COLLOQUE NATIONAL SUR LE TRAITEMENT DU SIGNAL
ET SES APPLICATIONS

NICE du 26 au 30 AVRIL 1977

DEUX PROGRAMMES D'AIDE A LA CONCEPTION DES FILTRES NUMERIQUES OPT-Z (optimisation binaire) et ANA-Z (analyse et simulation).

P. DUHAMEL et P. LE SCAN

THOMSON-CSF/DIS -33, rue de Vouillé- 75015-PARIS-FRANCE Service Recherche et Développement des Applications Scientifiques.

RESUME

Cette communication décrit deux programmes complémentaires d'aide à la réalisation de filtres numériques :

- le premier (OPT-Z) permet d'effectuer le passage d'une fonction de transfert théorique (c'est-à-dire calculée sans tenir compte des contraintes de réalisation) à une fonction de transfert "câblée", c'est-à-dire ayant des coefficients binaires.

Il s'agit donc d'un programme d'optimisation en variables discrètes qui, partant de l'arrondi des coefficients théoriques, recherche une configuration binaire des coefficients du filtre sur un nombre de digits donné, telle que des spécifications arbitraires d'amplitude soient respectées.

Les résultats fournis par ce programme OPT-Z sont comparés à ceux issus d'autres méthodes publiées par ailleurs (3, 5, 7, 17) et montrent la validité de l'approche choisie.

- le second (ANA-Z) permet d'analyser et de simuler toute machine numérique comportant uniquement des additionneurs, des retards et des multiplicateurs. C'est le cas, en particulier, des filtres (4, 20).

Cette approche a été rendue nécessaire par le nombre croissant de structures nouvelles proposées aux concepteurs de filtres (1, 11, 12, 13).

La description du filtre se fait alors élément par élément, en précisant à chaque noeud du circuit la taille du registre correspondant (défini par son bit le plus significatif, ou MSB, et son bit le moins significatif ou LSB).

Ce programme peut également, dans sa version actuelle, effectuer l'analyse en fréquence du filtre considéré.

Ces deux programmes sont destinés à être intégrés à un ensemble d'outils destiné à faciliter la conception des filtres numériques.

SUMMARY

This communication describes two complementary alforithms for computer aided design of digital filters.

The first one, OPT-Z, effects discrete optimization of the coefficients of digital filters implemented as direct, parallel or cascade structures, so that arbitrary magnitude specifications are met. The proposed algorithm is based on several aspects of discrete optimization (e. g. one-variable, two-variable and random search) and on the relation between DC gain and coefficients in a digital filter.

Several examples are provided and the results obtained for them are compared with those given by four other methods recently published (3, 5, 7, 17) and show the effectiveness of the approach.

The second one, ANA-Z, simulates any digital machine solely made up of adders, multipliers, and delays ; which is the case of digital filters (4, 20). This has become necessary due to the growing list of available digital filter topologies (1, 11, 12, 13). The description of the filter is then made element by element, indicating the size of the registers at each node (defined by its most significant bit MSB and its least significant bit LSB). This program at its present stage can also handle sinusoïdal steady-state analysis of the filter.

These two programs form the basis of an entire system for computer aided design of digital filters.

DEUX PROGRAMMES D'AIDE A LA CONCEPTION DES FILTRES NUMERIQUES OPT-Z (optimisation binaire) et ANA-Z (analyse et simulation).

I - INTRODUCTION

Une des premières étapes dans la réalisation d'un filtre numérique, une fois les coefficients de la fonction de transfert calculés, est de déterminer la valeur binaire des coefficients multiplicateurs qui seront effectivement câblés. En effet, la longueur de mots de ces multiplicateurs doit être réduire au maximum pour différentes raisons (coût, vitesse de fonctionnement, influence sur le bruit d'arrondi), bien que cette réduction entraîne des erreurs très importantes dans la réponse du filtre.

Le problème de la réduction de l'erreur sur la réponse en fréquence du filtre due à la quantification des coefficients sur un nombre de digits donnés peut se formuler comme un problème d'optimisation non-linéaire en variables discrètes.

De nombreuses méthodes utilisant des recherches aléatoires ont été proposées (2, 16, 17), ainsi qu'une approche basée sur une optimisation en variables booléennes (14). Plus récemment, des programmes utilisant une approche plus mathématique (méthodes dites "branch and bound") ont été publiées (3, 5). Une autre méthode heuristique a été proposée récemment (10) dans le cas de structures cascades.

Ce problème peut également être formulé d'une manière statistique : trouver des coefficients (obtenus avec une grande précision) tels que la réponse en fréquence du filtre ne dépassera pas des limites données quand les coefficients varient à l'intérieur de leur pas de quantification (6, 7). Néanmoins, une plus grande réduction des longueurs de mots peut encore être obtenue en optimisant les coefficients arrondis dans l'espace discret des paramètres (7).

La première partie de cette communication décrit les caractéristiques principales d'un nouvel algorithme, (programme OPT-Z), dont les résultats sont comparés à ceux fournis par d'autres auteurs (3, 5, 7, 17).

Une autre étape, dans la réalisation des filtres numériques, est de choisir la structure de câblage de celui-ci. En effet, les performances peuvent être gravement perturbées par le bruit de quantification des opérations internes au filtre, bruit qui varie énormément d'une structure de câblage à l'autre. Un nombre très important de telles structures ont été publiées (1, 11, 12, 13), puis, des auteurs ont préconisé des structures calquées sur celles des filtres passifs, en vue d'obtenir un bruit d'arrondi plus faible. Plus récemment (18, 19), sont apparues des structures minimisant le bruit de quantification.

On voit que le choix qui s'offre à l'utilisation est assez considérable. Il nous est donc apparu utile de pouvoir simuler des filtres numériques en décrivant leur structure, élément par élément, afin de s'affranchir de structures figées, et d'éviter une nouvelle programmation lors de l'apparition de nouveaux schémas (4 , 20).

La deuxième partie de cette communication décrit donc un programme (ANA-Z) permettant de simuler et d'analyser toute machine numérique comportant uniquement des additionneurs, des retards, et des multiplicateurs.

On discute ensuite des extensions possibles de ces deux programmes pour une aide efficace à la réalisation de filtres numériques.

II - OPTIMISATION BINAIRE (programme OPTZ).

A) Position du problème

Soit, par exemple, un filtre réalisé sous forme cascade (l'algorithme peut également traiter les formes directes et parallèles) ayant pour fonction de transfert l'expression suivante :

$$H(Z) = A \prod_{i=1}^{K} \frac{1 + a_i z^{-1} + b_i z^{-2}}{1 + c_i z^{-1} + d_i z^{-2}}$$

Il s'agit donc de déterminer les coefficients a_i, b_i, c_i et d_i de chaque cellule élémentaire soumis aux contraintes suivantes :

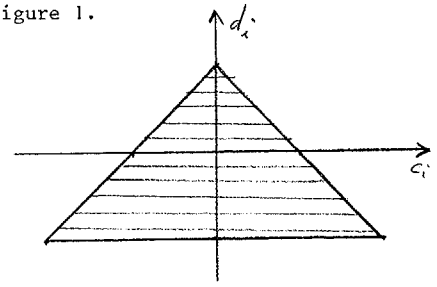
- $a_i . 2^{Q-Q_M}$, $b_i . 2^{Q-Q_M}$, $c_i . 2^{Q-Q_M}$, $d_i . 2^{Q-Q_M}$.

sont des nombres entiers dans lesquels 2^{Q_M} est la puissance de deux immédiatement supérieure à la valeur théorique du coefficient et Q le nombre de bits du registre correspondant.

- $G_m(F_j) \leqslant |H(f_j)| \leqslant G_M(f_j)$, j=1,... N.

dans laquelle G_m et G_M représentent les valeurs minimale et maximale d'amplitude à respecter

- le filtre obtenu doit être stable : le point (c_i, d_i) doit être dans la partie hachurée de la figure 1.



B) Description de l'algorithme

L'idée de base a été d'utiliser une méthode d'optimisation à pas d'exploration fixe ou multiple d'un pas donné (correspondant ici à la variation du digit de poids le plus faible sur chacun des coefficients).

Ceci nous a conduit à développer deux versions de ce programme, l'une autour d'une méthode de pas à pas, l'autre autour d'une méthode de Hookes-Jeeves (22).
Ces deux méthodes d'optimisation ont été adaptées au problème et modifiées selon les quatre points suivants:

- exploration simultanée sur deux composantes :

  la fonction d'erreur pouvant présenter de grandes variations autour d'un point pour un pas d'exploration faible, il paraît intéressant, lorsque la recherche en pas à pas n'améliore pas la fonction d'erreur , d'explorer l'espace discret des paramètres en modifiant deux coefficients en même temps ce qui équivaut à se déplacer en diagonale dans un plan (16).

- exploration aléatoire :

  une fonction de variables discrètes peut difficilement être considérée comme unimodale dans la région du minimum, surtout lorsque le nombre de digits devient faible (17) même si la fonction continue correspondante a de bonnes propriétés.

DEUX PROGRAMMES D'AIDE A LA CONCEPTION DES FILTRES NUMERIQUES OPT-Z
(optimisation binaire) ET ANA-A (analyse et simulation)

Ainsi lorsque la méthode de pas à pas (ou celle de Hooke et Jeeves) a amélioré la fonction d'erreur un déplacement est effectué dans la direction supposée du minimum à une distance qui est un multiple aléatoire du pas. Ceci permet de rendre l'algorithme moins sensible aux irrégularités de l'allure de la fonction d'erreur.

- Réponse fréquentielle et gain en continu: lorsque l'on fait varier un ou plusieurs des coefficients d'un filtre numérique, le gain en continu de celui-ci est modifié et la nouvelle valeur du gain peut ne plus être optimale vis-à-vis du gabarit. Ainsi à chaque modification d'un coefficient, l'allure de la courbe est translatée (gain en dB) ou dilatée (amplitude en linéaire) de manière à minimiser l'erreur (sans nouvelle analyse du filtre). Ceci est important dans la mesure où dans beaucoup de problèmes le gain en continu n'est pas quantifié.

- Nouvelles valeurs initiales: lorsque chacune des étapes de l'algorithme correspondant aux remarques précédentes n'ont pas amélioré la fonction d'erreur, un nouveau point de départ est recherché aléatoirement.

Une autre solution du problème posé peut être trouvée en modifiant les paramètres suivants :

* la puissance P à laquelle est élevée l'erreur dans le critère : $F = \sqrt[P]{\sum |\varepsilon_i|^P}$
* l'ordre dans lequel les coefficients sont optimisés (les meilleurs résultats sont généralement obtenus pour un ordre de sensibilité décroissant).
* le nombre maximum d'incréments aléatoire dans une direction.

Exemples d'application.

a. Filtre passe bas d'ordre 2
Fréquence d'échantillonnage : 10 kHz.

La constante A est quantifiée et il n'y a pas plus de 6 digits à droite de la virgule.

Spécifications d'amplitude (en gain linéaire) :
f = 0.900 (100), H(f) =1;
f = 1 000, H(f) =$1/\sqrt{2}$;
f = 1 200, H(f) =0;
f = 1 500, 5 000, (500), H(f) =0.

On a utilisé le point de départ de Suk et Mitra [17] ; c'est-à-dire :

A = 1, a = 0, b = 1,
c = 1, d = 5.

Paramètre	Suk-Mitra [17]	Bandler-Bardakjian-Chen [3]	Charalambous [5]	OPT-Z Version 1	Version 2
f d'erreur	.31535	.29059	idem	idem	idem
Nombre d'appels	139	574 (terminé à 1 030)	189 + 85 gradient (terminé à 2 502 gradient + 1 216	951 + 1 146 translations (terminé à 1 172 + 1 309 translations	607 + 759 translations (terminé à 889 + 1 101 translations

b. Un différentiateur d'ordre 2 décrit par Steiglitz [15]
Coefficients idéaux : A = .366;
a = -.32917379;
b = -.6708261;
c = .8593897;
d = .10210106.

Les fréquences f, sont : 0.. .05, .1,-----.95,1.
En fractions de la demi-fréquence d'échantillonnage.

Paramètre	Suk-Mitra [17]	Charalambous [3]	OPTZ Version 1	OPTZ Version 2
Fonction d'erreur	.64305 10^{-3}	.52233 10^{-3}	.52796 10^{-3}	.52233 10^{-3}
Nombre d'appels	72	1 880+1 007 gradient (terminé à 4 115+2 278 gradient)	145 + 23 dilatations (terminé à 509 + 39 dilatations)	618 + 87 dilations (terminé à 916 + 103 dilatations)

c. Filtre passe-bas elliptique réalisé en cascade. K = 4. (7)
Spécification en fréquence : .2dB de 0 à 1 000 Hz, 60dB de 1 200 à 8 000 Hz.
Fréquence d'échantillonnage : 16 000 Hz
Ce filtre sera codé avec 8 bits au plus à droite de la virgule, la constante n'étant pas quantifiée.

L'optimisation sera menée sur les a; c; d; et A jusqu'à l'obtention d'une solution qui passe dans le gabarit. (La fonction d'erreur finale est donc nulle dans les résultats suivants.)

Paramètre	Point de départ	A. Dubois 7	OPT-Z (version II)
a1.....	-1.777624	-1.77734375	-1.77734375
a2.....	-1.731973	-1.73046875	-1.7265625
a3.....	-1.536417	-1.53515625	-1.53125
a4.....	-.099236	-.09765625	-.09375
c1.....	-1.814702	-1.81640625	-1.81640625
c2.....	-1.767588	-1.765625	-1.76953125
c3.....	-1.70349	-1.6953125	-1.703125
c4.....	-1.640769	-1.63671875	-1.66015625
d1.....	.972861	.97265625	.97265625
d2.....	.902705	.8984375	.90234375
d3.....	.793502	.78515625	.79296875
d4.....	.683163	.679875	.69921875
Ondulation dans la bande passante		.14	.2
Nombre d'appels à la fonction d'erreurs...		?	159 + 851 translations

En diminuant l'ondulation maximum tolérée dans la bande passante, OPT-Z a donné les résultats suivants :

i	a_i	b_i	c_i	d_i
1.....	-1.77734375	1.	-1.81640625	.97265625
2.....	-1.7265625	1.	-1.76953125	.90234375
3.....	-1.53515625	1.	-1.70703125	.79296875
4.....	-.09765625	1.	-1.6328125	.67578125
Ondulations dans la bande : .1,6 Nombre final d'appels à la fonction : 1044, + 952 translations.				

Il apparait que la version II d'OPT-Z semble plus fiable que la première (elle mène au minimum d'une manière plus sûre, bien que menant parfois plus lentement à la solution).

Les comparaisons faites avec quatre autres méthodes (17 , 5 , 3 , 7) montrent que les résultats d'OPT-Z sont, au moins sur les exemples traités, sensiblement meilleurs que ceux fournis par d'autres méthodes heuristiques, et très semblables à ceux d'autres méthodes plus mathématiques. Un dernier exemple montre l'avantage d'une telle approche : un premier résultat peut être trouvé en peu de temps puis on peut rechercher d'autres solutions en changeant certaines données influant sur la marche de l'algorithme.

DEUX PROGRAMMES D'AIDE A LA CONCEPTION DES FILTRES NUMERIQUES OPT-Z
(optimisation binaire) et ANA-Z (analyse et simulation)

III - Analyse et simulation.

A- Position du problème

Il s'agit de pouvoir analyser (en fréquence) et simuler un circuit décrit élément par élément, en tenant compte des contraintes de la machine numérique qui effectuera l'opération de filtrage (nombre de digits pris en compte pour le calcul des opérations intermédiaires, type d'arithmétique utilisé, caractéristiques d'overflow).

Le but d'un tel programme est d'essayer d'atteindre les caractéristiques fréquentielles et temporelles d'un filtre numérique, en incluant dans ces dernières les problèmes posés par la réalisation proprement dite du filtre en arithmétique fixe.

B- Simulation

Cette partie du programme a été traitée d'une manière la plus générale possible, afin de pouvoir simuler le fonctionnement des dispositifs numériques les plus divers :

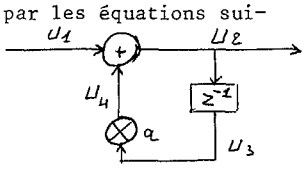
A chaque numéro de noeud rencontré dans la description du filtre, on associe un registre binaire (défini par LSB et MSB). En effet, quelle que soit la machine utilisée, le résultat d'une opération intermédiaire à l'intérieur du filtre devra être stockée à un moment ou à un autre. L'utilisateur est donc ainsi maître de la précision de chacune des variables intermédiaires.
Puis, la simulation est effectuée, séquentiellement (élément par élément) dans un ordre fixé par l'utilisateur comme dans le cas d'une machine numérique ne comportant qu'une unité arithmétique, avec la précision spécifiée en chacun des registres.

Les dépassements de capacités sont surveillés à tous les niveaux, et signalés par un message en cours de fonctionnement du programme. En effet, ils n'apportent pas tous des perturbations dans le fonctionnement du filtre : lors d'une série d'additions, en particulier, si le résultat final n'est pas en dépassement de capacité sur le registre considéré, peu importe qu'il y en ait eu lors des additions intermédiaires, le résultat sera cependant exact (arithmétique en complément à 2).

Le programme permet également l'initialisation des différents registres du circuit, ce qui permet d'effectuer une recherche des cycles limites dont le filtre peut être l'objet [20] .

C- Analyse en fréquence

Le filtre dont le schéma est représenté fig. ci-dessous peut se décrire par les équations suivantes :



$$U1 = x$$
$$U2 = U4 + U1$$
$$U3 = 3^{-1}\ U2$$
$$U4 = a\ U3$$

soit, en notation matricielle :

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & -1 \\ 0 & -2^{1} & 1 & 0 \\ 0 & 0 & -a & 1 \end{bmatrix} \begin{bmatrix} U1 \\ U2 \\ U3 \\ U4 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \times$$

il suffit donc de résoudre ce système d'équations en différentes fréquences (c. a. d. pour différentes valeurs de z^{-1}) pour observer le comportement fréquentiel du filtre entre l'entrée et n'importe quelle sortie.

Dans la pratique, cette analyse est effectuée suivant la méthode décrite dans [21], qui permet d'obtenir l'analyse fréquentielle entre 2 noeuds quelconques du circuit sans effectuer plusieurs résolutions.

IV - Conclusion.

Ces deux programmes constituent l'amorce d'un ensemble permettant de faciliter la mise en oeuvre de filtres numériques. Pour obtenir cet ensemble, les deux programmes précédents sont développés selon plusieurs axes :

. Calcul automatique des mises à l'échelle des variables internes au filtre en vue d'éviter les dépassements de capacité des registres correspondants, ceci en fonction des caractéristiques du signal d'entrée.

. Approximation la plus précise possible des caractéristiques temporelles et fréquentielles du bruit de quantification d'un filtre à structure quelconque.

. Caractérisation des différentes structures possibles de câblage d'un filtre ayant une fonction de transfert donnée (en fonction du bruit d'arrondi) en vue de l'obtention du meilleur compromis complexité ↔ performance pour une application donnée.

REFERENCES :

1 - R C AGARWAL and C.S. BURRUS : New recursive digital filter structures having very low sensitivity and roundoff noise, IEEE Transactions, Vol. CAS-22, n°12, décembre 1975.

2 - E. AVENHAUS : On the design of digital filters with coefficients of limited word length IEEE Trans.on Audio Electroacoustics, vol.AU-20 n°3, pp.206-212, August 1972.

3 - J.W. BANDLER, B.L. BARDAKJIAN, and J.H.K. CHEN Design of recursive digital filters with optimized word length coefficients : 8th Annual Princeton Conference on Information Sciences and systems, 28-29 March 1974.

4 - S.E. BELTER and S.C. BASS : Computer aided analysis and design of digital filters with arbitrary topology IEEE transactions, vol. CAS-22, n°10, octobre 1975

5 - C. CHARALAMBOUS and M.J. BEST : Optimization of recursive digital filters with finite word lengths IEEE Trans. on Acoustics, Speech and Signal Processing, vol.ASSP-22 n°6, pp.424-431, décember 1972.

6 - R.E. CROCHIERE : A new statistical approach to the coefficient word-length problem for digital filters IEEE Trans. On circuits and systems, vol. CAS-22, n°3, pp. 190-196, March 1975.

7 - H. DUBOIS : An algorithm for recursive digital filter approximation with quantized coefficient . Proceedings of the Florence Conference on Digital Signal Processing, 11-13 September 1975.

8 - P. DUHAMEL : Un algorithme pour la minimisation des effets de la quantification des coefficients d'un filtre numérique récursif (programme OPT-Z).
Revue technique THOMSON-CSF, Vol.8, n°2, juin 1976.

9 - P. DUHAMEL : An algorithm for the design of digital filters with finite word-length coefficients' to be presented at the 1977 IEEE International Conference on Acoustics, Speech and Signal Processing.

10 - A. HADJIFOTIOU and D.G. APPLEBY : Design of digital filters with severely quantized coefficients . The radio and Electronic Engineer, vol. 46, N°1, pp. 23-28, January 1976.

11 - S.K. MITRA and R. SHERWOOD, Canonic realizations of digital filters using the continued fraction expansion, IEEE Trans. Audio Electroacoust., vol. AU-20, pp.185-194, Août 1972.

12 - Digital ladder networks, IEEE Trans. Audio Electroacoust., vol AU-21, pp. 30-36,Fév.1973

13 - S. PARKER and S. HESS, Canonic realizations of second-order digital filters due to finite precision arithmetic, IEEE Trans. Circuit Theory vol. CT-19, pp. 410-413, Juillet 1972.

14 - W.H. STORZBACH : On the design of recursive digital filters with minimum coefficient word length. IEEE 1972 International Symposium on circuit Theory pp. 279-282.

15 - K. STEIGLITZ : Computer aided design of recursive digital filters , IEEE Trans. on Audio and Electroacoustics (special issue on digital filtering), vol. AU-18, pp. 123-129, Juin 1970.

16 - K. STEIGLITZ : Designing short-word recursive digital filters . Proceedings of the 9th Annual Allerton Conference on Circuit and System Theory, pp. 778-788, Octobre 1971.

17 - M. SUK and S.K. MITRA Computer-aided design of digital filters with finite word-lengths . IEEE Trans. on Audio and Electroacoustics, vol.AU-20 n° 5, pp. 356-363, décembre 1972.

18 - S.Y. HWANG : Minimum unit noise in the state space digital filtering. Proceedings of the 1976 IEEE International Symposium on Circuits and Systems pp. 352-355 Avril 1976.

19 - C.T. MULLIS and R.A. ROBERTS : Synthesis of minimum round noise off fixed point digital filters : IEEE transactions on circuits and Systems vol. CAS-23 n° 9, pp.551-562, Septembre 1976.

20 - A. LACROIX : Error Estimation of digital Filters with Arbitrary Structure and Arithmetic by Simulation. Proc. IEEE Int. Conf. on Acoustics, speech and Signal Processing, Philadelphia 1976, p. 533-536.

21 - S.S. LAWSON and A.G. CONSTANTINIDES : on the efficient analysis of digital filter Structures in the frequency domain.
Proceedings of the florence Conference on digital signal processing, Septembre 11-13, 1975.

22 - R. HOOKE et J.A. JEEVES : Direct search solution of numerical and statistical problem.(J.A.C.M., vol. 8 n° 2, avril 1961, P. 212-229).