N° Ordre : 10 / 2001 – M / EL

République Algérienne Démocratique et Populaire
Ministère de l'enseignement supérieur et de la recherche scientifique

Université des Sciences et la Technologie Houari BOUMEDIENE

جامعة هواري بومدين
للعلوم والتكنولوجيا
U S T H B

THESE

Présentée Par :

Lynda BOUCHEMAKH

POUR L'OBTENTION DU GRADE DE

MAGISTER EN ELECTRONIQUE

Option : Instrumentation et génie des systèmes

THEME

Synthèse de structures d'état efficaces et moins perturbées pour les filtres numériques récursifs

Soutenue le 17/10/2001 devant le jury :

M^me^ A. BELHADJ-AISSA	Maître de conférences (USTHB)	Présidente
M^me^ R. RAMDANI	Maître de Conférences (USTHB)	Directeur de thèse
M^r^ A. HOUACINE	Maître de conférences (USTHB)	Examinateur
M^r^ Y. SMARA	Maître de conférences (USTHB)	Examinateur

Dédicaces

A mes très chers parents pour leur amour, leurs sacrifices et leur compréhension sans limites.

A mes chères sœurs Hasna et Isma,

A mes chers frères Mohamed, Tarik et Adel,

A mon beau-frère Abdelkrim,

A mes petits neveux Raouf et Amine,

A ma jolie petite nièce Meriem,

A tous mes amis en particulier :

SAÏDA DOURIDJ, ZINEB OUDJEBOUR, SIHEM AÏT-DAOUD,SAMIA HAMITOUCHE, ASSIA BOUCIF, DALILA AKROUR,BACHIR LEHOUIDJ, CHERIFA BENNAKI, MOHAMED GUERMAZ, et ZOLA ALLAM.

A toute l'équipe de la première promotion PG Instrumentation et génie des systèmes de l'institut des techniciens supérieurs de l'USTHB.

A tous ceux qui me sont chers,

Je dédie ce travail

Lynda Bouchemakh

Remerciements

Je tiens à exprimer mes vifs remerciements à Madame Rabéa RAMDANI, mon Directeur de Thèse et Maître de conférences à l'ITS, pour son suivi continuel, pour les conseils qu'elle m'a prodigués ainsi que pour la patience dont elle a fait preuve tout au long de ces années.

Je remercie Monsieur Belkacem DERRAS, Maître de conférences à l'école polytechnique d'Alger pour m'avoir proposé un sujet de Recherche aussi passionnant.

Je remercie Madame A. BELHADJ-AISSA, Maître de conférences à la faculté de génie électrique pour l'honneur qu'elle me fait en présidant le Jury de cette Thèse.

Que Messieurs A. HOUACINE et Y. SMARA, Maîtres de Conférences à la faculté de génie électrique trouvent ici mes sincères remerciements pour avoir accepté aimablement d'examiner le travail et d'avoir bien voulu faire partie du Jury.

Je tiens à exprimer ma profonde gratitude à toute l'équipe d'enseignants de l'ITS.
Je cite tout particulièrement :

M^{r} F. BOUCHAFAA, M^{elle} A. BOUKHELIFA, Mr B. LEHOUIDJ
M^{r} H. MOULAI, M^{r} A. NACER, M^{r} R. OUSSAID, M^{r} S. TAHI,
M^{r} H. TEFFAHI, Mr ATTIF, et M^{r} A. TALHA

pour leur conseils, leur amabilité et leur entière disponibilité.

Je voudrais également exprimer toute ma reconnaissance et mes plus vifs remerciements à mes meilleures amies : M^{elle} M. DOURIDJ , M^{me} Z. OUDJEBOUR, et M^{elle} A. BOUCIF pour toute l'aide qu'elles m'ont prodiguée.

Que M^{elle} SEBAA Ouahiba, M^{rs}Kamel et Toufik du centre de calcul (Rectorat) trouvent ici mes sincères remerciements pour leur aide et leur patience

Je tiens à exprimer mes plus vifs remerciements à tous ceux qui ont contribué, directement ou indirectement à la réalisation de la présente Thèse.

S O M M A I R E

INTRODUCTION GENERALE

INTRODUCTION GENERALE

Introduction:

Il y a quelques années, la numérisation des filtres analogiques était seulement un moyen pour simuler leur comportement sur ordinateur, dans le but de tester leurs performances et de corriger leurs imperfections avant leurs réalisations analogiques définitives.

De nos jours le grand développement de la technologie des circuits intégrés numériques offre avec la complexité de l'arsenal mathématique du traitement numérique des signaux, la possibilité de réaliser et d'utiliser des filtres proprement numériques. Ces derniers se distinguent par leur grande fiabilité, par leur vitesse de traitement et par leur coût réduit.

La simulation de ces filtres avant leur implantation permet de résoudre les problèmes que peut poser leur réalisation matérielle ; en plus des problèmes classiques de l'échantillonnage, la représentation arithmétique,… il y a le problème de la précision avec laquelle doit s'opérer le filtrage du fait de l'utilisation de registres de tailles réelles et donc finies pour représenter les valeurs des paramètres des filtres et des signaux qui les parcourent ainsi que les valeurs résultant des diverses opérations mathématiques.

L'impact de la limitation de la longueur des mots sur la qualité du filtrage, et les effets néfastes qu'elle cause ont été la source de nombreuses investigations qui ont permis de prévaloir l'efficacité de la représentation des filtres numériques dans l'espace d'état [1]-[8].

Plusieurs recherches ont montré que l'implémentation d'un filtre numérique RII avec sa structure de forme directe donne une grande erreur d'arrondi. A cause de ce phénomène indésirable, l'implémentation d'une telle structure est toujours à éviter, quoique celle-ci représente une structure simple avec son nombre minimal de multiplications : (2N+1).

Nous savons que la fonction de transfert d'un filtre numérique récursif RII peut être représentée par un nombre infini de réalisations d'espace d'état. Cette flexibilité a été

exploitée par MULLIS & ROBERTS [9] et HWANG [10] pour trouver une réalisation d'état à bruit minimum.

Dans notre travail, nous proposons une nouvelle structure : **la tridiagonale globale** et nous montrons qu'elle représente un compromis entre la structure de forme directe et la structure à bruit minimum. Nous verrons qu'en plus de son gain de bruit d'arrondi acceptable et sa simplicité, la structure tridiagonale peut être considérée comme une structure modulaire car formée par des cellules régulières.

Organisation de la thèse:

Au Chapitre I, nous donnerons un aperçu général sur les filtres numériques récursifs représentés dans l'espace d'état ainsi que les différents problèmes rencontrés lors de la synthèse des filtres numériques et les erreurs d'arrondi dues à la longueur finie des mots binaires. Nous étudierons au chapitre II deux structures d'état : la structure canonique et la structure à gain de bruit minimal. Puis au chapitre III, nous présenterons une nouvelle structure : la structure d'état tridiagonale globale calculée soit par l'algorithme de Lanczos, soit par les transformations élémentaires. Différents exemples de simulations suivis par une interprétation des résultats seront donnés à la suite pour comparer les trois structures d'état étudiées. Nous clôturerons notre travail par une conclusion générale.

CHAPITRE I :

PROBLEMES RENCONTRES DANS LA REALISATION DES FILTRES NUMERIQUES

CHAPITRE I

PROBLEMES RENCONTRES DANS LA REALISATION DES FILTRES NUMERIQUES

I-1: Introduction:

Un dispositif de traitement du signal a pour fonction d'opérer sur un signal dans le but de satisfaire certaines exigences techniques (filtrage, modulation, analyse spectrale, modélisation …). L'opération de filtrage consiste à séparer les composantes d'un signal pour isoler les composantes utiles, et/ou pour affaiblir autant que possible les composantes parasites.

Le filtrage numérique est l'étude du filtrage des signaux mis sous forme numérique. Cette option se justifie par le développement actuel des circuits intégrés complexes (circuits VLSI), en particulier par celui des processeurs numériques programmables spécialisés dans le traitement du signal. Les applications sont très nombreuses en télécommunication, en technique audio, en génie biomédical, etc.

Après quelques concepts généraux sur les filtres numériques, nous développerons l'aspect performance des filtres et l'influence sur cette performance de la longueur finie des registres.

I-2: Concepts généraux et définitions

I-2-1 : Signal numérique :*[12]*

Un signal (à temps) discret résulte de l'échantillonnage périodique d'un signal (à temps) continu, ce que l'on note :

$$u(t) \rightarrow u(kT_e) \quad \text{(I.1)}$$

où T_e : période d'échantillonnage ; $fe=1/T_e$ désigne la fréquence d'échantillonnage.

Le traitement numérique d'un signal implique une représentation appropriée en virgule fixe ou en virgule flottante. Dans ce cas le signal à temps discret devient un signal numérique. Enfin pour alléger les notations, on admet que la période d'échantillonnage T_e est choisie égale à l'unité ; Un signal numérique sera donc noté simplement : u(k).

I-2-2 : Systèmes linéaires discrets invariants dans le temps (LIT) : [11],[12]

Les systèmes linéaires discrets invariants dans le temps (LIT) constituent un domaine très important de traitement numérique des signaux, qui est celui des filtres numériques à caractéristiques fixes. Ces systèmes se caractérisent par le fait que leur fonctionnement est régi par une équation de convolution. L'analyse de leur propriétés se fait à l'aide de la transformée en Z qui joue pour les systèmes discrets le même rôle que la transformée de Laplace ou la transformée de Fourrier pour les systèmes continus.

- Un système ***discret*** est un système qui convertit une suite de données discrètes d'entrée *u(k)* en une suite de données discrètes de sortie y(k*)*.

- Il est ***linéaire*** si la suite $Au_1(k)+ Bu_2(k)$ est convertie en la suite $Ay_1(k)+ By_2(k)$, où A et B sont des constantes.

- Il est ***invariant dans le temps*** si la suite $u(k-k_0)$ est convertie en la suite $y(k-k_0)$ $\forall$ k_0 entier

L'équation de convolution qui caractérise le système linéaire invariant dans le temps est :

$$y(k) = \sum_{i=-\infty}^{+\infty} u(i)h(k-i) \quad = \sum_{i=-\infty}^{+\infty} h(i)u(k-i) \quad = h(k) * u(k) \qquad \text{(I.2)}$$

où h(k) est la réponse impulsionnelle du système.

- Le système est dit ***causal*** si la sortie y(k) à $k=k_0$ ne dépend que des entrées pour $k \le k_0$ c'est à dire

$$h(k) = 0 \quad \text{pour} \quad k < 0$$
$$\text{et} \quad y(k) = \sum_{i=0}^{\infty} h(i)u(k-i) \qquad \text{(I.3)}$$

- Le système est dit ***stable*** si à toute entrée d'amplitude bornée lui correspond une sortie bornée :

$$\sum_{k} |h(k)| < \infty$$
$$et \quad h(k) = a^k \quad (k \geq 0) \text{ est stable pour } |a| < 1 \qquad \text{(I.4)}$$

Les systèmes discrets LIT sont composés des trois éléments de base suivants : (Figure I.1)

a) l'additionneur : dans lequel deux signaux d'entrée sont additionnés pour donner un signal de sortie $y(k)= u_1(k)+u_2(k)$.
b) Le multiplicateur, dans lequel un signal est multiplié par une constante : $y(k)= Au(k)$.
c) La cellule mémoire ou élément retard dans laquelle un signal u(k) est retardé d'un intervalle d'échantillonnage : $y(k)= u(k-1)$.

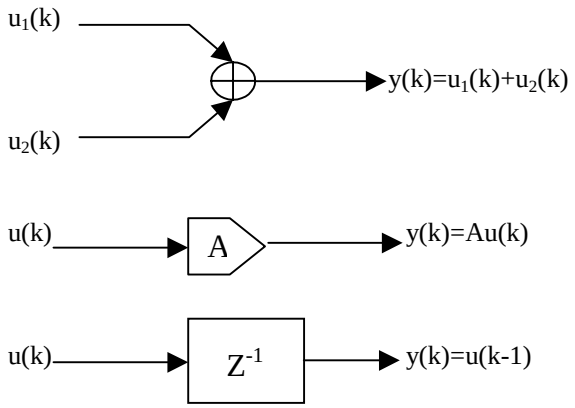


Figure I.1: Les trois éléments de base des systèmes LIT.

I.2.3 - Filtres numériques :[13]

On distingue 2 grandes classes de filtres numériques : les filtres à réponse impulsionnelle finie (R I F) appelés également filtres non récursifs et les filtres numériques à

réponse impulsionnelle infinie (R I I) ou filtres récursifs. Leurs principales propriétés sont données dans le tableau (I.1).[11],[20]

	Filtres R I F	**Filtres R I I**
Fonction de transfert	Possède uniquement des zéros	Possède des zéros et des pôles
Réponse en fréquence	Les méthodes de conception normales sont adaptées aux réponses en fréquence arbitraires ; par exemple aux filtres avec plusieurs bandes passantes, aux différentiateurs et aux filtres de caractéristique de fréquence spécifique dans la bande de transition.	Les méthodes de conception sont en général adaptées à la réalisation des filtre passe-bas, passe-haut, passe-bande et coupe-bande.
Stabilité	Filtre toujours stable	Filtre instable s'il y a des pôles à l'extérieur du cercle unité.
Sensibilité au bruit	L'état initial des cellules mémoires et tout signal interférant de courte durée peuvent affecter le signal sur une durée égale à la longueur de la réponse impulsionnelle.	L'état initial des cellules mémoires et tout signal interférant de courte durée peuvent affecter le signal de sortie sur une durée infinie.
Quantification	Les erreurs de calculs dues à la longueur finie des registres ne sont pas cumulatives et sont simples à analyser.	Les erreurs de calculs dues à la longueur finie des registres sont cumulatives. Les effets de la quantification, peuvent rendre le système instable, et peuvent aussi conduire à des oscillations indésirables telles que les cycles limites et les oscillations de dépassement.

Tableau I.1 : Propriétés des filtres R I F et R I I

I-3 :Effets de la longueur finie des registres dans les filtres numériques représentés en virgule fixe :

I-3-1 : Performance d'un filtre :

Lors de la conception de filtres discrets, on trouve généralement les valeurs des coefficients des filtres avec une très grande précision. Cependant dans la pratique, les valeurs sont stockées dans des registres de longueur finie (B bits).

Pour obtenir un filtre numérique réalisable dans lequel les coefficients ont une longueur de mots finie, il faut quantifier ces valeurs. Mais en faisant cela, on modifie la réponse fréquentielle, c'est à dire la position des pôles et des zéros du filtre.

Ces modifications peuvent être très importantes. Il peut arriver qu'après quantification, le filtre ne satisfasse plus les spécifications sur lequel le calcul des coefficients non quantifiés était basé. Dans des cas extrêmes, un filtre stable peut même devenir instable, si un pôle se décale d'une position à l'intérieur du cercle unité du plan Z vers une position à l'extérieur de celui-ci. [11],[14]

L'utilisation de registres de longueur limitée nous oblige à accorder une grande attention aux effets qu'elle occasionne au filtrage ; on peut citer :

1- Le bruit de la conversion analogique–numérique ou bruit de quantification.
2- Le bruit de calcul non corrélé.
3- L'inexactitude de la réponse du filtre due à la quantification des coefficients : c'est la sensibilité du filtre aux coefficients.
4- Les cycles limites dus aux effets non linéaires de la quantification du signal d'entrée.

Ces effets qui affectent la performance du filtre dépendent essentiellement :

- Du type d'arithmétique utilisée dans l'algorithme du filtre (arithmétique en virgule fixe ou en virgule flottante).
- Du type de la quantification utilisée pour réduire les mots à la longueur désirée (arrondi ou troncature).
- De la structure utilisée du filtre choisi.

I-3-2 : Représentation des nombres : [12]

Comme nous l'avons vu précédemment, dans la représentation numérique des signaux, la longueur des mots binaires est minimisée et cela pour des raisons de coût et de vitesse de calcul. Jusqu'à un certain temps, l'état de la technologie a privilégié pour les variables comme pour les coefficients, la représentation en virgule fixe qui simplifie beaucoup la réalisation des opérateurs arithmétiques [12]. Dans toute la suite de la thèse, c'est ce mode de représentation qui sera utilisé sauf indication contraire de notre part. Toutefois certains processeurs (Texas Instruments par exemple) utilisant une arithmétique en virgule flottante sont disponibles .

I-3-2-1 : Représentation en virgule fixe :[13],[16],[17]

Dans cette représentation, un nombre V fractionnaire et signé peut s'écrire sous 3 formes :

a) : En Signe et valeur absolue :

$$V = \begin{cases} 0 \bullet b_{-1}b_{-2}.....b_{-m} & \text{si } V \geq 0 \\ 1 \bullet b_{-1}b_{-2}.....b_{-m} & \text{si } V < 0 \end{cases} \quad ; \qquad \text{(I-5)}$$

$b_i = 0$ ou 1, i allant de 1 à m

b) : En complément à 1 :

$$V = \begin{cases} V & \text{si } V \geq 0 \\ 2 - 2^{-B} - |V| & \text{si } V < 0 \end{cases} \quad ; \qquad \text{(I-6)}$$

où B est la longueur du mot (Nombre de bits).

c) : En complément à 2 :

$$V = \begin{cases} V & \text{si } V \geq 0 \\ 2 - |V| & \text{si } V < 0 \end{cases} \qquad \text{(I-7)}$$

Parmi ces trois représentations, la première et la troisième sont les plus largement utilisées. L'opération d'addition (et de soustraction) est directe dans le cas du complément à 2.

La représentation en signe et valeur absolue convient mieux à la multiplication car elle se fait simplement par la multiplication bit par bit des valeurs absolues et par l'ajustement du bit de signe du résultat.

Généralement, la position de la virgule binaire est supposée être à la droite du premier bit le plus significatif. Ainsi la gamme des nombres représentables va de –1.0 à $1.0\text{-}2^{-(B-1)}$ où B est le nombre de bits. Il faut donc normaliser les signaux afin de respecter la gamme choisie.

I-3-2-2 : Représentation en virgule flottante :

Un nombre V s'écrit dans cette représentation sous la forme :

$$V = M \cdot 2^{e} \qquad \text{(I-8-a)}$$

Où M : est la mantisse signée et normalisée avec $\frac{1}{2} \le |M| < 1$ (I-8-b)

et e : positif ou négatif représente l'exposant.

Généralement le nombre de bits assignés à la mantisse est égal au trois quart du nombre de bits total du mot (M=3e) [12]

Le grand avantage de ce type de représentation est qu'il est possible de représenter une large gamme de nombres. Son inconvénient est que les opérations sont plus compliquées. Pour multiplier deux nombres, on doit multiplier les deux mantisses et additionner les exposants. Il faut alors s'assurer que le résultat remplit la condition (I.8-b). Ces opérations sont nettement plus compliquées que celles que l'on doit traiter en virgule fixe.

Après avoir choisi le type d'arithmétique, on peut adopter pour les opérations le principe d'arrondi ou la troncature sur le résultat.
L'arrondi convertit le résultat en nombre codé le plus proche, la troncature abandonne les bits de poids les plus faibles non représentatifs (Figure.I-2)
Un majorant e_m de l'erreur peut être donné dans le cas des nombres fractionnaires représentés en virgule fixe de la façon suivante :

Arrondi :

$$e_m = \frac{2^{-B}}{2} = 2^{-B-1}$$

Troncature :

$$e_m = 2^{-B}$$

B étant le nombre de bits

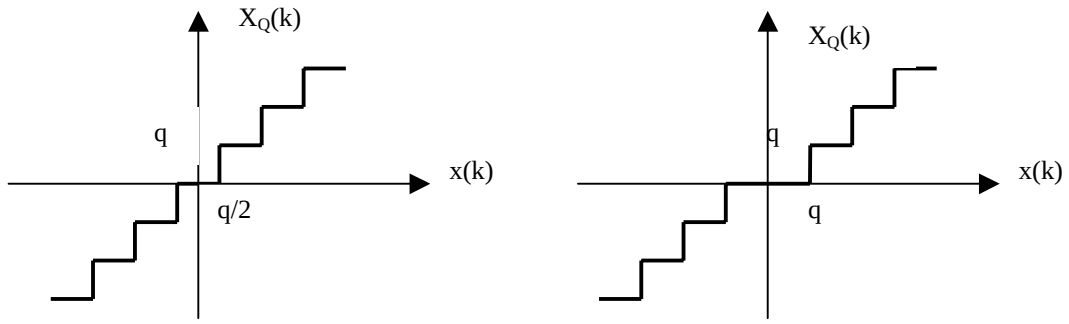
En ce qui concerne l'arithmétique à virgule flottante, l'erreur de troncature ou d'arrondi n'affecte que la mantisse. Si B désigne le nombre de bits de la mantisse alors un majorant de l'erreur est donné par :

Arrondi :

$$e_m = 2^{-B}$$

Troncature :

$$e_m = 2^{-B+1}$$



(a) Arrondi (b) Troncature

Figure I-2 :Caractéristiques de quantification

I-3-3 Conversion analogique – numérique : [12]

Dans la plupart des applications de filtrage numérique, le signal original à filtrer est de nature analogique ; c'est pourquoi un convertisseur analogique - numérique (CAN) est toujours une partie intégrante du filtre numérique (figure I.3).

Le signal d'entrée x(t) est filtré par un filtre passe bas analogique généralement très simple dont la fréquence de coupure est légèrement inférieure à la fréquence de Nyquist (fe/2). Son rôle consiste à éviter le repliement du spectre du signal échantillonné

Le CAN assure l'échantillonnage du signal continu et le codage des échantillons en une suite de nombres binaires de longueur finie. C'est ce codage qui est à l'origine des erreurs de quantification.

Le filtrage proprement dit est effectué par le processeur numérique, dont le modèle est en fait un algorithme de calcul.[12],[13]

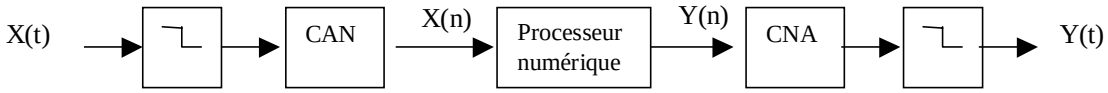


Figure I.3 : Chaîne de filtrage numérique

L'erreur entre un nombre réel « x » et sa représentation binaire finie « $[x_Q]$ » est appelée « bruit de quantification ». Elle est supposée aléatoire et est définie par :

$$e(k)= x(k) - x_Q(k) \qquad \text{(I-9)}$$

On schématise l'effet de la quantification à l'entrée d'un filtre numérique de réponse impulsionnelle h(k) par l'addition d'un signal de bruit e(k). (Figure I-4)

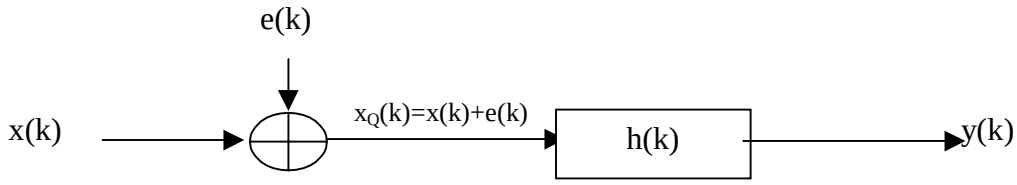


Figure I-4 : Schématisation de l'erreur de quantification

L'hypothèse de calcul formulée est que moyennant un pas de quantification suffisamment petit, la densité de probabilité associée est supposée uniformément répartie sur une plage de valeurs égales à q. On admet de plus que l'erreur commise n'est pas corrélée avec le signal.

Ces hypothèses permettent d'exprimer le lien entre les propriétés statistiques du signal quantifié et du signal de départ. De ce fait, l'erreur due à l'arrondi ou la troncature peut être statistiquement modélisée comme une source de bruit additive. La moyenne m_e et la variance σ_e^2 de ce bruit dépend de la distance entre deux nombres adjacents quantifiés. Si q représente le pas de quantification, on aura pour les deux cas d'approximation :

Arrondi :

$$\begin{cases} m_e = 0 \\ \sigma_e^2 = \dfrac{q^2}{12} \end{cases}$$

Troncature

$$\begin{cases} m_e = \dfrac{q}{2} \\ \sigma_e^2 = \dfrac{q^2}{12} \end{cases}$$

En résumé, l'effet de quantification d'un nombre réel par arrondi peut être simulé par l'ajout d'une source de bruit blanc, de variance $(q^2/12)$et de moyenne nulle, au signal d'entrée non quantifié comme montré dans la figure (I-4).

I-3-4 : Erreurs de dépassement ou d'overflow :[15]

Si l'amplitude d'un signal interne d'une réalisation en virgule fixe excède la gamme dynamique, un dépassement se produit et le signal de sortie sera distordu.

L'erreur causée par les dépassements est déterminée par la méthode utilisée pour représenter un nombre dépassant la gamme. Ce dernier peut être transformé par plusieurs façons, selon la caractéristique de dépassement choisie, parmi lesquelles on cite :

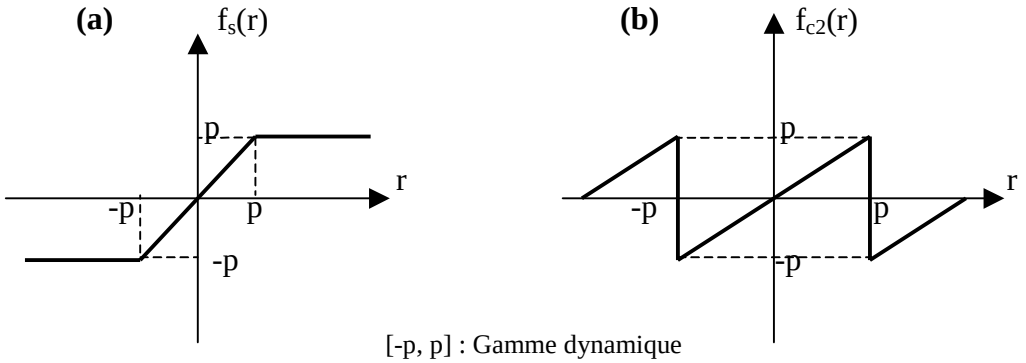
I-3-4-1 Caractéristique de complément à deux : [11]

Utilisée souvent dans la représentation en complément à deux. Elle consiste à attribuer à chaque nombre dépassant la gamme, le complément à deux du plus grand ou plus petit nombre représentable (Figure I-5-b).

Cette caractéristique est périodique et a pour but d'annuler les dépassements intermédiaires dans un accumulateur et d'obtenir des résultats de somme qui soient dans la gamme. [11],[15]

I-3-4-2 Caractéristique de saturation : [11],[13],[15]

Elle consiste à remplacer le nombre dépassant la gamme par le plus grand ou le plus petit nombre représentable (Fig.I-5-a). L'erreur de dépassement de cette caractéristique est inférieure à celle du complément à deux, mais elle est difficile à réaliser matériellement. [11],[15].



(a) Saturation (b) Complément à deux

Figure I-5 : Caractéristique de dépassement.

Lorsqu'un dépassement se produit à l'entrée, la propagation de l'erreur peut causer d'autres dépassements. Aussi, après un dépassement interne, la sortie du filtre peut devenir indépendante du signal d'entrée ; c'est l'auto-oscillation de dépassement. Les oscillations de dépassements sont causées par la non linéarité de la caractéristique de dépassement utilisée [13],[15].

Ces dépassements peuvent être minimisés en utilisant une normalisation (ou calibration) appropriée ou en élargissant la gamme dynamique des nombres ce qui aura pour effet l'augmentation du nombre de bits et donc l'augmentation de l'erreur de quantification.

I-3-5 : Oscillations de cycles limites

Dans l'étude du bruit de calcul, on suppose toujours que les valeurs des échantillons du signal d'entrée du filtre numérique sont de même ordre de grandeur que les différents multiples du pas de quantification q. Ce qui nous permet d'admettre que les échantillons du bruit de calcul sont statistiquement non corrélés aussi bien entre eux qu'avec la suite du signal d'entrée.

Cependant, le signal à l'entrée peut atteindre des valeurs faibles durant un certain temps, par conséquent, les erreurs de quantification tendent à devenir fortement corrélés et peuvent causer l'instabilité du filtre par l'apparition d'une auto-oscillation appelée « cycles limites ».[13],[16]

En effet, même si les conditions de stabilité sont remplies et que le signal à l'entrée soit absent, un signal périodique peut apparaître, généralement avec de faibles amplitudes. Ces oscillations tiennent du fait qu'en réalité le signal d'entrée n'est jamais nul en l'absence de données à l'entrée ; le signal d'erreur dû à la quantification des nombres avant la mise en mémoire est appliquée au filtre.

Une borne peut être obtenue pour ces oscillations sachant que le signal d'entrée e(k) possède en lui-même, dans le cas d'arrondi (avec un pas q), une borne donnée par :

$$|e(k)| = \frac{q}{2} \tag{I-10}$$

si la réponse impulsionnelle du filtre est h, les auto-oscillations sont limitées par :[12]

$$|y(k)| \le \frac{q}{2} \sum_{k=0}^{\infty} |h(k)| \tag{I-11}$$

Cette borne est en fait très large. Une estimation plus réaliste de l'amplitude des auto-oscillations est donnée par :

$$A_m = \frac{q}{2} \mathrm{Max}|H(\omega)| \tag{I-12}$$

où $H(\omega)$ est la réponse fréquentielle du filtre.

Il est à remarquer que les cycles limites peuvent être éliminés par l'utilisation de la troncature au lieu de l'arrondi.

I-3-6 : Erreurs de quantification à la sortie d'un filtre numérique R I I:

Pour réaliser un filtre numérique R I I, trois opérations de base sont nécessaires :

- La multiplication par des constantes (coefficients du filtre).
- L'accumulation des produits (i.e. leur addition)
- et le stockage en mémoire (dans des registres à longueur finie).

Les résultats des accumulateurs à l'intérieur du filtre doivent être éventuellement quantifiés car les multiplications augmentent le nombre de bits nécessaire à la représentation des produits. Une multiplication de deux nombres à B bits chacun engendre un nombre à 2B bits. Ce produit doit être quantifié à B bits. En supposant que la quantification du résultat est accomplie par arrondi, alors l'erreur résultante de la quantification de l'accumulation interne est appelée «bruit d'arrondi ».

Le modèle utilisé pour représenter le bruit d'arrondi est le même que celui utilisé dans le bruit de conversion analogique - numérique et est représenté dans la figure I-6.
Le quantificateur est remplacé par une source de bruit blanc de variance :

$$\sigma_e^2 = \frac{q^2}{12} \qquad \text{(I-13)}$$

où q représente le pas de quantification.

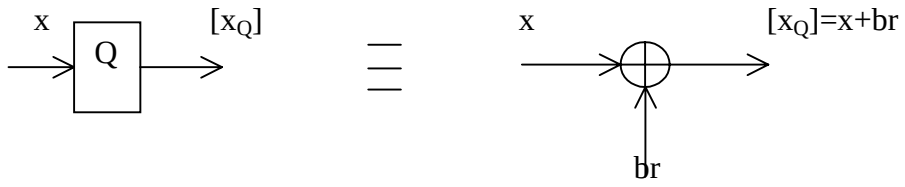
Le bruit d'arrondi est supposé uniformément distribué sur l'intervalle [-q/2 , q/2] avec une moyenne nulle.

Dans ce modèle nous supposons que :

(a) - Les différentes sources d'erreur formées par les différents accumulateurs sont non corrélées.

(b) - Chaque source de bruit est non corrélée avec l'entrée.

Le problème que nous discuterons par la suite est le suivant : Ayant une description Entrée/Sortie du filtre, la fonction de transfert par exemple, et avec le modèle de bruit d'arrondi de la figure I-6, quelle sera la structure d'état qui aura un bruit d'arrondi, dû à la quantification des produits internes, minimal à la sortie du filtre.



br : bruit de quantification
x :entrée non quantifiée, x_Q : entrée quantifiée

Figure I-6 : Modèle de la quantification interne des produits

Remarque: Nous n'avons pas pris en compte les erreurs de dépassement, car nous supposons que celles-ci sont très faibles après une normalisation appropriée des registres internes du filtre. Elles n'apparaissent donc pas dans l'expression du bruit de calcul.

I-3-6-1 : Calcul du bruit à la sortie du filtre :[15]

Avant de discuter comment normaliser un filtre numérique à virgule fixe, nous introduisons l'utilisation du modèle de la figure I-6.

On remarque sur la figure I-7qu'il existe 3 sources de bruit d'arrondi dans le filtre (filtre de second ordre). Elles occupent les nœuds A, B et C. Les nœuds A, B et le nœud C (registre de sortie) ont 3 sources de bruit chacun. Chaque produit est remplacé par son modèle de bruit équivalent.

La puissance totale du bruit à la sortie du filtre dû aux quantifications internes des produits est obtenue en sommant les puissances de bruit de sortie associées à toutes les sources de bruit interne.

En prenant comme exemple la figure I-7, considérons le bruit de sortie dû à la quantification au nœud A.

Soit g_A la réponse impulsionnelle qui lie le nœud A au nœud de sortie.

Alors la puissance du bruit de sortie dû aux sources de bruit blanc, e_{2A} et e_{3A} est :

$$\sigma_A^2 = \sigma_{e_{1A}}^2 \sum_{k=0}^{\infty} g_A^2(k) + \sigma_{e_{2A}}^2 \sum_{k=0}^{\infty} g_A^2(k) + \sigma_{e_{3A}}^2 \sum_{k=0}^{\infty} g_A^2(k)$$

$$\sigma_A^2 = 3\sigma_{e_A}^2 \sum_{K=0}^{\infty} g_A^2(K) = 3\sigma_{e_A}^2 \left\| g_A^2 \right\|_2^2 \qquad \text{(I-14)}$$

De même, si g_B est la réponse impulsionnelle qui lie le nœud B au nœud de sortie, la puissance du bruit de sortie dû à e_{1B} et e_{2B} est :

$$\sigma_B^2 = 3\sigma_{e_B}^2 \left\| g_B^2 \right\|_2^2 \qquad \text{(I-15)}$$

La puissance du bruit causé au nœud C par la quantification est :

$$\sigma_C^2 = 3\sigma_{e_C}^2 \qquad \text{(I-16)}$$

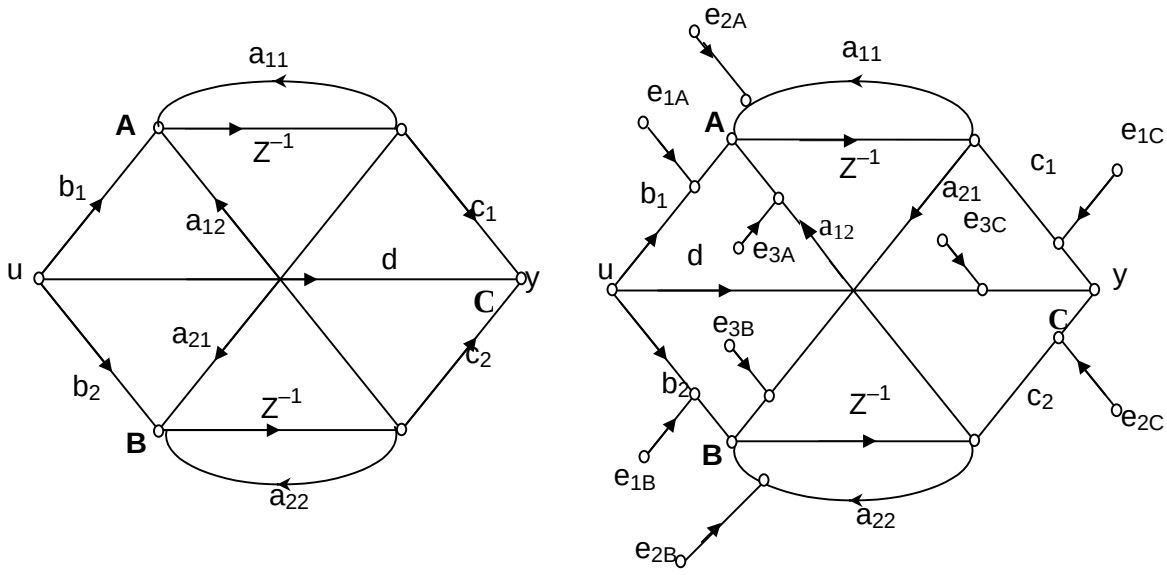


Figure I-7 : Modèle du bruit interne pour un filtre du second ordre.

La puissance du bruit d'arrondi total de sortie est la somme de σ_A^2, σ_B^2 et σ_C^2 car nous avons supposés que les 3 sources de bruit sont non corrélées. Il est donné par :

$$\sigma_{\text{total}}^2 = 3\sigma_e^2\left(1+\|g_A\|_2^2+\|g_B\|_2^2\right) \tag{I-17}$$

où $\sigma_{e_A} = \sigma_{e_B} = \sigma_{e_C} = \sigma_e$ représentent la puissance (ou variance) du bruit d'arrondi aux nœuds A, B et C respectivement. (σ_e donné par l'équation (I-13))

Dans le cas général, une structure d'espace d'état d'ordre n a (n+1) nœuds qui génèrent des bruits d'arrondis (ou de troncature). De ce fait, diverses multiplications sont effectuées pour le calcul de la $i^{\text{ème}}$ variable d'état, d'où divers bruits d'arrondis.
Si ν désigne le nombre de sources de bruits d'arrondi et si chacune est supposée être non corrélée avec les autres, alors le calcul de la $i^{\text{ème}}$ variable d'état produit un bruit d'arrondi en sortie donné par :

$$\sigma_i^2 = \nu_i\sigma_e^2\sum_{k=0}^{\infty} g_i{}^2(k) = \nu_i\sigma_e^2\|g_i\|_2^2 \tag{I-18}$$

où σ_e est donné par l'équation (I-13)

La puissance totale du bruit en sortie est alors donnée par :

$$\sigma_{\text{total}}^2 = \sigma_e^2\left[\sum_{i=1}^{n}\nu_i\|g_i\|_2^2+\nu_{n+1}\right] \tag{I-19}$$

où:

ν_i est le nombre de sources de bruit d'arrondi au nœud i, i=1…n.

g_i est la réponse impulsionnelle liant le nœud i au nœud de sortie

i = n+1 correspond au nœud de sortie avec $g_{n+1} = d$. (I-20)

Dans ce qui a précédé, pour l'écriture des équations (I-17) et (I-19), nous avons supposé que chaque produit était quantifié avant d'être sommé au niveau d'un nœud. C'est le cas des accumulateurs à simple précision.

Dans le but de réduire la puissance totale du bruit d'arrondi à la sortie du filtre, le résultat de l'accumulation des produits (i.e. leur somme) se fait dans un accumulateur double puis vient sa quantification. De ce fait, une seule erreur est commise au niveau du nœud.
Dans ce cas l'équation (I-15) devient :

$$\sigma_{total}^2 = \sigma_e^2 \left[\sum_{i=1}^{n} \nu_i \|g_i\|_2^2 + 1 \right] \qquad \text{(I-21)}$$

En résumé:

Accumulateur simple $\Rightarrow$ Longueur du registre = B bits $\Rightarrow$ $\nu_i = n+1$

Accumulateur double $\Rightarrow$ Longueur du registre = 2B bits $\Rightarrow$ $\nu_i = 1$

Le tableau I-2 établit la liste de quelques aspects importants de la limitation de la longueur des mots des résultats intermédiaires dans les filtres numériques récursifs [11].

	Quantification	**Dépassement**
Signal d'entrée x(k)=0	Cycles limites (Amplitudes relativement faibles).	Oscillations de dépassement (Amplitudes relativement élevées).
Signal d'entrée aléatoire	Bruit de quantification combiné avec le signal d'entrée.	Peu étudié a ce jour.
Caractéristique la plus adaptée	Pour le bruit de quantification : l'arrondi. Pour les cycles limites : la troncature en amplitude.	Saturation.

Tableau I-2: Effets de la longueur finie des registres

I-4 Conclusion:

Nous avons vu dans ce chapitre les effets de la longueur finie des registres et les effets de quantification des coefficients d'un filtre numérique implémenté en virgule fixe.

Ces effets causent un bruit appelé bruit de quantification à la sortie du filtre ainsi que des oscillations de cycles limites ou de dépassement.

Pour remédier aux problèmes d'oscillations de dépassement, il suffit de bien normaliser les résultats. En ce qui concerne les cycles limites l'approximation par troncature est la mieux adaptée. Enfin pour diminuer le bruit de quantification, il faut trouver une structure d'état adéquate qui puisse réduire le bruit de quantification. Ceci fera l'objet de l'étude du chapitre suivant.

CHAPITRE II

REPRESENTATION DES FILTRES NUMERIQUES DANS L'ESPACE D'ETAT

Chapitre II :
Représentation des filtres numériques dans l'espace d'état

II-1: Introduction:

Les filtres numériques peuvent être représentés par leur caractéristique entrée - sortie, c'est à dire leur fonction de transfert ou leur réponse impulsionnelle qui caractérisent complètement les propriétés et les spécifications du filtre à réaliser, c'est la représentation externe.

Mais pour décrire exactement l'opération de filtrage et les différentes opérations qui s'effectuent à l'intérieur du filtre, il est devenu nécessaire de concevoir une nouvelle représentation qui est la représentation d'état. La connaissance de la manière dont s'effectuent ces opérations dans le filtre permet de mieux analyser les erreurs dues à l'utilisation de registres de longueur finies et par conséquent de trouver la meilleure représentation répondant aux spécifications du filtre et présentant un meilleur rapport signal / bruit

II-2: Description externe des filtres numériques:

C'est la relation qui relie la suite des échantillons d'entrée {u(k)} à celle des échantillons de sortie {y(k)} en fonction des coefficients de pondérations du filtre.

$$y(k) = \sum_{i=0}^{n} b_i u(k-i) - \sum_{i=1}^{n} a_i y(k-i) \qquad \text{(II.1)}$$

où n représente l'ordre du filtre et les $(a_i, b_i, i = 1....n)$ sont les coefficients du filtre.

Cette équation permet de connaître la valeur de l'état présent de la sortie du filtre en fonction de celui des entrées présente et passées et sorties passées

Cette même relation peut-être donnée en utilisant la réponse impulsionnelle du filtre h(k) grâce à l'équation suivante :

$$y(k) = \sum_{i=0}^{\infty} h(i)u(k-i) = \sum_{i=-\infty}^{k} h(k-i)u(i) \quad \text{(dans le cas d'un filtre causal)} \qquad \text{(II.2)}$$

Les coefficients a_i de l'équation de récurrence (II.1) sont nuls pour les filtres à réponse impulsionnelle finie (R I F).

L'équation (II.2) n'est en fait que le produit de convolution entre les échantillons d'entrée u(k) et les coefficients de la réponse impulsionnelle du filtre h(k), c'est à dire :

$$y(k) = h(k) * u(k) \qquad \text{(II.3)}$$

L'analyse des systèmes discrets peut s'effectuer grâce à la transformée en Z. Cette dernière appliquée à (II.1) fournit la fonction de transfert du filtre numérique donnée par :

$$Y(Z) = H(Z).U(Z) \qquad \text{(II.4)}$$

avec $$H(Z) = \frac{\sum_{i=0}^{n} b_i Z^{-i}}{1 + \sum_{i=1}^{n} a_i Z^{-i}} \qquad \text{(II.5)}$$

H(Z) n'est autre que la transformée en (Z) de la réponse impulsionnelle du filtre donnée dans l'équation (II.2), donc la connaissance de l'une implique celle de l'autre.

II.3. Description interne où représentation d'état: [15]

Grâce à cette représentation, nous pouvons prévoir la sortie future ainsi que l'état futur exact du filtre, en connaissant à un instant donné l'état du filtre et la valeur de l'échantillon d'entrée.

L'état d'un filtre n'est autre qu'un ensemble de variables rangées dans un vecteur x(k) dit vecteur d'état lié à la sortie par l'équation :

$$y(k) = f[x(k), u(k)] \tag{II.6}$$

y(k) étant la sortie à l'instant (k), x(k) le vecteur d'état et u(k) l'entrée à ce même instant. Une relation similaire à (II.6) permet de trouver l'état futur, on obtient alors l'équation d'état suivante :

$$\begin{cases} x(k+1) = f_1[x(k), u(k)] \\ y(k) = f_2[x(k), u(k)] \end{cases} \tag{II.7}$$

f_1 et f_2 étant des relations linéaires à paramètres constants, on peut écrire :

$$\begin{cases} x(k+1) = Ax(k) + Bu(k) \\ y(k) = Cx(k) + Du(k) \end{cases} \tag{II.8}$$

où :

A est la matrice d'évolution $(n \times n)$

B est la matrice de commande $(n \times 1)$

C est la matrice d'observation $(1 \times n)$

D est le coefficient de transition directe de l'entrée à la sortie (1×1)

"n" étant l'ordre du filtre.

L'état du filtre à l'instant k en fonction de l'état initial x(0) est donné par:

$$x(k) = A^k x(0) + \sum_{i=1}^{k} A^{i-1} Bu(k-1) \tag{II.9}$$

II.4. Représentation canonique

Dite aussi structure directe, elle constitue la plus simple représentation d'un filtre, avec le minimum de coefficients de pondération à mémoriser [11], et très facilement tirée à partir de la fonction de transfert :

$$H(Z) = \frac{Y(Z)}{U(Z)}$$

qui elle même découle de l'équation (II.1).

Si on pose :

$$\frac{Y(Z)}{U(Z)} = \frac{N(Z)}{D(Z)} \quad et \quad W(Z) = \frac{U(Z)}{D(Z)}$$

on a : $$w(k) = u(k) - a_1 w(k-1) - \ldots\ldots\ldots - a_n w(k-n) \qquad \text{(II.9')}$$

ce qui donne :

$$y(k) = \sum_{i=0}^{n} b_i w(k-i) \qquad \text{(II.10)}$$

Les deux équations précédentes (II.9') et (II.10) donnent la représentation "canonique" montrée à la figure II–1, et écrite sous une forme plus explicite par le système d'équations suivantes:

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \\ \vdots \\ \vdots \\ x_n(k+1) \end{bmatrix} = \begin{bmatrix} 0 & 1 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & \ddots & & & & \vdots \\ \vdots & \vdots & & \ddots & & & \vdots \\ \vdots & \vdots & & & \ddots & & \vdots \\ \vdots & \vdots & & & & \ddots & \vdots \\ 0 & 0 & & & & & 1 \\ -a_n & \cdots & \cdots & \cdots & \cdots & -a_2 & -a_1 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ \vdots \\ \vdots \\ x_n(k) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \vdots \\ 1 \end{bmatrix} u(k) \qquad \text{(II.11a)}$$

$$y(k) = \left[(b_n - a_n b_0) \cdots\cdots\cdots (b_2 - a_2 b_0)(b_1 - a_1 b_0)\right] \begin{bmatrix} x_1(k) \\ x_2(k) \\ \vdots \\ \vdots \\ x_n(k) \end{bmatrix} + b_0 u(k) \qquad \text{(II.11b)}$$

Nous remarquons que les matrices composant le système (II.11) correspondant aux matrices A, B, C et D respectivement du système (II.8) contiennent le minimum d'éléments non nuls, ce qui fait la simplicité de la structure.

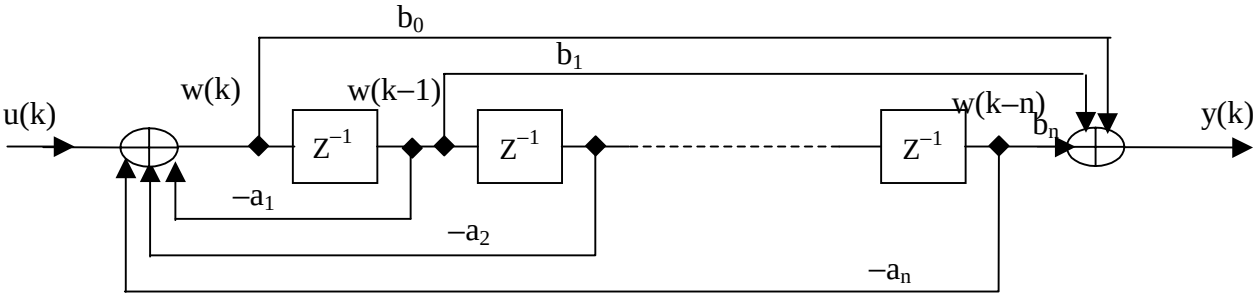


Figure II-1 :Structure canonique d'un filtre numérique

A partir de la représentation d'état nous pouvons aisément retrouver la fonction de transfert du filtre. Considérons le système (II.8) . Si nous lui appliquons une transformation dans le plan Z il devient :

$$\begin{cases} X'(Z) = AX(Z) + BU(Z) & \text{[III.12a]} \\ Y(Z) = CX(Z) + DU(Z) & \text{[III.12b]} \end{cases}$$

nous avons :

$$X(Z) = Z^{-1}X'(Z)$$

ce qui donne :

$$X(Z) = [ZI - A]^{-1}BU(Z) \qquad \text{(II.13)}$$

en substituant (II.13) dans (II.12b) nous obtenons :

$$Y(Z) = C[ZI - A]^{-1}BU(Z) + DU(Z) \qquad \text{(II.14)}$$

la fonction de transfert est alors donnée par :

$$H(Z) = C(ZI - A)^{-1}B + D \qquad \text{(II.15)}$$

Les pôles de la fonction de transfert ainsi obtenue, ne sont en fait que les racines du polynôme caractéristique de la matrice A tel que :

$$a(Z) = \det(ZI - A)$$

Ce sont donc les valeurs propres de la matrice A.

Pour assurer la stabilité du système, il suffit que les modules des valeurs propres de A soient inférieurs à l'unité. Ainsi en régime libre, c'est à dire entrée nulle, nous avons:

$$x(k) = A^k x(0)$$

qui tendra vers zéro quand k tend vers l'infini.

En régime établi , l'équation (II.9) devient pour les systèmes stables:

$$x(k) = \sum_{i=1}^{+\infty} A^{i-1} B\, u(k-i) \qquad \text{quand k tend vers l'infini} \qquad \text{(II.15')}$$

II.5. Transformation de coordonnées [15]

La fonction de transfert et la réponse impulsionnelle spécifient la relation entrée - sortie du filtre sans pour autant apporter une information sur sa structure interne.
Comme le vecteur d'état x(k) dans le système (II.8) n'est qu'un ensemble de n variables internes (variables d'état) permettant la détermination des matrices A, B, C et D, nous montrons dans ce qui suit que nous pouvons changer le système de coordonnées x(k) sans changer la fonction de transfert H(Z) ou la réponse impulsionnelle.
En appliquant une transformation T, matrice $(n \times n)$ non singulière, le nouveau vecteur d'état est donné par :

$$x'(k) = T^{-1} x(k) \qquad \text{(II.16)}$$

La substitution de (II.16) dans le système d'état (II.8) donne :

$$\begin{aligned} x'(k+1) &= T^{-1}\left[Ax(k) + Bu(k)\right] \\ &= T^{-1} A\, T\; x'(k) + T^{-1} B\, u(k) \end{aligned} \qquad \text{(II.17a)}$$

$$y'(k) = C\, T\; x'(k) + Du(k) \qquad \text{(II.17b)}$$

Les paramètres d'état se transforment alors comme suit:

$$(A,B,C,D) \xrightarrow{T} (A',B',C',D') = (T^{-1}AT, T^{-1}B, CT, D) \qquad \text{(II.18)}$$

Ce type de transformation conserve la fonction de transfert et la réponse impulsionnelle, en effet, l'équation (II.15) appliquée aux matrices transformées A', B', C' et D' donne :[15]

$$\begin{aligned} H'(Z) &= D' + C'(ZI - A')^{-1}B' = D + CT(ZI - T^{-1}AT)^{-1}T^{-1}B \\ H'(Z) &= D + CT(T^{-1}(ZI - A)T)^{-1}T^{-1}B = D + CTT^{-1}(ZI - A)^{-1}TT^{-1}B \\ H'(Z) &= D + C(ZI - A)^{-1}B = H(Z) \end{aligned} \qquad \text{(II.19)}$$

II.6. Matrice K et stabilité de Lyapunov du filtre [15]

Les descriptions d'état des filtres linéaires sont plus utiles quand il est nécessaire de calculer des quantités qui dépendent de la structure interne du filtre, et de se demander alors comment changent ces quantités lorsque la structure est modifiée. Quelques unes des quantités les plus importantes à calculer sont statistiques. Elles apparaissent quand le filtre est attaqué par une entrée aléatoire stationnaire au sens large.

On définit la matrice K comme étant la matrice de covariance d'état du filtre comme suit :

$$K = E[x(k)x^t(k)] \qquad \text{(II.20)}$$

la substitution de l'équation (II.15') dans (II.20) donne :

$$K = \sum_{l=0}^{\infty}\sum_{m=0}^{\infty}(A^l B)\Gamma_{uu}(l-m)(A^m B)^t \qquad \text{(II.22)}$$

avec :

$\Gamma_{uu}(\tau) = E\,[\,u(k)u(k+\tau)]$, fonction d'auto - corrélation de l'entrée.

Si l'entrée du filtre est un bruit blanc, nous avons:

$$\Gamma_{uu}(\tau) = \delta(\tau)$$

La double somme de (II.22) devient:

$$K = \sum_{l=0}^{\infty} (A^l B)(A^l B)^t \qquad \text{(II.23)}$$

A partir de l'équation précédente, nous remarquons que le calcul même da la matrice K impose ou exige la stabilité du filtre, c'est-à-dire, il faut que :

$$\lim_{l \to \infty} A^l = 0 \qquad \text{(II.24)}$$

d'où l'intime relation entre la stabilité du filtre et la matrice de covariance (K) qui vérifie donc l'équation de Lyapunov suivante :

$$K = AKA^t + BB^t \qquad \text{(II.25)}$$

La matrice K ne reste pas invariante à une transformation de coordonnées, en effet :
Reprenons la définition de l'équation (II.20) :

$$K = E[x(k)x^t(k)]$$

si on pose :

$$q(k) = T^{-1}x(k)$$

nous avons :

$$K' = E[q(k)q^t(k)]$$

alors

$$K' = E\left[T^{-1}x(k)x^t(k)T^{-t}\right]$$

d'où :

$$K' = T^{-1} K\, T^{-t} \tag{II.26}$$

Cette dernière équation qui montre l'influence d'une transformation de coordonnées sur la matrice K sera d'une grande importance dans le reste de notre étude.

II.7. Structures éliminant les oscillations de dépassement [15]

Les oscillations de dépassement se produisent si des vecteurs d'état augmentent en amplitude après une multiplication par la matrice A. Le rôle de la caractéristique de dépassement (fig. I-5) est de diminuer l'amplitude du vecteur d'état si celle ci dépasse la gamme. Des oscillations répétées ne peuvent se produire que si la multiplication par la matrice A augmente la grandeur du vecteur d'état.
Une condition nécessaire pour la production des oscillations de dépassement est que :

$| Ax | > | x |$; x variable d'état donnée

Pour cela, on peut trouver une autre structure d'état caractérisée par (A,B,C,D) pour laquelle la grandeur de x n'augmente pas du fait de la multiplication par A. La norme d'un vecteur est donnée par :

$$\|x\| = (x^t\, d\, x)^{1/2} \tag{II.27}$$

où d est une matrice diagonale positive.

La norme de la matrice A est donnée par :

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \max_{x \neq 0} \left[\frac{x^t A^t d\, A x}{x^t\, d\, x} \right]^{1/2} \tag{II-28}$$

Pour éviter les oscillations de dépassement pour une entrée nulle (ou constante), il suffit que :

$$\|Ax\| \leq r.\|x\| \qquad \text{avec} \quad 0 < r \leq 1 \tag{II-29}$$

Autrement dit, il suffit de trouver la matrice d telle que : [15]

$$Q = r^2 d - A^t d\, A^2 \qquad \text{(II-30)}$$

soit une matrice définie positive.

Parmi les structures qui satisfont la condition : Q définie positive on trouve les structures normales ($AA^t = A^t A$), de gain de bruit de calcul minimal, et en treillis [18]. Les formes directes (ou canoniques) sont susceptibles de produire des oscillations de dépassement qui dépendent de la position des pôles.

Exemple : Dans le cas des filtres du second ordre, si la matrice A(2x2)donnée par : $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ dont les valeurs propres λ vérifient : | λ | < 1, alors il existe une matrice diagonale définie positive d pour laquelle Q de (II-30) est définie positive, si et seulement si les conditions suivantes sont satisfaites :

$$\begin{array}{ll} a_{12}a_{21} \geq 0 & \\ \text{ou si} \quad a_{12}a_{21} < 0 & \text{alors} \quad \left|a_{11}\text{-}a_{22}\right| + \det(A) < 1 \end{array} \qquad \text{(II-31)}$$

II.8 Normalisation des filtres numériques en virgule fixe :

Pour éviter les dépassements dans les registres internes qui causent des erreurs importantes, on normalise convenablement la réalisation du filtre.

Normaliser revient à mettre toutes les valeurs numériques des variables internes dans une gamme appropriée à la réalisation matérielle. La gamme d'une variable interne est nécessairement limitée par le fait de l'utilisation des registres de longueur finie.

II.8.1 : Règles de normalisation :

Dans la représentation en virgule fixe, une variable interne v(k) telle que :

V(k)=(f*u)(k)

avec f(k) la réponse impulsionnelle entre l'entrée u(k) et la variable d'état v(k) ; Elle est bornée par le fait que l'entrée u(k) est limitée par :

$$|u(k)| \leq \Delta$$

ou Δ dépend du pas de quantification ($q = \Delta.2^{-B+1}$) et qui est généralement normalisé à 1 dans la représentation en virgule fixe.

La gamme des valeurs de V dépend ainsi de la nature de l'entrée u(k) et de la suite f(k). En effet, si un filtre caractérisé par sa fonction de transfert :

$$H(Z)=D(Z)+F(Z).G(Z)$$

Et par son diagramme bloc (fig.II-2a) :

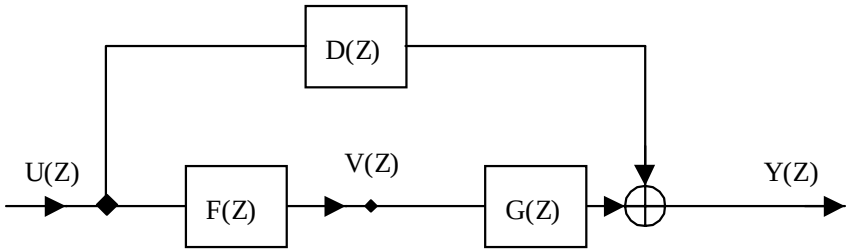


Figure II-2a : Filtre non normalisé

il est alors normalisé de la façon suivante (fig. II-2b)

De cette manière, H(Z) reste invariante et

$$|V'(k)| \leq \frac{1}{\beta} \sum_m |f(m)||u(k-m)|$$

et comme $|u(k)| \leq 1$

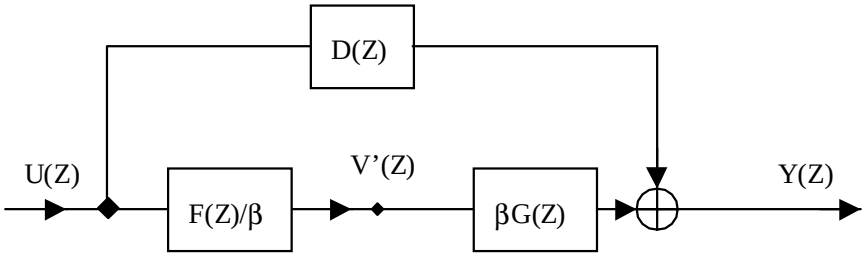


Figure II-2b : Filtre normalisé

alors $$|V'(k)| \le \frac{1}{\beta} \sum_m |f(m)| \qquad \text{(II-32)}$$

Donc β est choisi de façon à vérifier la relation (II-32). On peut choisir la norme L1 telle que :

$$\beta = \|f\|_1 = \sum_{l=0}^{+\infty} |f(l)|$$

mais étant donné que cette norme constitue une borne largement conservatrice de la gamme et que la norme L2 est :

$$\|f\|_2 = \left[\sum_{l=0}^{+\infty} \left| f^2(l) \right| \right]^{1/2} \le \|f\|_1$$

On peut donc choisir

$$\beta = \delta \|f\|_2$$

avec δ un paramètre choisi empiriquement afin d'obtenir une bonne représentation des valeurs de la variable V'(k) dans la gamme, d'éviter ainsi les oscillations dues aux amplitudes faibles (ou cycles limites) et de réduire les erreurs d'arrondi.

En conclusion, un filtre est normalisé s'il vérifie la contrainte de normalisation définie par une des règles de normalisation dont on cite :[15]

1- Normalisation par la norme L1 : $\|f\|_1 = 1$

2- Normalisation par la norme L2 : $\delta \|f\|_2 = 1$

Dans notre cas on opte pour la norme L2 qui permet à la fois de conserver la gamme et de réduire la probabilité des cycles limites et les erreurs d'arrondi par le choix d'une valeur adéquate de δ.

II.8.2 : Normalisation des paramètres d'état :

La normalisation d'un filtre par la norme L2 consiste à normaliser le vecteur d'état qui s'exprime par :

$$x(k) = \sum_{l=0}^{+\infty} A^l B u(k-1-l)$$

La réponse impulsionnelle de l'entrée à x(k) est alors :

$$f(k) = \begin{cases} 0 & k = 0 \\ A^{k-1}B & k > 0 \end{cases}$$

et donc $F(Z)=(Z\,I - A)^{-1}\,B$

f(k) est un vecteur nx1 à l'aide duquel on peut exprimer la matrice K de (II.20) pour une entrée bruit blanc normalisé comme suit :

$$K = \sum_{k=0}^{\infty} f(k)\, f^t(k) \qquad \text{(II-33)}$$

avec $K_{ij} = (f_i, f_j) = \sum_{k=0}^{\infty} f_i(k)\, f_j(k)$

donc $K_{ii} = \sum_{k=0}^{\infty} {f_i}^2(k) = \|f_i\|_2^2$

Donc pour normaliser ce filtre, il faut lui appliquer une transformation T diagonale pour laquelle la contrainte de normalisation (norme L2) soit :

$$\delta\sqrt{K'_{ii}} = 1 \qquad \text{i=1,2,…n} \qquad \text{(II-34)}$$

dans la nouvelle représentation d'état.

Donc il suffit de prendre

$$T_{ii} = \delta\sqrt{K_{ii}} \qquad i=1,2,\ldots,n \tag{II-35}$$

car d'après les équations (II-31), on a

$$K' = T^{-1}K\left(T^{-1}\right)^{t} = T^{-1}K\,T^{-t}$$

d'où $$K'_{ij} = \frac{K_{ij}}{T_{ii}T_{jj}}$$

$$K'_{ii} = \frac{1}{\delta^2}$$ d'où l'équation (II-34)

La transformation de normalisation T doit être appliquée aussi aux paramètres d'état (A,B,C,D) suivant la relation (II-18)

En résumé, nous donnons l'organigramme pour normaliser un filtre numérique (fig.II-3)

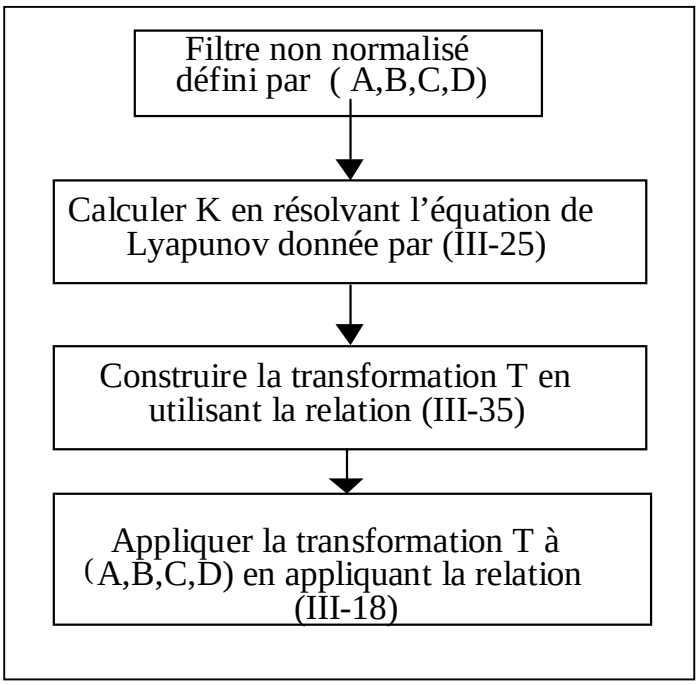


Figure II-3 : Organigramme de normalisation d'une structure d'état (A,B,C,D)

Nous remarquons que cette opération de normalisation a un grand intérêt dans la réalisation des filtres, car elle nous permet de réduire la probabilité de dépassement dans les registres internes.

Nous allons voir dans ce qui suit les effets de la normalisation sur le bruit de calcul interne.

Nous avons vu que la transformée de normalisation est sans effet sur la description externe c-a-d la fonction de transfert du filtre (d'après l'équation II-19), mais il n'en est pas de même pour le bruit de calcul interne.

La figure (II-4) montre l'influence du choix de la valeur du coefficient de normalisation δ de l'équation (II-34) sur la puissance des erreurs aussi bien de dépassement que du bruit interne de calcul [18].

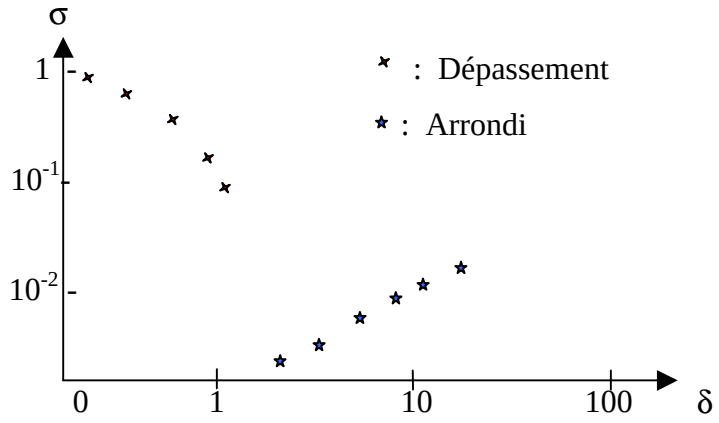


Figure II-4 : Influence du coefficient δ sur les erreurs de dépassement et de calcul interne

Nous remarquons que quand δ augmente, les erreurs de dépassement diminuent contrairement aux erreurs de quantification interne (arrondi). Cela peut être expliqué par ce qui suit :

A partir des équations (I-21) et (I.13) et en prenant un accumulateur double (ν_i=1), l'expression de la puissance totale du bruit en sortie est:

$$\sigma_y^2 = \frac{q^2}{12}\sum_{i=0}^{n}\|g_i\|^2 \qquad \text{(II-36)}$$

où $g_i(k)$ est la réponse impulsionnelle liant la i^{éme} variable d'état à la sortie, réciproquement à $f_i(k)$ qui est la réponse impulsionnelle liant l'entrée à la i^{éme} variable d'état.

d'où :

$$g(k)=\begin{cases}0 & k\le 0\\ CA^{k-1} & k>0\end{cases}$$

alors $G(Z)=C(Z\,I-A)^{-1}$

par dualité à la matrice de covariance K donnée par l'équation (II-33) utilisée pour réaliser la normalisation, nous définissons la matrice W par :

$$W=\sum_{k=0}^{\infty}g^{t}(k)g(k)=\sum_{k=0}^{\infty}\left(CA^{k}\right)^{t}\left(CA^{k}\right) \qquad \text{(II-37a)}$$

$$W=A^{t}W\,A+C^{t}C \qquad \text{(II-37b)}$$

Nous avons donc :

$$W_{ii}=\sum_{k=0}^{\infty}g_{i}(k)g_{i}(k)=\sum_{k=0}^{\infty}g_{i}(k)=\left\|g_{i}\right\|^{2}$$

D'où en remplaçant $W_{i\,i}$ dans (II-36), l'expression du bruit en sortie s'écrit en fonction des éléments diagonaux de la matrice W par :

$$\sigma_{y}^{2}=\frac{q^{2}}{12}\sum_{i=0}^{n}W_{ii}=\frac{q^{2}}{12}tr(W) \qquad \text{où tr : trace de la matrice} \qquad \text{(II-38)}$$

Si l'on applique une transformation T non singulière à W, la matrice transformée est donnée par :

$$W'_{i\,j}=T^{t}W\,T \qquad \text{(II-39)}$$

Après une transformation de normalisation T appliquée à une structure (A,B,C,D), tel que

$$T=diag[\,\delta\sqrt{K_{11}}\,,\delta\sqrt{K_{22}}\,,\ldots,\delta\sqrt{K_{nn}}\,] \qquad \text{(II-40)}$$

Les éléments de la matrice normalisée K' sont donnés par :

$$K'_{ij} = \frac{K_{ij}}{\sqrt{K_{ii}}\sqrt{K_{jj}}} \qquad \text{et} \qquad K'_{ii} = \frac{1}{\delta^2} \tag{II-41}$$

réciproquement les éléments de la matrice transformée *W'* sont donnés par :

$$W'_{ij} = W_{ij}\delta^2\sqrt{K_{ii}}\sqrt{K_{jj}} \quad \text{et} \quad W'_{ii} = \delta^2 W_{ii}K_{ii} \tag{II-42}$$

En remplaçant les expressions (II-41) et (II-42) dans l'expression de la puissance du bruit de sortie, nous obtenons :

$$\sigma_y^2 = \frac{q^2}{12}\sum_{i=1}^{n} W'_{ii} = \left(\frac{q^2}{12}\right)\delta^2 \sum_{i=1}^{n} W_{ii}K_{ii} \tag{II-43}$$

Cette équation montre la relation qui existe entre la puissance du bruit de sortie σ_y^2 avec le paramètre δ de normalisation et les paramètres du filtre non normalisé K_{ii} et W_{ii}.

On explique alors la figure (II-3) car d'après l'expression de la relation (II-43), le bruit de quantification par arrondi est proportionnel au carré du coefficient de normalisation δ. Donc le choix de δ doit être fait de manière adéquate.
L'expression du gain de bruit interne du filtre normalisé est donné par :

$$G = \delta^2 \sum_{i=1}^{n} W_{ii}K_{ii} \tag{II-44}$$

L'objectif essentiel de tout ce qui va suivre sera la minimisation de cette expression et donc de la quantité $\sum_{i=1}^{n} W_{ii}K_{ii}$ simplement par une transformation non singulière qui conserve les contraintes de normalisation.

II.9 : Structures à gain de bruit minimal :

La minimisation du gain de bruit G est relative à la variance de l'erreur de calcul à la sortie du filtre σ_y^2. Elle n'aura de sens que si l'on suppose que le facteur de normalisation a été convenablement choisi afin de rendre les dépassements peu probables et l'augmentation du bruit de calcul qui en découle négligeable.

Minimiser le gain de bruit d'un filtre numérique, décrit par ses paramètres d'état (A,B,C,D,K,W) revient à trouver une structure par une transformation T non singulière qui change les paramètres du filtre en :

$$(A',B',C',D',K',W') = (T^{-1}AT, T^{-1}B, CT, D, T^{-1}KT^{-t}, T^{t}WT)$$

de sorte que le gain de bruit soit minimal avec la contrainte de normalisation donnée en II-35.

Comme la quantité $\sum_{i=1}^{n} W_{ii}K_{ii}$ est invariante à une transformation diagonale, il est donc possible de minimiser d'abord le gain de bruit, ensuite d'appliquer une transformation de normalisation à la structure trouvée.

Actuellement, il existe deux méthodes de base permettant de construire les structures à gain de bruit minimal, la méthode de Mullis-Roberts [9] et la méthode de Hwang [10].

Bien que ces deux méthodes soient différentes du point de vue du traitement mathématique du problème, elles aboutissent chacune à un même gain de bruit minimal.

II.9.1 : Méthode de Mullis-Roberts [9]

Cette méthode traite le cas des registres de longueur égale. Elle consiste à rechercher le gain de bruit optimal à partir des conditions de son existence qui permettent d'établir la transformation réalisant la structure minimale.

Si $q = \Delta.2^{-B+1}$; $\Delta = 1$

est le pas de quantification et B le nombre de bits des registres considérés, alors l'expression de la variance (ou puissance) du bruit en sortie devient :

$$\sigma_y^2 = \frac{n(n+1)}{3}\left(\frac{\delta}{2^B}\right)^2\left[\frac{1}{n}\sum_{i=1}^{n} K_{ii}W_{ii}\right] + \frac{n+1}{3.2^{2B}}$$

Si n, δ, et B sont donnés, alors il suffit de minimiser le gain de bruit donné par :

$$\boxed{G = \sum_{i=1}^{n} W_{ii}K_{ii}}$$

A cet effet, on utilise la propriété des matrices K et W dans le cas où celles-ci sont réelles, symétriques et définies positives, l'inégalité suivante est vérifiée :

$$\left[\frac{1}{n}\sum_{i=1}^{n} K_{ii}W_{ii}\right] \geq M_a^2$$

$$où \quad M_a^2 = \frac{1}{n}\sum_{i=1}^{n} \mu_i$$

avec $\mu_1^2, \mu_2^2, ..., \mu_n^2$ sont les valeurs propres de la matrice produit KW et dont les racines carrées sont appelées les modes de second ordre [15]

Pour que la borne inférieure soit atteinte, il est nécessaire et suffisant de satisfaire les conditions suivantes : [9],[15]

1) $D_0^{-1}KD_0^{-1} = D_0WD_0$

où D_0 : matrice diagonale c-a-d :

$K = D'WD'$ telle que $D' = D_0^2$

2) $K_{ii}W_{ii} = K_{ii}W_{jj} \qquad \forall i, j = 1,2,...,n$

si K est normalisée (c-a-d $K_{ii} = \frac{1}{\delta^2}$) alors $W_{ii} = W_{jj} \qquad \forall i, j = 1,2,...,n$

La condition (1) implique que si $K' = D_0^{-1}KD_0^{-1} = W'$ alors $tr(K') = tr(W') = \sum_{i=1}^{n} \mu_i = n\,M_a$

La condition (2) implique l'égalité suivante :

$$K'_{ii}W'_{ii} = K_{ii}W_{ii} = K_{jj}W_{jj} = \text{constante} = \alpha \qquad \forall i, j = 1,2,...,n$$

Donc $tr(K') = n\, M_a = n\sqrt{\alpha}$ et $\alpha = M_a{}^2$

d'où $\sum_{i=1}^{n} K'_{ii} W'_{ii} = \alpha\, n = n M_a{}^2$

alors on a bien l'égalité $\left[\frac{1}{n}\sum_{i=1}^{n} K'_{ii} W'_{ii}\right] = M_a^2$

Dans ces conditions pour tout ε positif infiniment petit, il existe une transformation T_0 pour laquelle on a :

$$0 \le \frac{1}{n}\sum_{i=1}^{n} \left(T_0^{-1} K' T_0^{-t}\right)_{ii} \left(T_0^{t} K' T_0\right)_{ii} - M_a^2 < \varepsilon$$

Pour trouver la transformation T_0 , il suffit de diagonaliser les matrices K et W et d'appliquer une successions de rotations à la matrice diagonale W' afin de rendre ses éléments identiques. Le nombre de ces rotations dépend du choix de ε.

Le gain de bruit minimal obtenu est :

$$\boxed{G_{\min} = \frac{1}{n}\left[\sum_{i=1}^{n} \mu_i\right]^2} \qquad \text{(II-45)}$$

II-9-2 : Méthode de Hwang :[10]

Dans cette méthode il s'agit de construire la transformation *T* en maintenant la contrainte de normalisation

Soit *T* la transformation qui minimise le gain de bruit G et qu'on factorise en :

$$T = R\, S$$

avec *R* : Matrice orthogonale ($R\, R^t = I$)

S : Matrice définie positive

En diagonalisant *S*, la transformation *T* devient :

$$T = R\left(R_0 \Lambda {R_0}^t\right) = R_1 \Lambda {R_0}^t \qquad \text{où} \qquad \Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_n)$$

et R_0 : matrice des vecteurs propres de S. Elle est orthogonale

Le gain de bruit est alors :

$$G_1 = tr\left(T^t W_0 T\right) = tr\left(T^t T\, W_0\right)$$

$$= tr\left(\Lambda^2 R_1{}^t W_0 R_1\right)$$

$$= \sum_{i=1}^{n} \lambda_i{}^2 r_i{}^2$$

où $r_i{}^2 = \left(R_1{}^t W_0 R_1\right)_{ii}$

et T doit assurer la normalisation donc :

$$T^{-1} K_0 T^{-t} = \begin{bmatrix} 1 & & & x \\ & 1 & & \\ & & \ddots & \\ x & & & 1 \end{bmatrix} \quad \text{avec } \delta = 1 \quad \text{et x : élément non nul}$$

or d'après [19]

$$\det W_0 \le \prod_{i=1}^{n} r_i{}^2$$

et $$\det K_0 \le \prod_{i=1}^{n} \lambda_i{}^2$$

et sachant que la moyenne arithmétique est supérieure où égale à la moyenne géométrique, nous pouvons alors écrire :

$$\frac{1}{n}\left[\sum_{i=1}^{n}\lambda_i{}^2 r_i{}^2\right] \geq \left[\prod_{i=1}^{n}\lambda_i{}^2 r_i{}^2\right]^{1/n}$$

alors

$$G_1 = tr\left(T^t W_0 T\right) \geq n\left[\prod_{i=1}^{n} r_i{}^2\right]\left[\prod_{i=1}^{n}\lambda_i{}^2\right] \geq n\left[\det K_0 W_0\right]^{1/n} \qquad \text{(II-46)}$$

Comme les valeurs propres du produit K_0W_0 sont invariants à une transformation d'état T, cette borne inférieure de l'équation (II-46) est aussi invariante.

L'égalité dans (II-46) est réalisée si et seulement si :

1. K_0W_0 symétrique
2. $K_0{}^{-1}$ et W_0 sont équivalentes à une constante prés.

Pour minimiser le gain

a)- On normalise d'abord le système par T_0 :

$$K_1 = T_0^{-1} K_0 T_0^{-t} = \mathrm{I} \qquad \text{donc} \qquad T_0\, T_0^t = K_0$$

$$W_1 = T_0^t W_0 T_0$$

b)- On applique la matrice T décrite précédemment, donc :

$$G_1 = tr\left(T^t W_1 T\right) = tr\left(\Lambda^2 R_1{}^t W_1 R_1\right) = \sum_{i=1}^{n}\lambda_i{}^2 r_i{}^2$$

avec la contrainte

$$R_0\Lambda^{-2}R_0{}^t = \begin{bmatrix} 1 & & \mathrm{x} \\ & 1 & \\ & \ddots & \\ \mathrm{x} & & 1 \end{bmatrix} \quad \mathrm{x} : \text{élément non nul} \qquad \text{(II-47)}$$

dans ces conditions on obtient :

$$\sum_{i=1}^{n}\frac{1}{\lambda_i{}^2} = n$$
$$\prod_{i=1}^{n}\lambda_i{}^2 \geq 1 \qquad \text{(II-48)}$$

c)- En utilisant la fonction de Lagrange donnée par :

$$\mathcal{L}(\lambda) = \sum_{i=1}^{n} \lambda_i{}^2 r_i{}^2 + \alpha\left[\left(\sum_{i=1}^{n} \frac{1}{\lambda_i{}^2}\right) - N\right]$$

qu'on minimise par rapport à λ_i et α pour obtenir :

$$\lambda_i = \left(\frac{\sum_{i=1}^{n} r_i}{n.r_i}\right)^{1/2}$$ avec la contrainte donnée en (II-48)

donc le gain minimal est

$$G_{\min} = tr\left(T^t W_1 T\right) \geq \frac{1}{n}\left[\sum_{i=1}^{n} r_i\right]^2$$

Comme

$$\det\left(K_0 W_0\right) = \det\left(W_1\right) = \prod_{i=1}^{n} \mu_i{}^2$$

où μ_i sont les modes de second ordre de $K_0 W_0$ et

$$\det\left(R_1{}^t W_1 R_1\right) = \det\left(W_1\right)$$

Si $\left(R_1{}^t W_1 R_1\right)$ est une matrice symétrique définie positive, alors :

$$\left[\sum_{i=1}^{n} r_i{}^2\right] \geq \left[\sum_{i=1}^{n} \mu_i{}^2\right]$$

avec r_i et μ_i positifs pour tout *i*.

L'égalité est atteinte si la matrice $\left(R_1{}^t W_1 R_1\right)$ est diagonale, c'est à dire que $\left(R_1{}^t W_1 R_1\right)$ est la matrice R_1 des vecteurs propres de W_1, donc on a :

$$G_{\min} = \frac{1}{n}\left[\sum_{i=1}^{n} \mu_i\right]^2 \geq n\left[\det K_0 W_0\right]^{1/n}$$

La borne inférieure est atteinte si et seulement si :

$\mu_i{}^2 = \mu_j{}^2$ pour tout *i*,*j*=1,…,n

d)- On détermine R_0 sachant la contrainte (II-47). Etant donné que R_0 est une matrice orthogonale, elle est décomposable en facteurs de matrice de rotation élémentaires de la forme :

$$R_i = \begin{bmatrix} \mathrm{I} & \vdots & 0 & \vdots & 0 \\ \cdots & \cos\varphi_i & \cdots & \sin\varphi_i & \cdots \\ 0 & \vdots & \mathrm{I} & \vdots & 0 \\ \cdots & -\sin\varphi_i & \cdots & \cos\varphi_i & \cdots \\ 0 & \vdots & 0 & \vdots & \mathrm{I} \end{bmatrix} \qquad i = 1,....,n$$

(n-1) transformations sont nécessaires pour construire R_0, qui vérifie l'équation (II-47)

e)- Finalement on construit la transformation de minimisation T telle que :

$$T = T_0 R_1 \Lambda R_0{}^t$$

qu'on applique aux paramètres du filtre (A_0,B_0,C_0,D_0)

Algorithme de calcul des matrices K et W

Les matrices *K* et *W* vérifient les équations de Lyapunov et sont données par (II-25) et (II-37b).

La procédure qui calcule *K* calcule également *W* en remplaçant

A par A^t

et *B* par C^t

L'algorithme est le suivant :

1. Initialiser $F \leftarrow A$ et $K \leftarrow B.B^t$
2. Calculer $K \leftarrow FKF^t + K$
3. Faire : $F \leftarrow F^2$
4. Refaire (2) et (3) jusqu'à ce que $F=0$

La convergence de l'algorithme dépend de la position des pôles du filtre par rapport au cercle unité.

II.10 : Propriétés des structures à gain de bruit minimal

II.10.1 : Invariance par rapport à une transformation fréquentielle :

Une propriété fondamentale des structures à gain de bruit minimal est la conservation des modes de second ordre par une transformation fréquentielle. En effet, le bruit à la sortie des structures optimales est indépendant de la largeur de la bande passante des filtres comme il est montré à la figure. (II-5) pour deux types de filtres : Butterworth et Chebyshev.

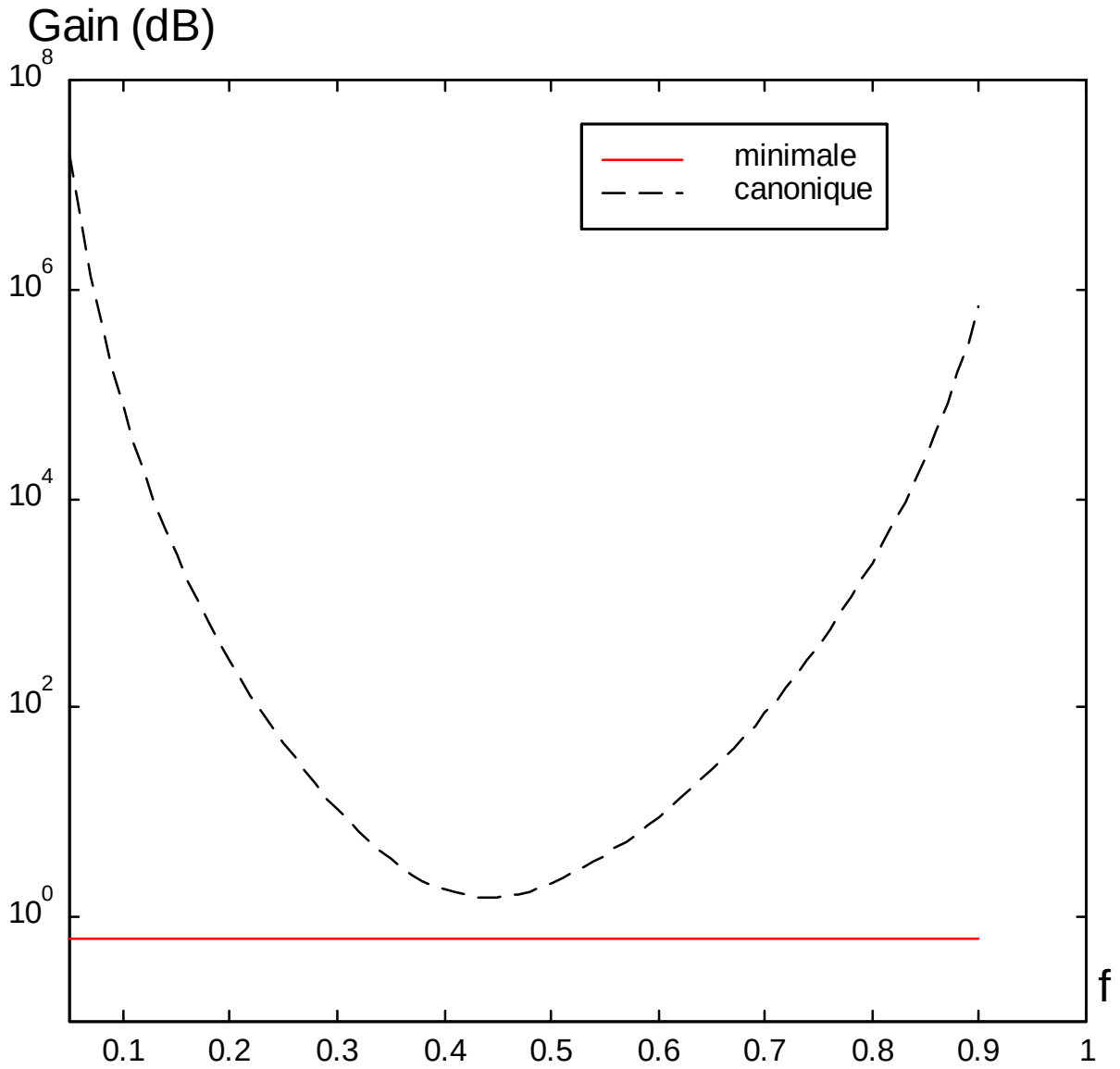


Figure II-5a : Variation du gain de bruit de calcul en fonction de la largeur de la bande passante du filtre de Butterworth d'ordre 5

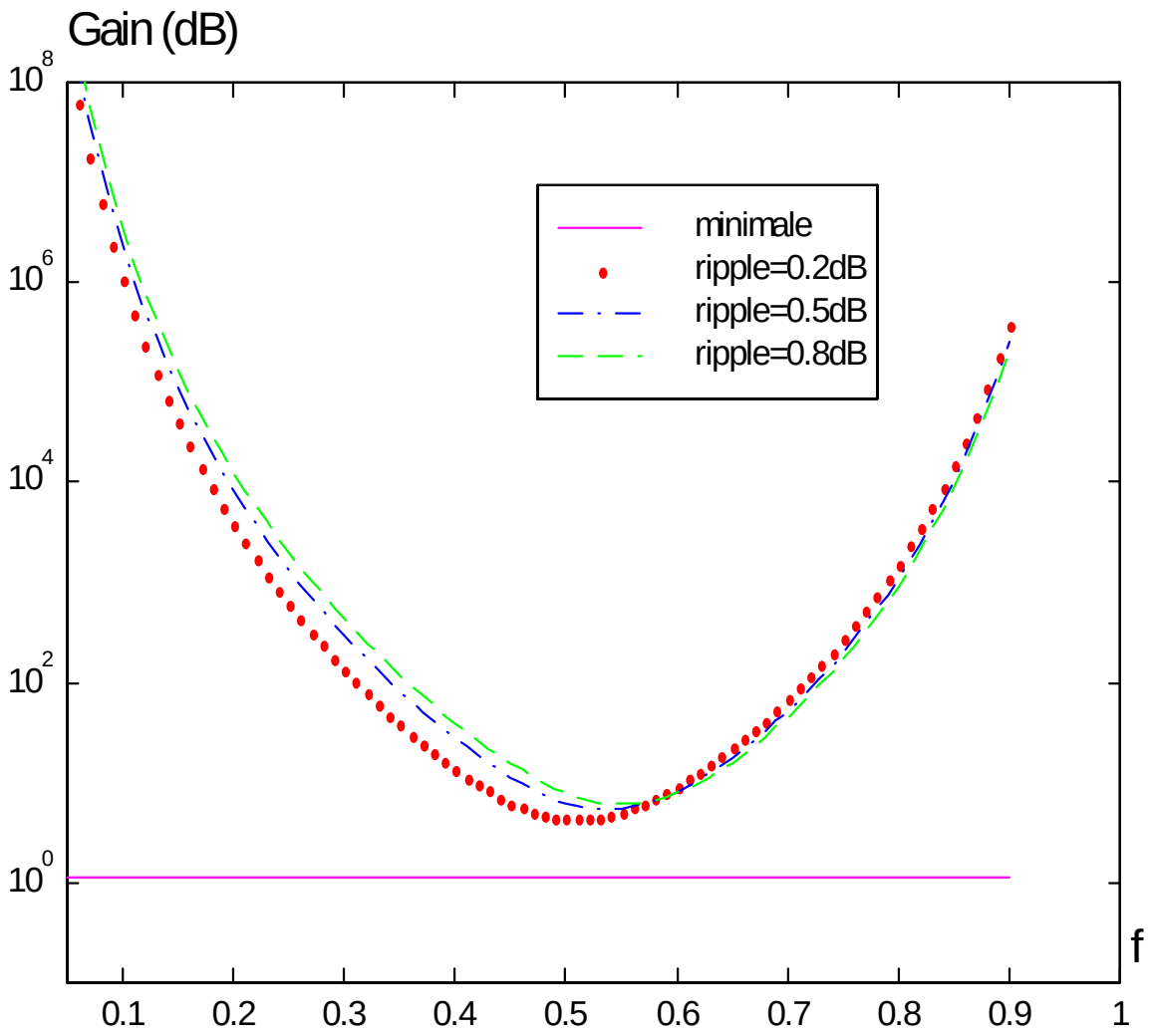


Figure II-5b : Variation du gain en fonction de la bande passante pour un filtre de Chebyshev d'ordre 5

II.10.2 : Gain de bruit minimal et performances des réalisations :

Bien que les structures optimales dans l'espace d'état donnent des gain de bruit de calcul minimal, leur réalisation pratique s'avèrent complexe lorsqu'il s'agit de filtres d'ordre n assez grand.

En effet, de telles structures requièrent $(n+1)^2$ multiplications pour calculer chaque échantillon de la sortie, ce qui constitue une augmentation de n^2 multiplications par rapport aux structures canoniques.

Pour réduire le nombre de multiplications dans les réalisations à bruit minimal, de multiples procédés ont été utilisés tel que :

- La décomposition d'un filtre d'ordre n en sections de structures optimales d'ordre 2 connectés en parallèle [20],[26], ou en cascade [3],[4],[5],[26] ce qui réduit le nombre de multiplications à (4n+1).
- Les structures à coefficients en puissance de 2 [1]
- Tridiagonalisation de la matrice A à (5n-1) multiplications [21]
- Tridiagonalisation de la structure d'état pour réduire les cycles limites [6]
- Les structures à faible bruit d'arrondi synthétisés par les algorithmes de Givens et de Housheulder [24]
- Structures d'état avec une rétroaction de l'erreur d'arrondi [27]

⋮

etc.

II.11 : Sensibilité aux coefficients :

Il est important dans les considérations pratiques de synthétiser des filtres dont la fonction de transfert présente une faible sensibilité aux variations de ses coefficients dues à la limitation de leur représentation binaire

Pour une fonction de transfert H(Z) décrite par le modèle d'état contrôlable et observable :

$$H(Z) = D + C(ZI - A)^{-1}B$$

le fait de représenter ses coefficients par des mots de longueur finies va faire dévier H(Z) des performances souhaitées.

Pour évaluer ces déviations dans l'espace d'état, on exprime les sensibilités de H(Z) par rapport à chaque composante des paramètres d'état individuellement par :

$$\begin{cases} Sa_{ij}(Z) = \dfrac{\partial H(Z)}{\partial a_{ij}} \\ Sb_{j}(Z) = \dfrac{\partial H(Z)}{\partial b_{j}} \\ Sc_{j}(Z) = \dfrac{\partial H(Z)}{\partial c_{j}} \end{cases}$$

que l'on peut écrire en fonction des fonctions de transfert F et G définies dans la section (II.8.2)

$$\begin{cases} Sa_{ij}(Z) = G_i(Z)F_j(Z) \\ Sb_j(Z) = G_i(Z) \\ Sc_j(Z) = F_j(Z) \end{cases}$$

où $F_i(Z)=[(Z\,I - A)^{-1}\,B]_i$ et $G_i(Z)=[C(Z\,I - A)^{-1}]_i$

et d'après les équations (II-33) et (II-37a), on définit une mesure de la sensibilité par :

$$M = Ma + Mb + Mc$$

avec

$$Mb = \frac{1}{2\pi j}\oint G^t(Z)G(Z)Z^{-1}dZ = tr(W) = \sum_{i=1}^{n} W_{ii}$$

$$Mc = \frac{1}{2\pi j}\oint F^t(Z)F(Z)Z^{-1}dZ = tr(K) = \sum_{i=1}^{n} K_{ii}$$

$$Ma = MbMc \geq \frac{1}{2\pi j}\oint \sum_{i=1}^{n}\sum_{j=1}^{n}\left|Sa_{ij}(Z)\right|^2 Z^{-1}dZ$$

$$Ma = tr(K)tr(W) = \sum_{i=1}^{n} K_{ii}\sum_{i=1}^{n} W_{ii}$$

Si l'on applique une transformation de normalisation telle que $tr(K')=n$

alors l'expression de la mesure de la sensibilité est :

$$M=(n+1)tr(W')+n$$

$$\boxed{M=(n+1)G'+n}$$

où G' est le gain de bruit de la nouvelle structure du filtre.

On remarque que la minimisation du gain de bruit G' entraîne celle de la mesure de la sensibilité M donc pour une structure à gain de bruit minimal nous avons :

$$\boxed{M = \frac{n+1}{n}\left[\sum_{i=1}^{n}\mu_i\right]^2 + n}$$

En conclusion, la minimisation du gain de bruit par une transformation T pour un filtre normalisé implique directement la minimisation de la mesure de la sensibilité.
En conséquence, une fonction de transfert calculée à partir des coefficients d'une structure optimale normalisée présentera le moins de distorsions possibles dans le cas de l'utilisation des registres de longueurs finies pour représenter les coefficients du filtre.

II.12 : Exemples d'expérimentation et résultats de simulation

Afin de comparer les performances de la structure minimale par rapport à la structure canonique, nous donnons quelques exemples de simulation d'un filtre numérique passe bas d'ordre 5 simulé sur ordinateur avec une arithmétique en virgule fixe . Toutes les simulations ont été faites avec le logiciel MATLAB 5.2 [32][33].

Exemple 1:

Nous prenons comme exemple le filtre de Butterworth de fréquence de coupure normalisée égale à 0.05
Le tableau II-1 donne les différentes valeurs de gain ainsi que la mesure de la sensibilité pour les deux types de structures considérées: canonique et minimale en fonction de l'ordre n du filtre.
Nous remarquons (tableau II-1 et Annexe) que pour un ordre élevé, le gain G et la mesure de la sensibilité M de la structure canonique sont très importants, alors que pour la structure minimale ils demeurent faibles.

Ordre du filtre :	**Structure Canonique**		**Structure minimale**	
n	**Gcan**	**Mcan**	**Gmin**	**Mmin**
4	9.03093	49.15468	0.55554	06.77770
6	247.72205	1740.05439	0.71213	10.98492
8	8019.68367	7215.15306	0.85972	15.73755
10	276762.16968	3044393.86655	1.00251	21.02765
12	9932605.91334	12912388.87354	1.14226	26.84946

Tableau II-1 : Gains de bruit G et sensibilité aux coefficients M pour un filtre de Butterworth (fréquence de coupure normalisée=0.05) en fonction l'ordre du filtre pour les structures canonique et minimale

Exemple 2 :

Une étude comparative entre les performances de la structure canonique et la structure minimale a été développée pour 3 types de filtres passe bas (Butterworth,, Chebyshev, et elliptique) d'ordre N=5 et de fréquence de coupure normalisée=0.05 (Tableau II-2).

Le signal d'entrée est la somme de 3 sinusoïdes d'amplitudes égales à l'unité et de fréquences : f1=15Hz, f2=40Hz et f3=50Hz (Fig. II-6a).

Le filtrage du signal d'entrée est fait en limitant la longueur des nombres représentant le signal de l'entrée, les coefficients du filtre et les registres de stockage et d'accumulation par nb bits. Cette quantification se fait par arrondi avec saturation. Pour évaluer les performances de chaque structure nous avons donné à la figure II-6b les sorties filtrées avec une précision qui est celle de l'ordinateur (c-à-d que l'on considère que le nombre de bits est infini). Ces trois sorties se superposent, ce qui revient a dire qu'elles sont identiques.

Pour chaque structure, sont présentés les effets de la longueur finie des registres sur la qualité du filtrage (figure II-7) et la sensibilité de la fonction de transfert (figure II-8) et donc de la réponse fréquentielle, à la précision des coefficients du filtre.

On effectue le filtrage pour un nombre de bits variant de 4 bits à 24 bits. La sortie pour ces cas est présentée par la figure II-7.

On constate que le filtrage est erroné lorsque le nombre de bits est :

- < à 18 bits → structure canonique
- < à 8 bits → structure minimale

La sinusoïde de la sortie est distordue et cela provient du fait que les coefficients deviennent de moins en moins représentables ;

On constate que lorsque le nombre de bits est :

- < à 12 bits → structure canonique
- < à 4 bits → structure minimale,

le signal en sortie est nul car tous les coefficients du numérateur de la fonction de transfert sont représentés par la valeur zéro.

Filtre de Butterworth		
Structures	Canonique	Minimale
Nb1	12	4
Nb2	18	8
Nb3	18	8
Filtre de Chebyshev (Ripple=0.2dB)		
Structures	Canonique	Minimale
Nb1	16	4
Nb2	20	10
Nb3	20	10
Filtre Elliptique (Rs=0.5dB, Rp=20dB)		
Structures	Canonique	Minimale
Nb1	14	4
Nb2	20	12
Nb3	20	12

Nb1 : Nombre de bits à partir duquel le filtre donne une sortie non nulle.

Nb2 : Nombre de bits à partir duquel le filtre donne une sortie où l'erreur est faible.

Nb3 : Nombre de bits pour lequel la fonction de transfert du filtre a une distorsion faible.

Tableau II-2: Comparaison des deux structures canonique, et minimale en fonction du nombre de bits pour trois types de filtres passe bas de fréquence de coupure normalisée égale à 0.05.

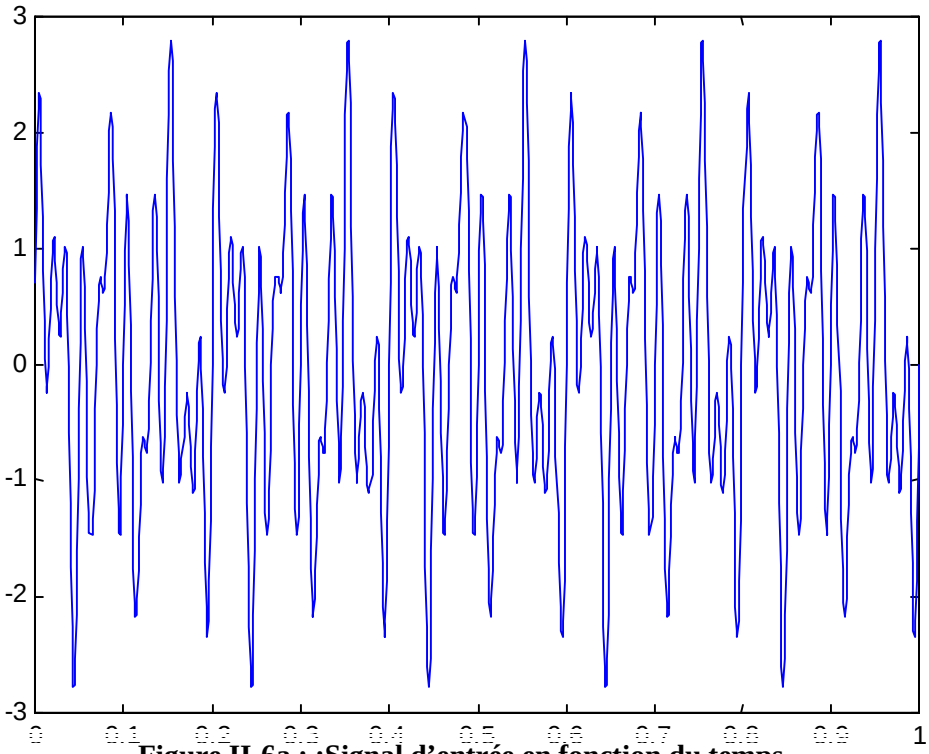


Figure II-6a : :Signal d'entrée en fonction du temps

e(k)=sin(2*pi*t*15)+sin(2*pi*t*40)+sin(2*pi*t*50)

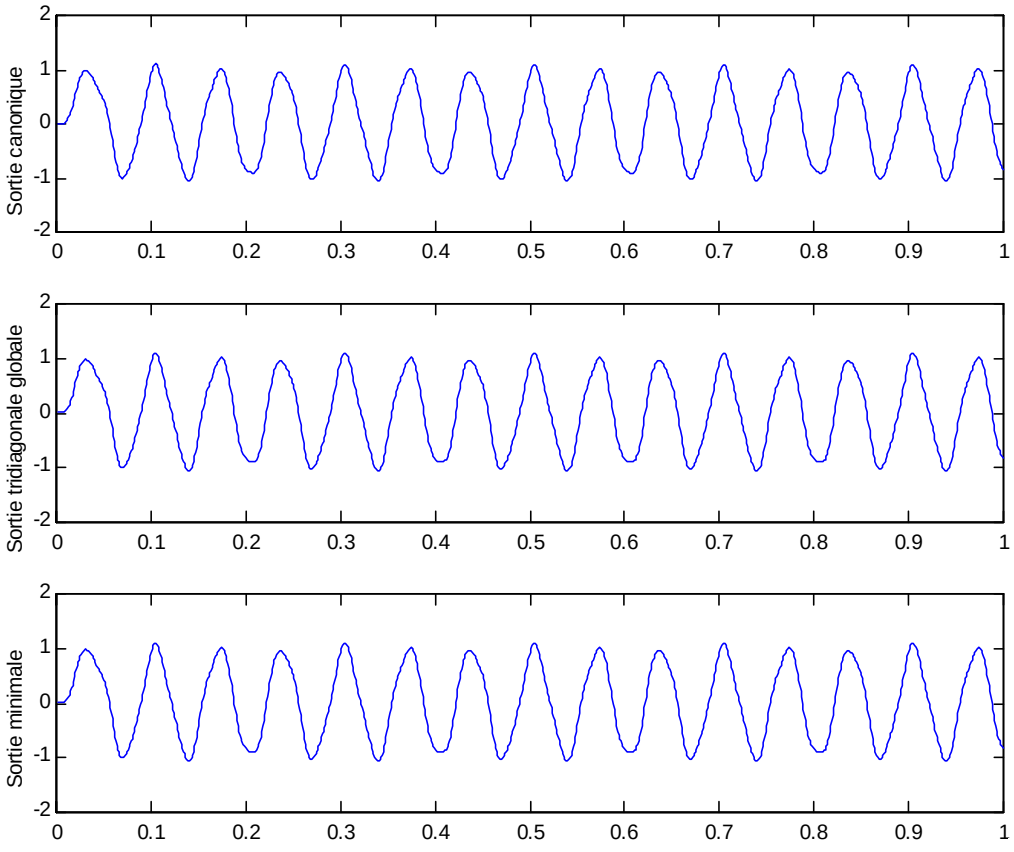
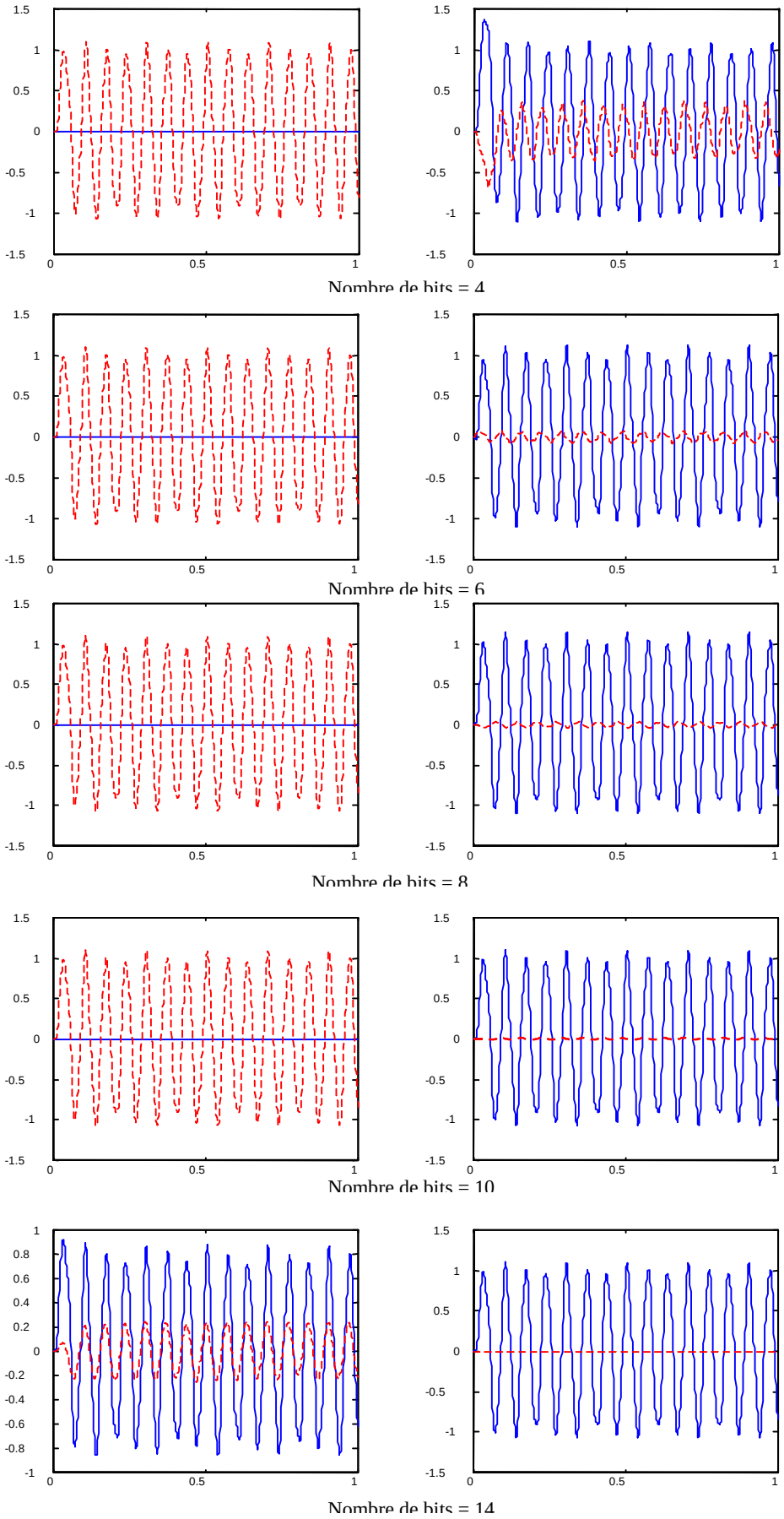
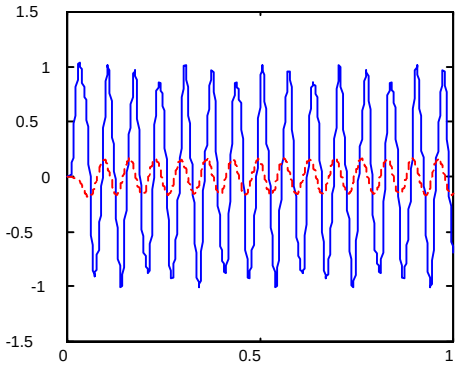


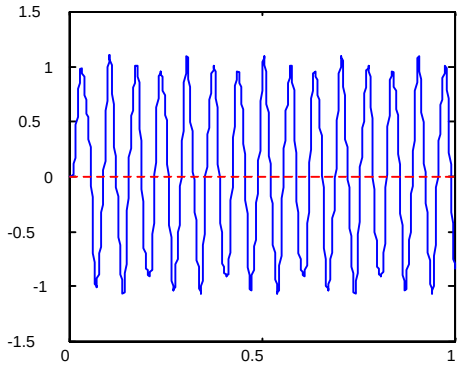
Figure II-6 b : Sortie filtrée pour les trois structures avec un nombre de bits infini



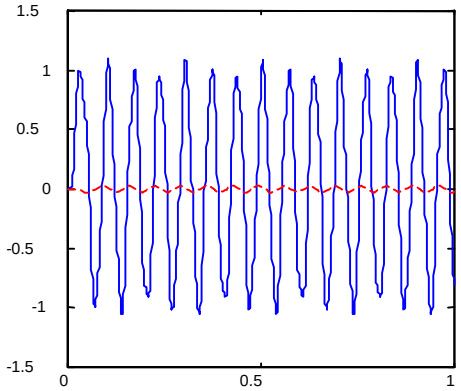
(a) Structure Canonique **(b) Structure Minimale**

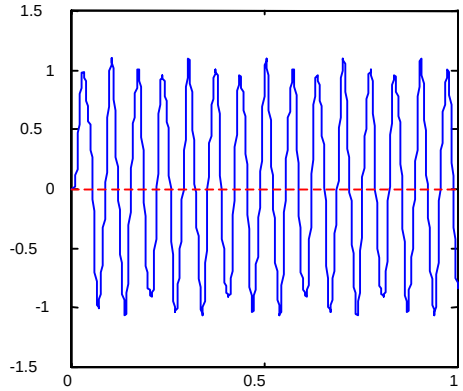
Figuree II-7a-: Sortie filtrée(___) et erreur d'arrondi (- -) par un Filtre de Butterworth d'ordre n=5 et de fréquence de coupure normalisée =0.05



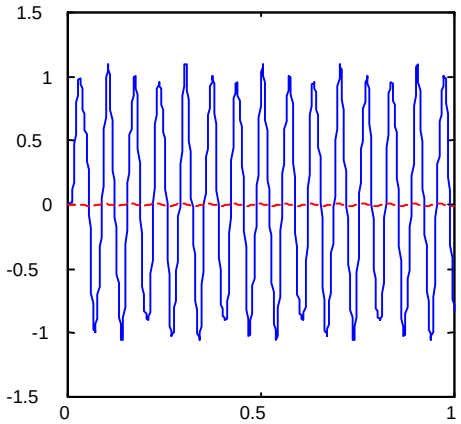


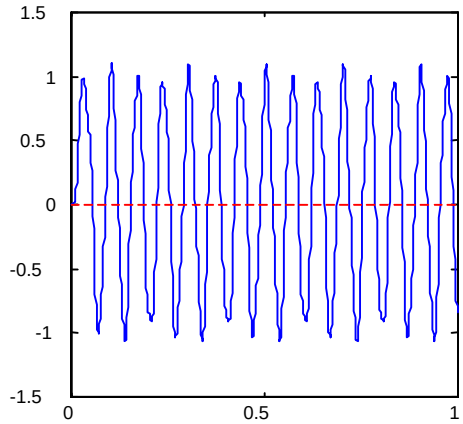
Nombre de bits = 16





Nombre de bits = 18



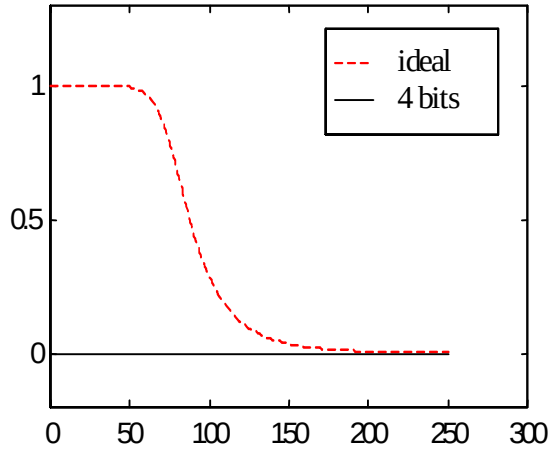


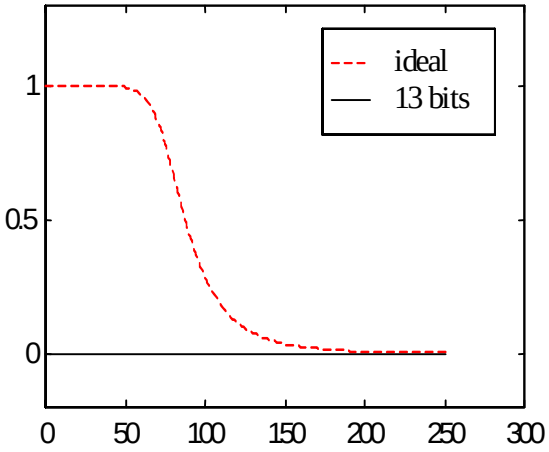
Nombre de bits = 20

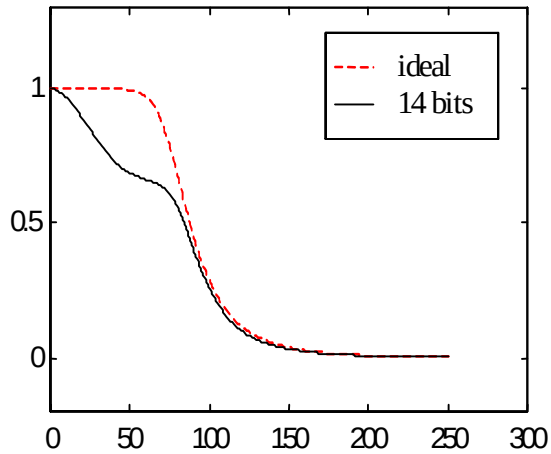
(a) Structure Canonique **(b) Structure Minimale**

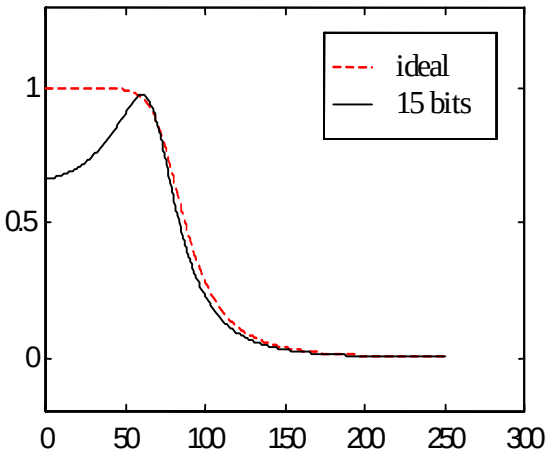
Figure II-7b: Sortie filtrée (___) et erreur d'arrondi (- - -) par un filtre de Butterworth d'ordre n=5 et de fréquence de coupure normalisée=0.05

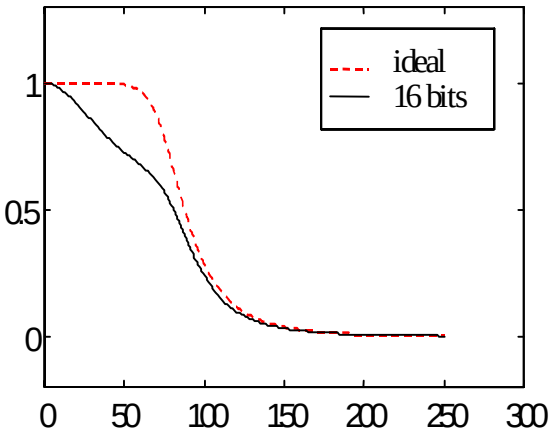
STRUCTURE CANONIQUE

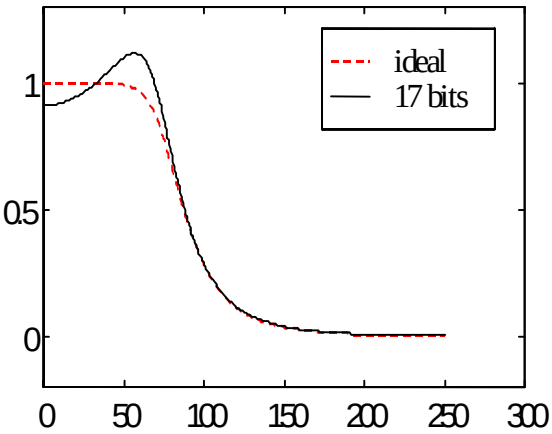


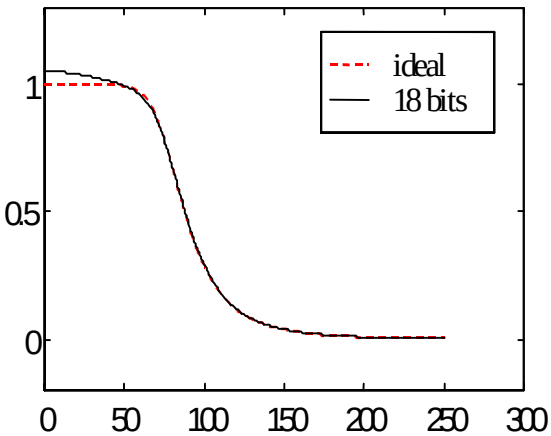


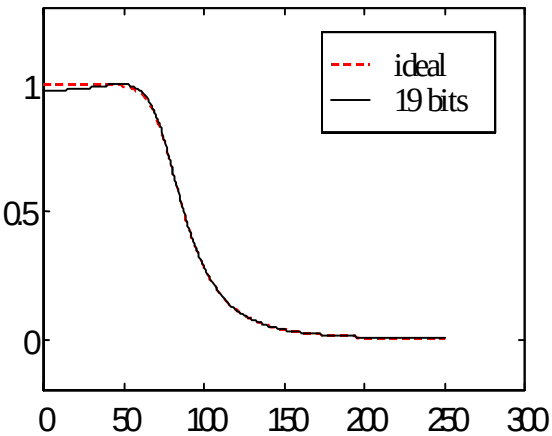




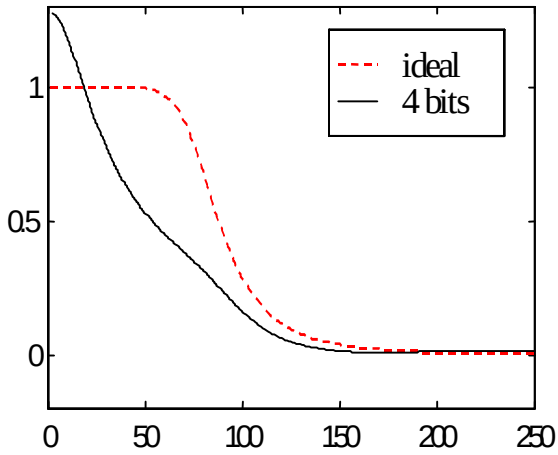


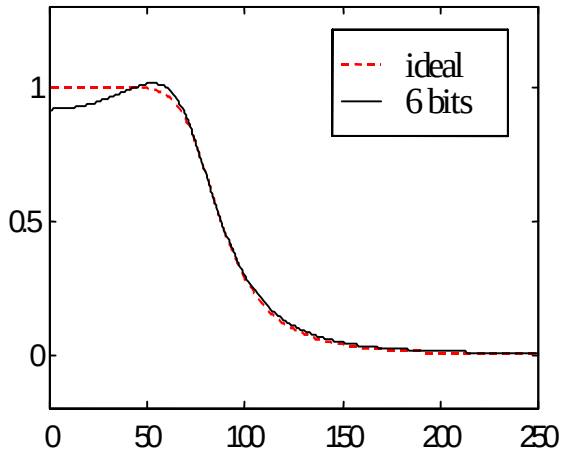


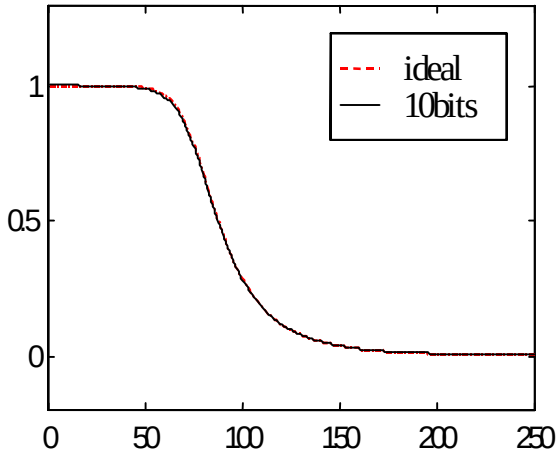


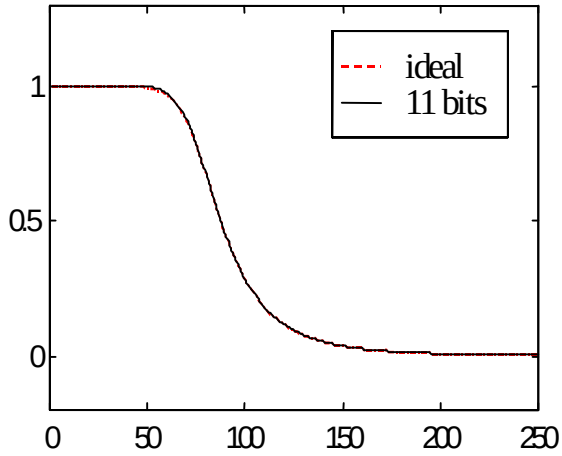


MINIMALE


ideal
4 bits
1
0.5
0
0
50
100
150
200
250


ideal
6 bits
1
0.5
0
0
50
100
150
200
250


ideal
10bits
1
0.5
0
0
50
100
150
200
250


ideal
11 bits
1
0.5
0
0
50
100
150
200
250

II.13. Conclusion

Ce chapitre a présenté quelques descriptions pour la conception des filtres numériques, et autres problèmes de traitement numérique du signal. Il est souvent nécessaire d'utiliser les description internes dans le but d'optimiser certain aspects du calcul. Les descriptions à variables d'état offrent une formulation mathématique qui permet l'étude des diverses structures, ce qui n'est pas faisable avec d'autres descriptions. L'importance des descriptions internes apparaît clairement dans l'étude de l'un des plus importants critères dans l'évaluation des structures, qui n'est autre que l'effet de la longueur finie des mots binaire sur les performances des filtres numériques. La minimisation de cet effet, comme nous l'avons vu dans ce chapitre, n'est possible que grâce à la représentation d'état qui permet de générer des structures optimales. De ce fait, nous utiliserons cette représentation d'état pour synthétiser une structure d'état ayant un gain de bruit et une complexité faibles. Il s'agit de la structure tridiagonale globale que nous allons découvrir au chapitre suivant.

CHAPITRE III

SYNTHESE DE STRUCTURES D'ETAT TRIDIAGONALES GLOBALES

CHAPITRE III

Synthèse de structures
d'état tridiagonales globales

III.1- Introduction

Dans le chapitre précédent, nous avons vu deux types de structures d'état de filtres numériques à réponse impulsionnelle infinie : la structure canonique et la structure minimale, ainsi que la relation qui existe entre les effets de la longueur finie des mots, la compléxité du filtre et la vitesse de calcul.

L'utilisation de l'espace d'état va nous permettre de trouver des structures de compromis qui auront une complexité et un gain de bruit faible par rapport à celui de la structure canonique.

Nous avons vu que pour une fonction de transfert donnée H(Z), nous pouvons trouver une infinité de structures d'état. Nous prenons en considération cette propriété pour trouver une structure d'état qui va non seulement avoir un gain d bruit de calcul inférieur à celui de la structure canonique mais de plus, cette structure sera modulaire (c-a-d faite avec des cellules identiques), et aura un nombre de multiplication inférieur à celui de la structure minimale. Nous avons pensé à tridiagonaliser la matrice bloc définissant une structure d'état donnée [25]. Cette opération est faite grâce a deux méthodes : l'une utilisant l'algorithme de Lanczos et l'autre utilisant les transformations élémentaires.

III-2 : Tridiagonalisation par l'algorithme de Lanczos :

Cet algorithme transforme une matrice A carrée quelconque en une matrice tridiagonale $\overline{A}$, en calculant une transformation T qui transforme A en $\overline{A}$ tel que

$$\overline{A} = T^{-1}AT \qquad \text{(III-1)}$$

Supposons que A est une matrice carrée (nxn), et $\overline{A}$ sa forme tridiagonale, alors il existe une transformation T non singulière tel que :

$$T^{-1}AT = \overline{A} = \begin{bmatrix} \alpha_1 & \gamma_1 & & & \\ \beta_1 & \alpha_2 & \ddots & 0 & \\ & \ddots & \ddots & \ddots & \\ & 0 & \ddots & \ddots & \gamma_{n-1} \\ & & & \beta_{n-1} & \alpha_n \end{bmatrix} \qquad \text{(III-2)}$$

On peut écrire T sous la forme :

$$T = [\psi(1). \quad . \quad . \quad \psi(n)] \qquad \text{(III-3a)}$$

et $$T^{-1} = [\rho(1). \quad . \quad . \quad \rho(n)]^t \qquad \text{(III-3b)}$$

où $\psi(i)$ et $\rho(i)$ sont la i^{ème} colonne et la i^{ème} ligne de T et T^{-1} respectivement

Alors

$$\rho(i)\psi(j) = \delta_{ij} = \begin{cases} 1 & si\ i = j \\ 0 & si\ i \neq j \end{cases} \qquad \text{(III-4)}$$

D'après l'équation (III-2), nous avons

$$AT = T\overline{A} \qquad \text{(III-5a)}$$

$$T^{-1}A = \overline{A}T^{-1} \qquad \text{(III-5b)}$$

de (III-5a), on peut écrire :

$$A[\psi(1)\cdots\psi(n)] = [\psi(1)\cdots\psi(n)]\begin{bmatrix} \alpha_1 & \gamma_1 & & & \\ \beta_1 & \alpha_2 & \ddots & 0 & \\ & \ddots & \ddots & \ddots & \\ & 0 & \ddots & \ddots & \gamma_{n-1} \\ & & & \beta_{n-1} & \alpha_n \end{bmatrix}$$

et d'après (III-5b), on a :

$$\begin{bmatrix} \rho(1) \\ \vdots \\ \rho(n) \end{bmatrix} A = \begin{bmatrix} \alpha_1 & \gamma_1 & & & \\ \beta_1 & \alpha_2 & \ddots & 0 & \\ & \ddots & \ddots & \ddots & \\ & 0 & \ddots & \ddots & \gamma_{n-1} \\ & & & \beta_{n-1} & \alpha_n \end{bmatrix} \begin{bmatrix} \rho(1) \\ \vdots \\ \rho(n) \end{bmatrix}$$

$$\Rightarrow \begin{cases} A\psi(1) = \alpha_1\psi(1) + \beta_1\psi(2) \\ A\psi(2) = \gamma_1\psi(1) + \alpha_2\psi(2) + \beta_2\psi(3) \\ A\psi(3) = \gamma_2\psi(2) + \alpha_3\psi(3) + \beta_3\psi(4) \\ \vdots \\ A\psi(n-1) = \gamma_{n-2}\psi(n-2) + \alpha_{n-1}\psi(n-1) + \beta_{n-1}\psi(n) \\ A\psi(n) = \gamma_{n-1}\psi(n-1) + \alpha_n\psi(n) \end{cases} \quad \text{(III-6a)}$$

et

$$\begin{cases} \rho(1)A = \alpha_1\rho(1) + \gamma_1\rho(2) \\ \rho(2)A = \gamma\beta_1\rho(1) + \alpha_2\rho(2) + \gamma_2\rho(3) \\ \rho(3)A = \beta_2\rho(2) + \alpha_3\rho(3) + \gamma_3\rho(4) \\ \vdots \\ \rho(n-1)A = \beta_{n-2}\rho(n-2) + \alpha_{n-1}\rho(n-1) + \gamma_{n-1}\rho(n) \\ \rho(n)A = \alpha_n\rho(n) + \gamma_{n-1}\rho(n) \end{cases} \quad \text{(III-6b)}$$

Nous pouvons réécrire ces deux systèmes d'équations comme :

$$\begin{cases} \beta_1\psi(2) = (A - \alpha_1 I)\psi(1) \\ \beta_2\psi(3) = (A - \alpha_2 I)\psi(2) - \gamma_1\psi(1) \\ \beta_3\psi(4) = (A - \alpha_3 I)\psi(3) - \gamma_2\psi(2) \\ \quad \vdots \\ \beta_{n-1}\psi(n) = (A - \alpha_{n-1} I)\psi(n-1) - \gamma_{n-2}\psi(n-2) \end{cases} \quad \text{(III-6c)}$$

et

$$\begin{cases} \gamma_1\rho(2) = \rho(1)(A - \alpha_1 I) \\ \gamma_2\rho(3) = \rho(2)(A - \alpha_2 I) - \gamma\beta_1\rho(1) \\ \gamma_3\rho(4) = \rho(3)(A - \alpha_3 I) - \gamma\beta_2\rho(2) \\ \vdots \\ \gamma_{n-1}\rho(n) = \rho(n-1)(A - \alpha_{n-1} I) - \gamma\beta_{n-2}\rho(n-2) \end{cases} \quad \text{(III-6d)}$$

d'où l'on déduit de (III-6a) et (III-6b) :

$$\begin{cases} A\psi(j) = \gamma_{j-1}\psi(j-1) + \alpha_j\psi(j) + \beta_j\psi(j+1) \\ \rho(j)A = \beta_{j-1}\rho(j-1) + \alpha_j\rho(j) + \gamma_j\rho(j+1) \\ \alpha_j = \rho(j)A\psi(j) \end{cases} \quad \text{(III-7a)}$$

et d'après (III-6c) et (III-6d), on a :

$$r_j = \beta_j \psi(j+1) = (A - \alpha_j I)\psi(j+1) - \gamma_{j-1}\psi(j-1)$$
$$P_j = \gamma_j \rho(j+1) = \rho(j)\left(A - \alpha_j I\right) - \gamma\beta_{j-1}\rho(j-1) \qquad \text{(III-7b)}$$

Le choix de β_j et γ_j est arbitraire. Nous pouvons par exemple prendre $\beta_j = \left\|r_j\right\|_2$ où β_j =1.

Alors $\gamma_j = P_j \psi(j+1)$

L'algorithme de Lanczos peut se résumer en quelques étapes :[22]

1) Initialiser les vecteurs $\psi(1)$ et $\rho(1)$ tel que :

$\rho(1)\ \psi(1)=1$ et $\|\psi(1)\|_2=1$

2) Pour j=1 à n-1 faire

$\alpha_j = \rho(j) A \psi(j)$

$r_j = (A - \alpha_j I)\psi(j+1) - \gamma_{j-1}\psi(j-1)$ avec $\gamma_0 \psi(0) = 0$

$\beta_j = \left\|r_j\right\|_2$

$\psi(j+1) = \dfrac{r_j}{\beta_j}$

$P_j = \rho(j)\left(A - \alpha_j I\right) - \beta_{j-1}\rho(j-1)$ avec $\beta_0 \rho(0) = 0$

$\gamma_j = P_j \psi(j+1)$

$\rho(j+1) = \dfrac{P_j}{\gamma_j}$

fin j

3) $\alpha_n = \rho(n) A \psi(n)$

fin algorithme.

Si nous avions choisi β_j =1 le processus de l'algorithme ne se serait pas modifié ; la seule étape à changer est de mettre β_j =1 au lieu de $\beta_j = \left\|r_j\right\|_2$.

III.2.1 : Structure d'état :

Si nous utilisons cet algorithme sur une structure d'état définie par le système d'équations donné en (II.8), où sur la matrice bloc S donnée par:

$$S = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

alors la nouvelle structure obtenue $\overline{S}$ sera tridiagonale et aura pour éléments les coefficients donnés en (III.2) tels que :

$$\overline{S} = \begin{bmatrix} T^{-1} & 0 \\ 0 & 1 \end{bmatrix} S \begin{bmatrix} T & 0 \\ 0 & T \end{bmatrix}$$

$$\overline{S} = \begin{bmatrix} \overline{A} & \overline{B} \\ \overline{C} & \overline{D} \end{bmatrix} = \begin{bmatrix} T^{-1}AT & T^{-1}B \\ CT & D \end{bmatrix} \qquad \text{(III.8)}$$

Le problème est de trouver T tel que :

$$\overline{S} = \begin{bmatrix} \alpha_1 & \gamma_1 & 0 & \cdots & \cdots & 0 \\ \beta_1 & \alpha_2 & \gamma_2 & \ddots & 0 & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & 0 & \ddots & \beta_{n-1} & \alpha_n & \gamma_n \\ 0 & \cdots & \cdots & 0 & \beta_n & D \end{bmatrix} \qquad \text{(III.8)}$$

cela implique que :

$$T^{-1}AT = \overline{A} = \begin{bmatrix} \alpha_1 & \gamma_1 & & & \\ \beta_1 & \alpha_2 & \ddots & 0 & \\ & \ddots & \ddots & \ddots & \\ & 0 & \ddots & \ddots & \gamma_{n-1} \\ & & & \beta_{n-1} & \alpha_n \end{bmatrix} \qquad \text{(III.9)}$$

$$T^{-1}B = \overline{B} = \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ \gamma_n \end{bmatrix} \qquad \text{(III.10)}$$

$$CT = \overline{C} = \begin{bmatrix} 0 & \cdots & \cdots & 0 & \beta_n \end{bmatrix} \qquad \text{(III.11)}$$

Sachant que T et T^{-1} sont donnés par (III.3), le calcul de $\overline{A}$ est fait de la même manière que précédemment.

Pour le calcul de $\overline{B}$ et $\overline{C}$, nous avons :

$$\overline{B} = T^{-1}B = \begin{bmatrix} \rho(1) \\ \vdots \\ \rho(n) \end{bmatrix} B = \begin{bmatrix} 0 & \cdots & 0 & \gamma_n \end{bmatrix} \qquad \text{(III.12)}$$

$$\overline{C} = C\begin{bmatrix} \psi(1) & \cdots & \psi(n) \end{bmatrix} = \begin{bmatrix} 0 & \cdots & 0 & \beta_n \end{bmatrix} \qquad \text{(III.13)}$$

Donc

$$\begin{bmatrix} \rho(1)B \\ \vdots \\ \rho(n)B \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \gamma_n \end{bmatrix} \qquad \text{et} \qquad \begin{bmatrix} C\psi(1) & \cdots & C\psi(n) \end{bmatrix} = \begin{bmatrix} 0 & \cdots & 0 & \beta_n \end{bmatrix}$$

d'où on obtient :

- $\rho(1)B = 0$ (III-14)

- $\rho(2)B = 0$

$$\rho(2) = \frac{1}{\gamma_1}\rho(1)\left(A - \alpha_1 I\right)$$

$$\Rightarrow \quad \rho(2)B = \frac{1}{\gamma_1}\rho(1)AB - \frac{\alpha_1}{\gamma_1}\underbrace{\rho(1)B}_{=0}$$

$$\Rightarrow \quad \rho(2)B = \frac{1}{\gamma_1}\rho(1)AB = 0$$

$$\rho(1)AB = 0 \qquad \text{(III.15)}$$

or $\rho(1)B = 0$ (III.14)

$$\Rightarrow \quad \rho(2)B = \frac{1}{\gamma_1}\rho(1)AB = 0$$

$$\Rightarrow \quad \rho(1)AB = 0 \qquad \text{(III.15)}$$

- $\rho(3)B = 0 \qquad et \qquad \rho(3) = \frac{1}{\gamma_2}\rho(2)\left(A - \alpha_2 I\right) - \frac{\beta_2}{\gamma_2}\rho(1)$

$$\Rightarrow \quad \rho(3)B = \frac{1}{\gamma_2}\rho(2)AB - \frac{\alpha_2}{\gamma_2}\underbrace{\rho(2)B}_{=0} - \frac{\beta_2}{\gamma_2}\underbrace{\rho(1)B}_{=0}$$

$$\Rightarrow \quad \rho(3)B = \frac{1}{\gamma_2}\rho(2)AB = 0$$

$$\Rightarrow \quad \rho(2)AB = 0 \qquad \text{(III.16)}$$

⋮
⋮
⋮

- $\rho(n)B = \gamma_n \qquad et \qquad \rho(n) = \frac{1}{\gamma_{n-1}}\rho(n-1)\left(A - \alpha_{n-1} I\right) - \frac{\beta_{n-2}}{\gamma_{n-1}}\rho(n-2)$

$$\Rightarrow \quad \rho(n)B = \frac{1}{\gamma_{n-1}}\rho(n-1)AB - \frac{\alpha_{n-1}}{\gamma_{n-1}}\underbrace{\rho(n-1)B}_{=0} - \frac{\beta_{n-2}}{\gamma_{n-1}}\underbrace{\rho(n-2)B}_{=0}$$

$$\Rightarrow \quad \rho(n)B = \frac{1}{\gamma_{n-1}}\rho(n-1)AB = \gamma_n$$

$$\Rightarrow \rho(n-1)AB = \gamma_n \cdot \gamma_{n-1} \qquad \text{(III.17)}$$

or

$$\rho(n-1)AB = \frac{1}{\gamma_{n-2}}\rho(n-2)A^2B - \frac{\alpha_{n-2}}{\gamma_{n-2}}\underbrace{\rho(n-1)AB}_{=0} - \frac{\beta_{n-3}}{\gamma_{n-3}}\underbrace{\rho(n-3)AB}_{=0}$$

d'où $\rho(n-1)AB = \frac{1}{\gamma_{n-2}}\rho(n-2)A^2B = \gamma_n \cdot \gamma_{n-1}$

alors $\rho(n-2)A^2B = \gamma_n \cdot \gamma_{n-1} \cdot \gamma_{n-2}$ (III.18)

Nous déduisons que :

$$\rho(1)A^{n-1}B = \gamma_n \cdot \gamma_{n-1} \cdot \gamma_{n-2} \cdots \gamma_1 = \prod_{i=1}^{n}\gamma_i$$

$$\text{d'où}\quad \begin{bmatrix}\rho(1)B\\ \rho(2)B\\ \vdots\\ \rho(n)B\end{bmatrix} = \rho(1)\begin{bmatrix}B & AB & \cdots & A^{n-1}B\end{bmatrix} = \begin{bmatrix}0 & \cdots & 0 & \prod_{i=1}^{n}\gamma_i\end{bmatrix} \qquad \text{(III.19)}$$

De la même façon pour l'équation (III.13), nous obtenons :

$$\begin{bmatrix}C\psi(1) & C\psi(2) & \cdots & C\psi(n)\end{bmatrix} = \begin{bmatrix}C\\ CA\\ \vdots\\ CA^{n-1}\end{bmatrix}\psi(1) = \begin{bmatrix}0\\ \vdots\\ 0\\ \prod_{i=1}^{n}\beta_i\end{bmatrix} \qquad \text{(III.20)}$$

- **Calcul de $\rho(1)$ et $\psi(1)$**

Pour trouver $\rho(1)$ et $\psi(1)$ nous avons à résoudre :

$$\rho(1)\mathcal{B} = \begin{bmatrix}0 & \cdots & 0 & \prod_{i=1}^{n}\gamma_i\end{bmatrix} \qquad \text{(III-21)}$$

et

$$\mathcal{C}\psi(1) = \begin{bmatrix}0\\ \vdots\\ 0\\ \prod_{i=1}^{n}\beta_i\end{bmatrix} \qquad \text{(III-22)}$$

avec $\mathcal{B}$ et $\mathcal{C}$: matrices de contrôlabilité et d'observabilité respectivement données par :

$$\mathcal{B} = \begin{bmatrix}B & AB & \cdots & A^{n-1}B\end{bmatrix} \qquad \text{(III-23)}$$

et

$$\mathcal{C} = \begin{bmatrix}C\\ CA\\ \vdots\\ CA^{n-1}\end{bmatrix} \qquad \text{(III-24)}$$

d'où

$$\rho(1) = \begin{bmatrix}0 & \cdots & 0 & \prod_{i=1}^{n}\gamma_i\end{bmatrix}\mathcal{B}^{-1} \qquad \text{(III-25a)}$$

et $$\psi(1) = \mathcal{C}^{-1}\begin{bmatrix} 0 \\ \vdots \\ 0 \\ \prod_{i=1}^{n}\beta_i \end{bmatrix} \qquad \text{(III-25b)}$$

on définit $\mathcal{B}^{-1}$ et $\mathcal{C}^{-1}$ par :

$$\mathcal{B}^{-1} = \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_n \end{bmatrix} \qquad et \qquad \mathcal{C}^{-1} = \begin{bmatrix} \xi_1 & \cdots & \xi_n \end{bmatrix} \qquad \text{(III-26)}$$

où η_i est la i^{ième} ligne de $\mathcal{B}^{-1}$ et ξ_i est la i^{ième} colonne de $\mathcal{C}^{-1}$

Nous avons donc d'après les équations (III-25) et (III-26) :

$$\rho(1) = \begin{bmatrix} 0 & \cdots & 0 & \prod_{i=1}^{n}\gamma_i \end{bmatrix}\begin{bmatrix} \eta_1 \\ \vdots \\ \vdots \\ \eta_n \end{bmatrix}$$

et $$\psi(1) = \begin{bmatrix} \xi_1 & \cdots & \cdots & \xi_n \end{bmatrix}\begin{bmatrix} 0 \\ \\ 0 \\ \prod_{i=1}^{n}\beta_i \end{bmatrix}$$

$$\Rightarrow \qquad \begin{cases} \rho(1) = \eta_n \prod_{i=1}^{n}\gamma_i \\ \psi(1) = \xi_n \prod_{i=1}^{n}\beta_i \end{cases} \qquad \text{(III-27)}$$

Sachant que d'après l'équation (III-4) on a : $\rho(1)\psi(1)=1$

Alors :

$$\left(\prod_{i=1}^{n}\gamma_i\right)\left(\prod_{i=1}^{n}\beta_i\right)\eta_n\xi_n = 1$$

$$\left(\prod_{i=1}^{n}\gamma_i\right) = \frac{1}{\eta_n\xi_n\left(\prod_{i=1}^{n}\beta_i\right)} \qquad \text{ou bien} \qquad \left(\prod_{i=1}^{n}\beta_i\right) = \frac{1}{\eta_n\xi_n\left(\prod_{i=1}^{n}\gamma_i\right)} \qquad \text{(III-28)}$$

Par conséquent, nous avons à initialiser soit les β_i ou les γ_i (avec i=1...n), pour pouvoir commencer la tridiagonalisation.

Supposons que l'on choisisse les β_i arbitrairement, alors nous pouvons calculer le terme produit donné par :

$$\left(\prod_{i=1}^{n}\gamma_i\right)=\frac{1}{\eta_n\xi_n\left(\prod_{i=1}^{n}\beta_i\right)}$$

Si nous posons :

$$P_\gamma=\left(\prod_{i=1}^{n}\gamma_i\right) \qquad et \qquad P_\beta=\left(\prod_{i=1}^{n}\beta_i\right) \tag{III-29}$$

alors $$P_\gamma=\frac{1}{\eta_n\xi_n P_\beta} \tag{III-30}$$

Donc connaissant P_γ, P_β, $\mathcal{B}$ et $\mathcal{C}$ nous pouvons résoudre le système d'équations donné par (III-25).

Nous pouvons voir que le seul fait de choisir les β_i (i=1...n) nous permet de définir complètement $\rho(1)$ et $\psi(1)$ avec bien sur la condition $\rho(1)\psi(1)=1$. Nous utilisons cette liberté de choix des β_i (i=1...n) pour optimiser la structure tridiagonale globale afin d'obtenir un bruit de calcul à la sortie du filtre faible.

III-2-2 : Gain de bruit de calcul à la sortie :

Nous avons vu précédemment, au chapitre II, que le gain de bruit de calcul pour une structure (A,B,C,D,K,W) est donné par :

$$G=\sum_{i=1}^{n}K_{ii}W_{ii}$$

Le gain de bruit de la structure tridiagonale globale est donc donné par :

$$G'=\sum_{i=1}^{n}K'_{ii}W'_{ii}$$

où $$K'=T^{-1}K\left(T^{-1}\right)^t \qquad et \qquad W'=T^tWT$$

d'où d'après (III-3) nous pouvons écrire :

$$K'=\rho(i)K\rho^t(i) \quad \text{et} \quad W'=\psi^t(i)W\psi(i) \tag{III-32}$$

d'où $$G'=\sum_{i=0}^{n}\left[\rho(i)K\rho^t(i)\right]\left[\psi^t(i)K\psi(i)\right] \tag{III-33}$$

Nous remarquons que le gain de bruit de la structure tridiagonale globale dépend des $\rho(i)$ et $\psi(i)$ (avec i = 1,...,n) qui eux même dépendent des paramètres β_i (i=1...n).
Nous pouvons dire alors que le gain de bruit est une fonction des paramètres β, c'est à dire ::

$$G' = f(\beta_1,..., \beta_n) \qquad \text{(III-34)}$$

et de ce fait l'optimisation des paramètres β_i va nous permettre d'optimiser le gain de bruit G'

III-2-3 : Optimisation du gain de bruit de la structure tridiagonale globale :

Nous donnons ci après la procédure de tridiagonalisation par l'algorithme de Lanczos avec optimisation du gain de bruit.

Algorithme :

1) Démarrer avec une structure d'état (A,B,C,D) canonique où minimale.
2) Calculer K et W : Matrices de covariance et de bruit respectivement.
3) Calculer $\mathcal{B}$ et $\mathcal{C}$ matrices de contrôlabilité et d'observabilité respectivement.
4) Calcul $\mathcal{B}^{-1}$ et $\mathcal{C}^{-1}$
5) Initialiser $\beta^{(0)} = [\beta_1,.....,\beta_n]^t$
6) - Calculer $P_\beta^{(k)} = \prod_{i=1}^{n} \beta_i$

   - Calculer $P_\gamma^{(k)} = \dfrac{1}{P_\beta^{(k)} \eta_n \xi_n}$, où ξ_n :dernière colonne de $\mathcal{C}^{-1}$

     et η_n : dernière ligne de $\mathcal{B}^{-1}$

   - Calculer $\psi^{(k)}(1) = \xi_N \prod_{i=1}^{N} \beta_i$ et $\rho^{(k)}(1) = \eta_N \left(\prod_{i=1}^{N} \gamma_i \right)$

   - **Exécuter l'algorithme de Lanczos :**
     * Pour j allant de 1 jusqu'à n-1 faire :
     - $\alpha_j = \rho(j) A \psi(j)$
     - $r_j = (A - \alpha_j I)\,\psi(j) - \gamma_{j-1}\,\psi(j-1)$; $\gamma_0\,\psi(0) \equiv 0$
     - $\psi(j+1) = r_j / \beta_j$
     - $P_j = \rho(j)(A - \alpha_j I) - \beta_{j-1}\psi(j-1)$; $\beta_0 \psi(0) \equiv 0$

- $\gamma_j = \mathrm{P}_j\, \psi(j+1)$
- $\rho(j+1) = \mathrm{P}_j / \gamma_j$

∗ fin j

- Calculer $\alpha_n = \rho(n) A \psi(n)$ et $\gamma_n = \dfrac{P_\gamma^{(k)}}{\prod_{i=1}^{n-1} \gamma_i}$

- Calculer le gain $G'_{(k)} = \sum_{i=1}^{n} \left([\rho(i)^{(k)} K \rho^t(i)^{(k)}] [\psi^t(i)^{(k)} W \psi(i)^{(k)}] \right)$

- Minimisation du gain de bruit :

  Si $\left| G'_{(k)} - G'_{(k-1)} \right| / G'_{(k)} < \varepsilon$ Alors Exit

  Sinon : Actualiser $\beta^{(k)} = [\beta_1, \ldots\ldots, \beta_N]^t$ en utilisant la méthode de Newton-Raphson [28] [34], et refaire depuis l'étape (6)

Fin de l'algorithme

III-3 : Tridiagonalisation par les transformations élémentaires :

Une autre méthode de tridiagonalisation est utilisée pour synthétiser des structures d'état de compromis. Nous utilisons des transformations élémentaires qui nous permettent de synthétiser un nombre important de structures d'état.

III-3-1 : Transformations élémentaires :[21][22]

Nous appelons transformation élémentaire, une matrice qui a des « 1 » sur la diagonale et un élément non nul en dehors de la diagonale, et des « 0 » partout ailleurs. Elle a la forme suivante :

$$P(\alpha, i, j) = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & \alpha & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 1 \end{bmatrix} \begin{matrix} \\ \rightarrow \text{ i}^{\text{éme}} \text{ ligne} \\ \\ \\ \\ \downarrow \\ \text{j}^{\text{iéme}} \text{ colonne} \end{matrix} \qquad \text{(III-42)}$$

Nous pouvons l'écrire sous la forme suivante :

$$P(\alpha, i, j) = I + \alpha\, e^i (e^j)^t \qquad \text{(III-43)}$$

Où I : est la matrice identité,

t : désigne la transposée

$e^i = (0 \quad \cdots \quad 0 \quad 1 \quad 0 \quad \cdots \quad 0)$

$\uparrow$ i$^{\text{iéme}}$ élément

Ce type de transformation nous permet d'avoir de nouvelles structures d'état. Ces transformations peuvent être utilisées comme moyen d'élimination de coefficients de structures d'état et ceci dans le but de simplifier des structures complexes.

Nous donnons dans ce qui suit quelques propriétés importantes concernant ces transformations élémentaires [21],[22] :

1. $P^{-1}(\alpha, i, j) = P(-\alpha, i, j)$
2. $P^t(\alpha, i, j) = P(\alpha, j, i)$
3. $\det P(\alpha, i, j) = 1 \quad \forall\, i, j \;\; (i \neq j)$
4. $P^k(\alpha, i, j) = P(k\alpha, i, j) \quad ; \quad k \in Z$

La méthode pour obtenir de nouvelles structures consiste à appliquer successivement des transformations élémentaires à une structure d'état (A,B,C,D) d'un filtre donné. Durant ce procédé, une transformation $P(\alpha, i, j)$ va modifier la i$^{\text{ième}}$ ligne et la j$^{\text{ième}}$ colonne de A, le i$^{\text{ième}}$ élément de B et le j$^{\text{ième}}$ élément de C.
Nous exploitons ces modifications dans A, B, et C pour éliminer le coefficient voulu par un choix approprié du paramètre α. Cette élimination va s'accomplir sans altérer les éléments qui ont déjà été éliminés.

III-3-2 : Application des transformations élémentaires

Soit une structure d'état définie par A, B, C, D, K, et W donnés par :

$$A=\begin{bmatrix} a_{11} & a_{12} & \cdots & & a_{1n} \\ a_{21} & a_{22} & \cdots & & a_{2n} \\ \vdots & & \ddots & a_{ij} & \vdots \\ & & & & \\ a_{n1} & a_{n2} & \cdots & & a_{nn} \end{bmatrix} \quad , \quad B=\begin{bmatrix} b_1 \\ \vdots \\ b_i \\ \vdots \\ b_n \end{bmatrix} \quad , \quad C=\begin{bmatrix} c_1 & \cdots & c_j & \cdots & c_n \end{bmatrix} \qquad \text{(III-44a)}$$

$$K=\begin{bmatrix} k_{11} & k_{12} & \cdots & & k_{1n} \\ k_{21} & k_{22} & \cdots & & k_{2n} \\ \vdots & & \ddots & k_{ij} & \vdots \\ & & & & \\ k_{n1} & k_{n2} & \cdots & & k_{nn} \end{bmatrix} \quad , \quad W=\begin{bmatrix} w_{11} & w_{12} & \cdots & & w_{1n} \\ w_{21} & w_{22} & \cdots & & w_{2n} \\ \vdots & & \ddots & w_{ij} & \vdots \\ & & & & \\ w_{n1} & w_{n2} & \cdots & & w_{nn} \end{bmatrix} \qquad \text{(III-45b)}$$

Si l'on applique la transformation donnée par les équations (III-42) et (III-43), alors les éléments de la structure d'état transformée (A',B',C',K',W') sont comme suit :

$$A'=\begin{bmatrix} a_{11} & \cdots & a_{1j}+\alpha\, a_{1i} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots & & \vdots \\ a_{i1}-\alpha\, a_{j1} & \cdots & -\alpha^2 a_{ji}+\alpha(a_{ii}-a_{jj})+a_{ij} & \cdots & a_{in}-\alpha\, a_{jn} \\ & & \vdots & & \\ a_{n1} & \cdots & a_{nj}+\alpha\, a_{ni} & \cdots & a_{nn} \end{bmatrix} \qquad \text{(III-46a)}$$

$$B'=\begin{bmatrix} b_1 \\ \vdots \\ b_i-\alpha\, b_j \\ \vdots \\ b_n \end{bmatrix} \leftarrow \text{i}^{\text{ième}}\text{élément} \quad ; \quad C'=\begin{bmatrix} c_1 & \cdots & c_j+\alpha\, c_i & \cdots & c_n \end{bmatrix} \qquad \text{(III-46b)}$$

$\uparrow$ j$^{\text{ième}}$ élément

$$K'=\begin{bmatrix} k_{11} & \cdots & k_{1i}-\alpha\, k_{1j} & \cdots & k_{1n} \\ \vdots & \ddots & \vdots & & \vdots \\ k_{i1}-\alpha\, k_{j1} & \cdots & \alpha^2 k_{jj}-2\alpha\, k_{ij}+k_{ii} & \cdots & k_{in}-\alpha\, k_{jn} \\ & & \vdots & & \\ k_{n1} & \cdots & k_{ni}-\alpha\, k_{nj} & \cdots & k_{nn} \end{bmatrix} \leftarrow \text{i}^{\text{ième}}\text{ligne} \qquad \text{(III-46c)}$$

$\uparrow$ i$^{\text{ième}}$ colonne

$$W' = \begin{bmatrix} w_{11} & \cdots & w_{1j} + \alpha\, w_{1i} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots & & \vdots \\ w_{j1} + \alpha\, w_{j1} & \cdots & \alpha^2 k_{ii} + 2\alpha\, w_{ij} + w_{jj} & \cdots & w_{jn} + \alpha\, w_{in} \\ & & \vdots & & \\ w_{n1} & \cdots & w_{nj} + \alpha\, k_{ni} & \cdots & w_{nn} \end{bmatrix} \leftarrow \text{j}^{\text{ième}}\text{ligne}$$

$$\uparrow \text{j}^{\text{ième}}\text{ colonne} \qquad \text{(III-46d)}$$

A partir de ces équations nous remarquons que l'application d'une transformation P(α,i,j) a (A,B,C,K,W) introduit les modifications suivantes à :

1. la i$^{\text{ième}}$ ligne et la j$^{\text{ième}}$ colonne de A
2. le i$^{\text{ième}}$ élément de B
3. le j$^{\text{ième}}$ élément de C
4. la i$^{\text{ième}}$ ligne et la i$^{\text{ième}}$ colonne de K
5. la j$^{\text{ième}}$ ligne et la j$^{\text{ième}}$ colonne de W

Alors, en choisissant une position (i,j) appropriée du coefficient α dans la transformation P(α,i,j), nous pouvons éliminer n'importe quel élément de A,B, ou C. C'est à dire, et plus clairement, l'élimination d'un coefficient dans (A,B,C) peut se faire en résolvant des équations de premier et/ou second ordre.

Par exemple, si l'on veut annuler , l'élément a'$_{1j}$ de A' donné par (III-46a), α sera la solution de l'équation de premier ordre suivante :

$$a_{1j} + \alpha a_{1i} = 0 \qquad \text{(III-47)}$$

qui implique :

$$\alpha = -\frac{a_{1j}}{a_{1i}} \qquad \text{(III-48)}$$

Ceci n'est possible que si : $a_{1i} \neq 0$

Mais si l'élément a éliminer est a'$_{ij}$ alors α doit satisfaire l'équation du second ordre suivante :

$$-\alpha^2 a_{ji} + \alpha\left(a_{ii} - a_{jj}\right) + a_{ij} = 0 \qquad \text{(III-49)}$$

qui implique :

$$\alpha = \frac{\left(a_{ii} - a_{jj}\right) \pm \sqrt{\left(a_{ii} - a_{jj}\right)^2 + 4a_{ij}a_{ji}}}{2a_{ji}} \qquad \text{(III-50)}$$

sous les conditions :

$$\left(a_{ii} - a_{jj}\right)^2 \geq -4a_{ij}a_{ji} \quad \text{et} \quad a_{ji} \neq 0 \qquad \text{(III-51)}$$

III-3-3 Optimisation du gain de bruit :

Le problème du bruit d'arrondi est un des plus importants effets parmi ceux de la longueur finie des registres. Pour cela nous prenons une attention particulière pour le minimiser.
En utilisant les transformations élémentaires, l'optimisation du gain de bruit peut être faite en appliquant une transformation de la forme :

$$T(\beta, i, j) = I + \beta e^i \left(e^j\right)^t \qquad \text{(III-52)}$$

où I : matrice identité et β : paramètre

avant le processus d'élimination. Le gain de bruit résultant dépendra du paramètre β et de sa position (i,j). Puisque le paramètre α dépend des éléments de A, B, C, alors l'expression du gain de bruit est une fonction de β. Nous montrons à la figure (III-1) la procédure d'optimisation du gain de bruit en utilisant la méthode des transformation élémentaires.

La modularité de cette structure tridiagonale globale est bien visible à la figure (III-2). Nous voyons bien qu'elle est formée par des cellules régulières et ce qui pour effet d'augmenter la vitesse de calcul. De plus cette structure est moins complexe que la structure à gain de bruit minimal, car elle possède (3n+1) multiplications.

Si une transformation élémentaire P(α,i,j) est appliquée à une structure d'état (A,B,C,D), alors d'après les équations de K' et W' données en (III-46c) et (III-46d), le gain de bruit devient :

$$G' = G + 2\alpha^2 w_{ii} k_{jj} + 2\alpha\left(w_{ij} k_{jj} - w_{ii} k_{ij}\right) \qquad \text{(III-53)}$$

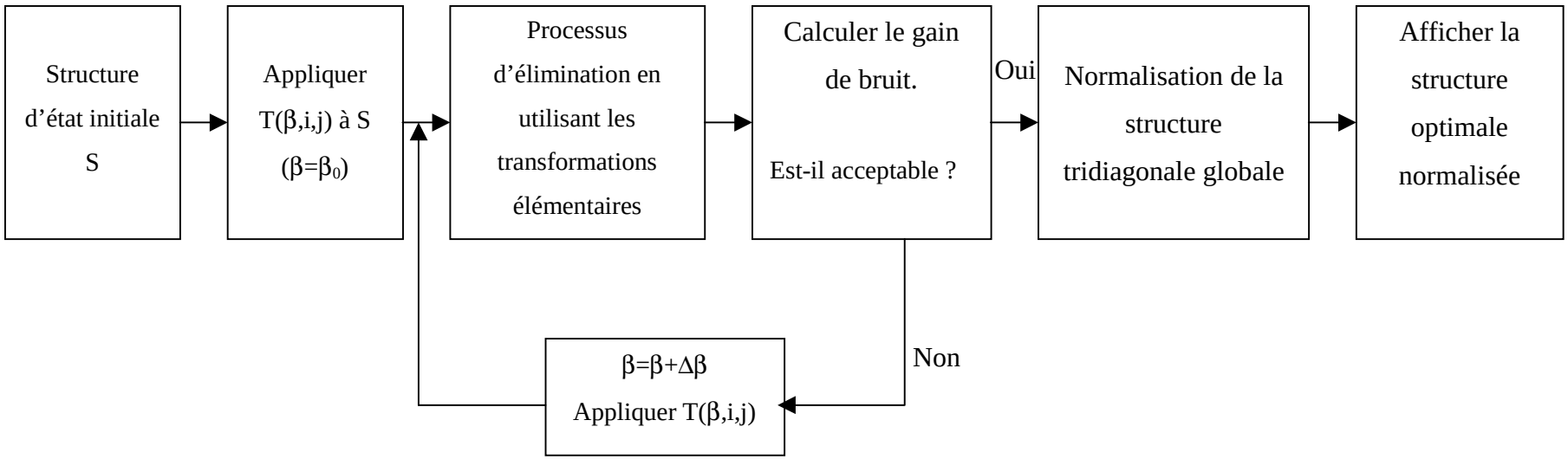


Figure III.1 : Processus d'élimination et optimisation du gain de bruit en utilisant la méthode des transformations élémentaires

où G est le gain de la structure (A,B,C,D).

Pour avoir le minimum du gain de bruit, il suffit que :

$$\frac{dG'}{d\alpha} = 0$$

ce qui donne comme valeur optimum de α :

$$\alpha = \frac{w_{ii}k_{ij} - w_{ij}k_{jj}}{2w_{ii}k_{jj}} \qquad \text{(III-54)}$$

La synthèse de la structure tridiagonale globale par la méthode des transformations élémentaires est donnée par l'algorithme suivant :

Algorithme d'élimination des coefficients par les transformations élémentaires:

Etape 1 : ***Initialisation***

Initialiser P à l'identité (P = I)

Etape 2 : ***Elimination des éléments du vecteur C***

Si c_n =0 alors aller à l'étape 3

Pour i = 1,..,n-1 faire

$\alpha(n,i) = -c_i / c_n$

Appliquer P(α,n,i) à A, B, C, K, et W

$\alpha(n,i) = 0$

fin i

Etape 3 ***: Elimination des éléments inférieurs à la sous diagonale de A***

Pour i = 2,…,n-1 faire

m=n-i+1

Pour j = 1,…,n-i faire

Si $a_{m+1,j} \neq 0$ alors $\alpha(m,j) = -a_{m+1,j} / a_{m+1,m}$

Appliquer P(α,m,j) à A, B, C, K, et W

$\alpha(m,j) = 0$

fin si

fin j

fin i

Etape 4 : ***Elimination des éléments du vecteur B***

Si $b_n = 0$ aller à l'étape 5

Pour i=1,n-1 faire

$\alpha(i,n) = -b_i / b_n$

Appliquer P(α,i,n) à A, B, C, K, et W

$\alpha(i,n) = 0$

fin i

Etape 5 *:* ***Elimination des éléments supérieurs à la sus diagonale de A***

L'élimination des éléments se fait colonne par colonne de la droite vers la gauche

Pour j=2,n-1 faire

M=n-j+1

Pour i=1,…,n-j faire

Si $a_{i,m+1} \neq 0$ alors $\alpha(i,m) = -a_{i,m+1} / a_{m,m+1}$

Appliquer P(α,i,m) à A, B, C, K, et W

$\alpha(i,m) = 0$

fin si

fin i

fin j

Fin de l'algorithme

Cet algorithme nous synthétise une structure tridiagonale globale. L'optimisation du gain de bruit est faite en utilisant la méthode décrite dans la figure (III-1).

Il ne faut pas oublier qu'a la fin du processus d'élimination des coefficients et après l'optimisation du gain de bruit, la transformation totale est :

$$Q = T \prod_{i=1}^{k} P_k$$

où k est le nombre de coefficients éliminés pour obtenir la structure tridiagonale globale

et P_i (i=1,…,k) les différentes transformations élémentaires appliquées à chaque étape de l'algorithme.

Remarques :

1) Les deux algorithmes proposés (Lanczos et transformations élémentaires) ont permis d'aboutir à la même structure tridiagonale globale dont le graphe de fluence est présenté en figure III-2.

2) Les différentes étapes pour synthétiser chacune des quatre structures d'état ci-après pour un filtre numérique d'ordre n sont résumées dans un diagramme (figure III-3) :

- ✓ La structure canonique
- ✓ La structure minimale
- ✓ La structure tridiagonale
- ✓ La structure tridiagonale globale

III.4 : Exemples d'expérimentations et résultats de simulations

Exemple 1 :

Nous avons simulé un filtre passe-bas de Butterworth de fréquence de coupure normalisée égale à 0.05 pour les 3 structures comparées : canonique, tridiagonale globale, et minimale. Le Tableau III-1 regroupe les valeurs de gains obtenus en fonction de l'ordre n du filtre [25].

Structures	**Ordre du filtre**		
	n=3	**n=5**	**n=6**
Canonique	1096,4186	19660017,8	2832520351,7
Tridiagonale globale	48,848631	103,671266	673,14589
Minimale	0,470495	0,63541415	0,7121316

Tableau IV-1 : Gains de bruit d'un filtre passe bas de Butterworth et de fréquence de coupure normalisée égale à 0.05

Ces valeurs montrent que, quelque soit l'ordre n du filtre, la structure tridiagonale globale donne une valeur intermédiaire du gain.

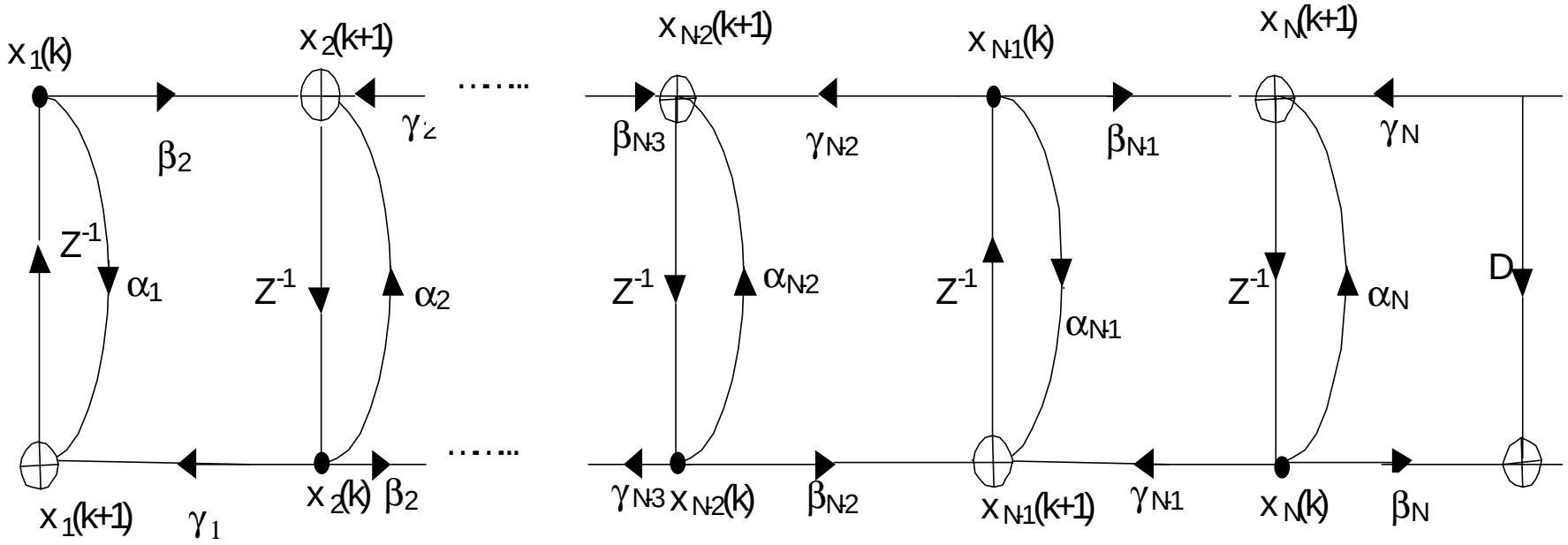


Figure IV-2: Graphe de fluence de la structure tridiagonale globale

$X_i(k)$ = vecteur d'état

α_i , β_i et γ_i = Coefficients de la structure tridiagonale globale

N :Ordre du filtre

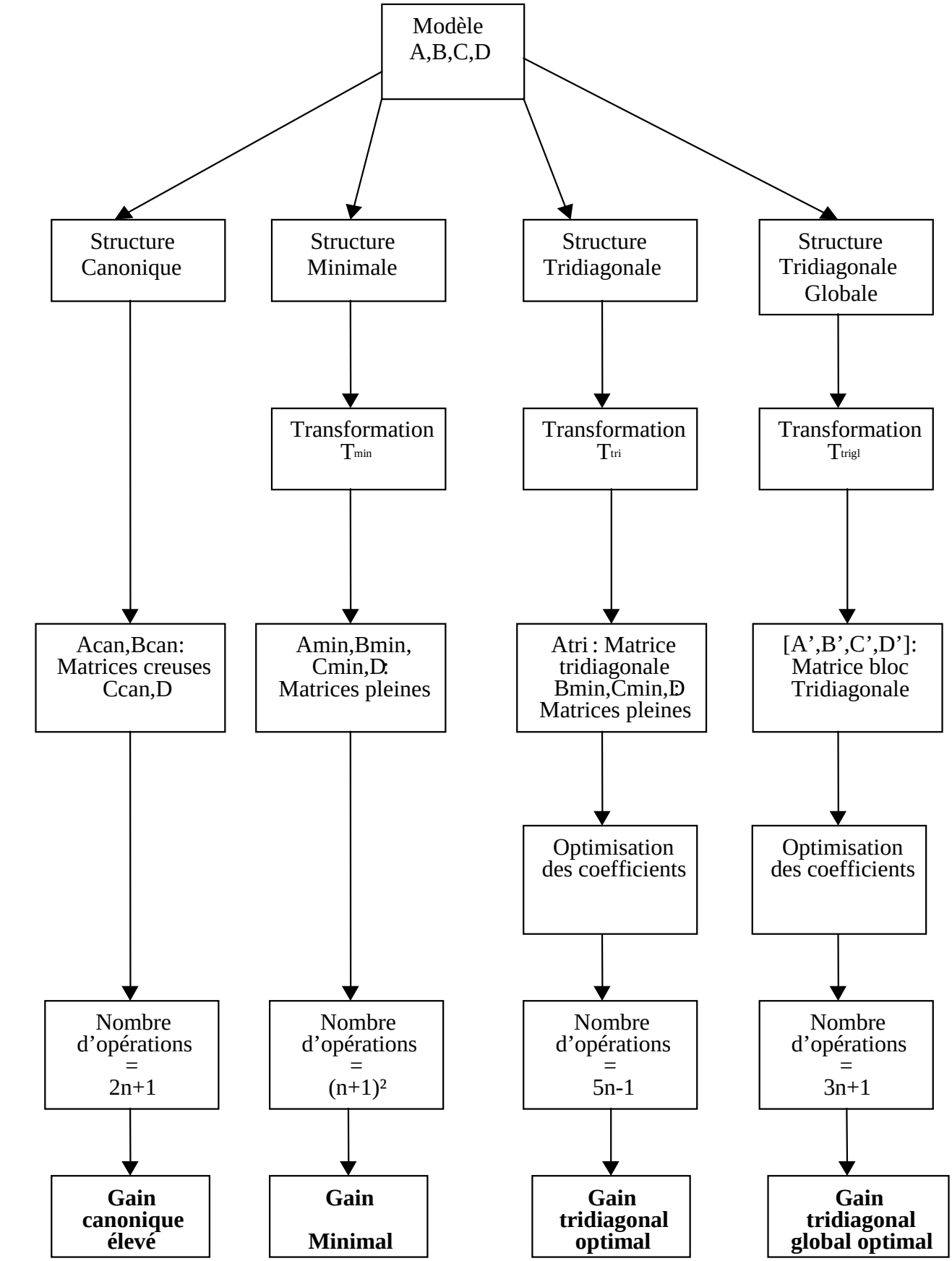


Figure III-3 : Diagramme de synthése des differentes structures de filtres numériques

La figure III-4 montre les variations du gain de bruit G' d'un filtre passe bas de Butterworth en fonction de l'ordre n et de la fréquence de coupure fc réalisé avec une structure tridiagonale globale [25].

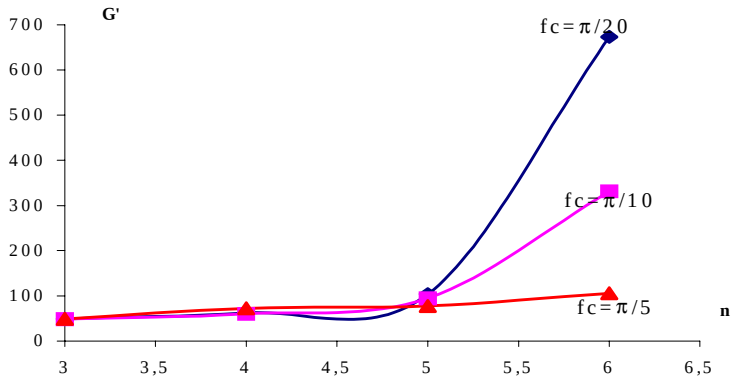


Figure III-4 : Gain de bruit de la structure tridiagonale globale en fonction de la fréquence de coupure fc et de l'ordre n pour un filtre passe bas de Butterworth

La figure III-4 montre que:

1) le gain de bruit G' de la structure tridiagonale globale est une fonction croissante de l'ordre n du filtre, quelque soit la fréquence de coupure fc,
2) plus la bande passante est étroite, plus le gain de bruit d'arrondi G' augmente.

Exemple 2 :

Nous prenons le même exemple de simulation que l'exemple 2 du chapitre II pour pouvoir faire une étude comparative entre les trois structures d'état à savoir : les structures canonique, minimale et tridiagonale globale.

Le même signal d'entrée donné précédemment à la figure (II-6a) sera filtré par 3 filtres passe bas, (Butterworth, Chebyshev, et elliptique), d'ordre 5 et de fréquence de coupure normalisée égale à 0.05 . Nous reprenons les résultats du chapitre II pour mieux comparer les résultats de simulation (Tableau III-2).

On effectue le filtrage pour un nombre de bits variant de 4 bits à 24 bits. La sortie pour ces cas est présentée par la figure (III-5) pour le filtre de Butterworth d'ordre 5.

Filtre de Butterworth			
Structures	Canonique	**Tridiagonale globale**	Minimale
Nb1	12	**10**	4
Nb2	18	**13**	8
Nb3	18	**13**	8
Filtre de Chebyshev (Ripple=0.2dB)			
Structures	Canonique	**Tridiagonale globale**	Minimale
Nb1	16	**12**	4
Nb2	20	**15**	10
Nb3	20	**15**	10
Filtre Elliptique (Rs=0.5dB, Rp=20dB)			
Structures	Canonique	**Tridiagonale globale**	Minimale
Nb1	14	**10**	4
Nb2	20	**16**	12
Nb3	20	**16**	12

Tableau III-2: Comparaison des trois structures canonique, minimale et tridiagonale globale en fonction du nombre de bits pour trois types de filtres d'ordre 5

Où : Nb1 : Nombre de bits à partir duquel le filtre donne une sortie non nulle.

Nb2 : Nombre de bits à partir duquel le filtre donne une sortie où l'erreur est faible.

Nb3 : Nombre de bits pour lequel la fonction de transfert du filtre a une distorsion faible.

On constate que le filtrage est erroné lorsque le nombre de bits est :

- < à 18 bits → structure canonique
- < à 13 bits → structure tridiagonale globale
- < à 8 bits → structure minimale

cela provient du fait que les nombres deviennent de moins en moins représentables à cause de la longueur finie des registres de stockage.

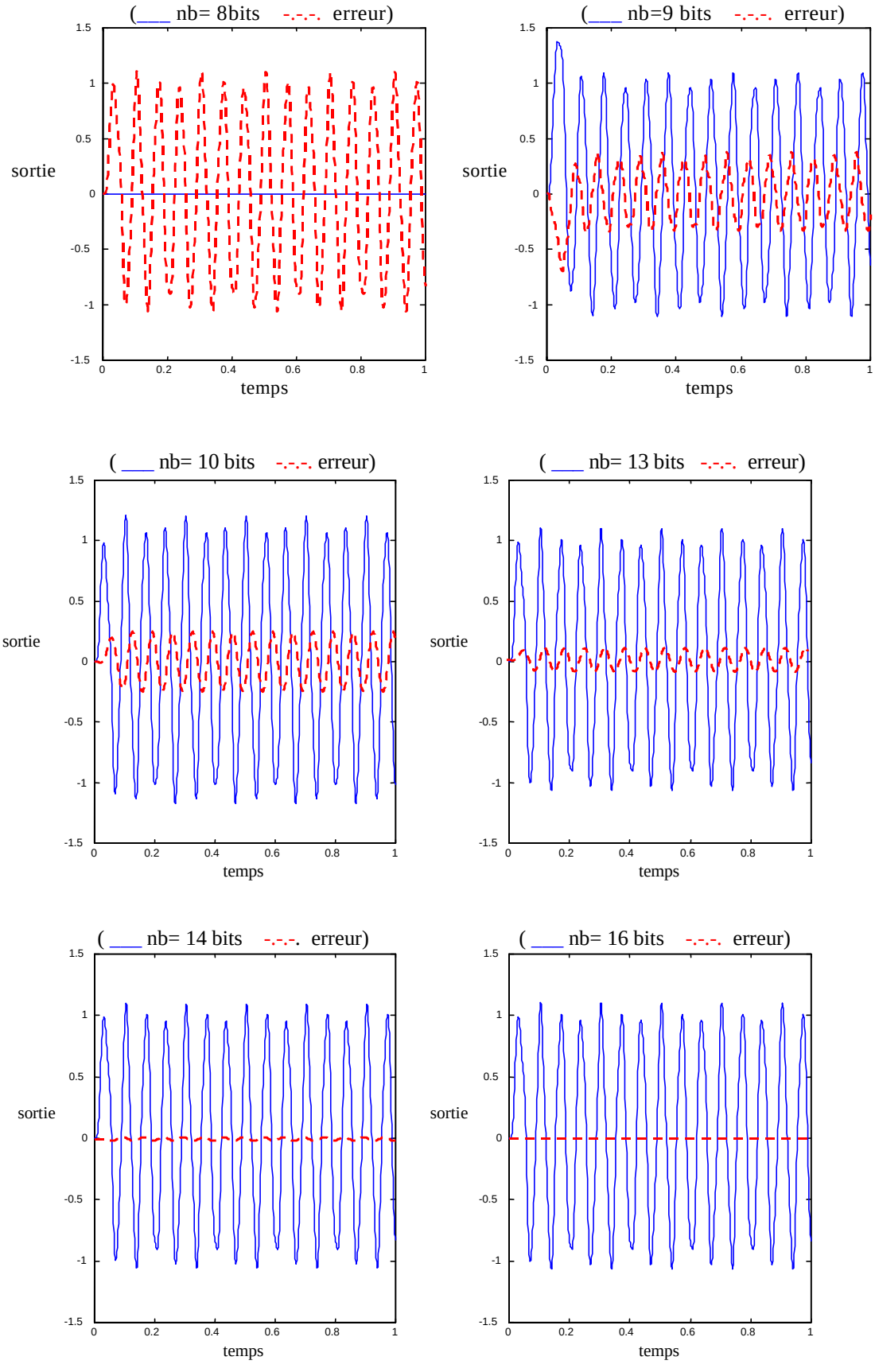
On constate (fig.III-5 et III-6) que lorsque le nombre de bits est :

- < à 12 bits → structure canonique
- < à 10 bits → structure tridiagonale globale
- < à 4 bits → structure minimale,

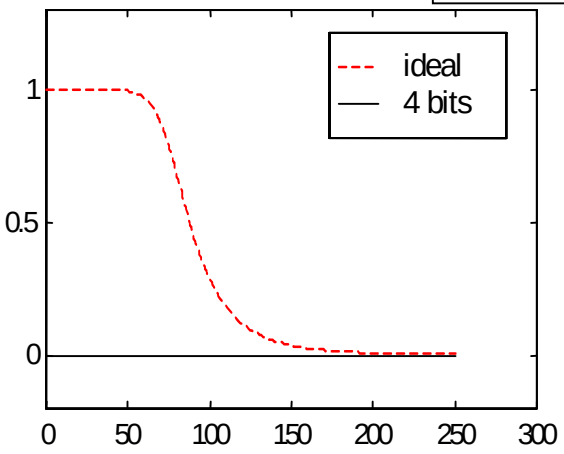
le signal en sortie est nul car tous les coefficients du numérateur de la fonction de transfert sont nuls .

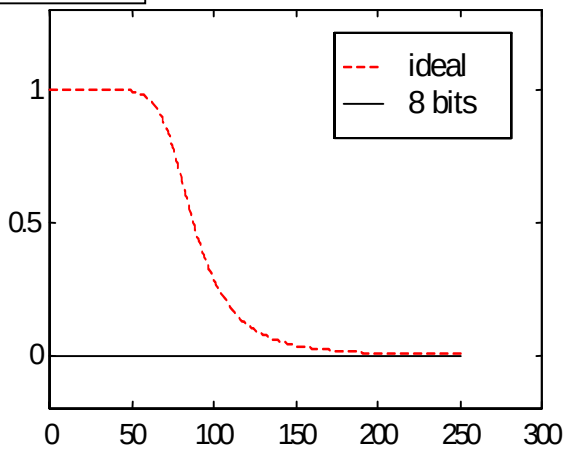
On constate que la structure tridiagonale globale donne une sortie filtrée meilleure que celle de la structure canonique pour le même nombre de bits.
Une nette amélioration de la fonction de transfert est constatée également pour la structure tridiagonale globale par rapport à la structure canonique.

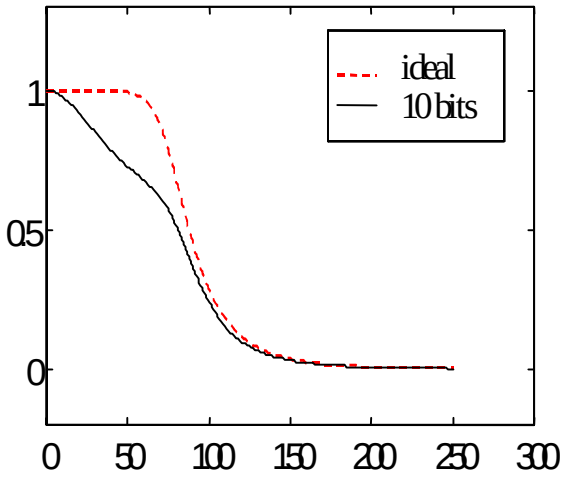
TRIDIAGONALE GLOBALE

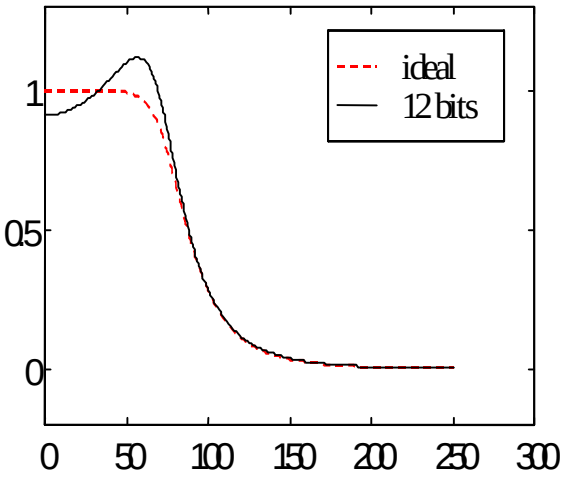


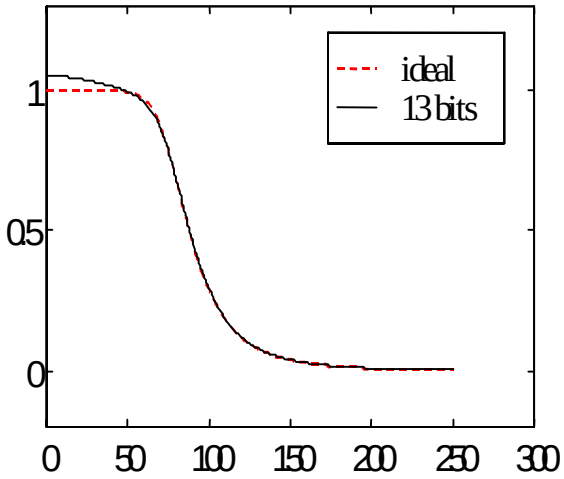
Tridiagonale globale

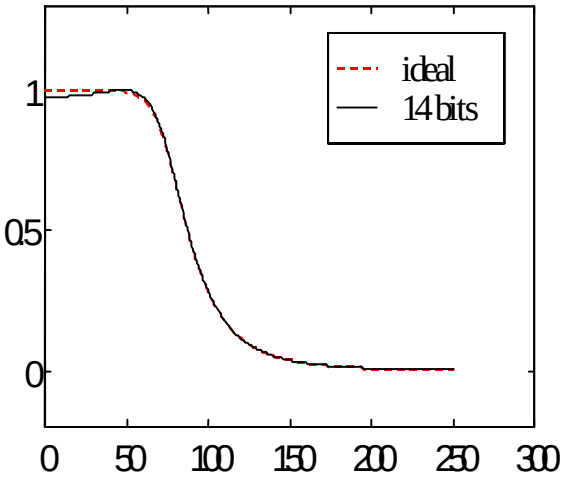












III.5 : Conclusion :

Nous avons construit deux algorithmes de synthèse de structures d'état tridiagonale globale pour un filtre numérique à réponse impulsionnelle infinie. Le premier algorithme se base sur la procédure de Lanczos pour tridiagonaliser une matrice quelconque que nous avons adapté à notre étude. Le deuxième algorithme utilise la méthode des transformations élémentaires qui éliminent les coefficients désirés.

Ces programmes ont été simulés sur PC avec une arithmétique en virgule fixe et avec une quantification par arrondi. Cela nous a permis de vérifier les résultats élaborés en théorie. Une comparaison entre les 3 structures : canonique, minimale et la tridiagonale globale a été établie en termes de sensibilité de la fonction de transfert à la quantification de ses coefficients pour chaque structure et de la qualité du filtrage.

Notre structure tridiagonale globale nécessite l'optimisation de ses coefficients durant le processus de tridiagonalisation. Nous avons remarqué que la première méthode basée sur l'algorithme de Lanczos convergeait lentement dés que l'ordre du filtre augmente ($n>6$). Nous avons donc choisi pour la simulation de filtres d'ordre élevé d'utiliser la seconde méthode où le phénomène de convergence de l'algorithme d'optimisation n'apparaît pas.

Beaucoup de travail reste encore à faire dans ce domaine, ne serait que pour garder une structure modulaire avec un gain de bruit plus faible. Pour cela une rétroaction de l'erreur de quantification [27], [29]-[31], pourra être faite sur une structure tridiagonale globale. Le but serait de réduire le gain de bruit d'arrondi sans toucher à la forme de la structure.

CONCLUSION GENERALE

Conclusion générale

Après avoir exposé les différentes distorsions et les effets non linéaires introduits par la quantification des coefficients et des résultats des opérations arithmétiques d'un filtre numérique R.I.I en virgule fixe, la représentation classique de ces filtres par leur fonction de transfert ou leur réponse impulsionnelle s'est révélée insuffisante pour décrire et analyser leur comportements surtout dans le cas de l'utilisation de mots de longueur finie.

La représentation d'état du filtre est donc utilisée pour apporter plus de détails et de précision sur le déroulement des opérations et sur la façon avec laquelle s'opère le filtrage.

Grâce à la propriété d'invariance de la fonction de transfert du filtre lors de changement de coordonnées dans l'espace d'état par une simple transformation non singulière, des structures à gain de bruit minimal , des structures normalisées et des structures tridiagonales globales sont possibles pour un même filtre numérique.

Pour synthétiser les structures à gain de bruit minimal, deux méthodes ont été utilisées (MULLIS-ROBERTS et HWANG). L'inconvénient de ces deux structures est leur complexité du point de vue du nombre d'opérations arithmétiques qui est de $(n+1)^2$ multiplications pour un filtre d'ordre n, alors que (2n+1) multiplications sont nécessaires pour les structures canoniques mais au prix d'un gain de bruit élevé.

La théorie de la tridiagonalisation d'une structure d'état nous a permis d'étudier des structures de compromis entre le gain de bruit et la complexité des réalisations car ces structures ne comportent que (3n+1) multiplications.

Deux algorithmes ont été pour cela élaboré et ont donné après simulation le même gain de bruit d'arrondi.

Le premier utilise l'algorithme de Lanczos qui nous permet de tridiagonaliser une matrice pleine. La synthèse de cette structure nécessite à chaque étape une optimisation du gain de bruit. La convergence de l'algorithme d'optimisation des coefficients devient très lente dés que l'ordre du filtre augmente. Pour cela une autre méthode a été élaborée . Elle s'appuie sur l'application de transformations élémentaires qui vont nous permettre d'éliminer les coefficients désirés. Une optimisation des paramètres de la transformation appliquée à la structure d'état à tridiagonaliser a été également nécessaire.

Nous avons montré par des exemples de simulation, les effets de la longueur finie des registres sur la sortie du filtre, ainsi que sur sa fonction de transfert.

Nous avons également montré que la structure tridiagonale globale réalise bien un compromis entre la structure canonique et la structure à gain de bruit minimal du point de vue nombre de multiplications, et gain de bruit.

Puisque le gain de bruit de la structure tridiagonale globale reste quand même important par rapport à la structure à gain de bruit minimal, la rétroaction de l'erreur d'arrondi sur une telle structure pourrait le faire diminuer considérablement. Elle devrait être faite sans modifier la forme tridiagonale de la structure. Cela peut faire l'objet de recherches ultérieures.

BIBLIOGRAPHIE

BIBLIOGRAPHIE

-.-

[1]- BOMAR, B.W . and HUNG, J. C : « Minimum roundoff noise digital filters with some power of two coefficients » IEEE. TRANS. Vol, CAS-31, N°10, 1984, pp.833-840

[2]- SMITH, L. M., and BOMAR, B.W. : « An Algorithm for constrained roundoff noise minimisation in digital filters with application to two-dimensional filters » IEEE. TRANS. 1988, pp.884-888

[3]- BARNES, C.W. : « On the design of optimal state-space realizations of second order digital filters » IEEE, TRANS. 1984, CAS-31, N°7, pp.602-608

[4]- BARNES, C.W. : « Computationally efficient second-order digital filter sections with low roundoff noise gain » IEEE, TRANS 1984, CAS-31, N°10, pp.841-847

[5]- JACKSON, L. R. : « Roundoff analysis for fixed point digital filters realized in cascade or parrallel form » IEEE TRANS. 1970, AU. 18, pp.107-122

[6]- SARCINELLI, F. and DINIZ, P. S. R. : « Tridiagonal state space digital filter structure » IEEE. Trans. 1990. CAS. Vol. 37 N°6, pp. 818-824

[7]- OPPENHEIM, A. V. and. SCHAFER, R. W : « Digital Signal Processing ». Englewood Cliffs. N.J. : Prentice-Hall, 1975

[8]- TAVSANOGLU, V. and THIELE, L. : « Optimal design of state space digital filters by simultaneous minimization of sensitivity and roundoff noise » IEEE TRANS. Circuits System, Vol.CAS-31, Oct 1984, pp.884-888.

[9]- MULLIS, C. T. and ROBERTS, R. A. : « Synthesis of minimum roundoff noise fixed point digital filters » IEEE Trans, Sept. 1976, CAS-23, pp 551-562

[10]- HWANG, S.Y. : « Minimum uncorrelated unit noise state space digital filtering » IEEE Trans. 1977, ASSP-25, pp.273-281

[11]- VAN DEN ENDEN, A.W.M. et VERHOECKX, N.A.M. : « Traitement numérique du signal. Une introduction. » Ed. Masson, Paris 1992

[12]- BOITE, R. et LEICH, H. : « Les filtres numériques. Analyse et synthèse des filtres unidimensionnels ». Ed. Masson, Paris 1990

[13]- BELLANGER,M. : « Traitement numérique du signal » Ed. Masson 1987.

[14]- AMBARDAR, A. : « Analog and digital signal processing » PWS foundation on publishing series . Boston 1995.

[15]- MULLIS, C. T. and ROBERTS, R. A. : « Digital Signal Processing » Addison Wesley, Reading, M.A, 1987.

[16]- RABINER, L.R. and GOLD, B. : « Theory and application of digital signal processing » Prentice Hall, Englewood Cliffs, NJ,1975

[17]- ANTONIOU, A. : « Digital Filters : Analysis and Design ». Mc Graw Hill, NY, 1979.

[18]- DINIZ, P.S. and, ANTONIOU, A. : « More economical state-space digital filter structures which are free of constant input limit cycles » IEEE. TRANS. ASSP, Vol ASSP-34, Août 1986

[19]- BARNES, C.W. : « Roundoff noise and overflow in normal digital filters » IEEE. TRANS. Circuits and Systems, Vol CAS-23, Mars 1979

[20]- SAADI-AHMED, N. : « Simulation des filtres numériques RII en virgule fixe » Projet de fin d'études, ENP 1991.

[21]- DERRAS, B. : « New efficient state-space structures for realisation of recursive digital filters ». Thesis for the Master of Science degree, University of Colorado, 1985.

[22]- GOLUB, G.H. and LOAN, F.V. : « Matrix computations. » Maryland, The John Hopkins University Press, 1984.

[23]- E. DURAND : « Solutions numériques des équations algébriques ». Tome 1. Ed Masson 1960.

[24]- LEHOUIDJ, B. : « Etude et simulation de structures de compromis pour les filtres numériques à réponse impulsionnelle infinie réalisés dans l'espace d'état » Thèse de Magister, USTHB 2001.

[25]- BOUCHEMAKH, L. et RAMDANI, R. : « Structure d'état tridiagonale globale à faible bruit d'arrondi pour la réalisation de filtres numériques récursifs ». The first Electric and Electronics Engineering Computing – Laghouat- Nov 2000, Partie1, pp131-136

[26]- ADDOU, D. : « Synthèse optimale des filtres numériques récursifs réalisés par des structures d'état globale et décomposée avec une arithmétique de calcul à virgule fixe » Thèse de Magister. ENP 1992

[27]- DJEBBARI, A. et TARGUI, B. : « Noise reduction in normal recursive digital filters using error feedback » Proceedings of 2nd CEA Algiers. Special Issue of AJOT- Vol2, Nov 1994.

[28]- RAO, S. R. « Engineering optimisation . Theory and practice ». Third edition. Prentice Hall, Englewood Cliffs,,1995

[29]- LAAKSO, T. I. : « Elimination of limit cycles in direct-form digital filters using error feedback » International journal of Circuit Theory and Applications, Vol 21, p141-163, 1993.

[30]- VAIDYANATHAN, P. P. : « On error spectrum shaping in state-space digital filters ». IEEE Trans. On Circuits and Systems, Vol CAS 32, N°1, January 1985

[31]- LAAKSO, T. I. and VALIMAKI, V. : « Energy based effective lenght of the impulse response of a recursive filter" »Proceedings of the 1998 IEEE International Conf. On ASSP (ICASSP'98), Vol. 3, pp1253-1256, May 1998.

[32]- LAPRESTE, J.T. : « Introduction à MATLAB ». Edition Ellipses,1999

[33]- MOKHTARI, M. : « MATLAB 5.2 & 5.3 et SIMULINK 2 & 3 pour étudiants et ingénieurs ».Ed. Springer., Paris, 2000.

[34]- LACASSAGNE, F. : « Etude et parallélisation de méthodes d'optimisation directes : Application à la programmation dynamique et au simplexe non linéaire ». Thèse de Doctorat, Laboratoire d'analyse et d'architecture des systèmes du CNRS, Université Paul Sabatier. Toulouse, 1994.

[35]- http://www.acoustics.hut.fi/~vpv/
ftp://ftp.ele.ufes.br/pub/publications

A N N E X E

Nous donnons, dans le tableau A-1, différentes valeurs de gain pour deux types de filtres : Butterworth et Chebyshev pour la structure minimale, en fonction de l'ordre n du filtre..

	Gain minimal			
Ordre du filtre n	**Butterworth**	**Chebyshev**		
		Ripple=0.2dB	**Ripple=0.5dB**	**Ripple=0.8dB**
2	0.3750	0.4224114504	0.4438491739	0.4565119962
3	0.4704949331	0.6091014243	0.6525679798	0.6765544435
4	0.5555412335	0.8108052296	0.8726294477	0.9053785614
5	0.6354141484	1.0221868940	1.0990621397	1.1387569759
6	0.7121315783	1.2399310736	1.3295044535	1.3750028857
7	0.7867035382	1.4622251762	1.5628142045	1.6133110633
8	0.8597278167	1.6879544856	1.7982671547	1.8531375459
9	0.9315819631	1.9163724954	2.0353714713	2.0941314155
10	1.0025142379	2.1469417699	2.2737924925	2.3360552449

Tableau A-1 : Gains de bruit minimal pour les filtres de Butterworth et Chebyshev en fonction de l'ordre n du filtre

Le tableau A-2 donne la mesure de la sensibilité pour une structure optimale et ce pour les filtres de Butterworth et Chebyshev pour la même fréquence de coupure (=0.05).

	Mmin			
Ordre du filtre N	**Buttrworth**	**Chebyshev**		
		Ripple=0.2dB	**Ripple=0.5dB**	**Ripple=0.8dB**
2	3.1250000	3.2672342	3.3315476	3.369536
3	4.8819796	5.4364056	5.6102720	5.7062176
4	6.7770615	8.0540260	8.3631470	8.526893
5	8.8124846	11.1331210	11.5943720	11.832542
6	10.984921	14.6795170	15.3065300	15.625020
7	13.293628	18.6978010	19.5025130	19.906488
8	15.737550	23.1915900	24.1844040	24.678237
9	18.315819	28.1637250	29.3537150	29.941314
10	21.027656	33.6163590	35.0117170	35.696607

Tableau A-2 : Mesure de la sensibilité des coefficients pour une structure minimale pour les filtres de Butterworth et Chebyshev en fonction de l'ordre du filtre (Fréquence de coupure=0.05)