tools and Applications. accepted.

[3] A. Lugmayr, S. Kalli, P. Rautavirta, and M. Oksanen. Component Based DVB-J Application Development in MHP based digiTV. IASTED International Conference: Applied Informatics (AI2001), Innsbruck, Austria. http://www.iasted.com/conferences/2001/austria/ai.htm.

[4] A. Lugmayr, C. Peng, and S. Kalli. A MHP Conform Software Implementation of a DVB Service Information Decoder. Multidimensional Mobile Communications (MDMC2001). accepted. http://www.pori.tut.fi/mdmc01.

[5] A. Lugmayr, S. Kalli. DVB over Wireless Networks: Converging Wireless Technology with digiTV. Multidimensional Mobile Communications (MDMC2001). accepted. http://www.pori.tut.fi/mdmc01

[8] Artur Lugmayr, Seppo Kalli. Transmission of DVB Service Information via Internet . In Next Generation Networks. 5th IFIP TC6 International Symposium, INTERWORKING 2000, Bergen, Norway, Proceedings. http://www.telenor.no/fou/om/konferanser/.

Flexible DSP platform for various workload patterns

Antti Pelkonen, Jussi Roivainen, Juha-Pekka Soininen

VTT Electronics
P.O.Box 1100 (Kaitoväylä 1), FIN-90571 Oulu
FINLAND
Tel. +358 8 551 2111, Fax. +358 8 551 2320
E-mail: {Antti.Pelkonen, Jussi.Roivainen, Juha-Pekka.Soininen} @vtt.fi

ABSTRACT

Future multimedia communication products require system chips that provide sufficient computing capacity and configuration flexibility for achieving interoperability between different communication systems and support for various multimedia signal processing standards. A flexible DSP platform that utilises pre-designed IP cores, such as DSP and RISC processors, advanced coprocessors for critical functions and configurable memory organisation is presented. ADSL, HiperLAN2 subset and MPEG2 decoding algorithms have been analysed as a basis of IHIP architecture design. The initial performance results look promising and it seems that the IP block based configurable architecture could provide satisfactory performance for various types of workloads.

1. INTRODUCTION

Future multimedia products must be capable to provide both high computational capacity and effective execution of different kinds of algorithms and applications. Algorithmic complexity of DSL (digital subscriber line), mobile and wireless communication standards are increasing heavily [1, 2]. Interoperability of different systems, such as UMTS (universal mobile telecommunication system) and WLAN (wireless local area network), and a need to support various compression and source coding standards, such as MPEG4 and MP3, set demanding requirements for processing platforms [3, 4, 5].

Integrated computer systems, e.g. system chips have been the kernels of telecommunication products. Integration capacity of System Chip technology is developing rapidly [6]. The capacity of a single chip doubles in every 18 months according to Moore's law and within next years, tens of complete computer systems can be integrated into a single ASIC. It will mean that a high-performance PC with memories can be scaled down to a single system on chip that has both capabilities to base-band and application processing.

The development of system chips is a huge effort even to best organisations. The reuse of existing designs and procurement of intellectual property make design more effective. The IP (intellectual property) based design requires standardised ways for interconnecting different virtual components [7]. Platform based design have been proposed as a solution for design complexity management [8]. Manufacturing integration platforms provide implemented hardware resources that can be used in the various applications. The application can be designed using hardware or software configuration or both [9].

The research hypothesis in this paper is that by combining ideas from configurable architectures and intellectual property based design it is possible to have effective platform architecture for various types of algorithms and applications. The key issues in the design of flexible DSP platform are the analysis of application characteristics and the evaluation of platform performance.

2. APPLICATION CHARACTERISTICS

There are basically three different categories of signal processing, when using the data timing characteristics for classification: stream, block and sporadic data processing.

In stream based processing, size of incoming data token is usually small (from few bits to few bytes), token arrive periodically or, at leas1t in some degree, predictably. Data is processed in many stages and SAR (segmentation and re-assembly) functions between stages are common. Number of operation for data token in each stage is usually quite small. Streaming data processing is be done most naturally in pipelined processing architecture, because the SAR functions can be implemented in between pipeline stages and pipelining is associated with high data throughput requirements. Due to high throughput requirements, stream based processing requires fast clock speed, small but fast buffer memories between stages and solid method of inter-process or inter-stage synchronisation. Examples of stream based processing in this paper are HiperLAN2 and ADSL base-band processing.

In block based processing data tokens to be processed are large, ranging from hundreds of bytes to thousands or even millions bytes. This causes number of requirements

for processing architecture. Data transfers happen in long bursts with inactivity times in between, which causes requirements for peak bandwidth of buses. Address space, memories, and memory bandwidth must also be large. Block based processing usually allow data parallel processing with MIMD (multiple instructions, multiple data) or SIMD (single instruction, multiple data) type of processors.

In Internet era, a hybrid form of streaming and block based data processing has emerged with the internet protocol based communications. Implementations of physical and data link layer manifest streaming characteristics, but network layer manifests block based processing characteristics with large packet sizes, large memory requirements associated with ordering and error detection codes for whole packet. Therefore current and especially future Internet infrastructure requires programmable signal processing platforms that are well suited for both types of processing.

Most difficult form of data processing is sporadic data processing. Size of data token is predictable but only worst case estimates and some statistics of data arrival rate is known. If hard real-time requirements are associated for this type of data processing, architecture must be designed for the worst case. This however may leave significant resources idle for most of the time. For achieving acceptable power consumption, chip should comprise of several parallel blocks where clocking of unneeded blocks can be switched off. Other solution is the use of adapting clock speed.

1.1. ADSL, HiperLAN2 and MPEG2

Case examples for flexible DSP platform are ADSL and HiperLAN2 processing and MPEG2 video decoding.

Programmable implementation of ADSL remote modem requires approximately 900-1300 MOPS (millions of operations in second) on typical DSP processor with MAC (multiply-accumulate) unit [1].

The estimated resources for HiperLAN2 modem in simplest case (6Mbit/s data-rate, BPSK modulation scheme) are around 600 MOPS for transmitter and 850 MOPS for receiver, excluding the Viterbi-decoding [4].

As block based processing example, MPEG-2 MP@ML video decoding at 30 frames per second sampled in YCrCb 4:2:0 and MP3 audio or MPEG2 audio is considered [5]. The most essential design parameter considering MPEG-2 video decoding is the high memory, memory bus and system bus bandwidth requirements. Raw uncompressed data stream bandwidth is 15 Mbits/s but internal communication results in huge bandwidth requirement. The amount of needed memory is at least 16 Mbit. The required processing resources estimation of the video decoding is estimated to be 620-750 MOPS [10]. This could, however, to be somewhat less since estimates were based on general-purpose processor and the multiplication was estimated to take 4 clock cycles. With VS56000 DSP, multiplication can be done in single clock cycle.

3. IHIP ARCHITECTURE

Case example presented in this paper is the IHIP (information highway interface platform) chip. The IHIP consists of RISC computer that is based on of Leon 32-bit SparcV8 RISC core designed by European Space Agency, and DSP cluster with four VS56000 24-bit fixed point DSPs and Viterbi and FFT coprocessors. The basic architecture is shown in Fig. 1.

In addition IHIP chip has a DMA unit, 16 Mbit on-chip memory and an interface to external bus. Components are connected with 32-bit AMBA (Advanced Microcontroller Bus Architecture) bus that supports pipelining bus accesses, split transactions, burst transactions and multiple bus-masters. Leon core can also access the memory blocks of DSP cluster in some operation modes.

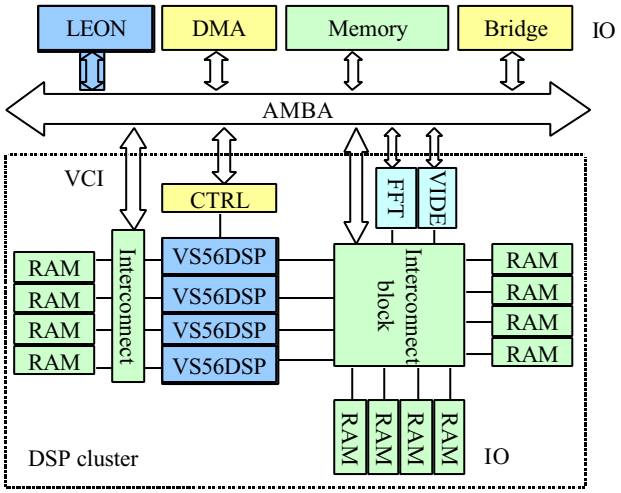


Fig. 1 IHIP Architecture

The components in the chip are introduced into the design as third party IP-blocks (intellectual property). These individual blocks are connected to the system bus via VSIA OCB (on-chip-bus) wrappers so the system bus can be modified or changed if needed without modifying the IP blocks and vice versa [7, 8].

In the DSP-cluster, the cores are connected with a configurable shared memory system (each core can access 64 kwords of memory per bus). There are four memory banks for each X and Y data buses and four 16kword segments in each bank. Each 16kword segment can be mapped to any processor or even for all processors at the same time. All processors can access to same piece of memory when needed or make a pipeline from one processor to another.

Two co-processors are needed in most computing intensive tasks of applications. For ADSL's FFT and IFFT there is an IP-block that can calculate 256-point complex to complex transformations. For the Viterbi-

decoding in HiperLAN2 receiver, there is a dedicated IP block.

The estimated processing capacity of the chip is at GOPS (giga operations per second) range. The chip is targeted to 0.13µm (or even more advanced) CMOS process technology that has ability to implement large on-chip DRAM memories. Target clock frequency is 200 MHz, which would provide 200 MIPS per DSP processor core and 200 MIPS for the Leon RISC core. The main on-chip DRAM memory would be 16Mbit and the used clock frequency would provide a peak bandwidth of 6400 Mbits/s for AMBA bus. The conceptual design and VHDL and SystemC implementation so far has been made without concerning too much on the technology constraints, as the chip is supposed to use future state-of-the-art technology process.

1.2. Configurable memory system

The configurable shared-memory system has three basic configurations, but the architecture allows others too.

- In parallel mode each DSP processor has private data and program memories.
- In pipelined mode memory is configured so that a core feeds a shared memory segment and another reads from the same segment. In Fig. 2, and example of pipeline configuration is shown. DMEM means data memory and PMEM program memory.
- In concurrent mode, common memory banks are mapped for all cores.

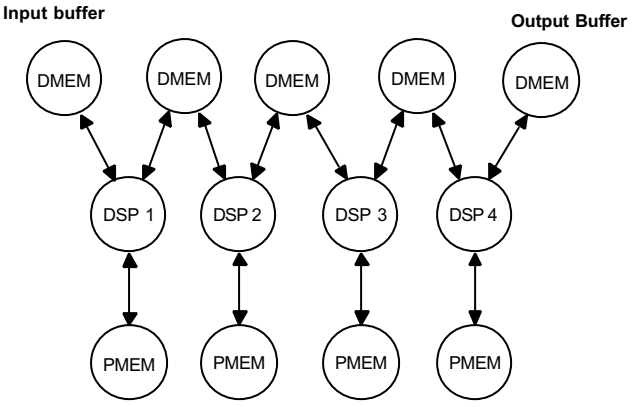


Fig. 2 DSP Cluster in pipeline mode.

Inter-processor synchronisation is achieved by reserving the 64-word peripheral memory area from DSP core's X-bus. These registers are used as semaphores or status registers between DSP cores and the Leon.

1.3. Operation modes and control

The basic idea in the control of IHIP architecture is that the LEON core controls DSP clusters operation, operation mode changes and configuration changes. There are four possible states:

- In the initialization state the basic system checks are performed.
- In operation mode change state the program and data are loaded to DSP cluster and system is configured according to operation mode
- In operation mode data is processed at DSP Cluster and RISC computer. During the operation the configuration of the memory system does not change.
- In configuration change mode, the memory configuration is changed. This used for transferring data between processing units.

The state of DSP cluster is controlled via the status registers and interrupts. The platform control must be explicitly programmed. The benefits are full control over the execution and possibility to fully exploit the capacity of DSP cluster.

1.4. Operation scenarios

In ADSL modem mode, one DSP core acts as transmitter and three DSP cores and RISC core are allocated to receiver side. The FFT block is also required for the receiver. In the receiver side two DSPs handle Reed-Solomon decoding, FFT-block handles fast Fourier transformation and RISC core and the remaining DSP handle the rest of the processing. The shared memory of the DSP cluster and FFT in the receiver side is configured as pipeline to allow easy data transfer and synchronisation between processing stages.

In HiperLAN2 case, two DSP cores are allocated on the transmitter side and two DSP cores, RISC core, FFT-block and Viterbi-decoder are allocated for the receiver side. The processing and memory bandwidth requirements are so high that additional FFT block might be needed, but this is confirmed in further simulations. The shared memory and co-processor blocks are configured as two parallel pipelines, one pipeline consisting of transmitting side DSP cores and possible additional FFT block and the other pipeline consisting of receiver side DSP cores, FFT-block and Viterbi-decoder.

In MPEG2 case DSP cores handle the IDCT (inverse discrete cosine transformation) and the RISC core handles other tasks. The shared memory of the DSP cores is configured as concurrent mode, where all cores can access same data.

4. EVALUATION OF ARCHITECTURE

Initial performance estimations suggest that ADSL and MPEG-2 fit in to the IHIP architecture easily. However if a Reed-Solomon decoder would be implemented as additional IP block it would take 400 MIPS of processing burden off the DSP cores. The HiperLAN2 might need additional FFT block to allow both receiver FFT and transmitter IFFT to be processed in hardware co-processor. These open questions should have an answer after further simulations are done.

The register-based configuration of DSP cores, the DSP shared memory and co-processors is very software friendly and intuitive and the programming of the chip is very similar to a single embedded RISC processor programming. The problematic software development for the DSP cluster is not yet tackled. For the DSP shared memory, the programmable switch matrix approach is proving to be very efficient solution to add flexibility to architecture.

The 16Mbit of on-chip memory is very challenging part of the design, although IBM provides up to 16Mbit of embedded DRAM memory blocks even today [11]. Also the clock frequency of 200 MHz for the DSP cores and the Leon core might not be feasible.

There is also an option of changing the processor cores to achieve better performance. The Leon RISC core could be exchanged for a more state-of-the-art ARM or MIPS processor core and 24-bit VS56000 DSPs could be changed to floating point or 32-bit fixed point cores. The current choice of cores is partly done because of the easy availability of VHDL implementations of the cores.

5. CONCLUSIONS

A flexible DSP platform that utilises pre-designed IP cores, such as DSP and RISC processors, advanced coprocessors for critical functions and configurable memory organisation is presented. The proposed architecture can be configured so that execution of various workload patterns is efficient so that unnecessary on-chip communication load can be minimised. The control solution for DSP cluster and configuration management approach seem to provide good basis for software development.

ADSL, HiperLAN2 subset and MPEG2 decoding algorithms have been analysed as a basis of IHIP architecture design. The initial performance results look promising and it seems that the IP block based configurable architecture could provide satisfactory performance for various types of workloads.

The implementation complexity of IHIP chip has not been studied yet. The next step is to implement development and verification platform for the architecture and to study in detail the operation of architecture and application mapping issues.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support given by the SciFi project and its funding organizations TEKES, Nokia Mobile Phones, Nokia Networks, Nokia Research Center, ABB Corporate Research, ABB Substation Automation, Elektrobit, Hantro Products, Synopsys Finland Smartech Group, VLSI Solution, DSLBit and Fincitec.

REFERENCES

[1] Wiese, B.R., Chow, J.S. "Programmable Implementations of xDSL Transceiver Systems", *IEEE Communications Magazine*, May, 2000, pages 114-119

[2] Starr, T., Cioffi, J.M. and Silvermann, P.J., Understanding Digital Subscriber Line Technology, Prentice Hall, 1999, 474 pages

[3] Kourtis, S., McAndrew, P. & Tottle, P. Technology Requirements of the 3GPP-TDD terminal, Proceedings of the 1st International Conference on 3G Mobile Communication Technologies, 27-29 March 2000, pages 89-93

[4] Khun-Jush, J., et al, "Structure and Performance of the HIPERLAN/2 Physical Layer". *Proceedings of Vehicular Technology Conference 5*, 19-22 September 1999, Amsterdam, Netherlands, p. 2667-2671.

[5] Parhi & Nishitani. Digital Signal Processing for Multimedia Systems, Marcel Dekker Inc, June 1999.

[6] International Technology Roadmap for Semiconductors, Edition 1999, World Semiconductor Council

[7] Virtual Component Interface Standard (OCB 2 1.0), March 2000, VSI Alliance

[8] Chang, H., et al. Survining the SOC Revolution – A Guide to Platform-Based Design, Kluwer Academic Publishers, 1999, 235 pages

[9] K. Keutzer, S. Malik, A. R. Newton, J. Rabaey and A. Sangiovanni-Vincentelli. "System Level Design: Orthogonolization of Concerns and Platform-Based Design", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Volume 19, Issue 12, December 2000. pages 1523 –1543

[10] Chang-Guo Zhou; Ihtisham Kabir; Leslie Kohn; Aman Jabbi; Rice, D.; Xio-Ping Hu "MPEG Video Decoding with the UltraSPARC Visual Instruction Set", CompCon '95 'Technologies for the Information Superhighway' Digest of Papers. 1995 Pages 470-477

[11] ASIC Cu-11 Gate array product brief, 2000, IBM, http://www.chips.ibm.com

Audio Restoration Using Sound Source Modeling

Paulo Esquef, Vesa Välimäki, and Matti Karjalainen

Helsinki University of Technology
Laboratory of Acoustics and Audio Signal Processing
P.O.Box 3000, FIN-02015 HUT, Espoo, Finland
esquef@acoustics.hut.fi

ABSTRACT

This paper presents new propositions to audio restoration and enhancement based on Sound Source Modeling (SSM). The main motivation is to take advantage of prior information of generative models of sound sources when restoring or enhancing musical signals. We describe a case based on the commuted waveguide synthesis algorithm for plucked string tones and devise a scheme to extend the bandwidth of guitar tones. Then, we study the de-hissing of guitar tones and propose a scheme in which the bandwidth extension method is applied as a post-processing stage to a spectral-based de-hissing procedure. According to our experiments, this is effective for the reduction of common side-effects associated with spectral-based de-hissing methods, such as musical noise and signal distortion.

1. INTRODUCTION

Signal modeling techniques have been widely used in audio restoration purposes. In these techniques, the analysis and synthesis parts of the processing only deal with the information available in the surface presentation of audio signals. However, audio analysis and synthesis can also consider how the sound elements are structured in the audio signal [1]. This kind of approach asks for better understanding of the human auditory perception, as well as deeper representations of sound sources, which, in fact, are important requirements for the actual challenges of the audio signal processing field, such as sound source recognition, sound source separation, automatic transcription and musical retrieval [2], content-based coding, and sound synthesis [3].

In principle, a content-based audio analysis would help to distinguish between a noise-like signal component to be preserved, and a degrading noise to be removed. This possibility could guide further choices of the signal components to be reconstructed in the synthesis part. Additionally, SSM allows taking advantage of previous knowledge of the model parameters associated with a high quality instrument sound to enhance the sound quality of a poorly recorded instrument. However, the practical usage of SSM is still limited to some specific cases, e.g., analysis and synthesis of monophonic instrument sounds.

In this paper, we show that it is possible to reconstruct the high frequencies either lost or severely degraded in the recording process, since high quality synthesis models for plucked-string tones are available, providing prior knowledge of their frequency content. Our problem is restricted to the synthesis stage, since only single acoustic guitar tones are considered. For the SSM of plucked strings, a simple *commuted waveguide synthesis* (CWS) algorithm is employed [4, 5]. This choice allows obtaining the model parameters by analyzing recorded tones [6]. The study presented here is divided basically in two parts: a proposition to extend the bandwidth of originally bandlimited guitar tones, and a de-noising scheme for guitar tones which mixes a traditional spectral-based de-hissing method and SSM.

2. STRING MODEL

The function of the vibrating string model is to simulate the generation of string modes after the plucking event. Considering an isolated string, its behavior can be efficiently simulated by the string model illustrated in Fig. 1, whose transfer function is given by

$$S(z) = \frac{1}{1 - z^{-L_i} F(z) H(z)}, \tag{1}$$

where L_i and $F(z)$ are, respectively, the integer and fractional parts of the delay line associated with the length of the string, L. $H(z)$ is called *the loop filter* and it is in charge of simulating the frequency dependent losses of the harmonic modes.

In this work, the loop filter is implemented as a one-pole low-pass filter with transfer function given by

$$H(z) = g \frac{1 + a}{1 + a z^{-1}}. \tag{2}$$

The magnitude response of the filter $H(z)$ must not exceed unity in order to guarantee the stability of $S(z)$. This constraint imposes that $0 < g < 1$ and $-1 < a < 0$.

The presence of the fractional delay filter, $F(z)$, is intended to provide a fine tuning of the fundamental frequency by precisely adjusting the length of the string. In this work, it is implemented as a fourth-order Lagrange

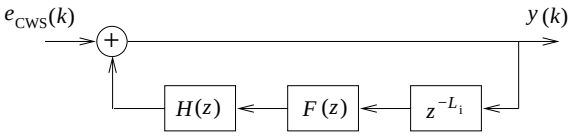


Fig. 1. Block diagram of the string model.

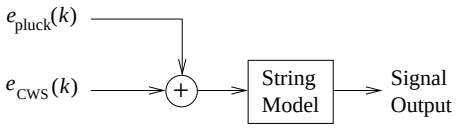


Fig. 2. Bandwidth extension scheme.

interpolator FIR filter [7]. In this configuration, the string-model transfer function, $S(z)$, is completely defined by the length of loop delay, L, the loop filter parameters, g and a.

The value of L is obtained by

$$L = \frac{f_s}{\hat{f}_0}, \tag{3}$$

where f_s the the sampling rate of the analyzed signal, and $\hat{f}_0$ is an estimate of the fundamental frequency of the tone.

The parameters of the loop filter are obtained by first estimating the decay rate of the harmonics. Then, the resulting loop gains are used as a target magnitude response for the loopfilter. A detailed description of the procedures used to estimate the string model parameters can be found in [6].

The excitation $e_{\mathrm{CWS}}(k)$, shown in Fig. 1, is obtained by inverse filtering the guitar tone through the previously estimated string-model.

3. BANDWIDTH EXTENSION OF GUITAR TONES

In this section, the problem of reconstruction of missing spectral information in guitar tones is addressed within the SSM approach. The connections between bandwidth extension and audio restoration appear in two cases: to overcome the intrinsic bandwidth limitations of old recording systems in capturing the audio source, and to reconstruct the spectral information lost during a de-noising procedure.

Let us consider a single guitar tone which was lowpass filtered in order to remove the high frequency harmonics, while preserving the fundamental frequency as well as a few harmonics. The first step of the bandwidth extension procedure is to estimate the string-model parameters [6]. Due to the simplicity of the string-model we are employing, and perceptual aspects [8], it is acceptable to analyze a similar fullband guitar tone to overcome the impossibility of estimating the decay rate of the missing harmonics.

Based on the CWS properties, the tone can be inverse filtered resulting in an excitation, $e_{\mathrm{CWS}}(k)$. If the analyzed tone is already lowpass filtered, the corresponding $e_{\mathrm{CWS}}(k)$ will have a lowpass characteristic as well. This means that we need to provide extra energy to the excitation in order to fully excite the string-model modes.

A possible way to achieve that consists of adding to the attack part of the excitation an artificially generated plucking event, $e_{\mathrm{pluck}}(k)$, as illustrated in Fig. 2.

A suitable option for $e_{\mathrm{pluck}}(k)$ is to generate an impulsive noise burst, for instance, by windowing a zero-mean Gaussian white noise sequence. However, it would be desired that the additional noise burst, composed with the filtered excitation, could emulate a typical spectral behavior of the attack part of an excitation corresponding to a full bandwidth tone. This can be achieved by coloring the noise burst sequence according to known information about typical spectral characteristics of guitar bodies.

The generation of the noise burst, which simulates a plucking event, is carried out as depicted in Fig. 3. The input sequence, $n(k)$, is a zero-mean Gaussian white noise sequence, the filter $E(z)$ is a coloring filter, whose magnitude response must approximate the spectral envelope of the very beginning of a full bandwidth excitation. The highpass filter $H_{\mathrm{hp}}(z)$ is optional and can be included to compensate for the unnecessary addition of energy within the effective bandwidth of the analyzed tone. The gain factor α controls the local signal-to-noise ratio (SNR) at the part of the excitation to be modified.

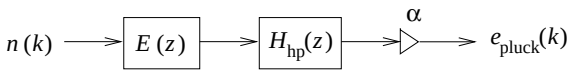


Fig. 3. Generation of the synthetic plucking event.

The capability of the previously described method to extend the bandwidth of guitar tones is illustrated in Fig. 4. In this example, a test signal consisting of an F_4 tone with fundamental frequency of 347 Hz, sampled at 22.05 kHz was used. The tone was lowpass filtered using a 101^{th} order equiripple FIR filter with cutoff frequency at 1 kHz, transition band of 1 kHz, and attenuation of 80 dB on the rejection band. Filter $E(z)$ was chosen as a second-order resonator tuned at 200 Hz. This frequency corresponds to the lowest mode of the top plate of the guitar body [9]. The radius of the poles was arbitrarily set to 0.8. With these parameters, the frequency response of $E(z)$ approximates the spectral envelope associated with the attack part of a full bandwidth excitation. The highpass filter $H_{\mathrm{hp}}(z)$ was not included. Finally, the noise burst was multiplied by a Hanning window of 600 samples, scaled, and added to the attack part of the excitation.

Based on informal listening tests, it was observed that coloring the noise burst has an important effect on the quality of the timbre of the resynthesized tone. The timbre of the resynthesized tone also varies depending on the power of the noise burst, which can be adjusted to produce a certain local SNR at the attack part of the excitation. Additional tests were performed on the same guitar tone but with bandwidth limited to 500 Hz and 3000 Hz. The obtained results were similar to that of the previous case.

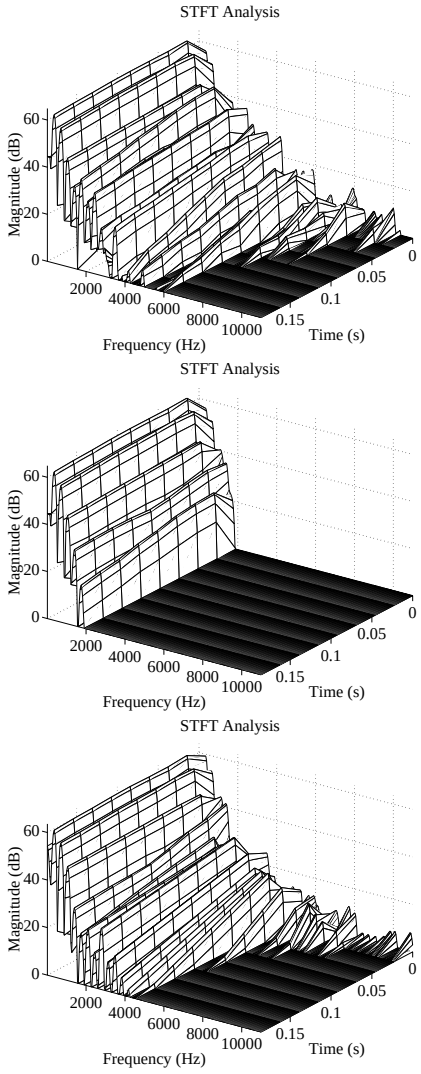


Fig. 4. Time-frequency analysis of the original tone (top), the lowpass filtered tone (middle), and the resynthesized tone (bottom).

4. SSM AND DE-NOISING OF GUITAR TONES

Usually, spectral-based de-hissing methods suffer from a difficult tradeoff between the reduction of the noise effects and the introduction of distortion in the restored signal [10, 11]. The results of the SSM-based bandwidth extension of guitar tones, described in Section 3, can be useful in the de-hissing problem. The hard tradeoff between noise reduction and preservation of the signal information can be softened on the grounds that the spectral content can be reconstructed afterwards if a sound source model and a synthesis algorithm are available for the analyzed signal.

A possible option to remove the noise effects from a guitar tone corrupted by zero mean white Gaussian noise is to de-hiss it through a spectral-based method using an overestimated value for the variance of the corrupting noise. However, the side-effect of this approach is to end up with an oversmoothed restored tone which lacks high frequencies. A remedy to the oversmoothing effect is to apply the SSM-based bandwidth extension to recover the signal information that was lost due to the aggressive de-hissing procedure.

The previous two-steps strategy was found to be effective to de-hiss guitar tones, as can be seen in Fig. 5. In this experiment, a zero mean white Gaussian noise sequence was added to the test guitar tone and its variance was adjusted to produce a global SNR of 20 dB. As can be seen in the top plot of Fig. 5, the noise masks the high-frequency harmonics of the tone.

The first step of the restoration procedure consisted of de-hissing the noisy signal through a Wiener filtering scheme, as described in [11]. Here, signal frames of 256 samples were used with an overlap of 50%.The noise variance was estimated in the frequency domain by taking the mean value of the upper quarter of the power spectrum. Additionally, a gain was assigned to the noise variance estimate. This gain, which hereafter will be called noise floor gain, worked as a control parameter for the amount of noise to be removed.

Considering the Wiener filter configuration and the test signal used in this experiment, it was found that a noise floor gain of 30 suffices to almost eliminate the residual noise effects in the restored signal. This can be verified in the middle plot of Fig. 5, as well as the strongly smoothed, i.e. lowpass filtered, characteristic.

The last step consisted of extending the bandwidth of the previously de-hissed guitar tone using the SSM-based scheme described in Section 3. The same approaches to estimate the string-model parameters and to generate the additional excitation signal were employed in this experiment. These choices were found to generate a restored tone whose timbre is similar to that of uncorrupted tone, without the annoying effects of the residual noise as can be seen in Fig. 5 (bottom). Sound examples are available at URL: http://www.acoustics.hut.fi/publications/papers/fs01-ssm/

5. CONCLUSIONS

In this paper, the enhancement of guitar tones was presented within a sound source modeling framework. First, it was shown how the reconstruction of spectral information in guitar tones can be attained by means of SSM techniques. Then, the SSM-based bandwidth extension scheme was applied as a post-processing stage after a traditional spectral-based de-hissing method. The obtained results for both the bandwidth extension and the de-hissing experiments demonstrate that the proposed schemes are effective in improving the perceptual quality of the restored tones.

Although showing some potential, the use of SSM for audio enhancement purposes is still restricted to special cases. Even if the attempt is to restore solo guitar music, SSM-based techniques face challenging tasks related to content-based representations of music. As an example, the separation of tones whose content overlaps both in time and frequency as well as the extraction of their musical features can be mentioned. Extensions to more

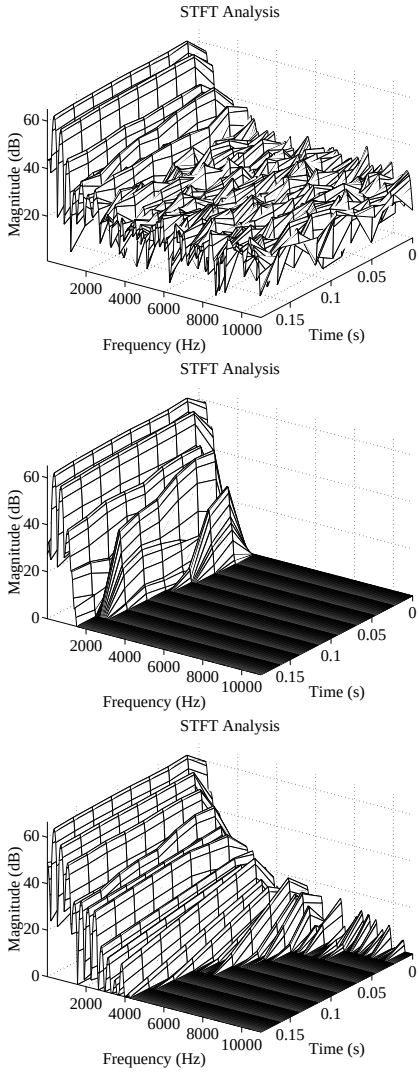


Fig. 5. Time-frequency analysis of the noisy tone (top), the de-hissed tone (middle), and the bandwidth extended tone (bottom).

general cases can be viewed as a multi-layered problem, which would include separation of more general musical elements in complex sound sources. On the synthesis side of the chain, the requirements are related to the development of model-based music synthesizers with more realistic sounds, and capable of simulating the playing features of real performances.

Finally, it is worth mentioning that SSM- and content-based audio processing is still in a youthful stage of development. However, as long as it develops into better ways to represent and recreate sound sources, performing audio enhancement within the SSM framework can lead to better results compared to those attained by using traditional techniques.

ACKNOWLEDGMENT

The work of P. Esquef has been supported by a scholarship from CNPq-Brazil and the Sound Source Modeling project of the Academy of Finland. V. Välimäki has been financed by a postdoctoral research grant from the Academy of Finland.

REFERENCES

[1] B. L. Vercoe, W. G. Gardner, and E. D. Scheirer, "Structured audio: creation, transmission, and rendering of parametric sound representations," *Proceedings of IEEE*, vol. 86, no. 5, pp. 922–940, 1998.

[2] A. Klapuri, "Automatic transcription of music," M.Sc. thesis, Tampere Univ. of Technology, Tampere, Finland, 1998, (Electronic version available at URL: http://www.cs.tut.fi/~klap/iiro/).

[3] T. Tolonen, V. Välimäki, and M. Karjalainen, "Evaluation of modern sound synthesis methods," Tech. Rep. 48, Helsinki Univ. of Technology, Lab. of Acoustics and Audio Signal Processing, Espoo, Finland, 1998, (Electronic version available at URL: http://www.acoustics.hut.fi/publications/).

[4] J. O. Smith, "Efficient synthesis of stringed musical instruments," in *Proc. Int. Computer Music Conf., ICMC'93*, Tokyo, Japan, 1993, pp. 64–71.

[5] M. Karjalainen, V. Välimäki, and Z. Jánosy, "Towards high-quality sound synthesis of the guitar and string instruments," in *Proc. Int. Computer Music Conf., ICMC'93*, Tokyo, Japan, 1993, pp. 56–63.

[6] V. Välimäki, J. Huopaniemi, M. Karjalainen, and Z. Jánosy, "Physical modeling of plucked string instruments with application to real-time sound synthesis," *J. Audio Eng. Soc.*, vol. 44, no. 5, pp. 331–353, 1996.

[7] T. I. Laakso, V. Välimäki, M. Karjalainen, and U. K. Laine, "Splitting the unit delay — tools for fractional delay filter design," *IEEE Signal Proc. Mag.*, vol. 13, no. 1, pp. 30–60, 1996.

[8] T. Tolonen and H. Järveläinen, "Perceptual study of decay parameters in plucked string synthesis," in *Proc. AES 109th Convention*, Los Angeles, California, USA, 2000, Preprint 5205.

[9] O. Christensen and B. B. Vistisen, "Simple model for low-Frequency guitar function," *J. Acoust. Soc. Am.*, vol. 68, pp. 758–766, 1980.

[10] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech and Audio Process.*, vol. 2, no. 2, pp. 345–349, 1994.

[11] S. J. Godsill and P. J. W. Rayner, *Digital Audio Restoration — A Statistical Model Based Approach*, Springer-Verlag, London, UK, 1998.

EQUALIZATION AND MODELING OF AUDIO SYSTEMS USING KAUTZ FILTERS

Tuomas Paatero and Matti Karjalainen

Helsinki University of Technology
Laboratory of Acoustics and Audio Signal Processing
P. O. Box 3000, FIN-02015 HUT
FINLAND
tuomas.paatero@hut.fi

ABSTRACT

This paper demonstrates the applicability of Kautz filters in audio signal processing. New methods for the choosing of Kautz filter poles are presented and utilized in two audio oriented applications.

1. INTRODUCTION

Frequency warping using allpass structures or *Laguerre filters* [7] has found increasingly applications in audio signal processing due to good match with the auditory frequency resolution [3, 8]. *Kautz filters* [6, 2] can be seen as a further generalization where each transversal element may be different, including complex conjugate poles. This enables arbitrary allocation of frequency resolution for filter design, such as modeling and equalization (inverse modeling) of linear systems.

After a brief theoretical background of implementation and design principles, we present two examples as case studies of using Kautz filters in modeling and inverse modeling of audio systems. In the first case we apply the method to loudspeaker response equalization. The second case deals with the modeling of guitar body impulse response.

2. KAUTZ FUNCTIONS AND FILTERS

For a given set of desired poles $\{z_i\}$ in the unit disk, the corresponding set of *rational orthonormal functions* is uniquely defined in the sense that the lowest order rational functions, square-integrable and orthonormal on the unit circle, analytic for $|z| > 1$, are of the form [9]

$$G_i(z) = \frac{\sqrt{1 - z_i z_i^*}}{z^{-1} - z_i^*} \prod_{j=0}^{i} \frac{z^{-1} - z_j^*}{1 - z_j z^{-1}}, \quad i = 0, 1, \ldots . \quad (1)$$

A Kautz filter is a finite weighted sum of functions (1), which reduces to a transversal structure of Fig. 1. Defined

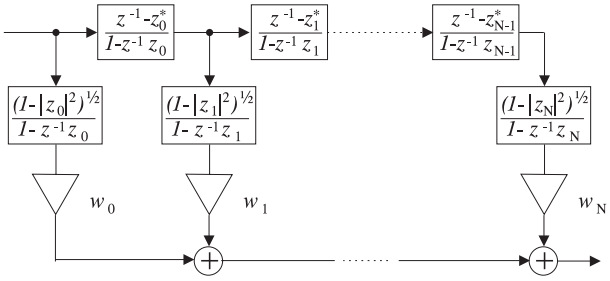


Figure 1: The Kautz filter. For $z_i = 0$ in (1) it degenerates to an FIR filter and for $z_i = a, -1 < a < 1$, it is a Laguerre filter where the tap filters can be replaced by a common prefilter.

in this manner, Kautz filters are merely a class of fixed-pole IIR filters, forced to produce orthonormal tap-output impulse responses. However, the fact that functions (1) provide a (*Fourier*) basis representation for any causal and finite-energy signal or system allows for linear-in-parameter models for many types of system identification and approximation schemes, including adaptive filtering, both for fixed and non-fixed pole structures. Here we address only the "prototype" least-square (LS) approach to approximation, implied by the orthonormal Fourier series expansion with respect to functions (1).

A Kautz filter produces real tap output signals only in the case of real poles. However, from a sequence of real or complex conjugate poles it is always possible to form real orthonormal structures. From the variety of possible solutions it is sufficient to use the intuitively simple structure of Fig. 2, proposed by Broome: the second-order section outputs of Fig. 2 are *orthogonal* from which an orthogonal tap output pair if formed [2]. Normalization terms are completely determined by the corresponding pole pair $\{z_i, z_i^*\}$ and are given by $p_i = \sqrt{(1-\rho_i)(1+\rho_i-\gamma_i)/2}$ and $q_i = \sqrt{(1-\rho_i)(1+\rho_i+\gamma_i)/2}$, where $\gamma_i = -2RE\{z_i\}$ and $\rho_i = |z_i|^2$ can be recognized as corresponding second-order polynomial coefficients. The construction works also for

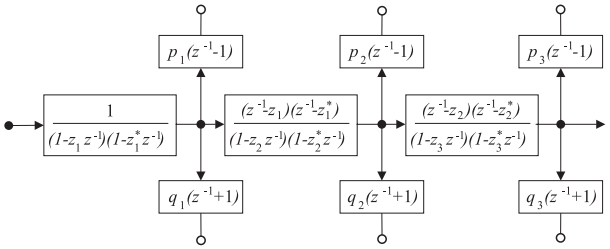


Figure 2: One realization for producing real Kautz functions from a sequence of complex conjugate pole pairs.

real poles but we use an obvious mixture of first- and second-order sections, if needed.

2.1. Kautz filter design

Kautz filter design can be seen as a two-step procedure involving the choosing of a particular Kautz filter (i.e., the poles) and the evaluation of the corresponding filter weights. For the latter, and in the case of a given target response $h(n)$ or $H(z)$, we use simply the Fourier coefficients, $c_i = (h, g_i) = (H, G_i)$, easily obtained by feeding the signal $h(-n)$ to the Kautz filter and reading the tap outputs $x_i(n) = G_i[h(-n)]$ at $n = 0$: $c_i = x_i(0)$. This implements convolutions by filtering and it can be seen as a generalization of rectangular window FIR design.

The contrast between the easy and well-defined model parameterization task and the complicated and non-linear model selection problem makes it tempting to use sophisticated guesses and random or iterative search in the pole position optimization. As a more analytic approach, the whole idea in the Kautz concept is how to incorporate desired *a priori* information to the Kautz filter. This may mean knowledge on system poles or resonant frequencies and corresponding time-constants, or indirect means, such as all-pole or pole-zero modeling. Furthermore, we have adopted a method proposed originally to pure FIR-to-IIR filter conversion [1], to the context of Kautz filter pole optimization. It resembles the iterative Steiglitz-McBride method of pole-zero modeling, but it genuinely and effectively optimizes (in the LS sense) the pole positions of a real Kautz filter, producing unconditionally stable and (theoretically globally) optimal pole sets for a desired filter order. In this paper we use the above *BU-method* as such or combined with, e.g., warped design or manual tuning of poles.

3. AUDIO APPLICATION EXAMPLES

We demonstrate the applicability of Kautz filter design in two different types of audio-oriented applications. The first one is the loudspeaker equalization task where frequency resolution is distributed both globally and locally. In the second case we use Kautz filters to model the body response of an acoustic guitar where the lowest frequencies are of primary interest.

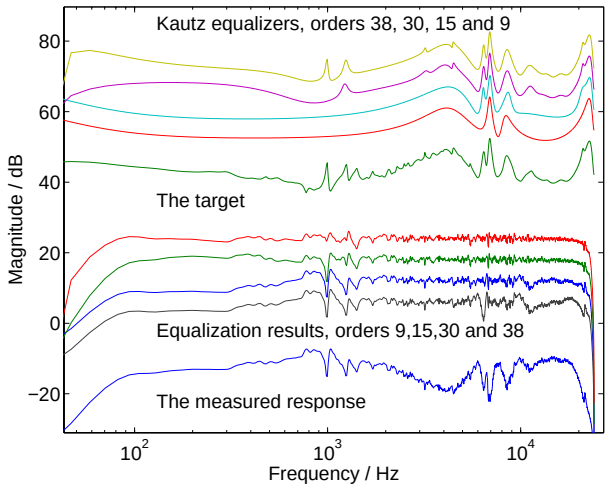


Figure 3: Kautz equalizers and equalization results for orders 9, 15, 30 and 38, compared to the measured loudspeaker response and the equalizer target response.

3.1. Example 1: Loudspeaker equalization

An ideal loudspeaker has a flat magnitude response and a constant group delay. Simultaneous magnitude and phase equalization would be achieved by modeling the response and inverting the model, or by identifying the overall system of the response and the Kautz equalizer, but here we demonstrate the use of Kautz filters in pure magnitude equalization, based on an inverted target response. The measured loudspeaker magnitude response and a derived equalizer target response are included in Fig. 3. The sample rate is 48 kHz.

As is well known, FIR modeling has an inherent emphasis on high frequencies on the auditorily motivated logarithmic frequency scale. Warped FIR (or Laguerre) [3] filters release some of the resolution to the lower frequencies, providing a competitive performance with 5 to 10 times lower filter orders than with FIR filters [4]. However, the filter order required to flatten the peaks at 1 kHz in our example is still high, of the order 200, and in practice Laguerre models up to order 50 are able to model only slow trends in the response. The proposed BU-method provides good pole sets for orders at least up to 40 and in Fig. 3 we have presented Kautz equalizers and equalization results for orders 9, 15, 30 and 38. For orders above 15, the BU-method produces poles really close to z=1 and omitting some of these poles actually tranquilize the low frequency region.

To improve the modeling at 1 kHz, we added three to four

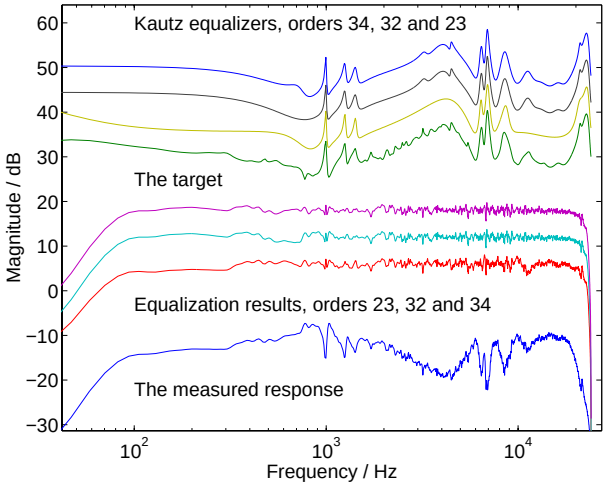


Figure 4: Kautz equalizers and equalization results for orders 23, 32 and 34, with combinations of manually tuned and BU-poles.

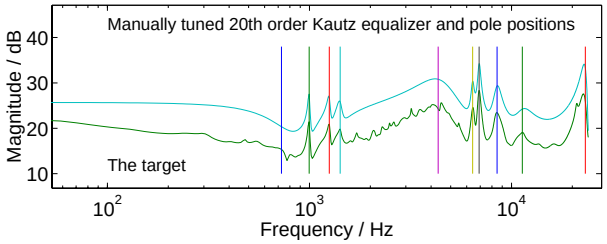


Figure 5: Manually tuned 20th order Kautz equalizer and the target magnitude response.

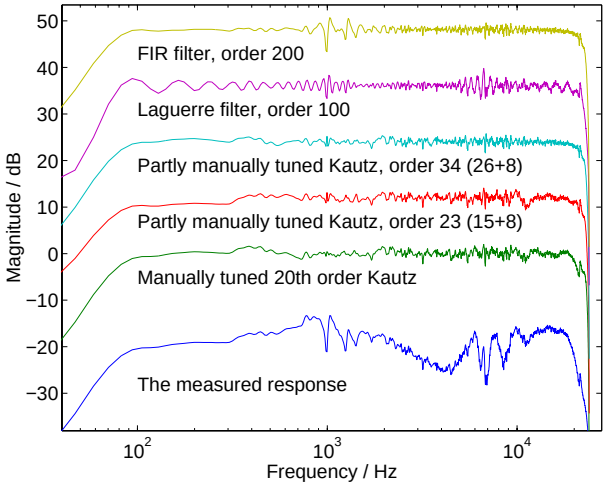


Figure 6: Comparison of FIR, Laguerre, and Kautz equalization results.

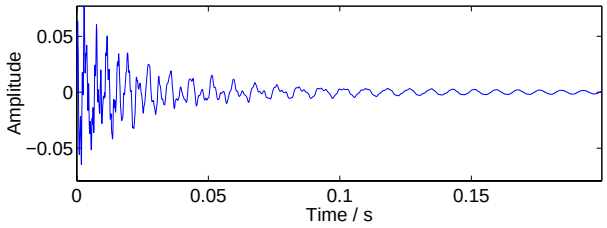


Figure 7: The measured impulse response of an acoustic guitar body.

manually tuned pole pairs to the BU-pole sets, corresponding to the resonances in the problematic area. Results for final filter orders 23, 32 and 34 are displayed in Fig. 4.

Finally, we abandon the pole sets proposed by the BU--method and try to tune 10 pole pairs manually to the target response resonances. The design is based on 10 selected resonances, represented with 10 distinct pole pairs, chosen and tuned to fit the magnitude response (Fig. 5).

A comparison of equalization results for some of the previous Kautz equalizers, and those achieved with FIR and Laguerre equalizers of orders 200 and 100, respectively, is presented in Fig. 6.

3.2. Example 2: Acoustic guitar body modeling

As another example of Kautz modeling we approximate a measured acoustic guitar body response sampled at 24 kHz (Fig. 7). The obvious disadvantage of a straightforward FIR filter implementation is that modeling of the slowly decaying lowest resonances requires a very high filter order. All-pole or pole-zero modeling are the traditional choices in improving the flexibility of the spectral representation. However, model orders remain problematically high and the basic design methods seem to work poorly. Perceptually motivated warped counterparts of all-pole and pole-zero modeling pay off, even in technical terms [5], but here we want to focus the modeling resolution more freely.

Figure 8 demonstrates that the BU-method is able to capture essentially the whole resonance structure. The Kautz filter order is 102 and the poles are obtained from a 120th order BU-pole set, omitting some poles close to $z = -1$. Lower-order models are achieved, e.g., by further pruning of the pole set.

Especially in this case of a target response dominated by the low-frequency part, we may compose very low order Kautz models with a combination of warping and BU-method: the BU-method is first applied to the *warped target response* [3] and then the poles are mapped back to the original frequency domain according to the *inverse allpass transformation*. In Fig. 9 are presented the magnitude responses of the attained Kautz models for orders 10, 16, 20 and 40,

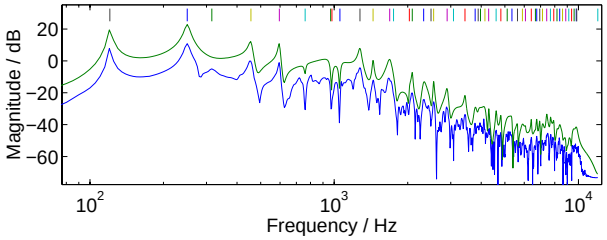


Figure 8: A 102th order Kautz model and the target magnitude response, and vertical lines indicating BU pole pair positions.

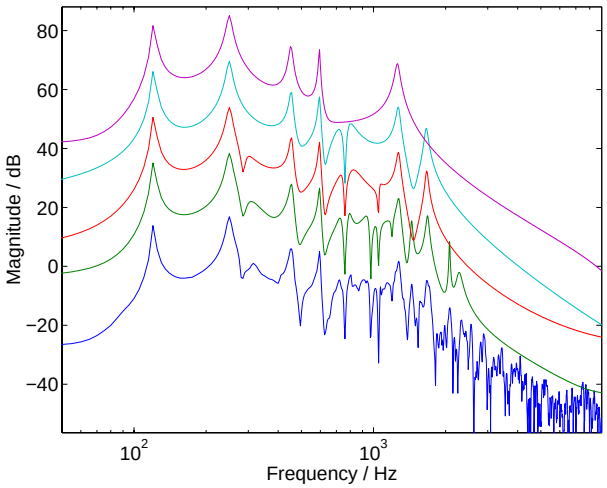


Figure 9: Displayed with offset from top to bottom, Kautz models of orders 10, 16, 20 and 40, and the target magnitude response.

where we used (allpass) warping parameter $\lambda = 0.7$. It is quite surprising that the BU-method found the five prominent resonances at model order 10, i.e., with exactly five complex conjugate pole pairs, in contrast to the unwarped case, where the required filter order is about 100.

Finally, in Fig. 10 we demonstrate that good fit to the five prominent resonances of the 10th order Kautz filter of Fig. 9 means also good match in the time-domain.

4. CONCLUDING REMARKS

We have demonstrated the potential applicability of Kautz filters in some typical audio signal processing tasks. They are found flexible generalizations of FIR and Laguerre filters, providing IIR-like spectral modeling capabilities with well-known favorable properties resulting from the orthonormality. A more detailed presentation of the underlying theory and the merely stated audio application results can be found in other related `~/publications` at

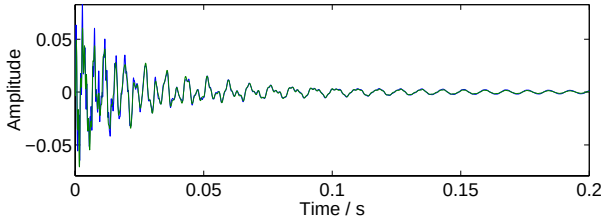


Figure 10: The impulse response of the 10th order Kautz filter compared to the measured response.

`http://www.acoustics.hut.fi` as well as MATLAB scripts and demos in `~/software/kautz`.

5. ACKNOWLEDGEMENTS

This work has been supported by the Academy of Finland as a part of project “Sound source modeling”.

6. REFERENCES

[1] H. Brandenstein and R. Unbehauen, “Least-Squares Approximation of FIR by IIR Digital Filters”, *IEEE Trans. Signal Processing*, vol. 46, no. 1, pp. 21–30, 1998.

[2] P. W. Broome, “Discrete Orthonormal Sequences”, *Journal of the Association for Computing Machinery*, vol. 12, no. 2, pp. 151–168, 1965.

[3] A. Härmä, M. Karjalainen, L. Savioja, V. Välimäki, U. K. Laine and J. Huopaniemi, “Frequency warped signal processing for audio applications”, *J. Audio Eng. Soc.*, vol. 48, no. 11, pp. 1011–1031, 2000.

[4] M. Karjalainen, E. Piirilä, A. Järvinen and J. Huopaniemi, “Comparison of loudspeaker equalization methods based on DSP techniques”, *J. Audio Eng. Soc.*, vol. 47, no. 1/2, pp. 15–31, 1999.

[5] M. Karjalainen and J. O. Smith, “Body modeling techniques for string instrument synthesis”, *Proc. Int. Computer Music Conf.*, Hong Kong, Aug. 1996, pp. 232–239.

[6] W. H. Kautz, “Transient Synthesis in the Time Domain”, *IRE Trans. Circuit Theory*, vol. CT–1, pp. 29–39, 1954.

[7] Y. W. Lee, *Statistical Theory of Communication*. John Wiley and Sons, New York, 1960.

[8] J. O. Smith and J.S. Abel, “Bark and ERB bilinear transform”, *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 6, pp. 697–708, 1999.

[9] J. L. Walsh, *Interpolation and Approximation by Rational Functions in the Complex Domain*, 2nd Edition. American Mathematical Society, Providence, Rhode Island, 1969.

Automatic Test Image Generation by Genetic Algorithms for Testing Halftoning Methods - Comparing Results using Wavelet Filtering

Timo Mantere and *Jarmo T. Alander*

University of Vaasa
Department of Information Technology and Production Economics
P.O. Box 700, FIN-65101 Vaasa
FINLAND
E-mail: *firstname.lastname*@uwasa.fi

ABSTRACT

This work introduces automatic test image generation by genetic algorithms for testing different halftoning methods. In general, the proposed method has potential in software test data generation. The goal was to reveal, if genetic algorithm is able to generate images that are difficult for the object software to halftone, in other words to find if some prominent characteristics of the original image disappear or ghost features appear due to the halftoning process. Using a Haar wavelet based fitness function did image comparison between the input image and the corresponding halftoned image.

1. INTRODUCTION

There does not seem to be much research in the field of test image evaluation. How to determine or create a good test image. What are the essential characteristics of a good test image? How to determine that a particular image is good for testing some specific image processing algorithm? More often than not researchers rely on commonly used but very limited test image sets. We encountered this problem, when we wanted to test the image-processing system we implemented for an ink jet marking machine [1, 2]. In our other study [3, 4, 5] we used genetic algorithms (GA) for software testing purposes. In this work, we try to combine the knowledge of these two previous studies in order to use GA for generating test images for halftoning methods. This study concentrates on finding how wavelets can be adapted to image comparison as essential component of the fitness function.

1.1. Genetic algorithms

Genetic algorithms [6] are optimization methods that mimic evolution in nature. They are simplified computational models of evolutionary biology. A GA forms a kind of electronic population, the members of which fight for survival, adapting as well as possible to the environment, which is actually an optimization problem. GAs use genetic operations, such as selection, crossover, and mutation in order to generate solutions that meet the given optimization constrains ever better and better. Surviving and crossbreeding possibilities depend on how well individuals fulfill the target function. The set of the best solutions is usually kept in an array called population. GAs do not require the optimized function to be continuous or derivable, or even be a mathematical formula, and that is perhaps the most important factor why they are gaining more and more popularity in practical technical optimization. The genetic algorithm (GA) in this study was written in Java. One of the advantages of Java is its easiness to use image handling procedures. However, the execution speed of Java programs may not be the best possible.

1.2. Dithering

Digital halftoning [7], or dithering, is a method used to convert continuous tone images into images with a limited number of tones, usually only two: black and white. The main problem is to do the halftoning, so that the bi-level output image does not contain artifacts, such as alias, moiré, lines or clusters, caused by dot placement [8]. The average density of the halftoned dot pattern should interpolate as precisely the original image pixel values as possible. Dithering methods include static methods, where each pixel is compared to a threshold value that is obtained e.g. from a threshold matrix, generated randomly or is a static median value. Depending on matrix this method can create both frequency or amplitude modulated halftones. There are also error diffusion methods, such as Floyd-Steinberg and Jarvis-Judge-Ninke coefficients. In these methods the rounding error of the current pixel is spread on those neighboring pixels, the bi-level value of which is not yet determined.

This study concentrates only on frequency modulated halftoning methods. The halftoning methods used here were Floyd-Steinberg (`FS`), and Jarvis-Judge-Ninke (`JJN`) error diffusions and thresholding with 16×16 ordered threshold matrix [7] (`THO`), and with GA optimized 16×16 threshold matrix [2] (`THG`). Also rounding (`NEAR`) the nearest bi-level tone (black/white) was used to compare results with, since it should lead rather poor halftone result and therefore lead worst results with each target function.

1.3. Haar Wavelet

The wavelet transforms [9] are signal processing operations that decompose signals into components at different frequency scales. A wavelet transform represents a sum of wavelets on different locations and scales. It is based on multiresolution analysis. The most well known and sim-

plest wavelet is Haar function (filter). The characteristic property of Haar function is sharp edges. Haar filter is special case of Daubechies filter family; it is actually order one Daubechies filter, and the only one of that orthonormal filter family that has explicit expression. Decomposition in the Haar basis eliminates high frequency terms when the input sequence is constant. Haar function is often used for images with high contrast of black and white; therefore, we can assume Haar function well suitable when applied to halftoned images.

1.4. Comparing the images

Comparing a dithered image with the original one is obviously a challenging problem. One cannot simply use pixel by pixel comparison, since dithered images usually have only two tones. The minimum difference by that measure would be achieved if every gray tone were rounded to the nearest tone (black or white), which in practice usually results in poor images. Better image comparison methods have been developed [7, 10].

In addition, a set of methods called inverse halftoning [7] has been developed. From these the perhaps most common is the low pass filtering method. In this method, images are first low pass filtered and the resulting images are then compared pixel by pixel. The problem with low-pass filtering is that the high frequencies will disappear and the images get a somewhat blurred overall appearance. However, this method is easy to implement and it enables pixel by pixel comparison. In a way the blurring by low pass filtering also resembles human eye perception: when we look the image from a distance the small details disappear and the visual observation of larger objects is averaged out from the small details.

If the images are not compared properly, the received evaluated difference between images may as well depend on the comparison method used as the actual difference between the images, i.e. the dithering methods used. Several fitness functions i.e. image comparison methods were tested in refs. [11, 12].

This study concentrates on using Haar wavelet in the comparison. Since the original and halftoned image represents the same image, Haar wavelet coefficients of them should be related. The lower frequency coefficients should be quite similar, since the average gray for both images should be approximately the same. The higher the frequency the more the coefficients are likely to diverge. A certain weight coefficient for each frequency scale is used to determine the significance of different level wavelet coefficient difference.

1.5. Related work

Wavelets have been applied for finding similarities on images i.e. image comparison in refs [13, 14]. In the previous studies on finding optimum halftone patterns the human eye modulation transform function [15] is considered the best method, while especially optimization speed may favor more simple methods.

Genetic algorithm were previously adapted to the dithering problem [16, 17]. For further references of GAs in image processing see e.g. bibliography [18] or book [19]. Image generation with GA is used at least in ref. [20]. Image generation for algorithm validation is represented in ref. [21]. GAs has previously been adapted to automatic software test data generation in several studies, see refs. [3–5] and references therein.

2. THE PROPOSED METHOD

This work is a continuation to that given in refs. [11, 12]. The image comparison in those papers were done using a) pixel by pixel comparison using low pass filtered images, b) tone difference between consecutive pixels in each image, c) the average density at the corresponding image areas, d) a hybrid of the three previous methods, and e) edge detector and comparing edge locations.

The GA runs as an independent program and optimizes parameter vectors which are used by an image generator to create images, which are further sent to the object software, that halftones it and returns the resulting image. The pixelgrapper reads pixels from both the test image and its halftoned transformation image and transmits 8-bit pixel arrays of both images to the fitness function evaluator. The difference between these images is used as the fitness function. GA generates new parameter vectors by using crossover and mutation, favoring those parent chromosomes that previously had gotten a high fitness value. Test images in this study were created by optimizing parameters, such as place, size and color of elementary graphical objects, like lines, rectangles, circles and ASCII characters, together with the background tiles and colors all encoded as one GA chromosome.

2.1. Implementation

The implementation used integer coded GA, where the chromosome consisted of total 79 parameters. From those the first seven parameters were for background, three of them break background into four segments and the other parameters determine the tone b of each background segment. This way one parameter does not dominate optimization. However, the background might still become monotone if one segments takes the whole space or the tone parameters b_n are equal. Next 70 parameters where divided into 10 groups of 7 parameters, each 7 parameter long group defines one elementary image object as follows:

1. Image object (line, rectangle, oval, ASCII character); for characters also the font style.
2. Object color.
3. Object starting point; x coordinate.
4. Object starting point; y coordinate.
5. Object length in x coordinate direction or character font size.
6. Object length in y coordinate direction or character font type.
7. Not used or the character value (only printable ASCII characters were used).

All objects are opaque and may cover partly or totally earlier created objects. Background is created first and then the other objects on it.

The generated image as such was still quite monotonous. Normal image usually has more variation between neighboring pixels. Our test image was further diversified by adding chaotic data with Verhulst [22] logistic equation: $x_i = ax_{i-1} \times (1 - x_{i-1})$. The chaotic data was used rather than random noise in order to control diversity and to keep the added noise repeatable. The last two parameters of the chromosome forms 16-bit value a for Verhulst function that was scaled to be a decimal number in range [2, 4]. The optimization process usually favored such chaos parameters that generated striped patterns rather than patterns that resemble uniform random noise.

The size of the generated image was selected to be 256×256 pixels, so that the values of most parameters would fit into eight bits. Population size was 50, elitism 40%, total 3050 evaluations (initial population + 100 generations) were done, uniform crossovers was used, and the mutation probability was 1%.

3. EXPERIMENTAL RESULTS

These experiments concentrated on finding a target function based on Haar wavelet coefficients. Target function 1 is given by equation (1). The notations used are the following: A = original image, B = halftoned image, S = wavelet coefficients, K represents different scales (levels), i and j are indices to the wavelet coefficients of particular level, W is weight coefficient for difference of wavelet coefficients, and X represents the threshold against which the wavelet coefficient are compared with. This target function compares images by counting the amounth of wavelet coefficients that differs from each other over some threshold value, which may be different for different frequency scales. In practice if the sum is small the images are similar to each other.

$$\sum_K \left(W_K \times \left\{ \sum_{i,j} \left| \left[A.S_K(i,j) - B.S_K(i,j) \right] \right| > X_K \right\} \right) \quad (1)$$

Haar wavelet coefficients for the original and halftoned images should be quite similar at large scales, however the difference between them tends to increase the higher the frequency scale. However, when comparing images the coefficient similarity for the lower frequency scale is more important.

Table 1 represents the maximal fitness values obtained using formula 1 for each test image set. Standard test image set (STD) contains 13 test images [23] {Airplane, Barbara, Bird, Boat, Bridge, Camera, Frog, Goldhill, Lenna, Mandrill, Peppers, Washsat, and Zelda} that are often used for expressing image processing systems. Random noise images (RN) contained 10 gray noise images generated by a random generator. One tone images (BCR) contains 256 possible 8-bit one tone gray images. Test images (GI) generated by genetic algorithm contained best values obtained from five different GA optimization runs.

Table 1. Best fitness values for different test image sets with target function 1.

	FS	JJN	THO	THG	NEAR
STD	280827	396177	324519	331283	518184
RN	268107	286279	371756	400798	402969
BCR	301234	348930	262144	259072	16386
GI	369576	471069	500874	465104	525869
Max(STD)	Peppers	Goldhill	Mandrill	Bridge	Frog
Max(BCR)	30, 225	60, 195	∈ [86, 170]	162	∈ [1, 254]

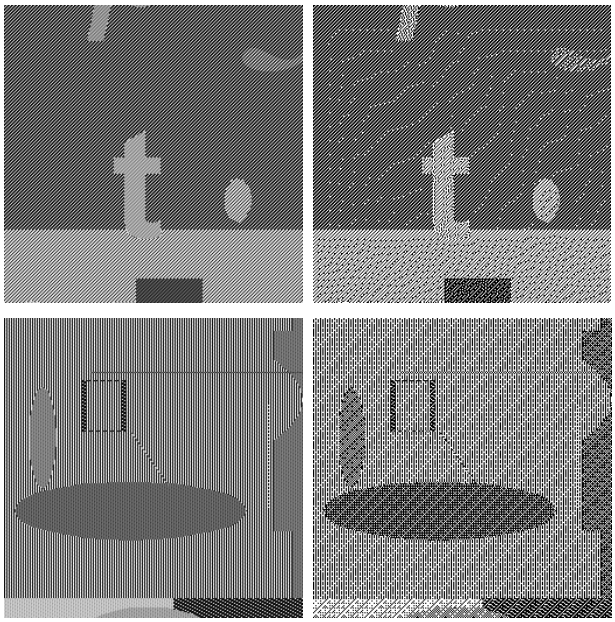


Fig. 1. An example of a GA generated test images and the corresponding halftoned images
a) Test image for FS. b) Dithered a.
c) Test image for THO. b) Dithered c.

With this target function GA was able to generate test image that resulted proportionally highest fitness value for each of the tested halftoning methods. This implies that our GA based image generator can optimize test images according to the given function.

From table 1 we can further see that according to the target function FS is best dithering method for all other test images, except for one images (BCR). What is notable here is that random noise is not the worst case for any halftoning method. If we want to analyze the test images generated by GA and evaluate what properties caused the high difference between the compared images we want them to have some properties that a human eye can observe. If our target function considered random noise the worst case we would not be able to generate any sensible test images. Random noise is still random noise after halftoning and human eye does not see that much difference when comparing original noise and halftoned one.

Figure 1 shows examples of GA generated test image. With FS and JJN the typical test images shows a kind of ghost worms c.f. (fig. 1b). With the ordered threshold matrix method the test images caused the kind of behavior where the vertical stripes caused by certain chaos parameter in the original image are changed into crosswise stripes when halftoned (fig. 1d).

4. CONCLUSIONS AND DISCUSSION

The results got seem to confirm that our genetic algorithm based image generator is capable of generating test images for testing different halftoning methods according to a given target function. With most cases, GA was able to generate test images that resulted highest difference score. In the cases were GA did not generated highest value it still reached very close to the highest value obtained with our static test image set.

In most cases, the halftoned images show some properties that evidently differ from the original images. This supports the proposal that wavelet based image comparison methods are worth considering.

4.1. Future

However there may not exist only one universal "right" way to compare halftoned images with originals. Different target functions may discover different kinds of dissimilarities between images. One future research alternative is to find a good set of comparison functions that together discovers all possible different types of dissimilarities.

The use of wavelets as a hybrid with other methods could be studied. The use of also other than Haar wavelets in image comparison could be studied. The possibilities of applying fuzzy logic to image comparison is under study.

GA coding can be improved. Integer coded GA may not be the most suitable for this problem. It is planned that future version will more freely create desired objects. More massive test runs may eliminate the bias of background tone dictation. The significance of other objects and their position in the image may be identified if we use static background tones and let other features settle.

However, so far the work is been mostly experimental, the goal has been to solve what this kind of optimization approach results in software testing, and how the method could be further improved.

After a satisfying fitness function has been found, the obvious application of the above testing method is automatic dithering method design. One GA generates halftone filters while the other GA tries to create the hardest test image for each filter. The best filter being the one where the hardest test image is closest to the original after dithering. In general, this kind of differential evolution based approach could be used in the design and testing of demanding software.

REFERENCES

[1] J. T. Alander, T. Mantere, and T. Pyylampi, Threshold matrix generation for digital halftoning by genetic algorithm optimization. In D. P. Casasent, ed., *Intelligent Systems and Advanced Manufacturing: Intelligent Robots and Computer Vision XVII: Algorithms, Techniques, and Active Vision, volume SPIE-3522*, Boston, MA, 1.-6. November 1998. SPIE, 1998, pp. 204-212.

[2] J. T. Alander, T. Mantere, and T. Pyylampi, Digital halftoning optimization via genetic algorithms for ink jet machine. In B. Topping, ed., *Developments in Computational Mechanics with High Performance Computing*, CIVIL-COMP Press, Edinburg, UK, 1999, pp. 211-216.

[3] J. T. Alander, T. Mantere, G. Moghadampour and J. Matila, Searching protection relay response time extremes using genetic algorithm – software quality by optimization. *Electric Power Systems Research 46*, 1998, pp. 229-233.

[4] J. T. Alander, and T. Mantere. Automatic software testing by genetic algorithm optimization, a case study. In C. Ryan and J. Buckley (eds.) *SCASE'99 - Soft Computing Applied to Software Engineering*, 11.-14.4.1999, Limerick, Ireland, 1999, pp. 1-10.

[5] T. Mantere, and J. T. Alander. Automatic Software Testing by Optimization with Genetic Algorithms Introduction to the Method and Consideration of the Possible Pitfalls. Submitted to *MENDEL 2001*, June 6-8, 2001, Brno, Check Republic, 5 pages..

[6] J. Holland, *Adaptation in Natural and Artificial Systems*. University of Michigan Press. Ann Arbor, MI, Reissued by The MIT Press, 1992.

[7] H. R. Kang, *Digital Color Halftoning*. SPIE Optical Engineering Press, Bellingham, Washington, & IEEE Press, New York, 1999.

[8] P. G. J. Barten, *Contrast Sensitivity of the Human Eye and Its Effects on Image Quality*. SPIE Optical Engineering Press, Bellingham, Washington, USA, 1999.

[9] J. C. Goswami and A. K. Chan. Fundamentals of Wavelets – Theory, Algorithms, and Applications. John Wiley & Sons.

[10] F. Nilsson, Objective quality measures for halftoned images. *Optics, Image, Science and Vision*, Volume 16, Number 9, September 1999, pp. 2151-2162.

[11] T. Mantere, and J. Alander. Automatic test image generation by genetic algorithms for testing halftoning methods. In D. P. Casasent, editor, *Intelligent Systems and Advanced Manufacturing: Intelligent Robots and Computer Vision XIX: Algorithms, Techniques, and Active Vision, volume SPIE-4197*, Boston, MA, 5.-8. November 2000, SPIE, Bellingham, Washington, pp. 297-308.

[12] T. Mantere, and J. Alander. Testing Halftoning Methods by Images Generated by Genetic Algorithms. Submitted to *Arpakannus*, 2001.

[13] M. K. Mandal, T. Aboulnasr, and S. Panchanathan. Image indexing using moments and wavelets. *IEEE Transaction on Consumer Electronics*, Vol. 42, No. 3, August 1996.

[14] J. Ze Wang, G. Wiederhold, O. Firschein, and S. Wei. Wavelet-based image indexing techniques with partial sketch retrieval capability. IEEE 1997.

[15] J. Sullivan, L. Ray and R. Miller (1991). Design of Minimum Visual Modulation Halftone Patterns. In IEEE Transactions on Systems, Man, and Cybernetics, *Vol. 21, No 1*, January/February 1991.

[16] N. Kobayashi, and H. Saito, Halftoning Technique Using Genetic Algorithms. In *Systems and Computers in Japan* **27**, 1996, pp. 89-97.

[17] J. Newbern, and V. M. Bove, Jr., Generation of blue noise arrays using genetic algorithm. In *Human Vision and Electronic Imaging II*, B.E. Rogowitz and T.N. Pappas, eds., vol SPIE-3016, SPIE, Bellingham, San Jose, CA, 10.-13. Feb. 1997, pp. 401-450.

[18] J. T. Alander, *An Indexed Bibliography of Genetic Algorithms in Optics and Image Processing*. Department of Information Technology and Production Economics. University of Vaasa. Report Series No. 94-1-OPTICS, 2000, Available ftp://garbo.uwasa.fi /cs/report94-1/gaOPTICSbib.ps.Z

[19] S. K. Pal, A. Ghosh, and M. K. Kundu, *Soft Computing for Image Processing*. Physica-Verlag, Heidelberg, New York, 1999.

[20] K. Sims, Artificial Evolution for Computer Graphics. *Computer Graphics (Siggraph '91 Proceedings)*, July 1991, pp.319-328.

[21] M. Miyojim, and H. Cheng, Synthesized images for pattern recognition. In *Pattern Recognition*, Vol. 28, No. 4, 1995, pp. 595-610.

[22] P. S. Addison, *Fractals and chaos: An illustrated course*. Bristol, Philadelphia, Institute of Physics Publishing, 1997.

[23] Michael Frydrych. Test images. Available http://www.it.lut.fi /research/ip/images/Standard/index.html

Note on Connections between Active Contours and Rayleigh Quotients

Jussi Tohka

Signal Processing Laboratory,
Tampere University of Technology,
P.O.Box 553, FIN-33101 Tampere
FINLAND
Tel. 358-3-3654508 Fax: +358-3-3653087
E-mail: jussi.tohka@cs.tut.fi
URL: http://www.cs.tut.fi/ ˜ jupeto

ABSTRACT

A global optimization approach to active contours is necessary if images to be analyzed have low signal to noise ratio. In this setting, it is reasonable to study global properties of energy functions to be optimized. A simple connection between internal energy functions of active contour models of a certain type and Rayleigh quotients is derived in this paper. The importance of Rayleigh quotients lies in the fact that they are related to eigenvalues of real symmetric matrices. As a consequence, one can study the internal energy of an active contour model with numerical routines that are designed for eigenvalue computations of real symmetric matrices.

1. INTRODUCTION

Deformable models [1] are widely used techniques in image analysis and processing. Particularly active contours [2], also termed snakes, have received a lot of attention. The idea behind snakes is to regularize edge-detection by imposing soft constraints on the shape of the contour to be extracted. This way it is possible to find a contour from a noisy image without knowing its exact shape or position. Active contours are frequently applied in medical image analysis [3], but also other applications exist [1].

To be more precise a snake is a curve with an associated energy function. A contour extraction from an image is formulated as the minimization of the energy function. The energy is divided into the internal energy and the external energy. The external energy is derived from image data. The internal energy depends only on the shape of the curve hence regularizing the often ill-posed problem.

The internal energy for the original snake-model [2] was not invariant to scaling of the curve in order to reduce sensitivity to initialization imposed by the applied local minimization technique. For most of the applications, this solution is not satisfactory, see for example [4], [5]. A possible solution is to minimize the energy globally and set hard constraints to ensure admissibility of the resulting curve. Normally, this requires the internal energy to be invariant to translation, rotation and scaling of the curve.

For implementation, it is convenient to approximate the curve by a polygon, which is completely described by its vertices. This simple representation is yet a powerful one. It permits one to incorporate detailed prior information about the expected shape of the target to be delineated in the internal energy of the snake [6], [7]. However, further analysis of the internal energy function is often omitted. The analysis of its global behaviour may prove to be important, especially as increased computation power and improved algorithms allow more efficient energy minimization. The intention here is to show that the internal energy of the snake can be interpreted as a Rayleigh quotient [8]. Rayleigh quotients relate to the eigenvalue problem for symmetric matrices for which there are a number of algorithms and software. Hence, the simple connection provides a fast way to obtain information about the specific snake model. Assumptions required are not prohibitive and many active contour models with little or no modification will satisfy them.

2. SNAKES AND RAYLEIGH QUOTIENTS

A snake is an ordered set of points $\mathbf{V} = \{\mathbf{v}_0, \ldots, \mathbf{v}_{N-1}\}$, where each *snaxel* $\mathbf{v}_i = [x_i, y_i]^T \in \mathbb{R}^2$. Only closed contours are considered and hence subscript arithmetic is modulo N. The energy of the snake is

$$E(\mathbf{V}) = \lambda E_{int}(\mathbf{V}) + (1 - \lambda) E_{ext}(\mathbf{V}), \qquad (1)$$

where E_{int} is the internal energy, E_{ext} is the external energy and $\lambda \in [0, 1]$ is the regularization parameter. The

internal energy is

$$E_{int}(\mathbf{V}) \quad (2)$$
$$= \frac{\sum_{i=0}^{N-1} E_{int}(\mathbf{v}_i|\mathbf{v}_0,\ldots,\mathbf{v}_{i-1},\mathbf{v}_{i+1},\ldots,\mathbf{v}_{N-1})}{l(\mathbf{V})}$$
$$= \frac{\sum_{i=0}^{N-1} ||\sum_{j=0}^{N-1} A_{i_j}\mathbf{v}_j - \mathbf{v}_i||^2}{\sum_{i=0}^{N-1}||\mathbf{v}_{i+1}-\mathbf{v}_i||^2},$$

where all A_{i_j} are 2×2 matrices (with the convention that $A_{i_i} = 0$). The purpose of the *normalization factor* $l(\mathbf{V})$ is to yield a scale invariant E_{int}. At least active contour models presented in [6] and [7] have internal energies, which can be written in a form (2).

If we now assume that the internal energy (2) is translation invariant, we can interpret it as a Rayleigh quotient. For this, let $\mathbf{s} = [\mathbf{v}_0^T,\ldots,\mathbf{v}_{N-1}^T]^T = [x_0,y_0,x_1,y_1,\ldots,x_{N-1},y_{N-1}]^T$. Now (2) can be written as:

$$E_{int}(\mathbf{V}) = \frac{||B\mathbf{s}||^2}{l(\mathbf{V})}, \quad (3)$$

where

$$B = \begin{bmatrix} -I & A_{0_1} & \cdots & & A_{0_{N-1}} \\ A_{1_0} & -I & A_{1_2} & \cdots & A_{1_{N-1}} \\ \vdots & & \ddots & & \vdots \\ A_{N-1_0} & \cdots & & & -I \end{bmatrix}$$

and I is the 2×2 identity matrix. However, this is still not what we are after; $l(\mathbf{V})$ can be zero even if $\mathbf{s}$ is not. Therefore, recalling that the internal energy is translation invariant, we assume $\mathbf{v}_0 = \mathbf{0}$ without any loss of generality. Let $\mathbf{z} = [x_1,y_1,\ldots,x_{N-1},y_{N-1}]^T$. Now

$$\begin{aligned} ||B\mathbf{s}|| &= ||B[0,0,\mathbf{z}^T]^T|| \\ &= ||[\mathbf{b}_1,\ldots,\mathbf{b}_{2N}][0,0,\mathbf{z}^T]^T|| \\ &= ||[\mathbf{b}_3,\ldots,\mathbf{b}_{2N}]\mathbf{z}|| = ||\hat{B}\mathbf{z}||. \end{aligned}$$

The normalization factor $l(\mathbf{V})$ is a quadratic form:

$$l(\mathbf{V}) = \sum_{i=2}^{N-2} ||\mathbf{v}_{i+1}-\mathbf{v}_i||^2 + ||\mathbf{v}_2||^2 + ||\mathbf{v}_{N-1}||^2 = \mathbf{z}^T L\mathbf{z},$$

where L is $2N-2\times 2N-2$ non-singular matrix. The matrix L is also positive definite and hence there is a positive definite matrix $\sqrt{L}$ such that $\sqrt{L}^2 = L$ [8, Thm. 2.14.2]. The introduction of a new variable $\mathbf{w} = \sqrt{L}\mathbf{z}$ gives the interpretation of (2) as a Rayleigh quotient

$$E_{int}(\mathbf{V}) = \frac{\mathbf{w}^T(\sqrt{L}^{-1})^T\hat{B}^T\hat{B}\sqrt{L}^{-1}\mathbf{w}}{\mathbf{w}^T\mathbf{w}}. \quad (4)$$

3. PROPERTIES OF RAYLEIGH QUOTIENTS

The next theorem connects the Rayleigh quotient (4) and the internal energy (2) to the eigenvalues of the real symmetric matrix $(\sqrt{L}^{-1})^T\hat{B}^T\hat{B}\sqrt{L}^{-1}$, see [8] Theorems 3.2.1 and 3.3.1 and Exercise 1 at page 111.

Theorem 1 *Let R be real and symmetric square-matrix. The Rayleigh quotient $R(\mathbf{x}) = \frac{\mathbf{x}^T R\mathbf{x}}{\mathbf{x}^T\mathbf{x}}$ is stationary at, and only at, the eigenvectors of the matrix R. At an eigenvector ξ, $R(\xi) = \mu$, where μ is the associated eigenvalue. Moreover $\mu_1 = \max R(\mathbf{x}), \mu_0 = \min R(\mathbf{x})$, where μ_1 is the greatest eigenvalue and μ_0 is the least eigenvalue of the matrix R.*

Define $\mathbf{V}+\mathbf{W} = \{\mathbf{v}_i + \mathbf{w}_i | i = 0,\ldots,N-1\}$, where $\mathbf{V},\mathbf{W}$ are snakes with N snaxels. Note that when snakes are taken as vectors of $\mathbb{R}^{2N}$ their addition is simply vector addition. The scalar multiplication in $\mathbb{R}^{2N}$ corresponds to the scaling of snakes. If we now set $E_{int}(\mathbf{V}) = \mu_0$ if $l(\mathbf{V}) = 0$, where μ_0 is the least eigenvalue of the related matrix $(\sqrt{L}^{-1})^T\hat{B}^T\hat{B}\sqrt{L}^{-1}$ we obtain a corollary to the Theorem 1.

Corollary 1 *Let μ_0 be the least eigenvalue of the matrix $(\sqrt{L}^{-1})^T\hat{B}^T\hat{B}\sqrt{L}^{-1}$ related to $E_{int}(\mathbf{V})$. Let the multiplicity of μ_0 be K. Then the set of snakes of minimum internal energy $\mathcal{V} = \{\mathbf{V} : E_{int}(\mathbf{V}) = \mu_0\}$ is a vector space. Moreover, if $\mathbf{V}_j, j = 1,\ldots,K$, are K snakes corresponding to K linearly independent eigenvectors associated with μ_0, a basis for $\mathcal{V}$ is*

$$\{\mathbf{V}_j | j = 1\ldots K\} \cup \{\mathbf{X},\mathbf{Y}\},$$

where $\mathbf{X} = \{[1,0]^T,\ldots,[1,0]^T\}$ and $\mathbf{Y} = \{[0,1]^T,\ldots,[0,1]^T\}$.

The proof of the Corollary is given in the Appendix. Of course, while performing actual computations, one normally does not want to find a contour whose length is zero. However, the above Corollary is still a useful one. For example, it is applied in the Section 4.

4. EXPERIMENTS

As an example two particular internal energy functions are analyzed by computing eigenvalues and eigenvectors of the related matrices. The internal energy functions are

$$E_{int}^1(\mathbf{V}) = \frac{\sum_{i=0}^{N-1}||\mathbf{v}_i - \frac{1}{2}(\mathbf{v}_{i-1}+\mathbf{v}_i)||^2}{l(\mathbf{V})},$$

$$E_{int}^2(\mathbf{V}) = \frac{\sum_{i=0}^{N-1}||\mathbf{v}_i - \hat{\mathbf{v}}_i||^2}{l(\mathbf{V})},$$

where

$$\hat{\mathbf{v}}_i = \frac{1}{2}(\mathbf{v}_{i-1}+\mathbf{v}_i + \tan\frac{\pi}{N}R_{90}(\mathbf{v}_{i-1}-\mathbf{v}_{i+1}))$$

and R_{90} is 90 degrees rotation matrix. The function E_{int}^1 is the discretized version of the curvature term of the internal energy of the original snake model [2]. It has been normalized by $l(\mathbf{V})$ for scale invariance. The function E_{int}^2 is from [6]. Symbols $M_i, i = 1,2$, are used when referring to the matrix $(\sqrt{L}^{-1})^T\hat{B}^T\hat{B}\sqrt{L}^{-1}$ corresponding the function E_{int}^i.

Table 1: Minima and maxima of the two energy functions when the number of snaxels is varied. Minima of E^2_{int} are always zero.

N	min E^1_{int}	max E^1_{int}	max E^2_{int}
20	0.0245	1	1.0251
21	0.0222	0.9944	1.0170
30	0.0109	1	1.0110
31	0.0102	0.9974	1.0077
50	0.0039	1	1.0040
51	0.0038	0.9991	1.0029
100	0.0010	1	1.0010
101	0.0010	0.9998	1.0007

Numerical computations were performed by Matlab 5.3 (Mathworks, Natick, MA, U.S.). It uses the EISPACK routines [9] for eigenvalue calculations. Square roots of matrices L were also computed by Matlab. For this, it applies the Parlett-algorithm described in [10, p.384]. The properties of the two internal energy functions that will be presented are based on numerical simulations. Some of these ought to be taken with caution. For example, it is possible to make an error when stating results concerning multiplicities of eigenvalues. We may not notice that two eigenvalues are not equal if they are very close to each other.

Minima and maxima of the both energy functions for several values of N are listed in Table 1. As can be seen from Table 1, their ranges tended to $[0, 1]$ as N increased. The least eigenvalue of M_1 had multiplicity 4. Snakes $\mathbf{V}^j, j = 1, \ldots 4$, corresponding some four linearly independent eigenvectors were related by a linear transformation, i.e. $\mathbf{V}^i = T\mathbf{V}^j = \{T\mathbf{v}^j_k | k = 0, \ldots, N-1\}$, where $T : \mathbb{R}^2 \to \mathbb{R}^2$ is a linear transformation. Now, noting that the curves $\mathbf{V}^j, j = 1, 2, 3, 4$, all had a shape of an ellipse, by Corollary 1 it follows that all minimum energy curves of E^1_{int} are ellipses. Curves corresponding to all the other eigenvalues of M_1 were self-intersecting and hence classified as inadmissible solutions to the problem. Also from the shape of these curves it was clear that all linear combinations of them were also self-intersecting.
For the function E^2_{int} the curve of minimal energy is, by the construction, circle. Our simulation verified the fact. Moreover, since the multiplicity of the least eigenvalue of M_2 was 2, we can conclude that circle is the only minimum energy curve of E^2_{int}. However, E^2_{int} had also other admissible curves as stationary points. Some of these are shown in Fig. 1.

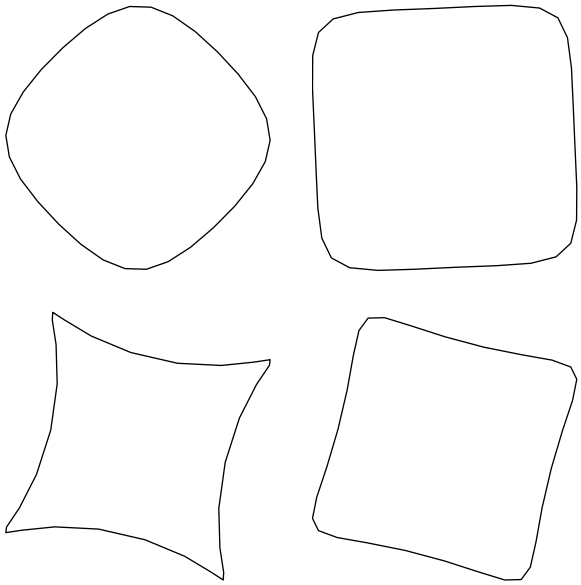


Figure 1: Few curves for which E^2_{int} is stationary when $N = 30$.

5. DISCUSSION

We have shown how to interpret a scale and translation invariant internal energy of a snake as a Rayleigh quotient. The approach is quite general. For example, the snake models from [6] and [7] can be seen to satisfy our assumptions. The only real restriction of our approach is the choice of normalization factor. Also normalization factors that do not permit the Rayleigh quotient interpretation can of course be used. However, further studies and discussions about the meaning of the form of the normalization factor are beyond the scope of this paper.

Rayleigh quotients relate to eigenvalues of the real symmetric matrices. Because the symmetric eigenvalue problem is well-studied, the connection allows one to analyze global properties of the internal energy functions of the snake models. Here minima, maxima and stationary points of two internal energy functions were found by using the derived connection. Another function had also admissible, i.e. non-intersecting, curves as stationary points in addition to the ones of minimal energy. This is an interesting result, because it clearly demonstrates a disadvantage of gradient descent techniques for the optimization in the framework of active contours.

APPENDIX

The proof of Corollary 1 is presented. Snakes $\mathbf{V}_j, j = 1, \ldots, K$, belong to $\mathcal{V}$ by Theorem 1. By assumption that if $l(\mathbf{V}) = 0$ then $E_{int}(\mathbf{V}) = \mu_0$, also $\mathbf{X}, \mathbf{Y} \in \mathcal{V}$. Since the (algebraic) multiplicity and the geometric multiplicity of an eigenvalue of a real symmetric matrix are equal [8], $l(\mathbf{V}_j) \neq 0$ and the first snaxel of $\mathbf{V}_j$ is zero for each j, the set $\mathcal{B} = \{\mathbf{V}_j | j = 1 \ldots K\} \cup \{\mathbf{X}, \mathbf{Y}\}$ is linearly independent.

Now let $\mathbf{W} \in \mathcal{V}$ be arbitrary. Then also $\mathbf{V} = \mathbf{W} - x_0\mathbf{X} - y_0\mathbf{Y} \in \mathcal{V}$, where x_0 (resp. y_0) is the x-coordinate (y-coordinate) of the first snaxel of $\mathbf{W}$. Furthermore $\mathbf{v}_0 = \mathbf{0}$. Since Rayleigh quotients are differentiable where defined

and μ_0 is the minimum of the Rayleigh quotient corresponding to E_{int}, from Theorem 1 it follows that there is an eigenvector associated with μ_0 that corresponds to $\mathbf{V}$. Hence, $\mathbf{W}$ belongs to a space spanned by $\mathcal{B}$. Assume now that $\mathbf{W}$ is an arbitrary element of the space spanned by $\mathcal{B}$. Then $\mathbf{W}$ is obtained by a translation from some linear combination of $\mathbf{V}_j, j = 1, \ldots, K$. It follows that $E_{int}(\mathbf{W}) = \mu_0$ and the proof is completed.

ACKNOWLEDGMENT

The author is financially supported by *Tampere Graduate School of Information Science and Engineering* (TISE).

REFERENCES

[1] A.K. Jain, Y. Zhong, and M-P. Dubusson-Jolly, "Deformable template models: A review," *Signal Processing*, vol. 71, no. 2, 1998.

[2] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321 – 331, 1988.

[3] T. McInerney and D. Terzopoulos, "Deformable models in medical image analysis: A survey," *Medical Image Analysis*, vol. 2, no. 1, 1996.

[4] S.R. Gunn, *Dual Active Contour Models for Image Feature Extraction*, Ph.D. thesis, University of Southampton, 1996.

[5] C. Xu and J.L. Prince, "Snakes, shapes and gradient vector flow," *IEEE Transactions on Image Processing*, vol. 7, no. 3, 1998.

[6] S.R. Gunn and M.S. Nixon, "A robust snake implementation: a dual active contour," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 1, 1997.

[7] K.F. Lai and R.T. Chin, "Deformable contours - modelling and extraction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 11, 1995.

[8] P. Lancaster, *Theory of Matrices*, Academic Press, 1969.

[9] B. T. Smith, J. M. Boyle, J. J. Dongarra, B. S. Garbow, Y. Ikebe, V. C. Klema, and C. B. Moler, *Matrix Eigensystem Routines - EISPACK-Guide*, vol. 6 of *Lecture Notes in Computer Science*, Springer Verlag, 1976.

[10] G.H. Golub and C.F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, 1st edition, 1983.