

FIXED-POINT-COEFFICIENT FIR FILTERS AND FILTER BANKS: IMPROVED DESIGN BY RANDOMIZED QUANTIZATIONS

Heute, U., Srivastav, A., Sauerland, V., Kliewer, J.

Chair for Circuit & System Theory / Chair for Discrete Optimization
Faculty of Engineering, Christian-Albrechts Universität, D-24143 Kiel, Germany

ABSTRACT

Frequency-selective, linear-phase FIR filters are considered, as single systems and within analysis-synthesis filter banks. They are usually designed, in the single-channel case, to fulfill tolerances in the Chebychev sense, or, in near-perfect-reconstruction filter banks, to minimize a reconstruction-error measure. If hardware is limited, fixed-point coefficient quantization is needed. It causes, in general, tolerance violations or a larger reconstruction error. Discrete re-optimization may help. A recent technique, able to handle also large filter orders, is successfully applied and newly extended to filter banks. Even better are randomized strategies, introduced and examined in the mathematical-optimization community over the past 15 years; especially, randomized rounding is very effective. Thereby, good results are found for both single-system and filter-bank designs. We further introduce a new random sub-set selection within the above re-optimization. Like randomized rounding, it allows a trade-off between computational effort and solution quality. Clear improvements over deterministic heuristics are obtained by both randomized algorithms.

1. INTRODUCTION

1.1 Linear-Phase FIR Filters

FIR filters are defined by their finite-length impulse response $h(k)$, $k \in \{0, 1, \dots, n\}$; n denotes the filter order, $(n+1)$ the filter length. The corresponding transfer function is of the all-zero type $H(z) = \sum_{k=0}^n h(k) \cdot z^{-k}$. Choosing $h(k)$ symmetrically leads

to strictly linear-phase systems. Without restriction of generality, we consider a real-valued filter with even order $n=2N$ and even-symmetry, i.e., $h(n-k) = h(k)$, $k \in \{0, 1, \dots, N\}$, whose frequency response is a purely real, even function $H_o(\Omega)$ in

$$H(e^{j\Omega}) = e^{-jN\Omega} \cdot \left[h(N) + 2 \sum_{k=0}^{N-1} h(k) \cdot \cos((N-k) \cdot \Omega) \right] \\ \doteq e^{-jN\Omega} \cdot H_o(\Omega), \quad (1)$$

except for a linear phase term. This type of systems is of interest for frequency-selective filters, including filter banks; also odd-order systems are possible here; their description may be transformed into the above form. Odd-symmetry systems are used in other applications not covered here.

1.2 Design of Frequency-Selective Filters

The choice of $h(k)$ defines the shape of $H_o(\Omega)$. For a low-pass filter (or high-pass etc.), a rectangular shape is desired, to be approximated by $H_o(\Omega)$ according to a suitable error criterion.

Mostly, a Chebyshev solution is appropriate: The cosine polynomial of (1) fulfills a tolerance scheme with a minimized maximum deviation reached by the maximum possible number of “equal error ripples” (see Fig. 1 for a low-pass). This solution is found efficiently by means of the well-known program due to Parks and McClellan [1], using the Remez exchange algorithm [2] and included in available software packages [3, 4]. A priori, this program takes prescribed edge frequencies $\{\Omega_p, \Omega_s\}$ and minimizes the stop-band ripple δ_s , given an order n and a prescribed ratio $R \doteq \delta_p / \delta_s$. This case will be dealt with here, although there are well-known formulae [5, 6, 7] also for the order n needed for a complete tolerance prescription $\{\delta_1, \delta_2, \Delta\Omega \doteq \Omega_s - \Omega_p\}$.

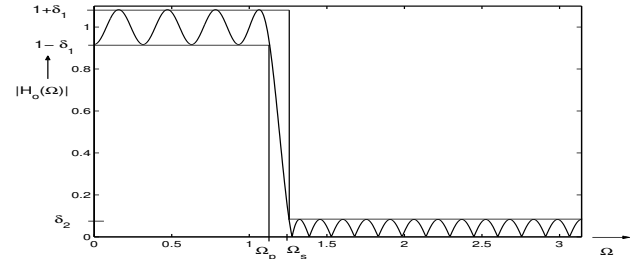


Figure 1: Example of an equal-ripple FIR-filter design.

1.3 Design of Filter-Bank Prototype Filters

The combination of a single narrow low-pass filter in a so-called poly-phase network (PPN) with a subsequent spectral transformation (DFT or DCT, implemented efficiently by an FFT) realizes a filter bank with M pass-bands (“channels”) of identical “prototype-filter” shape shifted to equispaced frequency points by the transform-inherent modulations. Using a DCT leads to a real-valued “cosine-modulated” system. The latter is of interest in the following, especially in the combination of a spectral decomposition followed by a recombination, with some application-specific manipulation in the spectral domain, like noise or echo reduction, or compression (see Fig. 2). Without any intermediate spectral changes, the “analysis-synthesis” filter bank should reproduce the input signal, at least approximately, according to

$$\tilde{x}(k) \approx x(k - k_o), \quad (2)$$

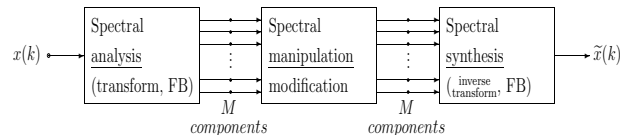


Figure 2: Spectral analysis-synthesis system with internal application-dependent spectral modification

with some delay k_o being allowed. This requires that the M neighbouring channels’ frequency responses add to a constant value ≈ 1 . Furthermore, to keep the data rate and computational load low, the transformation is calculated only once every r sampling intervals, with $r \leq M$, corresponding to a down-sampling of the

narrow-band channel signals; after up-sampling and filtering on the synthesis side, aliasing components may remain. The case $r=M$ is termed “critical sampling”. There are “perfect-reconstruction” (PR) systems and designs [8, 9], with complete alias cancellation and making (2) an equality, but with some restrictions on the prototype shapes. With more freedom, “near-perfect reconstruction” (NPR) designs are found (e.g., [10, 11]), with some deviation of the total system’s frequency response from 1.0 and some aliasing left. A suitable design will keep both errors at some minimum, e.g., in terms of a weighted sum of the mean-square linear distortions and the aliasing power. The critical-sampling NPR case with the iterative design proposed in [12] was used in our work, with equal weights for both errors.

1.4 Fixed-Point Realization

Except for special cases (like very narrow-band filters) the “direct form” is the standard realization of an FIR filter (see Fig. 3): The impulse-response values $h(k)$ are also the realization coefficients.

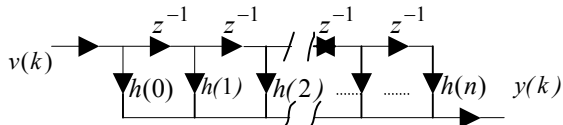


Figure 3: Direct-form (or “2-nd canonical” or “tapped delay-line”) realization of an FIR filter.

There are “PPN+transformation”-type filter banks employing the same structure in their PPN section [13], as to be used here.

2. COEFFICIENT QUANTIZATION

2.1 Possibilities

After the design, all coefficients are given with “infinite” (i.e., computer, floating-point) precision. If a hardware realization of low cost is required, a fixed-point representation of all data is still favourable. Here, we shall deal with the effect of fixed-point filter coefficients only; the treatments of input-signal quantization noise as well as, in the filter-bank case, FFT-coefficient quantization are well-known (see, e.g., [14, 15, 16]). Fixed-point coefficients, be it the set $\{h(k)\}$ or the lifting coefficients, implemented after a normalization to the range ± 1 , by one sign bit and $w_c - 1$ bits “behind the comma”, deviate from the original values. So, the frequency response $H_o(\Omega)$ deviates from its optimized shape, both in a single filter, where the tolerances are violated, and in a filter bank, where linear and aliasing distortions increase, in general. The additional deviation from the actually required ideal behaviour is random-like, for sufficiently large filter orders n , if the coefficients are simply rounded to w_c bits. An analysis of the error statistics lead to statistical estimations of the minimum possible wordlength $w_{c,min}$ and the correspondingly necessary, increased order $n' > n$ for given tolerances [17, 18], as applied in an iterative design of FIR filters with finite-word-length coefficients [19]. An analysis of (linear and aliasing) error bounds for cosine-modulated filter banks was published recently [20]. It is quite clear that an additive *random* error may cause a system to be close to – or far away from the required behaviour. Especially, in [20], a strong variance of the resulting quality was observed (see Fig. 4). So, in many cases, there is a potential of better solutions still with quantized coefficients, either with less deviation with the same w_c and n' , or with the same error at lower word-length and / or order.

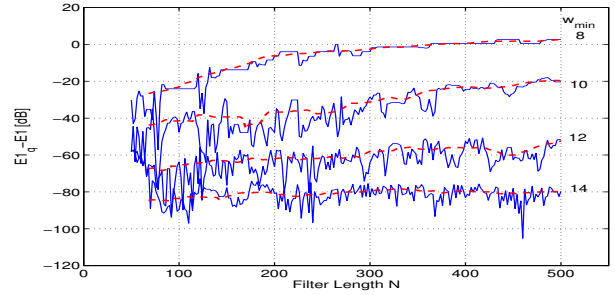


Figure 4: Difference of the mean-square linear distortions E_{1Q} after and E_1 before coefficient quantization for varying filter lengths and word-lengths, in a 16-channel system (from [20]).

2.2. Discrete Optimization

There are various tools for an optimal choice of discrete coefficients for the above-named criteria, e.g., mixed integer-linear programming [21, 22], simulated annealing [23], or genetic algorithms [24]. These methods normally suffer from a high time-complexity [25]. In all cases, a rounded version of a floating-point set would normally be used as a good start. But still, only relatively low filter orders like $n \approx \dots 100\dots$ can be handled. Therefore, various “local-search” [26] and “branch and bound” methods have been proposed, though without a breakthrough, since in the cosine polynomial of (1), for a high order N , the sensitivity of *many* coefficients is equally large at any frequency, due to the cosine factors. But for the single-filter, tolerance-scheme case, a modification of an algorithm due to [27] was applied in [25] for re-optimization successfully, also for long filters (like $n = \dots 1000\dots$). In [27], a fast algorithm is introduced, finding maximum-norm bounded least-squares solutions with infinite-precision coefficients. An equal-ripple solution with unquantized values $h(k)$ is found iteratively from the L_2 -error design with a varying bound on the maximum norm; [25, 28] apply this iteration successively to suitably chosen coefficient subsets during a one-by-one quantization strategy. Results of this approach are included in Tab. 1 and 2, for comparison with our approach to be discussed next.

2. RANDOMIZED STRATEGIES

3.1 Randomized Rounding

A very fast approach to solve large integer-linear programs are randomized algorithms, using structural properties of the underlying linear-programming relaxation (LP), as introduced in the last 15 years. Among them, randomized rounding has been most successful approximating optima of integer-linear programs [29, 30]. We formulate the algorithm for filter design as follows:

For a value range ± 1 , the quantization interval is $Q = 2^{-w_c}$. In naïve rounding, a value $h \in [(i-1/2) \cdot Q, (i+1/2) \cdot Q)$ is rounded to $h_Q \doteq i \cdot Q$ in a “hard decision”. Randomized rounding, instead, takes a “soft decision”: Before applying the above rule, a value $Q \cdot \text{sign}(h - iQ)$ is added to h with probability $p_{add} \doteq |h - iQ| / Q$, while with probability $p_r = 1 - p_{add}$ rounding without addition happens. So, with probability p_{add} , the next larger or smaller quantized value will result, rather than the naively rounded one. A single trial may yield larger errors – but in a large ensemble of randomized designs, also better versions are found, see examples included in Tabs. 1 and 2.

3.2 Randomized Subset Selection

A discrete local-search technique would try to find a “most influential” rounded coefficient $h_Q = i \cdot Q$ and check whether an improvement results with the replacement

$$h_Q = i \cdot Q \rightarrow h'_Q \doteq (i \pm m) \cdot Q, \text{ with } m = 1 \text{ or } m = 2.$$

Alternatively, with all other coefficients kept at their quantized values, this element can be re-optimized continuously, according to the given criterion, and then be rounded again. The latter idea is not directly applicable to a Chebychev design, as the programs of [3, 4] deal with the frequency response, a priori, finding all coefficients together at the end. But the approach of [27] minimizes an MSE, a priori, with a linear constraint concerning the maximum absolute error, which may be done with just one coefficient. This is applied in [25] to re-optimize the shrinking subset of non-quantized coefficients, following a deterministic choice of the elements in each step. Positive results are reported for filter orders up to $n \approx 600 \dots 1000 \dots$

Parameters							
n	50	62	86	106	200	400	600
Ω_p	$\pi/8$	$\pi/2$	$\pi/5$	$\pi/4$.192	.192	.192
Ω_p	$\pi/4$	$3\pi/5$	$\pi/4$	$\pi/3$.208	.208	.208
$R = \delta_1 / \delta_2$	500	500	100	100	1	1	1
$a_s / dB \doteq -20 \lg \delta_2$	82.9	86.3	63.7	96.0	34.3	58.8	82.2
w_c	15	16	12	18	8	12	16
$(a_s)_{naive \text{ round.}}$	80.1	79.2	59.1	90.7	30.7	51.5	74.7
$(a_s)_{rand. \text{ round.}}$	80.3	83.1	59.7	92.7	31.4	53.1	75.5
$(a_s)_{reopt. [25]}$	81.0	82.7	61.4	93.3	32.7	55.5	77.9
$(a_s)_{rand. \text{ select.}}$	81.6	85.5	62.4	94.8	33.2	56.3	78.8

Tab. 1: Parameters and design qualities of several examples.

As observed in [25], deterministic re-optimization techniques suffer from a loss of performance for rounding the last non-quantized coefficients. Rounding a remaining set of coefficients at once can mitigate this effect. At this point, we combine the benefits of re-optimization and randomization in a new strategy: The restriction to just one deterministic solution may again be broken up. After rounding all $(N+1)$ coefficients first, a sub-set $\{h_Q(k)\}$ with $\rho \leq N+1$ elements is dealt with, according to

$$k \in I \doteq \{k_1, k_2, \dots, k_\rho\},$$

with ρ and I being chosen randomly; $\{h_Q(k)\}$ is replaced by a continuously re-optimized set $\{\tilde{h}(k), k \in I\}$ which then undergoes a rounding again. This is done for a suitable number of trials, always continuing with the best solution found so far. So, the approach of [25] is generalized and randomized.

4. RESULTS

4.1 Single-Filter Design

With the same computation time, designs result with a similar quality as in [25]; as the design can now, however, be repeated ad libitum with new, random, solutions, at the cost of computation time, better results are possible. This is demonstrated in Tab. 1: For filters of orders $n = 50 \dots 600$ and varying word-

lengths w_c , equal-ripple designs were found first. The resulting loss of stop-band attenuation after (“naïve”) rounding is obvious, as well as the growing enhancements when randomized rounding, a deterministic re-optimization, and a randomized multi-coefficient re-optimization are applied. The last result, following the proposal of Sec. 3.2, comes quite close to the unquantized solution’s quality.

4.2 Prototype-Filter Design

The same strategies can also be applied to filter banks. An initial (“infinite precision”) solution is found by means of the method in [12]. The objective function

$$E \doteq \alpha \cdot E_1 + (1 - \alpha) \cdot E_2$$

is minimized. Here, the mean-square value of the linear distortion is described by

$$E_1 \doteq \int_{\Omega=0}^{\pi/M} |H_o^2(\Omega) + H_o^2(\Omega - \pi/M) - 1|^2 \cdot d\Omega.$$

Parameters and errors						
M	4	4	8	8	16	16
L	32	48	64	96	128	192
$E \cdot 10^7$	72.1	0.29	38.6	0.16	20.1	0.09
$E_1 \cdot 10^7$	29.8	0.023	11.3	0.017	4.06	0.015
$E_2 \cdot 10^7$	42.2	0.26	27.3	0.14	16.1	0.071
w_c	9	14	9	14	9	13
$E_q \cdot 10^7$ <i>naive round</i>	1280	3.35	598	0.33	800	4.12
$(E_{1q})_{naive \text{ round.}}$	1230	3.08	571	0.18	781	4.04
$(E_{2q})_{naive \text{ round.}}$	50	0.27	27	0.15	18.4	0.08
$E_q \cdot 10^7$ <i>reopt.[25]</i>	154	0.425	60.7	0.195	33.9	0.126
$(E_{1q})_{reopt.}$	99	0.151	31.4	0.050	15.5	0.042
$(E_{2q})_{reopt.}$	55	0.274	29.2	0.145	18.4	0.084
$E_q \cdot 10^7$ <i>rand.sel.</i>	88.4	0.352	58.3	0.170	25.0	0.158
$(E_{1q})_{rand.sel.}$	35.6	0.077	27.6	0.025	6.29	0.076
$(E_{2q})_{rand.sel.}$	52.7	0.275	30.7	0.145	18.7	0.082
$E_q \cdot 10^7$ <i>rand.round</i>	88.4	0.326	52.1	0.168	24.2	0.118
$(E_{1q})_{rand.rd.}$	35.6	0.048	21.8	0.021	4.41	0.026
$(E_{2q})_{rand.rd.}$	52.7	0.278	30.4	0.147	19.8	0.092

Tab. 2: Error figures E, E_1, E_2 before quantization with designs for various channel numbers M and prototype lengths L , and corresponding terms E_q, E_{1q}, E_{2q} after naïve rounding, deterministic and randomized re-optimization, and randomized rounding. All errors are normalized to 10^{-7} .

Here, equal filters in both analysis and synthesis are assumed, and the aliasing power, due to limited stop-band attenuation of the prototype, is denoted by

$$E_1 \doteq \int_{\Omega=\Omega_s}^{\pi} H_o^2(\Omega) \cdot d\Omega.$$

For example, a prototype low-pass filter of order $n=127$ may be wanted for a system with $M=16$ channels. With $\alpha \doteq 0.5$, i.e.,

equal weights for both error types, $E = 2.01 \cdot 10^{-6}$, with $E_1 = 0.40 \cdot 10^{-6}$ and $E_2 = 1.61 \cdot 10^{-6}$ are achieved, prior to quantization. Now, a wordlength of only 9 Bits is required. After naïve rounding, we have a much larger linear distortion described by $E_{1q} = 78.1 \cdot 10^{-6}$, a little more aliasing with $E_{2q} = 1.84 \cdot 10^{-6}$, and a much higher total error figure $E_q = 80.0 \cdot 10^{-6}$. A deterministic re-optimization, with the same word-length, following the approach of [25, 27, 28] as adapted to filter banks, reduces the linear distortion term strongly to $E_{1q} = 1.55 \cdot 10^{-6}$, the total error to $E_q = 3.39 \cdot 10^{-6}$. Introducing a random selection into this algorithm yields a further improvement, with $E_q = 2.50 \cdot 10^{-6}$, which, here, is minimally worse than the randomized-rounding solution, yielding $E_q = 2.42 \cdot 10^{-6}$. Tab. 2 shows this example, with more details, in the fifth column, together with a few other cases.

3. CONCLUSIONS AND FURTHER WORK

Obviously, naïve rounding leads to huge error enlargements. For a single filter, 3...8 dB of the stop-band attenuation are lost (corresponding to a tolerance enlargement by 40...100...%), and in a filter bank, factors up to 40 occur, in terms of the objective function. Deterministic re-optimization leaves losses up to 4.3 dB and factors around 1.5...2.1, respectively, while both randomized strategies are able to reduce this deterioration substantially again: Only 1.3...3.4 dB are still lost, and the remaining error-increase factor lies between 1.05 and 1.35, in our examples.

Also obviously, the randomized-selection strategy is slightly less successful in the filter-bank application than for single filters. A possible reason is seen in our ad-hoc choice of the statistics for the index sub-set selection. Further work is carried out to optimize this choice, also with regard to the computational effort: Presently, we choose our randomized techniques to spend a 50-fold computation time compared to the deterministic algorithm.

REFERENCES

- [1] L.R. Rabiner, J.H. McClellan, T.W. Parks, "FIR Digital Filter Design Techniques Using Weighted Chebyshev Approximation", *Proc. IEEE*, vol. 63, pp. 595-610, April 1975.
- [2] E.Y. Remez, "General Computational Methods of Tchebycheff Approximation", Kiev, At. En. Comm. Translation 4491, pp. 1-85, 1957.
- [3] L.R. Rabiner, J.H. McClellan, T.W. Parks, "FIR Linear-Phase Filter Design Program", in *Programs for Digital Signal Processing*, pp. 5.1.1-5.1.13, IEEE Press, New York, 1979.
- [4] "REMEZ - Compute the Parks-McClellan optimal FIR filter design", in MATLAB Software, Signal-Processing Toolbox.
- [5] O. Herrmann, L.R. Rabiner, D.S.K.Chan, "Practical Design Rules for Optimum FIR Low-Pass Digital Filters", *Bell Syst. Techn. J.*, vol. 52, pp. 769-799, July/Aug. 1973.
- [6] "REMEZORD - Parks-McClellan optimal FIR filter order estimation", in MATLAB Software, Signal-Processing Toolbox.
- [7] U. Heute, "Necessary Degree for Equal-Ripple FIR Filters", *Proc. IASTED Int. Symp. App. Sig. Process. Dig. Filt.*, Paris, pp. 9-12, 1985.
- [8] T. Karp, N.J. Fliege, "MDFT Filter Banks with Perfect Reconstruction", *Proc. IEEE Int. Sympos. Circ. & Syst. (ISCAS) 1995*, 1995, Seattle, USA, April 1995.
- [9] R.D. Koilpillai, P.P. Vaidyanathan, "Cosine-Modulated FIR Filter Banks Satisfying Perfect Reconstruction", *IEEE Trans. Sig. Process.*, vol.SP-40, pp. 770-783, 1992.
- [10] J. Kliewer, "Simplified Design of Linear-Phase Prototype Filters for Modulated Filter Banks", *Proc. EUSIPCO 1996*, Trieste, Italy, pp. 1191-1194, 1996.
- [11] T. Nguyen, "Near-Perfect Reconstruction Pseudo-QMF Banks", *IEEE Trans. Sig. Process.*, vol. 42, pp. 64-76, 1994.
- [12] H. Xu, W.S. Lu, and A. Antoniou, "Efficient Iterative Design Method for Cosine-Modulated QMF Banks", *IEEE Trans. Sig. Process.*, vol. 44, pp. 1657-1668, 1996.
- [13] P. Vary, U. Heute, "A Short-Time Spectrum Analyzer with Polyphase Network and DFT", *Sig. Proc.*, vol. 3, pp. 55-65, 1980.
- [14] S. K. Mitra, *Digital Signal Processing - a Computer-Based Approach*. McGraw Hill, New York 1998.
- [15] U. Heute, "Results of a Deterministic Analysis of FFT Coefficient Errors", *Signal Processing*, vol.4, pp. 321-331, 1981.
- [16] U. Heute, "The Impact of FFT Coefficient Errors on Polyphase Filter Banks", *Signal Processing*, vol. 7, pp. 119-133, 1984.
- [17] U. Heute, "Necessary and Efficient Expenditure for Non-Recursive Digital Filters in Direct Structure", in *Proc. Europ. Conf. Circ. Th. Des. (ECCTD) 1974*, London, UK, Sept. 1974, IEE Publ. 116, pp. 13-19.
- [18] U. Heute, "Recent Developments in Digital FIR Filtering", in *Main Lects, Summer. Symp. Circ. Th.*, Prague, 1977, pp. 14-27.
- [19] U. Heute, "A Subroutine for Finite-Wordlength FIR Filter Design", in *Programs for Digital Signal Processing*, pp. 5.4.1-5.4.20, IEEE Press, New York, 1979.
- [20] T. Rusc, U. Heute, "Nearly-Perfect Reconstruction Filter Banks with Critical Sampling: Achievable Quality with Quantized Coefficients", *Proc. UkrObraz*, Kiev, Ukraine, 2004, pp. 237-240.
- [21] Y. C. Lim, "Efficient Special-Purpose Linear Programming for FIR-Filter Design", *IEEE Trans. Acoust., Speech & Sig. Process.*, vol. ASSP-31, pp. 963-968, 1983.
- [22] H. Qi, U. Heute, "Cascaded FIR Filter with Discrete Coefficients", in *Proc. Eur. Conf. Circ. Th. Des. (ECCTD) 1985*, Istanbul, Turkey, pp. 183-186.
- [23] E. J. Diethorn, D. C. Munson, "Finite-Wordlength FIR Digital-Filter Design Using Simulated Annealing", in *Proc. IEEE Int. Sympos. Circ. & Syst. (ISCAS) 1986*, pp. 217-220.
- [24] M. Haseyama, D. Matsuura, "A Filter-Coefficient Quantization Method with Genetic Algorithm, Including Simulated Annealing", *IEEE Sig. Process. Lett.*, vol. 13, pp. 189-192, 2006.
- [25] G. Evangelista, "Design of Optimum High-Order Finite-Wordlength Digital FIR Filters with Linear Phase", *Signal Processing*, vol. 82, pp. 187-194, 2002.
- [26] D. M. Kodek, "An Algorithm for the Design of Optimal Finite-Wordlength FIR Digital Filters", in *Proc. IEEE Int. Sympos. Circ. & Syst. (ISCAS) 1980*, pp. 304-308.
- [27] J. W. Adams, "FIR Digital Filters with Least-Squares Stopbands Subject to Peak-Gain Constraints", *IEEE Trans. Circ. & Syst.*, vol. CAS-39, pp. 376-388, 1992.
- [28] G. Evangelista, "About the Design of Digital Systems for Asynchronous Data-Rate Conversion", Notes on Signal Processing, no. 1, Ed. H. Goeckler (Diss. Ruhr-Univ. Bochum, in German), Shaker, Aachen, Germany, 2001.
- [29] P. Raghavan, C. D. Thompson, "Randomized Rounding: a Technique for Provable Good Algorithms and Algorithmic Proofs. *Combinatorica*, vol. 7, pp. 365-374, 1988.
- [30] A. Srivastav, P. Stangier, "Algorithmic Chernoff-Hoeffding Inequalities in Integer Programming. *Disc. Appl. Math.*, vol. 57, pp. 255-269, 1995.