Quantization effects in the polyphase N-path IIR structure.

Artur M. Krukowski
Richard C.S. Morling
Izzet Kale

School of Informatics

Quantization Effects in the Polyphase N-Path IIR Structure

Artur Krukowski, Richard Charles Spicer Morling, *Member, IEEE,* and Izzet Kale, *Member, IEEE*

***Abstract*—Polyphase IIR structures have recently proven themselves very attractive for very high performance filters that can be designed using very few coefficients. This, combined with their low sensitivity to coefficient quantization in comparison to standard FIR and IIR structures, makes them very applicable for very fast filtering when implemented in fixed-point arithmetic. However, although the mathematical description is very simple, there exist a number of ways to implement such filters. In this paper, we take four of these different implementation structures, analyze the rounding noise originating from the limited arithmetic wordlength of the mathematical operators, and check the internal data growth within the structure. These analyses need to be done to ensure that the performance of the implementation matches the performance of the theoretical design. The theoretical approach that we present has been proven by the results of the fixed-point simulation done in Simulink and verified by an equivalent bit-true implementation in VHDL.**

***Index Terms*—Digital filters, dynamic range analysis, polyphase IIR structure, quantization effects, quantization noise, roundoff noise.**

I. Introduction

ONE OF the important implementation issues, which have to be considered during the design of the architecture for any type of filter, is the storage requirement for the internal calculations. The size of the memory has to be such that it does not cause the loss of precision due to rounding effects of the results of internal multiplications and summations. It happens very often, especially for IIR filters having a feedback loop, that even if the input and output samples are limited to unity and are represented with w-bits, the internal values might have values well above one (even infinitely large values for unity allpass coefficient) and might require many more bits than w in order to provide a reliable output. Additionally, there may be a big difference between internal values. The result of one summation can be below unity; the output of the other one may be very large. This makes for many problems in the implementation, as it would require varying the position of the decimal point

$$H(z) = \frac{1}{N}\sum_{n=0}^{N-1} A_n(z^{-N})z^{-n}$$

$$A_n(z^{-N})\prod_{k=1}^{K_n} A_{n,k}(z^{-N}) = \prod_{k=1}^{K_n}\frac{\alpha_{n,k}+z^{-N}}{1+\alpha_{n,k}z^{-N}}. \qquad (1)$$

Manuscript received May 29, 2001; revised September 23, 2002.

The authors are with the Applied DSP and VLSI Research Group, Department of Electronic Systems, University of Westminster, London, U.K.

Digital Object Identifier 10.1109/TIM.2002.808032

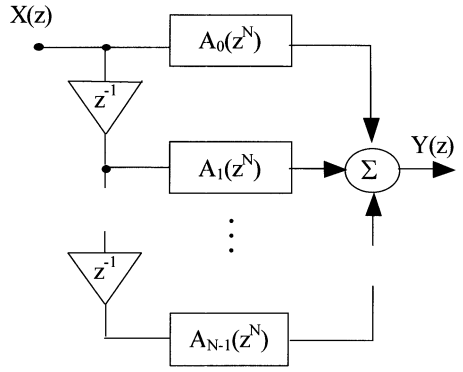

Fig. 1. General polyphase structure.

The general polyphase IIR structure [1]–[3] is given in Fig. 1 and described by (1). It incorporates Nth-order allpass subfilters, $A_i(z)$, that can be implemented both in floating-point and in fixed-point arithmetic using any combination of structures shown in Fig. 2. Because of the small number of calculations required per filter order and very high performance, such a structure is very attractive for fixed-point implementation like for decimation filters for high-accuracy A/D converters [4]. It is advantageous over the floating-point one in terms of calculation speed, area on the integrated circuit, and the total power consumption. The disadvantage is that the fixed-point implementation is subject to such effects as quantization noise, caused by quantizing the result of multiplications and summations to the internal arithmetic wordlength, and limit-cycle oscillations—repetitive flipping of the least significant bit (LSB)—caused by the chosen quantization scheme. The effect of the quantization noise can be reduced at the time of designing the filter. For example, for the lowpass filter design, knowing the passband and stopband ripples, the arithmetic wordlength will be chosen such that the estimated overall quantization noise falls below the level of filter ripple. It will be shown later in this paper how this noise characteristic can be determined, assessed and applied in the filter design. The second quantization effect can be limited by the appropriate selection of the quantization (loss of precision) schemes and the choice of the internal arithmetic wordlength.

Using any type of fixed-point binary arithmetic with a uniform quantization step size, whether it is signed binary, one or two's complement, will cause errors due to product quantization and due to register overflow after additions. In the filtering operation, where a number of multiplications and additions are undertaken, the error will accumulate, causing in the most drastic cases wrong filtering results. Therefore, proper care has to be

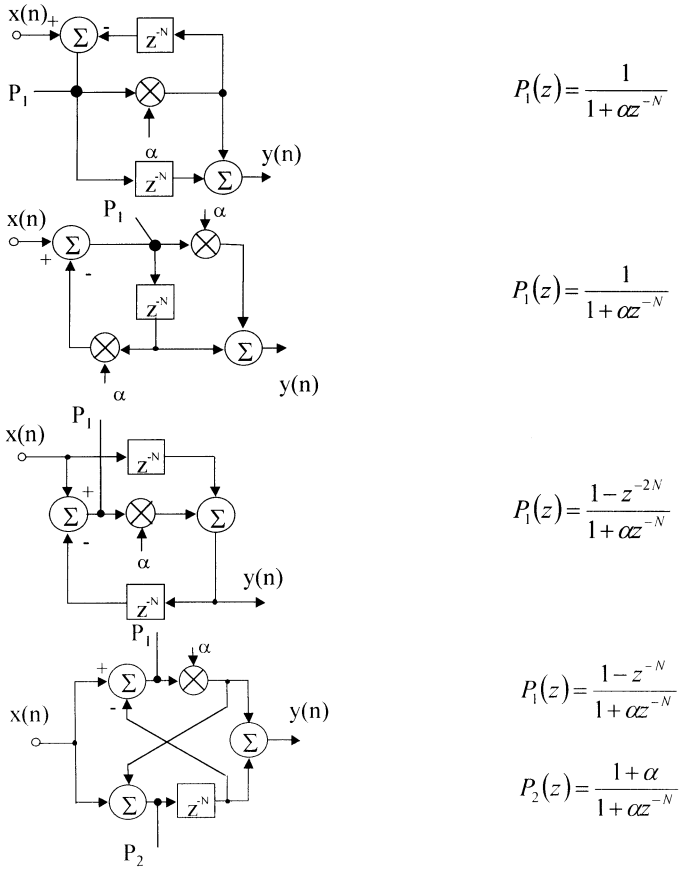

Fig. 2. Peak gains for different allpass structures.

taken to avoid or minimize the effect of the arithmetic quantization error. The error due to product quantization is caused by the fact that a product of two fixed-point numbers with N_1 and N_2 bits, respectively, will yield a new number that requires $N_1 + N_2 - 1$ bits for its nonerror representation. On the other hand, an addition of the same two numbers yields a result that requires at most $\max(N_1, N_2) + 1$ bits. In the digital filter implementation, the signal values are multiplied with the filter coefficients, and the result is usually constrained back to the original signal wordlength. It is not feasible to allow the increase in the number of bits at every product calculation, especially when a recursive filter structure is used, as the wordlength would keep growing too rapidly toward infinity after successive multiplications. It is commonly used in the literature to model the product quantization as an additive disturbance

$$Q[x(n)] = x(n) + e(n) \tag{2}$$

where $Q[.]$ denotes the quantization operation on the argument signal and $e(n)$ is the disturbance due to quantization. Assuming that the quantization noise for the w-bit data path length is caused by variation of the $\Delta = \text{LSB} = 2^{-w}$, then to a first approximation the quantization error can be modeled as a white noise source with power of $P_e = \Delta^2/12 = 2^{-2w}/12$ [5], [6]. Such an approximation is used very often in modeling quantization effects [6], [7] even if it does not consider such effects like the loss of precision, when the data needs to be represented with a smaller number of bits than is required for keeping it in full precision, and the correlation between the quantization noise and an input signal. The method of modeling the quantization process more accurately by considering the loss of precision mechanisms can be found in [8]. The correlation between the quantization noise and the structure of the input signal is explained in [9], [10]. There are a number of different quantization schemes: truncate, round to infinity, round to zero, round, and convergent round. These differ in dc offset (bias) and number of loss bits to consider. For most of the quantization schemes, except for the simple truncation, the LSB and some of the discarded bits need to be considered to derive the final quantized value. This requires more complicated hardware to perform the quantization. For applications when dc bias is detrimental, the loss of precision should be carefully considered. There are only two quantization schemes that are dc bias free: round-to-zero and convergent-round. This is due to the results of the quantization being rounded up and down with, statistically, equal probability.

II. Peak Gains Along Allpass Filter Structures

When looking at the whole structure of the polyphase filter in Fig. 1, it can be seen that it does not have any feedback. The block, which makes the polyphase structure an IIR one, creating local feedback loops, is the allpass filter $A_i(z)$. For the purpose of our analysis we made an assumption that the input signal, just like the filter coefficients, is less than or equal to one. Such an approach, where there are no integer bits and the whole data is the mantissa, simplifies the analysis of both the values of the internal calculation and assessment of the quantization noise.

When the input signal is limited to unity, then the output of the allpass filter is also limited to unity. Therefore, the summation of two such filters would require one extra bit to represent the result of the addition without loss of precision. In most practical cases, application of the structure in Fig. 1 is used with the number of paths equal to the power of two, i.e., $N = 2^n$ with n being an integer. In such a case, the output scaling by N can be realized as a simple n-bit right-shift operation, moving the fractional point by n bits. Regarding the Nth-order allpass filters themselves, they can be, in principle, implemented in at least four different ways, as shown in Fig. 2. They differ in terms of the number of summations and multiplications as well as the arrangement of their numerator and denominator parts—which one is first and which one is second.

The important distinction between the structures is the value of the results of the internal multiplications and summations, which influences the memory size requirements. These results are dependent not only on the value of the allpass coefficient, but also on the frequency of the input signal. In order to assess how big the internal value may become (its dynamic range), transfer functions of internal calculations (TFIC) have to be determined for each allpass filter structure between the input of the filter (limited to unity) and the input to each of the delayers (implemented as a memory). In other words, this has to be done only for outputs of each adder. As filter coefficients are less than one for the polyphase structure, multipliers will not contribute to the increase of the peak internal value. The dynamic range analysis can be done without considering the quantization effects

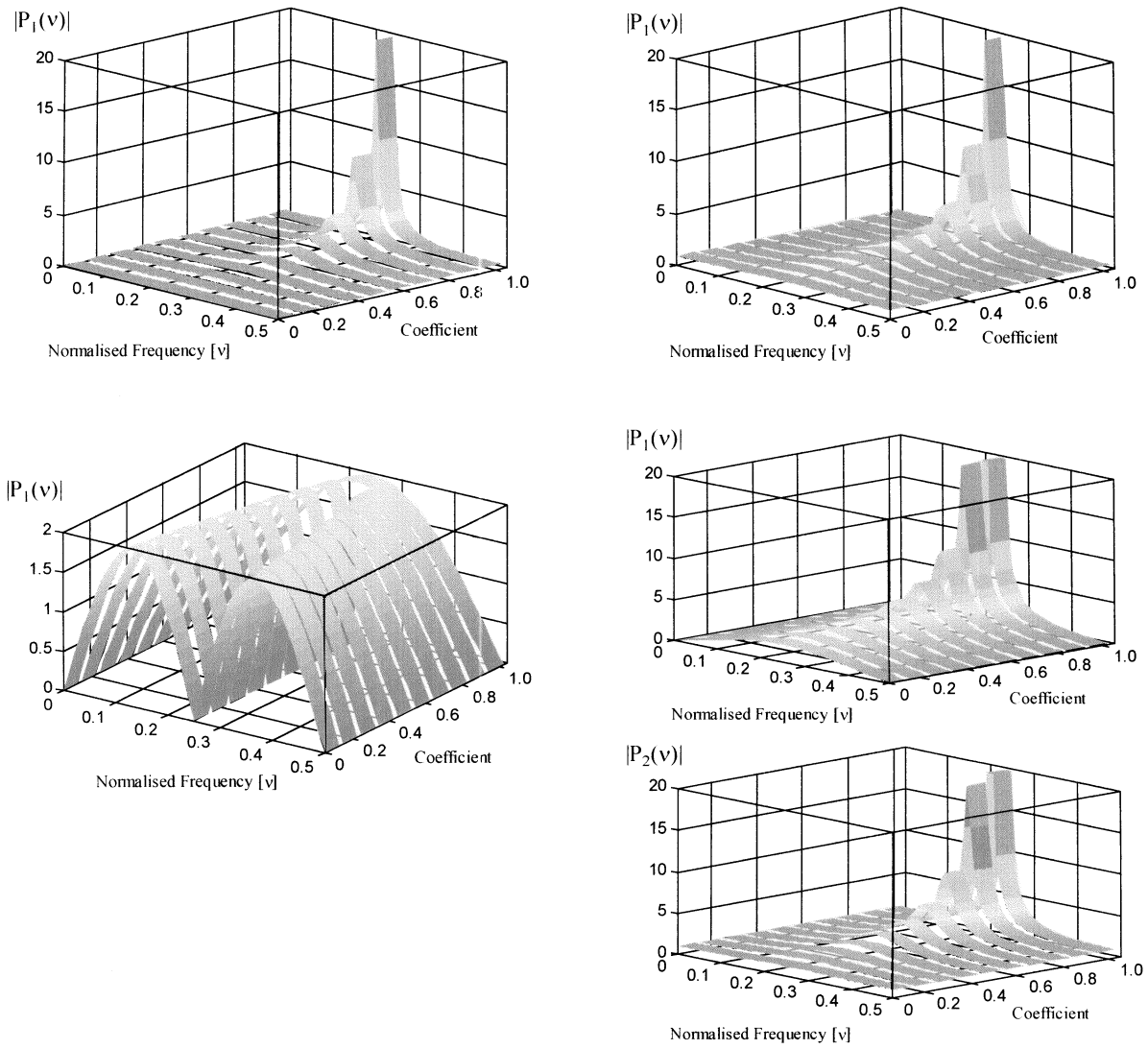

Fig. 3. Peak gains for coefficient values $0 < \alpha < 1$ for allpass structures from Fig. 2 ($N = 2$).

because of the several order of magnitude differences in value between the one and the other. The transfer functions in question have been determined for all of the allpass filter structures and are presented in Fig. 3.

Those transfer functions have been evaluated for a range of coefficient values between zero and one, and the resulting magnitude responses are shown in Fig. 3. These plots clearly illustrate how only one structure [Fig. 2(c)] has TFIC magnitude responses limited to a finite value of two for any frequency and independent of the coefficient value. All the other ones are peaking at half-Nyquist ($\nu = 0.25$) to very large numbers. This means for such structures that, if the input signal has frequency components close to half-Nyquist, then the results of the internal multiplication and summations may well have very large values, causing problems with different positioning of the fractional point in different parts of the system or inaccuracy of the final result.

The verification of the theoretical analysis has been done by applying various input signals to the fixed-point model of the polyphase filter designed using the fixed-point blockset from Simulink™: single tones at various frequencies, impulse, wideband signals as well as noise sources and speech. The simulation results matched the theory. The maximum values at the points of interest were below limits given in Fig. 2, derived for an impulse input.

Only one structure, in the numerator–denominator (N–D) arrangement, does not suffer from such effects. For the dc, half-Nyquist, and Nyquist frequencies, the memory will store very small numbers. The maximum values are for two frequencies: $\nu = 0.125$ and $\nu = 0.375$. This can be taken care of by increasing the size of the memory by one additional integer guard bit. The advantage of this allpass structure is that it makes it easy to create higher order allpass filters, simply by cascading a number of such structures together (Fig. 4), sharing one delayer between each pair of allpass sections.

III. Quantization Noise Due to Rounding of Arithmetic

In the polyphase filter, like in any other filter, quantization has to be performed on the result of any arithmetic operation. This is because any such operation requires more bits to represent the result than is required for each of the operands. If the

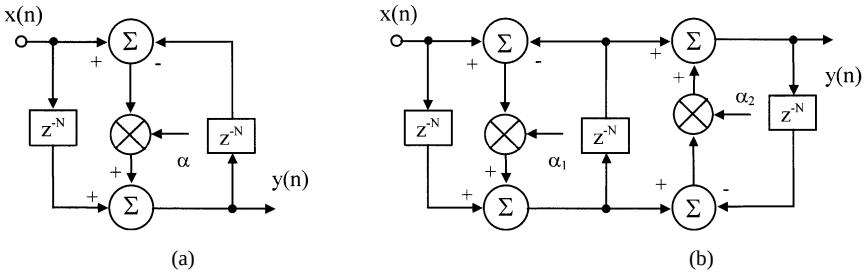

Fig. 4. (a) Using a one-coefficient allpass section (b) to form higher order allpass filters.

wordlength were always to be adjusted to store the data in full precision, this would be impractical, as there would soon be too many bits required to be stored in the available memory. Therefore, the wordlength of the internal data, w, has to be chosen, and the result of any arithmetic operation has to be constrained back to w_a using the quantization scheme chosen from the ones shown in the previous section, as appropriate for the given application. The quantization effects in allpass filters have been studied in different publications [5], [10]. This paper concentrates on the analysis of the special case of the polyphase structure detailed in Fig. 1, incorporating allpass filters as shown in Fig. 2.

The quantization operation may cause a disturbance to the result of the arithmetic operation. For normal filtering operations, such a quantization disturbance can usually be successfully considered as white noise and modeled as an additive noise source at the point of the arithmetic operation with the quantization step equal to the LSB of the internal data, $LSB = 2^{-w}$. This certainly is not the case for zero-valued or constant input signals. However, modeling the quantization has—in most cases—the purpose of determining the maximum noise disturbance in the system. Hence, even if the additive quantization noise model gives overestimated values of the noise for very specific signals, this fact does not decrease the usefulness of the approach. After the shape of the quantization noise power spectral density (NPSD) is found, it can be used to identify regions that might cause overloading or loss of precision due to arithmetic noise shaping; also the required input signal scaling and the required internal arithmetic wordlength can be estimated for a given noise performance. The standard methods of estimating the maximum signal level at a given node are L1-norm (modulus of the impulse response—worst-case scenario), L2-norm (statistical mean-square), and L∞-norm (peak in frequency domain giving the effect of the input spectral shaping). These norms can be easily estimated for the given node from the shape of the NPSD.

The quantization noise injected at each adder and multiplier, originally spectrally flat, is shaped by the noise shaping function (NSF), $H_N(z)$, calculated from the output of the filter to the input of each of the noise sources, i.e., to the output of each of the arithmetic operators. These functions were calculated for all of the allpass filter structures from Fig. 2 and are shown in Fig. 5. The shapes of the nontrivial of the NFS are shown in Fig. 6.

The accumulated quantization NPSD transferred to the output, $S_{ee}(\nu)$, is obtained by shaping the uniform NPSD from each of the quantization noise sources by the square of

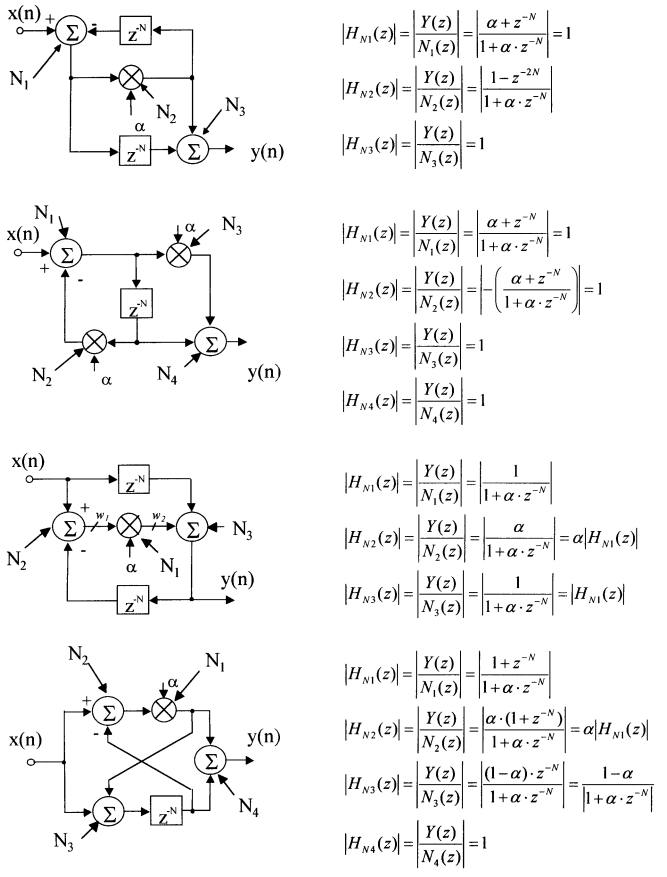

Fig. 5. Quantization noise-shaping functions for selected test points of the allpass structures.

the magnitude of the NFS corresponding to the given noise injection point and can be described by (3)

$$S_{ee}(\nu) = \sum_k |H_{Nk}(\nu)|^2 S_Q = \sum_k |H_{Nk}(\nu)|^2 \frac{2^{-2w_a}}{12}. \quad (3)$$

The results show that all structures perform in a way very distinct from the other ones. Structure (a) has the best performance at dc, half-Nyquist ($\nu = 0.25$), and Nyquist ($\nu = 0.5$), where the NPSD falls toward minus infinity. Its two maxima are symmetric about $\nu = 0.25$ and independent of the coefficient value. The peaks are distant from $\nu = 0.25$ for small coefficient values and approaches it as the coefficient increases. Structure (b) has uniform noise spectral distribution as all the arithmetic operations are either at the filter input—then noise is shaped by the allpass characteristic of the whole filter—or at its

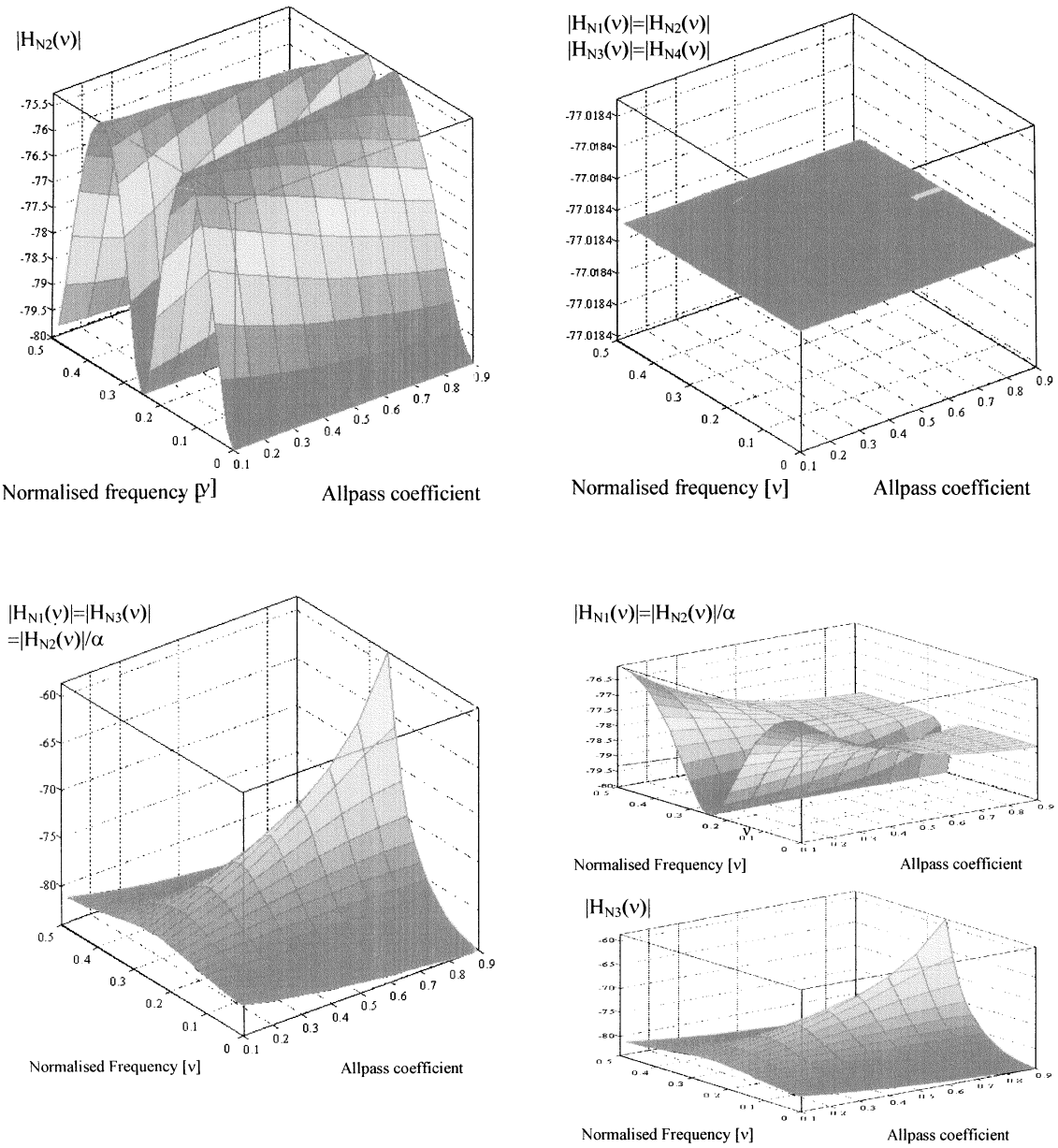

Fig. 6. Shapes of NFS for coefficient values $0 < \alpha < 1$ for allpass structures from Fig. 2 ($N = 2$).

output. Structure (d) also has a minimum at $\nu = 0.25$. Its average noise power level decreases as the value of the allpass coefficient increases. Structure (c), the best from the point of view of the required guard bits, has its maximum at $\nu = 0.25$ going toward infinity for coefficient values approaching one. This effect is a result of the denominator of the Nth-order allpass filter causing the poles of the filter to move toward the unit circle at normalized frequencies of $\nu = 2\pi k/N$, $k = 0 \cdots N-1$, for the coefficient α approaching one. If there is no counter effect of the numerator, like for the case of $P_1(z)$ for structure (c) and $|H_{N2}(z)|$ for structure (a), then the function goes to infinity. Even though structure (c) goes to infinity at $\nu = 0.25$ for $\alpha = 1$, it has the lowest average noise power from all the structures. This structure has a big advantage in terms of the number of required guard bits and ease of cascading a number of them into higher order allpass filters. If the filter coefficients approach one, then the increase in quantization noise power could be countered with few additional bits. Using other structures would only replace the problem of dealing with an increase in the quantization noise with the problem of having to increase the number of guard bits required to deal with an increase of the peak gains.

The NPSD of the quantization noise at the output of the polyphase structure can be calculated as the sum of the NPSD at the output of all allpass filters in the filter scaled by the $1/N$ factor, N being the number of paths. If the filter is cascaded with another filter, the NPSD of the first one will also be shaped by the square of the magnitude of the second filter.

The verification of the theoretical analysis was done in Simulink™ by comparing the results from the fixed-point implementation with the floating-point equivalent that incorporated quantization effects modeled as additive white noise

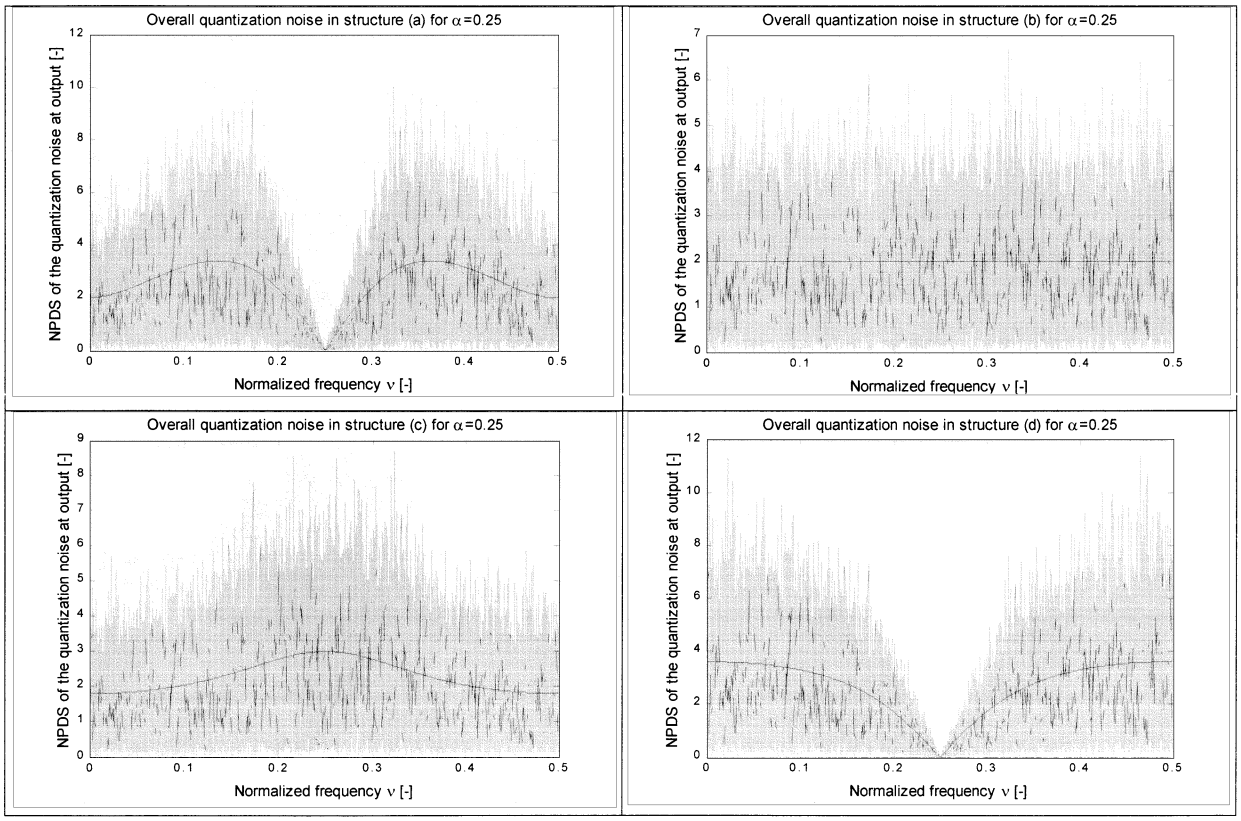

Fig. 7. Output quantization noise normalized to P_e for allpass structures from Fig. 2 ($\alpha = 0.25$).

sources. The intention was to check the correctness of the theoretical equations by applying the white noise sources instead of quantization and by performing the quantization after addition and multiplication (rounding and truncating) to verify the shaping of the quantization noise and its level both for white input noise sources and real-life signals. The shape of the output quantization noise accumulated from all arithmetic elements for a wide-band input signal assuming, for simplicity, no correlation between the noise sources, is shown for all considered allpass structures in Fig. 7. The solid curve indicates the theoretical NSF that is very well matching the median of the quantization noise (curves lying on top of each other). The quantization noise power increase calculated for the given coefficient $\alpha = 0.25$ was 8.5 dB for structure (a), 6 dB for structure (c), 7.3 dB for structure (d), and 9 dB for structure (b).

It is clear that the quantization "noise" differs from the assumed white noise characteristic. However, the approximation still holds with an accuracy of around 5–10% depending on the structure of the input signal. An example of more accurate modeling of the quantization noise caused by arithmetic operations can be found in [8]. The arithmetic quantization noise certainly decreases the accuracy of the filter output. The value of the arithmetic wordlength has to be chosen such that the quantization noise power is smaller than the stopband attenuation of the filter and the stopband ripples. In certain cases, the design requirements have to be made more stringent to allow some unavoidable distortion due to the arithmetic wordlength effects. For the case of decimation filters for the $\Sigma\Delta$ based A/D converters, the quantization noise adds to the one originating from the modulator. In such a case, each stage of the decimator has to be designed so that it filters out this noise as well.

The verification of the peak gain analysis was performed by applying single-tone signals at the characteristic frequencies—where functions from Fig. 2 have their extremes—and by using wideband signals to make sure that the estimates are accurate. The experimental results confirmed the theoretical calculations. The results of the simulation for the white noise input signal of unity power are given in Fig. 8. The simulation was performed for a white noise input signal of unity power in order to have a uniform gain analysis across the whole range of frequencies. The theoretical shape of the gain is shown by a solid line that is very closely matching the median value of the signal at the test points.

IV. Conclusion

In this paper, we have presented the analysis of the N-path polyphase IIR structure in view of its application for fixed-point implementation. We have shown the different ways of implementing the allpass filters that are the only recursive elements in the structure and therefore very sensitive to internal overflows. We identified structure (c) from Fig. 2 to be the best for full-band filtering. Any other structure can also be used, but the usefulness is limited to signals having no spectral components at 0.25 of the sampling frequency or its closest vicinity. Otherwise internal overflow may occur. The same structure (c) also performs very well in terms of the quantization noise injection.

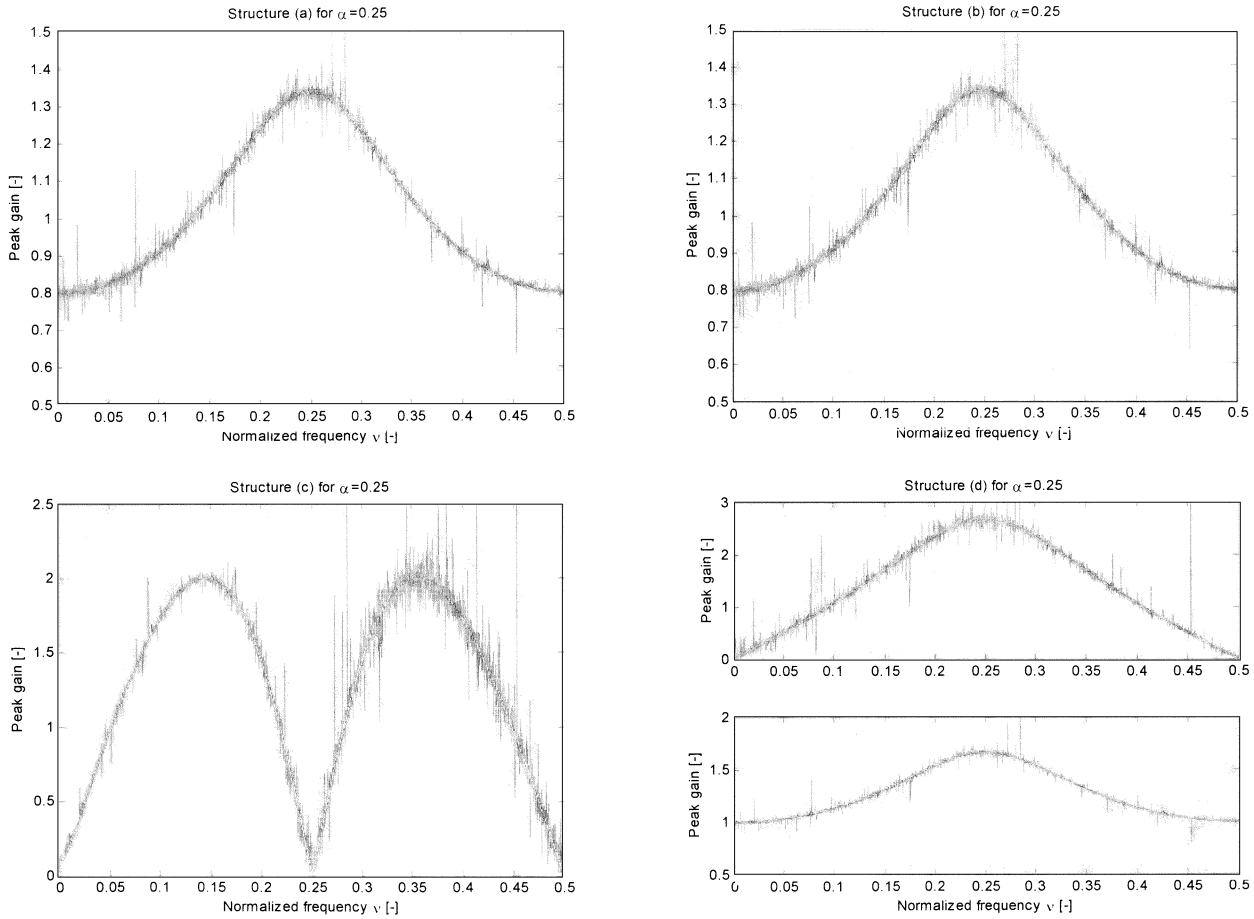

Fig. 8. Peak gains for white noise input for allpass structures from Fig. 2 ($\alpha = 0.25$).

Even though it peaks in its transition band, it has the lowest injection in the filter passband and stopband, in the parts of the filter that are the most important. In most of the cases the transition band is not that important anyway. If the noise performance is important in the transition band of the filter, then the structure (a) from Fig. 2 could be a choice. In such a case, the noise has its minimum in the transition band as well as at dc and at the half sampling frequency. If it is preferable to have the same noise injection throughout the frequency range, then structure (b) from Fig. 2 could be the choice.

REFERENCES

[1] M. Renfors and T. Saramäki, "Recursive Nth-band digital filters—Part II: Design of multistage decimators and interpolators," *IEEE Trans. Circuits Syst.*, vol. CAS-34, pp. 40–51, Jan. 1987.

[2] F. Harris, M. d'Oreye de Lantremange, and A. G. Constantinides, "Digital signal processing with efficient polyphase recursive all-pass filters," in *Proc. IEEE Int. Conf. Signal Processing*, Florence, Italy, Sept. 4–6, 1991.

[3] R. A. Valenzuela and A. G. Constantinides, "Digital signal processing schemes for efficient interpolation and decimation," *Proc. IEE*, pt. G, vol. 130, no. 6, pp. 225–235, 1983.

[4] I. Kale, R. C. S. Morling, and A. Krukowski, "A high-fidelity decimator chip for the measurement of sigma–delta modulator performance," *IEEE Trans. Instrum. Meas.*, vol. 44, Oct. 1995.

[5] P. P. Vaidyanathan, "On coefficient-quantization and computational roundoff effects in lossless multirate filter banks," *IEEE Trans. Signal Processing*, vol. 39, pp. 1006–1008, Apr. 1991.

[6] R. C. S. Morling and I. Kale, "Dynamic range of allpass filter structures," in *Proc. IEEE Int. Symp. Circuits Syst.*, vol. 4, Phoenix, AZ, May 2002, pp. 433–436.

[7] J. Holmberg, L. Harnefors, and S. Signell, "Quantization noise analysis of wave digital and lossless digital integrator allpass/lattice filters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 3, May 1999, pp. 1133–1136.

[8] H. Jing, L. Gang, X. Xibin, and Y. Yan, "Estimation for the quantization noise spectrum of linear digital filter," in *Proc. Int. Conf. Commun. Technol.*, vol. 1, 2000, pp. 184–187.

[9] F. Hartwig and A. Lacroix, "Roundoff noise analysis on the basis of an improved floating point error model," in *Proc. IEEE Int. Symp. Circuits Syst., Connecting the World*, vol. 2, 1996, pp. 133–136.

[10] J. Kauraniemi and T. I. Laakso, "Roundoff noise analysis of modified delta operator direct form structures," in *Proc. IEEE Int. Symp. Circuits Syst.*, vol. 4, 1997, pp. 2365–2368.

Artur Krukowski was born in Warsaw, Poland, in 1967. He received the M.Sc. degree in instrumentation and measurement from the Warsaw University of Technology, Warsaw, in 1992, and the M.Sc. and Ph.D. degrees in DSP from the University of Westminster, London, U.K., in 1993 and 1999, respectively.

In 1993, he joined the staff of the University of Westminster working as an Academic Researcher, and, since 1999, as a Post Doctoral Researcher in advanced DSP Systems in the applied DSP and VLSI Research Group, and, in 2001, he became a permanent member of the research staff. His areas of interest include multirate digital signal processing for telecommunication systems, digital filter design, and their efficient low-level implementation, integrated circuit design, digital audio broadcasting, teleconferencing, and internet technologies for teaching.

Richard Charles Spicer Morling (M'79) was born in Rochford, U.K., in 1947. He received the B.Sc. (honors) degree in physics from the Polytechnic of Central London, London, U.K., in 1971, and the Ph.D. degree in information engineering from the City University, London, in 1989.

He worked in the field of television for Decca Records, Ltd. and, during a period at the Imperial College of Science and Technology, London, was engaged in the design of electromyographic equipment for the migraine trust. He joined the staff of the University of Westminster, London, in 1971, where he is currently the Chair of the Department of Electronic Systems. His research activities have included packet-switching local area networks, discrete-time signal processing structures, aircraft breaking systems, design methodologies for integrated circuit design, sigma–delta conversion techniques, and teaching methods in electronic engineering. He is currently working on the low-power integrated circuit design for mobile communication systems and hearing aids.

Izzet Kale (M'88) was born in Akincilar, Cyprus. He received the B.Sc. (honors) degree in electrical and electronic engineering from the Polytechnic of Central London, London, U.K., the M.Sc. degree in the design and manufacture of microelectronic systems from Edinburgh University, Edinburgh, U.K., and the Ph.D. in techniques for reducing digital filter complexity from the University of Westminster, London.

He joined the staff of the University of Westminster in 1984, and he has been with them since. He is currently Professor of applied DSP and VLSI systems, leading the applied DSP and VLSI Research Group at the University of Westminster. His research and teaching activities include digital and analog signal processing, silicon circuit and system design, digital filter design and implementation, and A/D and D/A sigma–delta converters. He is currently working on efficiently implementable, low-power DSP algorithms/architectures and sigma–delta modulator structures for use in the communications and biomedical industries.