分析 IIR 滤波器值域与精度的高效算法

庞 宇[1],贺志龙[1],王绍全[1],王骏超[1],高 翔[1],吴 玮[2]
(1.重庆邮电大学光电工程学院,重庆 400065;2.四川大学电子信息工程学院,四川成都 610065)

摘 要: 值域与精度分析是高级综合的重要步骤.虽然过去已提出了不少方法试图解决这两个问题,但针对无限冲击响应滤波器(Infinite Impulse Response,IIR)来说,这些方法要么过高估计数值要么无法处理任意阶的反馈电路.对于给定输入值范围与误差界限的 IIR 滤波器,我们提出了一个高效的启发式算法来解决值域与精度分析.该算法能用于优化整数和分数的比特宽度分配,获得优化的电路面积.实验结果证明了所提出的算法具有快速收敛性与鲁棒性,由于高阶 IIR 滤波器能分解为低阶结构的滤波器,因此该算法能高效的处理任意阶 IIR 滤波器.

关键词: 无限冲击响应滤波器; 值域; 精度; 最大误差; 比特宽度
中图分类号: TN492, TN495 **文献标识码:** A **文章编号:** 0372-2112 (2012)09-1752-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2012.09.009

Efficient Methods of Range and Precision Analysis for IIR Filters

PANG Yu[1], HE Zhi-long[1], WANG Shao-quan[1], WANG Jun-chao[1], GAO Xiang[1], WU Wei[2]
(1. *College of Electronic Engineering, Chongqing University of Posts and Telecommunications, Chongqing* 400065, *China*;
2. *College of Electronics and Information Engineering, Sichuan University, Chengdu, Sichuan* 610065, *China*)

Abstract: Analysis of range and precision is always an important task for high level synthesis. Although several researches have been dedicated to these two problems, in the case of an infinite impulse response (IIR) filter, conventional approaches either constitute major overestimations or cannot handle arbitrary order feedback circuits. In this paper we focus on these problems and propose one efficient heuristic method for range and precision analysis of such circuits. When the input and error bounds are given, the minimum integer and fractional bit-widths can be allocated to satisfy the error bound, which can obtain smaller circuit area. The experimental results prove that the proposed method has fast convergence and robustness. Because a high order IIR can be decomposed to low order IIRs, the method can efficiently handle IIRs with arbitrary orders.

Key words: IIR; range; precision; maximum error; bit-width

1 引言

作为一种常用的表达实数的数据格式,定点数表示在低功耗芯片中扮演了至关重要的作用.定点数算术在低功耗设计、嵌入式系统和可编程逻辑器件中具有重要地位.定点数表示由两组数据流组成,一组表达整数,另一组表达分数,由虚拟的小数点将他们分开.定点数表示如下:

字长 = IB(Integer Bit-width) + FB(Fractional Bit-width) (1)

IB 与 FB 分别代表整数比特宽度与分数比特宽度.字长作为 IB 与 FB 的和,通常代表变量宽度.在 FPGA 和 ASIC 设计中,因为可随意指定定点数字长,因此对定点数表示的研究包含了整数与分数两方面的内容,分别对应了值域与精度分析的任务.定点数电路的值域与精度分析是高级优化与验证的重要任务.在优化流程上,值域与精度的计算使高效确定变量的整数比特宽度与分数比特宽度成为可能.而且,在高级抽象验证中,值域分析能够证明中间变量或输出变量根据其整数比特长度是否会发生溢出,同时,精度分析利用分数比特长度来验证一个实现是否能满足给定的误差界限.工程师提出了多种方法针对非反馈电路来完成值域与精度分析.基于仿真的动态分析法[1~4]是一种数值处理的常用方法.虽然动态分析法理论上能找到精确的值域,但效率较低,限制了此方法的广泛应用.作为替代方法,静态分析能平衡结果精确性与运行时间.区间算术 IA (Interval Arithmetic) 是计算数值界限的常用静态算法,但不可避免的导致很不精确的结果.仿射算术 AA (Affine Arithmetic)是 IA 的一种衍生算法.在 AA 中,数值区间被表示

收稿日期:2011-06-26;修回日期:2012-06-11
基金项目:国家自然科学基金(No.61102075);重庆市自然科学基金(No.CSTC2011BB2142);重庆市教委科学技术研究项目(No.KJ120507)

成一些本源变量的线性组合,这些变量代表了数据中的不确定源或在计算中产生的近似.文章[5]使用了这种静态分析法来计算值域与精度.文献[6~8]的作者采用了一种谱技术—算术变换,来研究在泰勒级数这种非精确表示中的精度问题.另外文献[9]介绍了更一般的方法—SAT-Modulo(SMT)理论,用以计算值域.SMT 首先利用 IA 计算得到的不精确结果,然后通过插入约束条件进行优化,通过该算法获得的结果比使用 AA 算法获得的结果更精确.而且,相对于传统算法,SMT 能处理特殊的算术操作比如除法.基于此种理论框架,SMT 引擎能通过检查可满足性来证明或反驳给定表达式的边界有效性.

以上这些方法的主要缺陷在于它们只能处理正向数据通路而不能处理带有反馈的无限循环电路.在定点数表示的 DSP 电路中经常存在反馈,因此这类电路的值域与精度分析仍然是一个相当大的难题.对于无限冲击响应滤波器(IIR)这类典型的具有反馈的电路,研究者们提出了一些方法来解决值域与精度分析的问题.Harju[10]研究了量化的特殊预测 IIR 多项式行为.Milic[11]提出了使用并联的两个全通网络、具有较少乘法器的能保证精度的 IIR 滤波器电路结构.Carletta[12]提出了一个分析框架用于确定系数的比特长度.此技术能估计输出值界限并分析截断误差.根据计算结果,在定点数硬件中整数和分数比特长度被确定以避免溢出及保证精度.Diniz[13]使用并联的 2 阶 IIR 滤波器构成了自适应频域滤波器,具有快速收敛性、鲁棒性和较小均方差的优点.这些方法的主要缺陷不仅在于会导致过高估计值域与精度,还在于其使用的追踪系数量化影响的敏化分析只适用于 2 阶 IIR 滤波器.因此,高阶 IIR 滤波器只能分解成很多 2 阶滤波器使得精度运算能被独立应用到这些分解的滤波器上,根据三角不等式,这会造成数值过高估计并导致优化流程低效.

为了解决上述问题,本篇论文针对任意阶 IIR 滤波器,以最小的过度估计代价计算定点反馈算术电路的值域与精度并提出一个启发式算法.在给定输入与误差界限下,该算法能高效的分配整数与分数比特宽度.

2　背景理论与基本定义

IIR 滤波器具有反馈电路,如下定义:

$$z[n]=\sum_{i=0}^{P} b_i x[n-i]+\sum_{j=1}^{Q} a_j z[n-j] \qquad (2)$$

P 是前向通路的阶数,b_i 代表前向通路的系数;Q 是后向通路的阶数,a_i 代表后向通路的系数.

定理 1　式(2)所表示的 IIR 滤波其能被展开成:

$$z[n]=c_n x[n]+c_{n-1}x[n-1]+\cdots+c_1 x[1] \qquad (3)$$

这里,$c_i(i=1,2,\cdots,n)$是常数值.计算 c_i 的复杂度为 $O(n)$.

证明　使用式(2)代替反馈变量并展开该表达式可以得到式(3).因为迭代表达式中 $a_j, b_i, x[n], z[n]$ 的次数为 1,所以计算 c_i 的复杂度是线性的.

例 1　考虑一个 IIR 滤波器 $z[n]=3x[n]-0.4z[n-1]+0.2z[n-2]$.第一个展开表达式为 $z[1]=3x[1]$,第二个展开表达式为 $z[2]=3x[2]-1.2x[1]$,第三个展开表达式为 $z[3]=3x[3]-1.2x[2]+1.08x[1]$.所以对于 $n=3$,(c_3,c_2,c_1)分别等于$(3,-1.2,1.08)$.式(2)含有的反馈可能导致系统不稳定,意味着随着迭代次数增加,输出界限增加(或减少)到正无穷(负无穷).以下的定义涉及到该问题.

定义 1　有界输入-有界输出(Bounded-Input Bounded-Output,BIBO)稳定性.如果对于任意有界输入,输出总是有界,那该系统称为有界输入-有界输出稳定[14].

我们用符号 B_i 表示有界输入,B_o 表示有界输出.

3　值域分析与精度分析

定义 2　符号 $B_o[n]=(B_{olow}[n],B_{ouppp}[n])$表示经过 n 次迭代的输出界限,其中 $B_{olow}[n]$表示 $B_o[n]$的下边界,$B_{ouppp}[n]$表示 $B_o[n]$的上边界.其定义如下:

$B_{ouppp}[n]=\max\{\max(z[n]),\max(z[n-1]),\cdots,\max(z[1])\}$,$B_{olow}[n]=\min\{\min(z[n]),\min(z[n-1]),\cdots,\min(z[1])\}$

这里,$\max(z[n])/\min(z[n])$代表公式(3)给出的第 n 次迭代的输出最大值与最小值.

定理 2　$B_{ouppp}[n]$和 $B_{olow}[n]$是关于迭代次数"n"的单调增和单调减函数,当 n 趋近于无穷时,得到的值就是精确的上限与下限.

证明　根据定义 2,当 n 增大到无穷时,$B_o[n]$趋近精确的边界值.

定义 3　两个不同迭代 n 和 n-W(W 是常数)的边界差值定义为:

$$\Delta[n,W]=|B_{ouppp/low}[n]-B_{ouppp/low}[n-W]|$$

推论 1　边界差值 $\Delta[n,W]$是关于迭代次数"n"的单调减函数,且$\{\Delta[n,W]\to 0 \mid n\to\infty\}$.

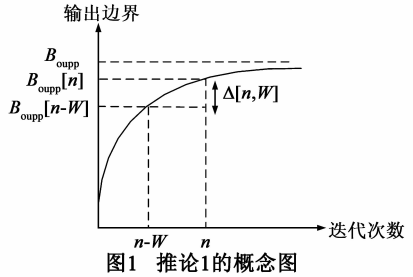

图1　推论1的概念图

图 1 描述了 $B_{ouppp}[n]$的概念.当 n 增长时,$B_{ouppp}[n]$是单调增函数,$\Delta[n,W]$是单调减函数.参数 W 在计算

B_{oupp}扮演重要角色,尤其是在高阶滤波器中. W 值应该设置得远大于滤波器阶数.而且,W 表示迭代观察窗口(从 n-W 到 n),因此设置更大的 W 值能得到更准确的结果.基于以上讨论,只要选择合适的参数 ε 和 W(ε 是收敛分辨率,W 是收敛窗口尺寸),就能找到一个迭代值 n_0 使 $\Delta[n_0, W] < \varepsilon$ 且 $B_{oupp}[n_0] \approx B_{oupp}$.在本文提出的算法中,设置 $W = 100$,这是因为几乎所有的 IIR 滤波器阶数都小于 20.在值域分析中是为了找到整数边界,所以 ε 设置为"1".因此这两个参数分别设置为 $\varepsilon = 1$ 和 $W = 100$.同时将这两个参数作为收敛条件,能导致快速收敛与鲁棒估计.精度计算能分析特定的实现是否能满足给定的误差边界,而且也能对所有的变量找到适合的分数比特宽度.下式定义了最大误差 e_{max}:

$$e_{max} = \max|z_{fixed} - z| \tag{4}$$

z 和 z_{fixed}分别代表 IIR 滤波器的浮点数与定点数实现.该算法同时能找到合适的分数比特宽度以满足给定的误差边界 E,即 $e_{max} < E$.

本文中,为了简化起见,我们考虑所有定点数表达形式的输入/系数/输出变量具有相同的分数比特宽度 FB.基于以上讨论及考虑到最坏量化错误,可将式(2)中定点实现 $z[n]$表示如下:

$$z_{fixed}[n] = \sum_{i=0}^{P}[(b_i - e_{bi}) \times (x^{-i} - e_{in}) - e_{tr}] + \sum_{j=1}^{Q}[(a_j - e_{aj}) \times z_{fixed}^{-j} - e_{tr}] \tag{5}$$

其中,e_{bi}和 e_{aj}是常系数 b_i 和 a_j 的量化错误,且基于舍入效应可知$|e_{aj}| \leqslant 2^{-FB-1}$和$|e_{bj}| \leqslant 2^{-FB-1}$. e_{in}是输入变量 x 的截断误差且 $e_{in} = x - x_{fixed}$.此参数总为正值且 $e_{in} = 2^{-FB}$.而且,每次乘法后,会产生一个正值截断误差 e_{tr},因为执行乘法运算时,结果的分数比特宽度会增加到 2FB,而所有寄存器的分数比特宽度限制到 FB.在最坏情况下,$e_{tr} = 2^{-FB}$.带有错误源常数值的公式(5)表示了一个具有过高估计不精确度的定点电路.此表示形式仍然是输入变量的线性函数,根据定理 1 可以展开为:

$$z_{fixed}[n] = d_n x[n] + d_{n-1}x[n-1] + \cdots + d_1 x[1] + d_0 \tag{6}$$

使用式(3)和式(6)可得到线性不精确函数 z_e:

$$z_e[n] = z_{fixed}[n] - z[n] = e_n x[n] + \cdots + e_1 x[1] + d_0 \tag{7}$$

这里,$e_i = d_i - c_i (i = 1, \cdots, n)$.式(5)中寻找 e_{max}的不精确度运算的主要目标可以重声明为在式(7)中寻找 $z_e[n]$的边界值.

4 整数比特宽度与分数比特宽度分配的启发式算法

图 2 解释了计算输出值域与分配分数比特宽度的算法,其输入是前向通路系数 b、反馈系数 a、滤波器输入值边界 $B_i = [x_{low}, x_{upp}]$与误差界限 E.在步骤 2 循环开始,计算每次迭代产生的边界值.子程序 Expand 根据定理 1 展开 $z[n]$并计算新生成的系数 $\boldsymbol{C}_n = [c_n, c_{n-1}, \cdots, c_1]$.步骤 4 计算在第 n 次迭代的最大和最小输出值(标记为 Z_{max}和 Z_{min}).根据式(3),$z[n]$是输入变量的线性组合,所以能计算出精确结果.步骤 5 比较计算出这次迭代的边界值和上次迭代出的边界值,如果当前值大,就保留当前值,否则就用上次的值替代.如果边界差值$|\Delta[n, W]|$小于"$\varepsilon = 1$",循环终止并返回值域.完成值域计算后,收敛参数 ε 重设置为远小于 E 的值,如 $\varepsilon = E/1000$.这是因为 ε 直接影响到最终结果并且在精度计算中,其边界值远小于范围中的边界值.从步骤 8 中设置的初始值开始,迭代搜索不断循环以寻找合适的分数比特宽度.算法在步骤 12 和 13 分别利用两个子例程 Expand 和 Expand_fixed 展开 $z[n]$和 $z_{fixed}[n]$.注意展开 $z[n]$是基于定理 1 并返回式(3)中系数向量 $\boldsymbol{C}_n = [c_n, c_{n-1}, \cdots, c_1]$.另一方面,展开式(6)中的 $z_{fixed}[n]$需要 FB 作为输入用以计算量化与截断误差.此子例程返回式(6)中的系数向量 $\boldsymbol{C}_{n\_fixed} = [d_n, d_{n-1}, \cdots, d_1, d_0]$.步骤 14 采用式(7)计算系数向量 $\boldsymbol{C}_e = \boldsymbol{C}_n - \boldsymbol{C}_{n_fixed} = [e_n, \cdots, e_1, d_0]$来表达 $z_e[n]$的展开形式.步骤 15 计算第 n 次迭代时 $z_e[n]$的最大值,标记为 Z_{e_max}.如果 Z_{e_max}的值大于 E(步骤 16),意味着 FB 的值不够大所以 while 循环终止.另一方面,步骤 17 刷新错误上边界,步骤 18 检查收敛条件.如果该条件不满足,算法增大 n 值并继续循环,否则达到收敛并返回最大不精确度 $e_{max} = B_{e_upp}[n]$与最终 FB 的值.

图 2 所示的算法适用于直接型 IIR 滤波器.然而,它很容易被修改以用于其它类型的 IIR 滤波器.比如对于由几个直接型滤波器节构成的并联型 IIR 滤波器,我们只需对于所有的并联型滤波器节运行 Expand 和 Expand_fixed 这两个子例程,则 z_e 等于 $z_e[n] = \sum_{i=1}^{N}(z_{fixed(i)}[n] - z_{(i)}[n])$.此处 $Z_{fixed(i)}$是第 i 节直接型滤波器的定点实现,$z_{(i)}$是相应的非量化模型.

得到值域后,根据如下公式找到合适的整数比特宽度:

$$IB = ceiling(\log_2(\max(|r_{low}|, |r_{upp}|))) + a \tag{8}$$

这里 $a = \begin{cases} 2, & \mathrm{mod}(\log_2(\max(|r_{low}|, |r_{upp}|)), 1) = 0 \\ 1, & \text{其他} \end{cases}$

r_{low}和 r_{upp}代表了计算出的值域 R_o 的下界与上界.

例 2 一个稳定的滤波器表达式为 $z[n] = 2x[n] + 0.1x[n-1] - 0.4x[n-2] - 0.1z[n-1] + 0.46z[n-2] - 0.08z[n-3]$,输入值边界为$[-100, 100]$.经过 21 次迭代和 121 次迭代,输出值边界分别为$[-208.07,$

208.07]和[- 208.53,208.53].因为边界差值小于 1,循环终止,返回值域[- 209,209],同时根据式(8)确定整数比特宽度为 9.

```
IIR_Range_Prec(b,a,B_i,E)
{ /* 输入：前向通路系数b=[b0, b1, b2, …, bp]，后向通路系数a=[a0, a1, a2, …, aq]，变量值域Bi=[x_low, x_upp] ，误差界限E;
     输出：值域Ro, 分数比特宽度FB, 最大不精确度e_max； */
1.  W=100; ε =1; n=1; // 收敛参数
2.  while (|B_oupp/low[n]-B_oupp/low[n-W] ≥ ε)
3.  { C_n= Expand (b, a, c_n-1, … , c_n-c1);   // C_n=[c_n, c_n-1,… , c1]
4.    z_max = Σ_{i=1}^{n} max(c_i x_low , c_i x_upp);    z_min = Σ_{i=1}^{n} min(c_i x_low , c_i x_upp)   //Z_max/min=第n次迭代的最大/最小值
5.    B_olow[n]= min{B_olow[n-1], z_min};   B_oupp[n]= max{B_oupp[n-1], z_max};   //B_oupp/low[n] =经过n次迭代的上界/下界值
6.    n++; }
7.  R_O = ceiling(B_O[n]); }
8.  ε=E/1000, FB= 4, converge = 0;   // 选择ε<<E, 搜索最佳FB的初始值
9.  while (converge = 0)
10.  { FB++; n=1;
11.  while
12.  { c_n= Expand (b, a, c_n-1, … , c_n-Q);    // c_n= [c_n,c_n-1, … , c1] w.r.t Eqn. (3)
13.    c_n_fixed= Expand_fixed(b, a, C_n-1_fixed,, … , C_n-Q_fixed, FB);   //  c_n_fixed =[d_n,… , d1, d0], w.r.t Eqn. (6)
14.    c_e= c_n_fixed -c_n;    // c_e= [e_n,… , e1, e0] w.r.t (7)
15.    z_e_max = |Σ_{i=1}^{n} (max{e_i * x_max , e_i * x_upp}) + e_0|   //  Z_e_max= 第 n 次迭代的最大错误.
16.    if (Z_e_max > E)   break;    //FB数值不满足要求
17.    B_e_upp[n] = max{Z_e_max, B_e_upp[n-1]}    // B_e_upp[n]=n次迭代后的误差上边界
18.    if (|B_e_upp[n] - B_e_upp[n-W]| < ε)   { converge = 1; break; }
19.    n++; } }
20.  e_max= B_e_upp[n];
21.  return (R_O, e_max, FB);}
```

图2　计算优化的整数与分数比特宽度的高效启发式算法

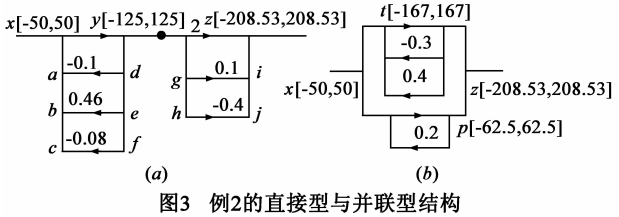

图3　例2的直接型与并联型结构

我们将实验运行至 1000 次迭代,所得到的边界值并无改变,证明结果的准确性以及该启发式算法的鲁棒性.使用 AA 算法得到的值域为[- 347,347],可见我们的结果比通过 AA 算法获得的结果精确得多.需要指出的是,值域计算并不依赖于具体的数据通道只依赖于滤波器表达式.在 IIR 滤波器不同的实现中,中间变量的整数比特宽度能很容易的被输入边界和输出边界确定.图 3(*a*)和(*b*)描述了直接型结构与并联型结构.

考虑图 3(*b*)所示的并联结构.给定的误差边界 $E=0.1$,参数 ε 设置为 $E/1000=0.0001$.当 FB = 11 时,Z_{e_max}小于 E.此时算法发现经过 120 次迭代后最大不精确度 $B_{e_upp}[120]$为 0.0347,而 $B_{e_upp}[220]$是 0.03474.$B_{e_upp}[120]$与 $B_{e_upp}[220]$的差异 0.03474 - 0.0347 = 0.00003 < ε,因此算法停止循环,最大不精确度 $e_{max}=B_{e\_upp}[220]$,且最优 FB = 11.

由于高阶 IIR 滤波器能分解为以 2 阶为基础的低阶结构滤波器,因此该算法能高效的处理任意阶 IIR 滤波器.

5　实验结果

本节我们将根据几个测试向量评估提出的值域与精度分析算法的鲁棒性与高效性,采用 matlab 编写并运行在 Intel 2.8 GHz Pentium 4 与 2GB 内存的机器上,使用 XP 系统.

为了获得合适的测试向量,我们使用 matlab 中的 fdatool 工具,根据一些典型指标,比如采样频率和 3dB 带宽等,同时产生直接型(DR)和并联型(PRL)结构的任意阶 IIR 滤波器.测试向量如下:

向量 # 1(3 阶):

直接型:$z[n]=2x[n]+0.1x[n-1]-0.4x[n-2]-0.1z[n-1]+0.46z[n-2]-0.08z[n-3]$

并联型:$z[n]=z_1[n]+z_2[n]$,$z_1[n]=x[n]-0.2z_1[n$

$-1]$, $z_2[n]=x[n]-0.3z_2[n-1]+0.4z_2[n-2]$

向量#2(3阶):

直接型: $z[n]=0.44x[n-1]+0.362x[n-2]+0.02x[n-3]-0.4z[n-1]-0.18z[n-2]+0.2z[n-3]$

并联型: $z[n]=z_1[n]+z_2[n]$, $z_1[n]=0.24x[n-1]+0.4z_1[n-1]$, $z_2[n]=0.2x[n-1]+0.25x[n-2]-0.8z_2[n-1]-0.5z_2[n-2]$

向量#3(4阶):

直接型: $z[n]=0.3564x[n]-1.1147x[n-1]+1.7712x[n-2]-0.9531x[n-3]+0.2575x[n-4]+3.3553z[n-1]-4.3439z[n-2]+2.557z[n-3]-0.5771z[n-4]$

并联型: $z[n]=z_1[n]+z_2[n]$

$z_1[n]=0.11x[n]-0.1041x[n-1]+0.11x[n-2]+1.58z_1[n-1]-0.6469z_1[n-2]$

$z_2[n]=0.2426x[n]-0.426x[n-1]+0.2426x[n-2]+1.7753z_2[n-1]-0.892z_2[n-2]$

向量#4(5阶):

直接型: $z[n]=-0.0526x[n]+0.3179x[n-1]-0.3795x[n-2]-0.0994x[n-3]+0.2591x[n-4]+2.9609z[n-1]-3.5088z[n-2]+1.6939z[n-3]-0.0268z[n-4]-0.2111z[n-5]$

并联型: $z[n]=z_1[n]+z_2[n]+z_3[n]$

$z_1[n]=0.13x[n]-0.267z_1[n-1]$, $z_2[n]=0.31x[n]+0.365x[n-1]+1.4826z_2[n-1]-0.827z_2[n-2]$, $z_3[n]=-0.4926x[n]+0.286x[n-1]+1.7452z_3[n-1]-0.9561z_3[n-2]$

向量#5(6阶):

直接型: $z[n]=-1.5608x[n-1]+3.07x[n-2]+3.07x[n-3]-1.5608x[n-4]+x[n-5]+2.547z[n-1]-4.2203z[n-2]+4.3179z[n-3]-3.0547z[n-4]+1.3498z[n-5]-0.3168z[n-6]$

并联型: $z[n]=z_1[n]+z_2[n]+z_3[n]$

$z_1[n]=x[n]+1.488x[n-1]+x[n-2]-1.157z_1[n-1]-0.451z_1[n-2]$, $z_2[n]=-0.2375x[n-1]+x[n-2]+0.6z_2[n-1]-0.9388z_2[n-2]$, $z_3[n]=0.1629x[n]+x[n-2]+0.79z_3[n-1]-0.7483z_3[n-2]$

表 1 对几个 IIR 滤波器测试向量执行提出算法的实验结果

向量	类型	值域迭代		精度迭代		值域		e_{max}		IB/FB	值域运行时间(s)		精度运行时间(s)	
		情况 1	情况 2	情况 1	情况 2	情况 1	情况 2	情况 1	情况 2		情况 1	情况 2	情况 1	情况 2
1	DR	128	2560	135	2700	−417,417	−417,417	0.0214	0.0214	9/14	1.2	13	5.01	25.7
	PRL			140	2800			0.0695	0.0695	9/12			3.73	28.6
2	DR	116	2320	127	2540	−104,104	−104,104	0.0503	0.0503	7/11	1.06	12.8	3.59	25.6
	PRL			125	2500			0.0754	0.0754	7/10			2.91	31.3
3	DR	317	6340	261	5220	−11801, 11801	−11801, 11801	0.0255	0.0255	14/23	2.18	35.8	10.6	57.5
	PRL			189	3780			0.0272	0.0272	14/16			2.55	35.5
4	DR	458	9160	489	9780	−2965, 2965	−2965, 2965	0.0691	0.0691	12/23	3.01	47.8	20	119.6
	PRL			480	9600			0.0896	0.0896	12/17			18	115.4
5	DR	463	9260	311	6220	−37323, 37323	−37323, 37323	0.0093	0.0093	16/24	0.66	10.7	2.87	37.5
	PRL			380	7600			0.0777	0.0777	16/18			9.89	136.1

第一个实验探索算法的收敛性.表1展示了运算结果.对于所有测试向量,最大误差界限设置为 $E=0.1$,输入值域 B_i 为[−100,100].图2和4中提出的算法用于计算值域(然后分配整数比特宽度)和最大不精确度 e_{max}(然后分配分数比特宽度)使条件 $e_{max}<E$ 满足.如前所述,经过 n 次迭代后,最大值域和最大不精确度能被严格计算出来.因此只要增加迭代次数使 n 趋近无穷就能使结果更加精确.表1报告了两种情况.情况1显示算法的执行情况.假设在情况1收敛时需要 k 次迭代,为了强调情况1获得结果的鲁棒性,在情况2中,我们增加迭代次数到 $20k$ 来对比情况1.表1的"值域/精度迭代"栏报告情况1和情况2中的迭代次数.可以看出,对于所有的测试向量,与情况1比较,情况2收敛到相同的值域和 e_{max}.而且,为了提供更加清晰的结果,针对测试向量#4(需要最多的迭代次数才能收敛),我们同时观察迭代次数(从 $n=1$ 到 $n=1000$)与 $z[n]$的最大值(图2步骤4中的 Z_{max})的关系.图4显示了该结果.需要注意的是,$z[n]$的最大值关于"n"并不是严格单调递增,虽然总的来说呈现出递增行为.实际上在一些初始化迭代中,$z[k]$的最大值小于 $z[k-1]$的最大值.然

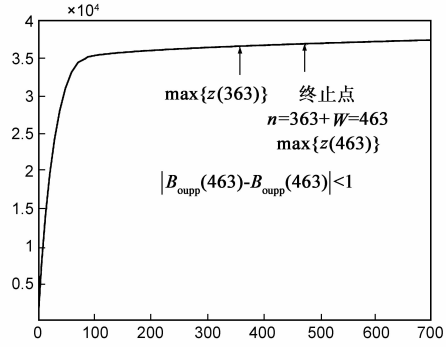

图4 测试向量#5的z[n]的最大值与n的关系

而,定理 2 中提到的最大边界 $B_{upp}[n]$关于"n"总是单调递增.图 5 中 $z[n]$的第一个最大整数值出现在 $n=363$,收敛迭代次数为 $n=363+W=463$.在随后的迭代中,$z[n]$的最大整数值和最大边界 $B_{ouppp}[n]$保持不变.

在第二个实验中,采用不同的误差界限($E=1$,$E=0.01$,$E=0.001$),对于 # 4 和 # 5 测试向量的直接型和并联型结构,运行精度分析算法,输入值域均为[-100,100].表 2 显示了运行结果.就面积而言,并联型实现要好得多.像以前的实验一样,表 2 中所有测试向量 e_{max}和分数比特宽度的鲁棒运算仅只需要几秒钟.表 2 中值得注意的一点是,我们减小输入的错误界限 E 时,算法收敛需要的迭代次数没有明显增加.这意味着对于具有任意错误界限的一个特定的滤波器结构,精度算法以大概相同的时间收敛到鲁棒结果.

表 2　对给定的任意错误界限 E 计算分数比特宽度

向量	错误界限"E"	迭代		e_{max}		FB		# 门		时间(s)	
		DR	PRL	DR	PRL	DR	PRL	DR	PRL	DR	PRL
3	1	484	480	0.647	0.6052	18	16	17045	11624	7.95	18.1
	0.01	460	480	0.0039	0.0066	26	22	30879	18773	10.5	23.6
	0.001	411	489	0.00016271	0.00074901	29	26	37101	24355	11.4	24.9
4	1	348	389	0.2386	0.9691	20	14	20274	10728	2.63	11.1
	0.01	394	369	0.0093	0.0056	24	21	27861	18134	2.41	12.1
	0.001	348	358	0.00032518	0.00037939	29	25	39596	23958	2.95	17.6

最后一个实验对本文提出的算法与文[12]中的方法进行比较.表 3 结果基于两个测试向量和误差界限 $E=0.01$ 与非对称输入值域[0,127].使用文[12]中的方法运算得到的值域导致了额外的整数比特与分数比特.在表 3 中,情况 1(例 2)在文[12]的方法下需要 17 比特来表示分数,而我们的算法仅需要 15 比特.结合整数比特和分数比特,使用 Xilinx ISE v.11 工具在 FPGA 上综合这两个滤波器,我们算法得到的实现能减少 10%的逻辑门,并能提供更优的最大不精确匹配值 e_{max}.

考虑表 3 中第二个测试情况(测试向量 # 1,一个 3 阶滤波器),文献[12]中的方法不能用于精度计算,这是因为此方法对于二阶滤波器,当系数量化错误被插入到系统中时,使用敏化分析来跟踪零极点扰动.该分析只对二阶滤波器有效,而不能用于高阶滤波器.

表 3　比较我们提出的算法与文献[10]中的算法

向量	参考	值域	e_{max}	IB/FB	# 门
例 2	[12]	(-632,632)	0.0049	11/17	15729
	本文	(-417,417)	0.0026	10/15	13258(节约 10%)
1	[12]	(-530,530)	-	11/-	-
	本文	(-115,415)	0.0098	10/16	5747

6　结论与未来工作

本论文提出了高效算法用以分析 IIR 滤波器的值域与精度.过去的传统方法不适用于高阶滤波器或导致过高估计.为了克服这些局限,本篇论文首先给出了 IIR 滤波器值域与精度问题的相关定义与定理,在此基础上提出了启发式算法能用于值域与精度的鲁棒计算并能分配合适的整数比特和分数比特.实验结果证明了算法的收敛性与鲁棒性.未来的工作将延伸此方法,使其能涵盖其余的误差度量比如最小均方差和信噪比,提出除了截断之外其他的缩放技巧以及处理非线性反馈电路.

参考文献

[1] A Gaffar, O Mencer, W Luk, P Cheung. Unifying bit-width optimization for fixed-point and floating-point designs [A]. IEEE Symposium on Field-Programmable Custom Computer [C]. Napa, USA: IEEE CS Press, 2004. 79 - 88.

[2] A Nayak, M Haldar, A Choudhary, P Banerjee. Precision and error analysis of matlab applications during automated synthesis for FPGAs [A]. Proceeds of DATE [C]. Munich, Germany: IEEE CS Press, 2001. 722 - 728.

[3] C Shi, R Brodersen. Automated fixed-point data-type optimization tool for signal processing and communication systems [A]. Proceedings of Design Automation Conference [C]. San Jose, USA: IEEE CS Press, 2004. 478 - 483.

[4] K Kum, W Sung. Combined word-length optimization and highlevel synthesis of digital signal processing systems [J]. IEEE Transaction on CAD, 2001, 20(8): 921 - 930.

[5] D U Lee, A Gaffar, R C Cheung, O Mencer, W Luk, G Constantinides. Accuracy-guaranteed bit-width optimization [J]. IEEE Transaction on CAD, 2006, 25(10): 1990 - 2000.

[6] Y Pang, K Radecka, Z Zilic. Arithmetic transforms of imprecise datapaths by Taylor series conversion [A]. Proceedings of IEEE International Conference on Electronics, Circuits and Systems [C]. Nice, France: IEEE CS Press, 2006. 696 - 699.

[7] Y Pang, K Radecka. Optimizing imprecise fixed-point arithmetic circuits specified by Taylor series through arithmetic transform [A]. 45th ACM/IEEE DAC [C]. San Jose, California, USA: IEEE CS Press, 2008. 397 - 402.

[8] Y Pang, K Radecka, Z Zilic. Optimization of imprecise circuit represented by Taylor series and real-valued polynomials [J]. IEEE Transactions on Computer-Aided Design, 2010, 29(8), 1177 – 1190.

[9] A Kinsman, N Nicolici. Finite precision bit-width allocation using SAT-modulo theory [A]. IEEE DATE [C]. Nice, France: IEEE CS Press, 2009. 1106 – 1111.

[10] T Harju. Finite wordlength implementation of IIR polynomial predictive filters [A]. Instrumentation and Measurement Technology Conference [C]. Ottawa, Canada: IEEE IMS Press, 1997. 60 – 65.

[11] L D Milic, M D Lutovac. Design of multiplierless elliptic IIR filters with a small quantization error [J]. IEEE Transactions on Signal Processing, 1999, 47(2), 469 – 479.

[12] J Carletta, R Veillette, F Krach, Z Fang. Determining appropriate precisions for signals in fixed-point IIR filters [A]. Design Automation Conference [C]. San Jose, USA: IEEE CS Press, 2003. 656 – 661.

[13] P S Diniz, J E Cousseau, A Antoniou. Improved parallel realisation of IIR adaptive filters [J]. IEE Proceeding of Circuits, Devices and Systems, 1993, 140(5), 322 – 328.

[14] J G Proakis, D G Manolakis. Digital Signal Processing: Principles, Algorithms, and Applications [R]. New Jersey: Prentice-Hall, 1996.

作者简介

庞 宇 男,1978 年 10 月出生于四川省泸州市,2010 年毕业于加拿大 McGill 大学获博士学位.现为重庆邮电大学光电工程学院副教授、硕士生导师.在 IEEE 上发表学术论文 20 余篇.主要研究方向为集成电路设计与验证、数字信号处理、短距离无线通信.

E-mail: pangyu@cqupt.edu.cn

贺志龙 男,1987 年 2 月出生于湖南省邵阳市,现为重庆邮电大学通信工程学院硕士研究生.主要研究方向为短距离无线通信.

E-mail: zlon0228@126.com

王绍全 男,1988 年 8 月出生于云南省丽江市,现为重庆邮电大学通信工程学院硕士研究生.主要研究方向为通信信号处理.

王骏超 男,1990 年出生于新疆乌鲁木齐市,现为重庆邮电大学光电工程学院本科生,主要研究方向为逻辑综合.

高 翔 男,1988 年 10 月出生于重庆市,现为重庆邮电大学通信工程学院硕士研究生.主要研究方向为集成电路设计.

吴 玮 男,1975 年 5 月出生于天津市,现为四川大学电子信息学院副教授.主要研究方向为数字信号处理.