Conception des filtres numériques et analyse des erreurs résultant de leur réalisation en arithmétique fixe

Subramaniam Sankar

THESE

présentée

A L'UNIVERSITE SCIENTIFIQUE ET MEDICALE DE GRENOBLE

pour obtenir

LE TITRE DE DOCTEUR-INGENIEUR

par

Subramaniam SANKAR

INGENIEUR IISc. INDE

Conception des filtres numériques et analyse des erreurs résultant de leur réalisation en arithmétique fixe.

Soutenue le devant la Commission d'examen

JURY

MM. L. BOLLIET		Président
	J.L. LACOUME	Examinateurs
	R. GARIOD	
	C. MACCHI	Invité

UNIVERSITE SCIENTIFIQUE
ET MEDICALE DE GRENOBLE

LISTE DES PROFESSEURS

Président : Monsieur Michel SOUTIF
Vice-Président : Monsieur Gabriel CAU

PROFESSEURS TITULAIRES

MM.	ANGLES D'AURIAC Paul	Mécanique des fluides
	ARNAUD Georges	Clinique des maladies infectieuses
	ARNAUD Paul	Chimie
	AUBERT Guy	Physique
	AYANT Yves	Physique approfondie
Mme	BARBIER Marie-Jeanne	Electrochimie
MM.	BARBIER Jean-Claude	Physique expérimentale
	BARBIER Reynold	Géologie appliquée
	BARJON Robert	Physique nucléaire
	BARNOUD Fernand	Biosynthèse de la cellulose
	BARRA Jean-René	Statistiques
	BARRIE Joseph	Clinique chirurgicale
	BEAUDOING André	Pédiatrie
	BENOIT Jean	Radioélectricité
	BERNARD Alain	Mathématiques Pures
Mme	BERTRANDIAS Françoise	Mathématiques Pures
MM.	BESSON Jean	Electrochimie
	BEZES Henri	Chirurgie générale
	BLAMBERT Maurice	Mathématiques Pures
	BOLLIET Louis	Informatique (IUT B)
	BONNET Georges	Electrotechnique
	BONNET Jean-Louis	Clinique ophtalmologique
	BONNETAIN Lucien	Chimie minérale
	BONNET-EYMARD Joseph	Pathologie médicale
	BONNIER Etienne	Electrochimie Electrométallurgie
	BOUCHERLE André	Chimie et Toxicologie
	BOUCHEZ Robert	Physique nucléaire
	BOUSSARD Jean-Claude	Mathématiques Appliquées
	BRAVARD Yves	Géographie
	BRISSONNEAU Pierre	Physique du Solide
	BUYLE-BODIN Maurice	Electronique
	CABANAC Jean	Pathologie chirurgicale
	CABANEL Guy	Clinique rhumatologique et hydrologie
	CALAS François	Anatomie
	CARRAZ Gilbert	Biologie animale et pharmacodynamie
	CAU Gabriel	Médecine légale et Toxicologie
	CAUQUIS Georges	Chimie organique
	CHABAUTY Claude	Mathématiques Pures
	CHARACHON Robert	Oto-Rhino-Laryngologie
	CHATEAU Robert	Thérapeutique
	CHENE Marcel	Chimie papetière
	CHIBON Pierre	Biologie animale
	COEUR André	Pharmacie chimique
	CONTAMIN Robert	Clinique gynécologique
	COUDERC Pierre	Anatomie Pathologique
	COUMES André	Radioélectricité
	CRAYA Antoine	Mécanique
Mme	DEBELMAS Anne-Marie	Matière médicale
MM.	DEBELMAS Jacques	Géologie générale
	DEGRANGE Charles	Zoologie
	DEPORTES Charles	Chimie minérale
	DESRE Pierre	Métallurgie
	DESSAUX Georges	Physiologie animale

MM.	DODU Jacques	Mécanique appliquée
	DOLIQUE Jean-Michel	Physique des plasmas
	DREYFUS Bernard	Thermodynamique
	DUCROS Pierre	Cristallographie
	DUGOIS Pierre	Clinique de Dermatologie et Syphiligraphie
	FAU René	Clinique neuro-psychiatrique
	FELICI Noël	Electrostatique
	GAGNAIRE Didier	Chimie physique
	GALLISSOT François	Mathématiques Pures
	GALVANI Octave	Mathématiques Pures
	GASTINEL Noël	Analyse numérique
	GAVEND Michel	Pharmacologie
	GEINDRE Michel	Electroradiologie
	GERBER Robert	Mathématiques Pures
	GERMAIN Jean-Pierre	Mécanique
	GIRAUD Pierre	Géologie
	KAHANE André	Physique générale
	KLEIN Joseph	Mathématiques Pures
	KOSZUL Jean-Louis	Mathématiques Pures
	KRAVTCHENKO Julien	Mécanique
	KUNTZMANN Jean	Mathématiques Appliquées
	LACAZE Albert	Thermodynamique
	LACHARME Jean	Biologie végétale
	LAJZEROWICZ Joseph	Physique
	LATREILLE René	Chirurgie générale
	LATURAZE Jean	Biochimie pharmaceutique
	LAURENT Pierre	Mathématiques Appliquées
	LEDRU Jean	Clinique médicale B
	LLIBOUTRY Louis	Géophysique
	LONGEQUEUE Jean-Pierre	Physique nucléaire
	LOUP Jean	Géographie
Mle	LUTZ Elisabeth	Mathématiques Pures
	MALGRANGE Bernard	Mathématiques Pures
	MALINAS Yves	Clinique obstétricale
	MARTIN-NOEL Pierre	Seméiologie médicale
	MAZARE Yves	Clinique médicale A
	MICHEL Robert	Minéralogie et Pétrographie
	MOURIQUAND Claude	Histologie
	MOUSSA André	Chimie nucléaire
	NEEL Louis	Physique du Solide
	OZENDA Paul	Botanique
	PAUTHENET René	Electrotechnique
	PAYAN Jean-Jacques	Mathématiques Pures
	PEBAY-PEYROULA Jean-Claude	Physique
	PERRET René	Servomécanismes
	RASSAT André	Chimie systématique
	RENARD Michel	Thermodynamique
	REULOS René	Physique industrielle
	RINALDI Renaud	Physique
	ROGET Jean	Clinique de pédiatrie et de puériculture
	DE ROUGEMONT Jacques	Neurologie
	SANTON Lucien	Mécanique
	SEIGNEURIN Raymond	Microbiologie et Hygiène
	SENGEL Philippe	Zoologie
	SILBERT Robert	Mécanique des fluides
	SOUTIF Michel	Physique générale
	TANCHE Maurice	Physiologie
	TRAYNARD Philippe	Chimie générale
	VAILLANT François	Zoologie
	VALENTIN Jacques	Physique Nucléaire
	VAUQUOIS Bernard	Calcul électronique
Mme	VERAIN Alice	Pharmacie galénique
M.	VERAIN André	Physique
MM.	VEYRET Paul	Géographie
	VIGNAIS Pierre	Biochimie médicale

PROFESSEURS ASSOCIES

MM. CHEEKE John	Thermodynamique
GILLESPIE John	I.S.N.
ROCKAFELLAR Ralph	Mathématiques appliquées
WOHLFARTH Erich	Physique du solide

PROFESSEURS SANS CHAIRE

MM. BELORIZKY Elie	Physique
BENZAKEN Claude	Mathématiques appliquées
BERTRANDIAS Jean-Paul	Mathématiques appliquées
BIAREZ Jean-Pierre	Mécanique
Mme BONNIER Jane	Chimie générale
MM. CARLIER Georges	Biologie végétale
COHEN Joseph	Electrotechnique
DEPASSEL Roger	Mécanique des Fluides
DURAND Francis	Métallurgie
GAUTHIER Yves	Sciences biologiques
GAUTRON René	Chimie
GIDON Paul	Géologie et Minéralogie
GLENAT René	Chimie organique
HACQUES Gérard	Calcul numérique
IDELMAN Simon	Physiologie animale
JANIN Bernard	Géographie
JOLY Jean-René	Mathématiques pures
JULLIEN Pierre	Mathématiques appliquées
Mme KAHANE Josette	Physique
MM. MAYNARD Roger	Physique du solide
MULLER Jean-Michel	Thérapeutique
PERRIAUX Jean-Jacques	Géologie et minéralogie
PFISTER Jean-Claude	Physique du solide
PIERY Yvette	Physiologie animale
POULOUJADOFF Michel	Electrotechnique
REBECQ Jacques	Biologie (CUS)
REVOL Michel	Urologie
REYMOND Jean-Charles	Chirurgie générale
ROBERT André	Chimie papetière
SARRAZIN Roger	Anatomie et chirurgie
SARROT-REYNAULD Jean	Géologie
SIBILLE Robert	Construction Mécanique
SIROT Louis	Chirurgie générale
Mme SOUTIF Jeanne	Physique générale
MM. VIALON Pierre	Géologie
VAN CUTSEM Bernard	Mathématiques appliquées
ZADWORNY François	Electronique

MAITRES DE CONFERENCES ET MAITRES DE CONFERENCES AGREGES

Mle AGNIUS-DELORD Claudine	Physique pharmaceutique
ALARY Josette	Chimie analytique
MM. AMBLARD Pierre	Dermatologie
AMBROISE-THOMAS Pierre	Parasitologie
ARMAND Yves	Chimie

MM.	BEGUIN Claude	Chimie organique
	BILLET Jean	Géographie
	BLIMAN Samuel	Electronique (EIE)
	BLOCH Daniel	Electrotechnique
Mme	BOUCHE Liane	Mathématiques (CUS)
MM.	BOUCHET Yves	Anatomie
	BOUVARD Maurice	Mécanique des Fluides
	BRODEAU François	Mathématiques (IUT B)
	BRUGEL Lucien	Energétique
	BUISSON Roger	Physique
	BUTEL Jean	Orthopédie
	CHAMBAZ Edmond	Biochimie médicale
	CHAMPETIER Jean	Anatomie et organogénèse
	CHERADAME Hervé	Chimie papetière
	CHIAVERINA Jean	Biologie appliquée (EFP)
	COHEN-ADDAD Jean-Pierre	Spectrométrie physique
	COLOMB Maurice	Biochimie médicale
	CONTE René	Physique
	COULOMB Max	Radiologie
	CROUZET Guy	Radiologie
	CYROT Michel	Physique du solide
	DELOBEL Claude	M.I.A.G.
	DUSSAUD René	Mathématiques (CUS)
Mme	ETERRADOSSI Jacqueline	Physiologie
MM.	FAURE Jacques	Médecine légale
	FONTAINE Jean-Marc	Mathématiques Pures
	GENSAC Pierre	Botanique
	GIDON Maurice	Géologie
	GRIFFITHS Michaël	Mathématiques Appliquées
	GROULADE Joseph	Biochimie médicale
	GUITTON Jacques	Chimie
	HOLLARD Daniel	Hématologie
	HUGONOT Robert	Hygiène et Médecine préventive
	IVANES Marcel	Electricité
	JALBERT Pierre	Histologie
	JOUBERT Jean-Claude	Physique du Solide
	KRAKOWIAK Sacha	Mathématiques appliquées
	KUHN Gérard	Physique
	LACOUME Jean-Louis	Physique
Mme	LAJZEROWICZ Jeannine	Physique
MM.	LANCIA Roland	Physique atomique
	LE JUNTER Noël	Electronique
	LEROY Philippe	Mathématiques
	LOISEAUX Jean-Marie	Physique Nucléaire
	LUU DUC Cuong	Chimie Organique
	MACHE Régis	Physiologie végétale
	MAGNIN Robert	Hygiène et Médecine préventive
	MARECHAL Jean	Mécanique
	MARTIN-BOUYER Michel	Chimie (CUS)
	MICHOULIER Jean	Physique (I.U.T. "A")
	MICOUD Max	Maladies infectieuses
	MOREAU René	Hydraulique (INP)
	NEGRE Robert	Mécanique
	PARAMELLE Bernard	Pneumologie
	PECCOUD François	Analyse (IUT B)
	PEFFEN René	Métallurgie
	PELMONT Jean	Physiologie animale
	PERRET Jean	Neurologie
	PERRIN Louis	Pathologie expérimentale
	PHELIP Xavier	Rhumatologie

MM. RACHAIL Michel	Médecine interne
RACINET Claude	Gynécologie et obstétrique
RAYNAUD Hervé	M.I.A.G.
RENAUD Maurice	Chimie
RICHARD Lucien	Botanique
Mme RINAUDO Marguerite	Chimie macromoléculaire
MM. ROMIER Guy	Mathématiques (IUT B)
SHOM Jean Claude	Chimie Générale
STIEGLITZ Paul	Anesthésiologie
STOEBNER Pierre	Anatomie pathologique
VEILLON Gérard	Mathématiques Appliquées (INP)
VOOG Robert	Médecine interne
VROUSSOS Constantin	Radiologie

MAITRES DE CONFERENCES ASSOCIES

MM. CRABLEE Pierre	C.E.R.M.O.
CURRIE Jan	Mathématiques appliquées
YACOUD Mahmoud	Médecine légale

CHARGES DE FONCTIONS DE MAITRES DE CONFERENCES

Mme BERIEL Hélène	Physiologie
Mme RENAUDET Jacqueline	Microbiologie

Fait le 1.10.73

A V A N T - P R O P O S

Le travail présenté dans cette thèse a été effectué au Laboratoire d'Electronique et de Technologie de l'Informatique du Centre d'Etudes Nucléaires de GRENOBLE grâce au soutien du Centre Régional des Oeuvres Universitaires et du L.E.T.I.

Je tiens à remercier Monsieur le Professeur BOLLIET, Professeur à l'Université Scientifique et Médicale de GRENOBLE, d'avoir bien voulu me faire l'honneur de diriger ma thèse et de présider ce jury.

Je désire exprimer ma reconnaissance à Monsieur le Professeur LACOUME, Directeur du Centre d'Etude des Phénomènes Aléatoires et Géophysiques de GRENOBLE, qui a bien voulu examiner mon travail et de faire partie du jury.

Je tiens à exprimer ma gratitude à Monsieur CORDELLE, Directeur du Laboratoire d'Electronique et de Technologie de l'Informatique, d'avoir bien voulu m'accueillir dans son laboratoire.

Que Monsieur GARIOD, Chef du Laboratoire de Mesure, Controle et Traitement Electronique, soit assuré de ma reconnaissance pour l'encouragement et l'intérêt qu'il a porté à mon travail et pour sa participation au jury.

Je désire exprimer ma reconnaissance à Monsieur le Professeur MACCHI à l'Institut de Programmation, Université de PARIS VI, pour avoir bien voulu participer au jury.

Je tiens à remercier Monsieur MONGE pour ses conseils précieux et pour sa cordiale sympathie tout au long de ce travail.

Que Monsieur DELCROIX soit assuré de ma reconnaissance pour avoir mis à ma disposition les moyens matériels nécessaires à la réalisation de cette thèse.

Que mon collègue, Monsieur THOMAS, trouve ici l'expression de mon amicale reconnaissance pour nos discussions fructueuses.

Je tiens également à remercier Monsieur RUQUET pour l'aide précieuse qu'il m'a apportée dans la mise au point des programmes de simulation.

TABLE DES MATIERES

Page

INTRODUCTION

Depuis peu de temps, l'intérêt du filtrage numérique ne cesse de grandir. Les raisons en sont les suivantes :

- le coût des circuits digitaux ne cesse de diminuer, rendant ainsi les modules de calcul numériques compétitifs par rapport aux modules analogiques.

- l'amélioration constante des performances des circuits digitaux, en particulier la vitesse des opérations, permet de réaliser des filtres opérant en temps réel.

- la possibilité d'augmenter considérablement la précision dans les calculs. En effet, la précision, dans les systèmes numériques, dépend directement du nombre de bits des registres. On peut, simplement en augmentant la longueur de ceux-ci, augmenter la précision à volonté. Cela est beaucoup plus simple que l'obtention de composants analogiques plus précis.

- aucun changement dès paramètres d'un filtre numérique au cours du temps (par opposition au vieillissement des composants analogiques) et aucun problème d'adaptation d'impédance.

- facilité de réalisation des filtres numériques dans la gamme des très basses fréquences, gamme pour laquelle les filtres analogiques sont mal adaptés.

- grande flexibilité des filtres numériques. En effet, pour réaliser des filtres différents, il suffit de changer le jeu des coefficients correspondants.

- possibilité de détermination de filtres complexes ayant des gabarits fréquentiels quelconques à l'aide des programmes d'optimisation puissants.

Néanmoins, le filtrage numérique pose un certain nombre de problèmes tant au stade de la conception des filtres qu'au stade de leur "implémentation" dans les petits calculateurs ou dans les opérateurs cablés spécifiques.

Au stade de la conception, le problème est le suivant :

- si le filtre numérique est dérivé de son équivalent analogique, il faut analyser avec soin dans chaque cas particulier, le type et la nature de la transformation mathématique nécessaire à ce passage.

Au stade de "l'implémentation", on rencontre les problèmes suivants :

- erreur résultante sur le gabarit du filtre par suite de la quantification des paramètres du filtre.

- dégradation du signal à la sortie, due aux opérateurs arithmétiques à dynamique limitée qui réalisent l'algorithme de filtrage choisi.

Le travail présenté ici se situe dans le cadre général du traitement du signal et comporte trois objectifs :

- étude d'un programme d'optimisation et utilisation de celui-ci pour la conception des filtres numériques récursifs.

- étude et mise au point d'un programme d'optimisation pour les filtres non récursifs utilisant l'algorithme de la transformée de Fourier rapide (F F T [1]).

- l'étude des erreurs dûes à la réalisation des filtres numériques récursifs en arithmétique virgule fixe.

Nous avons étudié les problèmes de précision liés à l'utilisation de ces algorithmes de filtrage numérique dans les petits calculateurs ou dans des opérateurs cablés spécifiques. Ces machines sont caractérisées par des unités arithmétiques travaillant en virgule fixe et sur des mots de longueur réduite (8 à I6 bits par exemple).

Nous avons développé un modèle statistique pour le bruit d'arrondi des opérations arithmétiques. Nous avons déduit, pour estimer la variance de ce bruit, des expressions plus générales que celles actuellement existantes. Ces expressions permettent une comparaison aisée des différentes formes de réalisation et en particulier, elles permettent pour une application donnée, de choisir la forme de réalisation qui rend minimum les bruits de calcul.

Des vérifications expérimentales de ces expressions ont été faites à l'aide d'un ensemble de traitement du signal comprenant un petit calculateur.

1 FAST FOURIER TRANSFORM

C H A P I T R E I

GENERALITES SUR LA THEORIE DES FILTRES NUMERIQUES LINEAIRES

La théorie des réseaux linéaires est basée sur les propriétés electriques des selfs, des condensateurs et des résistances. Celles-ci conduisent à la description du réseau, moyennant certaines lois (la loi d'Ohm, les lois de Kirchoff), par l'intermédiaire d'équations différentielles linéaires. L'analyse et la synthèse de tels réseaux sont facilitées par l'emploi de la transformée de Laplace et la notion de la fonction de transfert. Les filtres linéaires analogiques qui réalisent des gabarits fréquentiels voulus, sont conçus à l'aide de la théorie de l'approximation §6§. Cette théorie conduit à des fonctions de transfert du filtre voisines du gabarit désiré. Ces fonctions s'expriment sous la forme d'un rapport de deux polynômes rationels en p (p étant la variable de Laplace). Il est utile de signaler qu'il existe une vaste gamme de polynômes d'approximation (par exemple Chebycheff, Bessel, Butterworth, Lerner, Elliptique, etc ...)

Les filtres numériques linéaires qui appartiennent à la classe des systèmes linéaires discrets, sont régis par des équations linéaires aux différences à coefficients constants. La réalisation de tels filtres implique l'utilisation d'ordinateurs ou bien d'ensembles spécialisés à logique et arithmétique cablées. Nous les appellerons indifféremment, les machines numériques. De telles machines qui réalisent les équations aux différences ne peuvent travailler que sur un signal d'entrée échantillonné et quantifié en amplitude. Nous allons considérer uniquement le cas de l'échantillonnage périodique. Donc, le vocable "filtre numérique" signifie pour la suite un filtre numérique causal régi par des équations linéaires aux différences, dont les paramètres (les coefficients) sont invariants dans le temps, et, dont le signal d'entrée est échantillonné avec une période fixe de T secondes.

L'étude des processus échantillonnés est facilitée par la transformée en z. La transformée en z est un outil mathématique pour l'étude des équations linéaires aux différences et à coefficients constants. La définition ainsi que quelques propriétés (que nous utiliserons) de cette transformée, sont rappelées en annexe **A** .

I.1 Le filtre numérique défini comme un opérateur linéaire

On peut considérer le filtre numérique comme un opérateur linéaire transformant une séquence d'entrée / $x(n)$/ en une séquence de sortie / $y(n)$/ sans se soucier de l'équivalence entre le signal continu fonction du temps et le signal échantillonné (en temps et en amplitude) §2§ §3§. Dans la section I.3 on étudiera plus en détail cette équivalence.

Pour un système linéaire discret, la séquence d'entrée $/x(n)/$ et la séquence de sortie $/y(n)/$ sont reliées par l'équation linéaire aux différences à coéfficients constants de forme suivante :

$$y(n) + b_1 y(n-1) + \ldots b_M y(n-M) = a_0 x(n) + a_1 x(n-1) + \ldots + a_N x(n-N) \qquad (I.1)$$

c'est à dire qu'à l'instant (nT) la valeur de la sortie est fonction de l'entrée présente et d'une combinaison linéaire des entrées et des sorties précédentes.

Si l'on prend la transformée en z de (I.1) terme à terme, en appliquant l'équation (A.1) et (A.3) nous avons :

$$Y(z) \left(1 + \sum_{i=1}^{M} b(i) z^{-i}\right) = X(z) \sum_{i=0}^{N} a(i) z^{-i} \qquad (I.2)$$

nous avons alors : $$Y(z) = H(z)\ X(z) \qquad (I.3)$$

où $H(z)$ est rationnelle en z^{-1} et représente la fonction de transfert discrète du filtre.

On peut aussi définir la réponse impulsionnelle de la manière suivante :

soit la séquence $/\delta(n)/$ telle que $\delta(n) = 1$ pour $n = 0$ et $\delta(n) = 0$ pour $n \neq 0$. On l'appelle séquence impulsionnelle.

La réponse du filtre à cette séquence est définie comme la réponse impulsionnelle / h(n)/ du filtre. / h(n)/ est une séquence définie pour tout n entier positif, puisque le filtre est causal. / h(n)/ est nulle pour n négatif. Pour tout autre signal d'entrée / x(n)/ ayant pour transformée en z H(z), la réponse du filtre est d'après (I.2) et (I.3) :

$$y(n) = \sum_{j=0}^{\infty} x(j)\ h(n-j) \qquad (I.4)$$

Cette équation est une convolution discrète du signal / x(n)/ et de la réponse impulsionnelle / h(n)/. En prenant la transformée en z de deux membres de (I.4), on a :

$$Y(z) = \sum_{n=0}^{\infty} \left(\sum_{j=0}^{\infty} x(j)\ h(n-j)\right) z^{-n} \qquad (I.5)$$

Avec l'hypothèse de convergence uniforme de la série (I.5) on peut intervertir les indices de sommation. D'où :

$$Y(z) = \sum_{j=0}^{\infty} y(j) \left(\sum_{n=0}^{\infty} h(n-j)\ z^{-n}\right) \qquad (I.6)$$

Enfin, en utilisant la propriété de retard (A.3), on a :

$$Y(z) = \sum_{j=0}^{\infty} x(j)\ z^{-j}\ H(z)$$

$$Y(z) = X(z)\ H(z) \qquad (I.7)$$

Ce qui montre que la transformée en z de la réponse impulsionnelle peut être interprêtée comme une fonction de transfert similaire au cas de la transformée de Laplace de la réponse impulsionnelle du filtre analogique.

L'équation (I.1) nous donne l'équation de récurrence suivante :

$$y(n) = a_0 x(n) + a_1 x(n-1) + \ldots a_N x(n-N) - b_1 y(n-1) \ldots - b_M y(n-M) \qquad (I.8)$$

C'est sous cette forme que l'on réalise des filtres numériques, soit par programmation directe, soit par un ensemble câblé. Si tous les b(i) sont nuls, c'est à dire s'il n'y a pas de contre-réaction (feed-back), le filtre est dit non-récursif. Sa réponse impulsionnelle est finie et il est toujours stable. On peut démontrer que ces filtres ont pour fonction de transfert des polynômes trigonométriques §4§. En effet, prenons l'exemple suivant :

$$Y(n) = \frac{x(n) + x(n-1)}{2} \qquad (I.9)$$

La transformée en z des deux membres de cette équation est :

$$Y(z) = \frac{1}{2} x(z) \cdot (1 + z^{-1}) \qquad (I.10)$$

La fonction de transfert discrète est donc :

$$H(z) = \frac{Y(z)}{X(z)} = \frac{1}{2} (1 + z^{-1}) \qquad (I.11)$$

On peut définir la réponse fréquentielle de ce filtre (voir A.3) pour

$z = \exp(j\omega T)$ $\qquad j = \sqrt{-1}$

$$H(\exp(j\omega T)) = \frac{1}{2}(1 + \exp(-j\omega T)) \qquad (I.12)$$

$$\left|H(\exp(j\omega T)\right|^2 = \frac{1}{4}(2 + 2\cos\omega T) = \frac{1}{2} + \frac{1}{2}\cos\omega T \qquad (I.13)$$

La réponse en amplitude au carré est tracée fig. I.1. Nous voyons que ce filtre se comporte comme un filtre passe-bas, sa largeur de bande étant de l'ordre d'une octave. Pour obtenir des gabarits fréquentiels ayant des caractéristiques différentes (bande de transition étroite ou faible largeur de bande, etc ...) on est amené à utiliser un grand nombre de termes dans (I.9) ou à introduire des termes récursifs.

Si, dans l'équation (I.8) au moins un des b(i) est différent de zéro, le filtre est dit récursif. Ces filtres ont une réponse impulsionnelle de durée infinie. Si tous les a(i) pour $i \geqslant 1$ sont nuls, le filtre est dit purement récursif ou auto-régressif. §5§

A un gabarit donné correspondent trois types de filtres, non récursifs, auto-régressifs, ou récursifs (qui comportent les **a(i)** et les **b(i)** . Cependant pour un gabarit donné, il est préférable de retenir un filtre récursif, car l'approximation par un filtre non récursif nécessite un nombre élevé de coefficients a(i). Les schémas-bloc correspondant à ces trois types de réalisation sont montrés fig. I.2 à I.4.

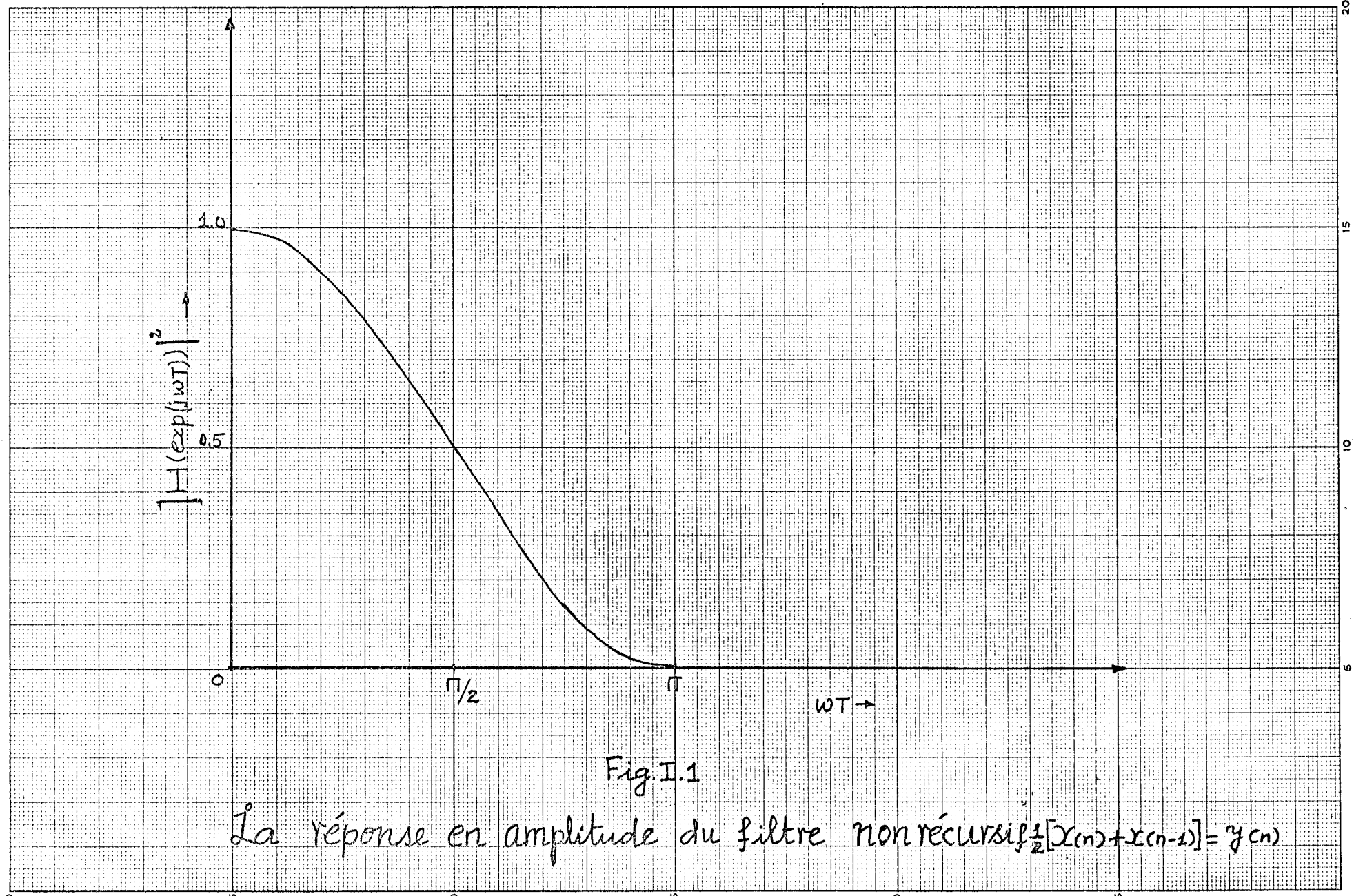
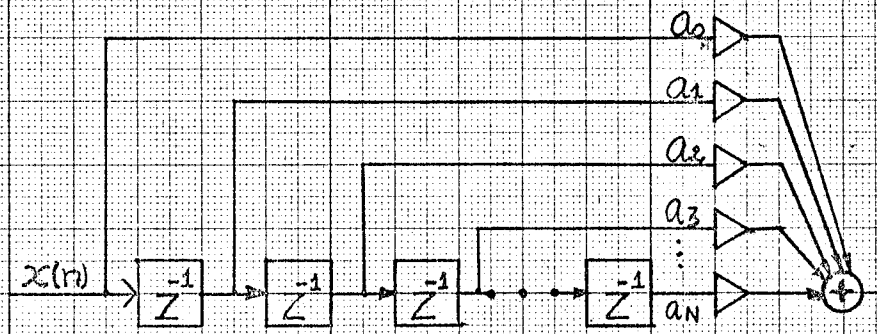


Fig. I.1

La réponse en amplitude du filtre non récursif $\frac{1}{2}[x(n)+x(n-1)]=y(n)$



Les a_i peuvent représenter la R.I spécifiée du filtre. Voir l'équation (III. 2).

fig.I.2

le schéma bloc du filtre nonrécursif

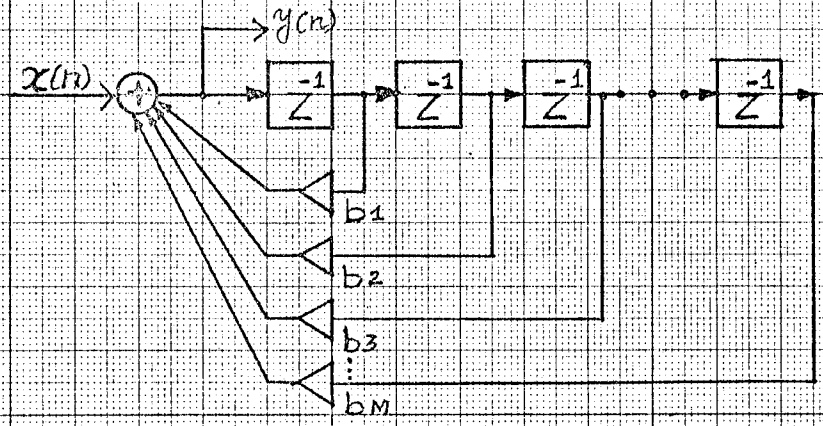


fig.I.3

le schéma bloc du filtre purment récursif

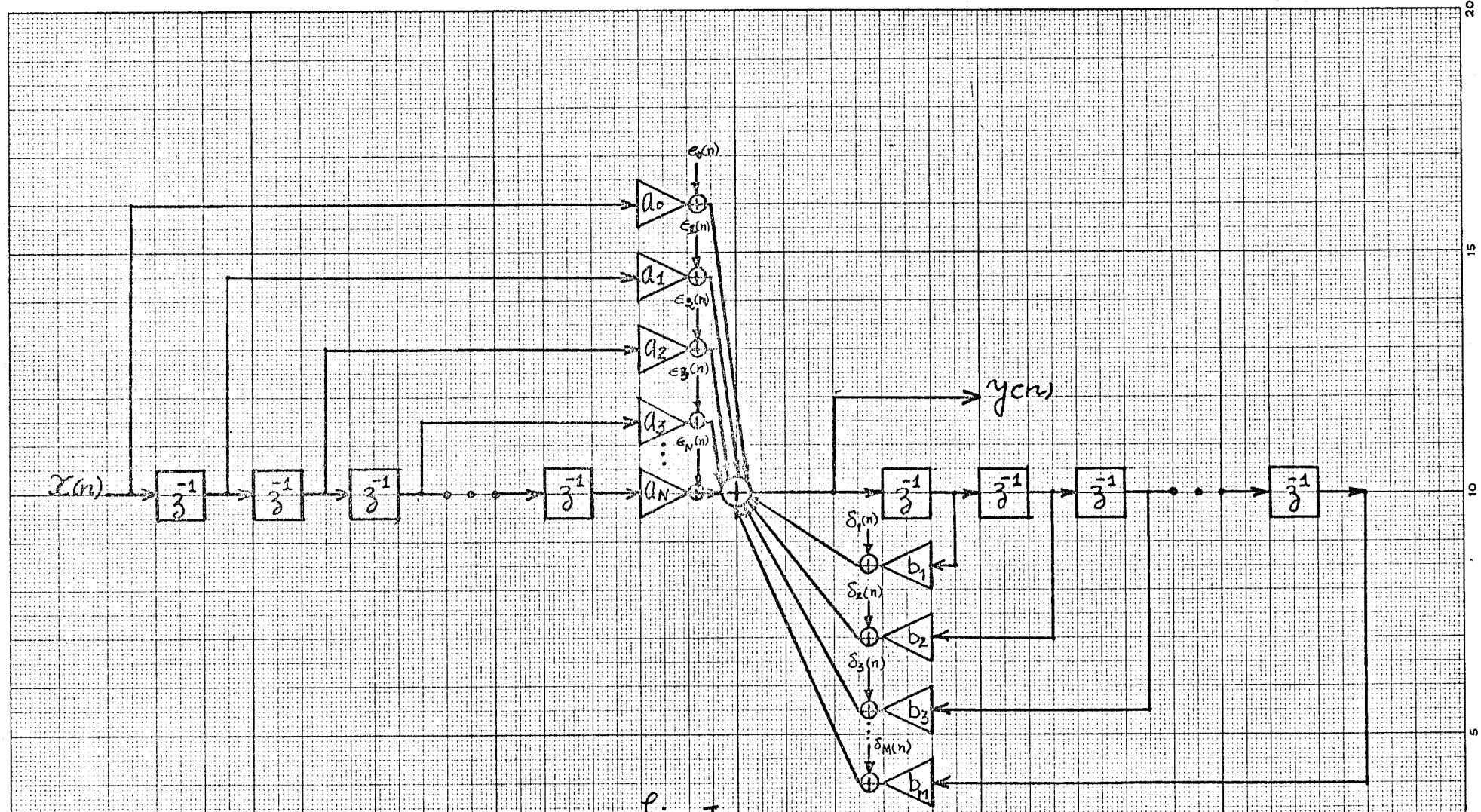


fig. I.4

La réalisation de filtre récursif en forme non-factorisée (forme Directe)

I.2 Réponse fréquentielle du filtre numérique :

La réponse fréquentielle du filtre linéaire analogique défini par la fonction de transfert $H(p)$ est déterminée par le module $|H(p)|$ et la phase de $H(p)$ quand la variable p se déplace le long de l'axe imaginaire ($p=j\omega$). Dans le cas du filtre numérique ayant la fonction de transfert $H(z)$, on peut définir la réponse fréquentielle comme le module et l'argument de $H(z)$ quand z se déplace sur le cercle unité. En effet, soit une séquence /Sin nb/ appliquée à l'entrée d'un tel système. La transformée en z d'une telle séquence est :

$$Z\ /\sin nb/ = \frac{z^{-1} \sin b}{(1 - e^{jb} z^{-1})(1 - e^{-jb} z^{-1})} \qquad (I.14)$$

Cette fonction en z^{-1} a deux pôles (complexes conjugués) sur le cerc le unité $z = \exp(\pm jb)$. Si l'on admet que $H(z)$ est rationnel en z avec tous ses pôles à l'intérieur du cercle unité, la transformée en z de la sortie /y(n)/ est aussi une fonction rationnelle en z et admet donc une décomposition en fractions simples.

On a :

$$Y(z) = H(z)\ X(z) = \frac{z^{-1} \sin b\ H(z)}{(1 - e^{jb} z^{-1})(1 - e^{-jb} z^{-1})}$$

$$= \frac{A}{(1 - e^{jb} z^{-1})} + \frac{B}{(1 - e^{-jb} z^{-1})}$$

où $A = e^{-jb} \sin b \; H(e^{jb})$

et $B = e^{jb} \sin b \; H(e^{-jb})$

en posant $\text{Sin } b = \frac{e^{jb} - e^{-jb}}{2j}$ et simplifiant, on a :

$$Y(z) = \frac{H(e^{jb})}{(1 - e^{jb} z^{-1}) 2j} + \frac{H(e^{-jb})}{(1 - e^{-jb} z^{-1})} + \frac{e^{j2b} H(e^{-jb})}{(1 - e^{-jb} z^{-1}) (2j)}$$

$$- \frac{e^{-j2b} H(e^{jb})}{(1 - e^{+jb} z^{-1}) (2j)}$$

Les deux derniers termes de cette équation représentent la réponse transitoire. Donc, en régime permanent, on a :

$$H(z) = \frac{H(e^{jb})}{(1 - e^{jb} z^{-1}) (2j)} + \frac{H(e^{-jb})}{(1 - e^{-jb} z^{-1}) (-2j)}$$

soit :

$$H(z) = \frac{\text{Re } (H(e^{jb})) \; z^{-1} \sin b + \text{Im } (H(e^{jb})) \; (1 - z^{-1} \cos b)}{(1 - e^{jb} z^{-1}) \; (1 - e^{-jb} z^{-1})}$$

La transformée inverse de cette expression est :

$$Y(n) = \text{Re } (H(e^{jb})) \sin b + \text{Im } (H(e^{jb})) \cos nb$$

qui peut se mettre sous la forme :

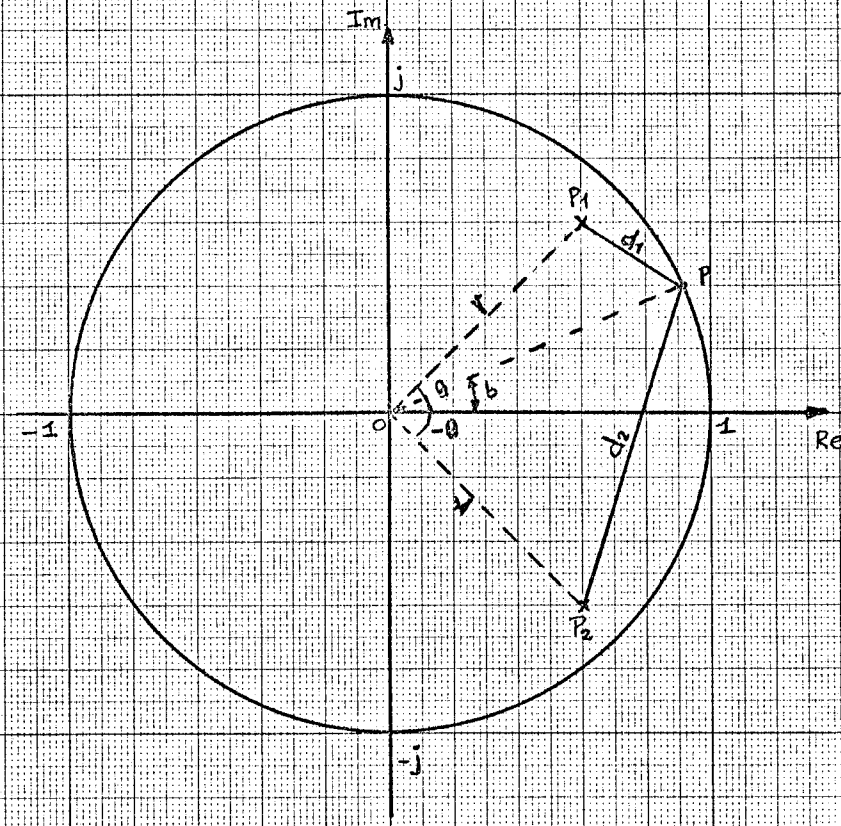
$$Y(n) = (H(e^{jb}) \sin (nb + \text{Arg. } H(e^{jb})) \qquad \text{(I.15)}$$

Ceci montre que le module de $H(z)$ pour $z = e^{jb}$ (c'est à dire, sur le cercle unité et à la même fréquence que la sinusoïde d'entrée) représente le facteur de transmission, (ou la réponse en amplitude) et l'argument de $H(z)$ (à la même fréquence) représente le déphasage de la sortie par rapport à l'entrée. Ainsi donc, la réponse de $H(z)$ pour toutes les fréquences comprises entre $0 \leq b \leq \pi$ peut être évaluée en substituant e^{jb} à z dans $H(z)$. Ceci est analogue aux systèmes continus où l'évaluation est faite pour $p = j\omega$ (sur l'axe imaginaire).

La fig. (I.5) montre la représentation dans le plan z d'un circuit résonant numérique. La sélectivité fréquentielle du filtre apparaît quand on utilise les méthodes géométriques pour calculer le module et la phase de la fonction de transfert pour un signal d'entrée sinusoïdal de fréquence donnée. Plus les pôles du filtre se rapprochent du cercle unité, plus il est sélectif. Voir aussi à ce sujet la référence §31§.

I.3 Echantillonnage du signal et équivalence entre les systèmes continus et discrets :

Jusqu'ici, on a considéré les systèmes échantillonnés en tant que tel, sans aucun rapport avec les systèmes continus. On examinera brièvement, dans cette section, les problèmes relatifs à l'échantillonnage d'un signal continu et à la numérisation des fonctions de transfert continues.



P_1, P_2 : pôles complexes conjugués

$P_1 = r\exp(j\theta)$

$P_2 = r\exp(-j\theta)$

P : le point courant à $\exp(jb)$

$$|H(\exp(jb))| = \frac{1}{d_1 \cdot d_2}$$

fig. I.5

Un résonateur Numérique

On peut considérer le processus d'échantillonnage préiodique d'un signal continu $f(t)$ comme une multiplication par un train d'impulsions §2§. Ceci donne lieu à un train d'impulsions pondérées en amplitude par la valeur du signal $f(t)$ aux instants d'échantillonnage. On le désigne par $f^*(t)$. On a :

$$f^*(t) = f(t) \sum_{n=0}^{\infty} \delta(t-nT) \qquad (I.16)$$

qui peut se mettre sous la forme suivante compte tenu de la propriété de la "fonction" dirac δn

$$f^*(t) = \sum_{n=0}^{\infty} f(nt) \quad \delta(t-nT) \qquad (I.17)$$

Dans cette équation, quand on prélève $f(t)$ à l'instant d'une discontinuité, on suppose que $f(nT)$ vaut $f(nT^+)$, c'est-à-dire la valeur juste après la discontinuité.

La transformée de Laplace de (I.17) est donc :

$$F^*(p) = \sum_{n=0}^{\infty} f(nT)\, e^{-npT} \qquad (I.18)$$

Cela nous montre que la transformée en z du signal échantillonné est obtenu à partir de la transformée de Laplace du train d'impulsions en effectuant le changement de variable suivant :

$$F(z) = \overset{*}{F}(p) \Big|_{e^{pT} = z} \qquad (I.19)$$

La transformée de Laplace de l'équation (I.17) nous donne le produit de convolution en p suivant :

$$F(p) = \frac{1}{2\pi j} \int_C \frac{F(s)}{1 - \exp(-(p-s)T)} \cdot ds + \frac{1}{2} f(0^+) \qquad (I.20)$$

où C est un contour qui englobe toute la partie à droite des singularités de F(s), et, à gauche des singularités du train d'impulsions. Si l'on suppose, pour simplifier, que F(s) a tous ses pôles simples dans le demi plan gauche de s on peut fermer le contour à gauche, ce qui nous donne §2§ :

$$F^{*}(p) = \sum_{\substack{\text{poles de} \\ F(s)}} (\text{résidus de } \frac{F(s)}{1 - \exp(-(p-s)T)}) \qquad (I.21)$$

ou bien, quand on le ferme à droite, cela donne :

$$F^{*}(p) = \frac{1}{T} \sum_{k=-\infty}^{+\infty} F(p+jk\frac{2\Pi}{T} + \frac{1}{2} f(0^{+})) \qquad (I.22)$$

L'équation (I.21) est équivalente à la décomposition de F(z) en fractions simples, alors que (I.22) nous démontre l'effet de pliage du spectre dû à la contribution des termes de F(s) en dehors de la bande primaire §2§. En effet, la transformée en z ne donne pas une relation bi-univoque entre le signal discret et le signal continu.

Par exemple, quand on échantillonne une sinusoïde de fréquence f_0 on ne peut pas distinguer celle-ci des sinusoïdes de fréquences $(f_0 + \frac{k}{T})$ où k est un entier positif. Une approche pour faire face à ce problème de "pliage du spectre" consiste à considérer seulement les signaux à bandes limitées, c'est à dire ceux dont l'amplitude du spectre décroît très rapidement au voisinage de $\frac{1}{2T}$.

Dans ces conditions, la relation suivante est bi-univoque entre l'axe imaginaire et le cercle unité :

$$z = \exp(j\omega T), \quad |\omega| < \frac{\Pi}{T} \qquad (I.23)$$

Pour les mêmes raisons, quand on échantillonne la fonction de transfert continue d'un filtre H(p), en utilisant la transformée en z, et pour éviter les erreurs de "pliage", on suppose que le comportement asymptotique de H(p) est de la forme suivante :

$$\lim_{p\to j\infty} H(p) = \lim_{p\to j\infty} \frac{k}{(p/j\omega_c)^n}, \; n>0 \qquad (I.24)$$

où ω_c = la fréquence de coupure.

Cette hypothèse n'est valable que lorsque n est grand ou bien quand $\omega_c << \frac{\omega_e}{2}$. Si, au moins, l'une de ces conditions n'est pas satisfaite, le filtre échantillonné approchera mal le gabarit du filtre analogique H(p). Donc, pour l'échantillonnage des filtres continus dont la fréquence de coupure ω_c est une fraction appréciable de $\frac{\omega_e}{2}$ (filtre numérique dit de large bande) ou ceux du type stop-bande ou passe-haut, la transformation en z peut introduire des erreurs importantes. Kaiser §6§ suggère d'utiliser la transformation "bande limitée" qui applique tout le plan complexe p dans une bande horizontale d'un autre plan complexe p_1. Elle est définie comme :

$$p = \frac{2}{T} \text{th} (p_1 T/2) \qquad (I.25)$$

où $p_1 = \alpha + j\nu$. Nous avons aussi, pour $p=j\omega$

$$j\omega = \frac{2}{T} \text{th} \left(\frac{j\nu T}{2}\right) = j \frac{2}{T} \text{tg} \left(\frac{\nu T}{2}\right) \qquad (I.26)$$

Quand ω varie de $-\infty$ à $+\infty$, ν varie de $-\frac{\pi}{T}$ à $\frac{\pi}{T}$. Puisque tout le plan complexe p est appliqué dans une bande horizontale du plan p_1. Le "pliage du spectre" est éliminé. Cette relation est bi-univoque.

D'autre part, la substitution $z = e^{p_1 T}$ dans (I.25) nous donne :

$$p = \frac{2}{T} \left(\frac{1-z^{-1}}{1+z}\right) \qquad \text{(I.27)}$$

Donc, il suffit d'effectuer cette substitution de p dans $H(p)$ pour obtenir la fonction de transfert discrète du filtre numérique équivalent $H'(z)$. On a :

$$H'(z) = H(p)\Big|_{p=\frac{2}{T}\frac{(1-z^{-1})}{(1+z^{-1})}} \qquad \text{(I.28)}$$

$H'(z)$ est aussi une fonction rationnelle en z^{-1} de même ordre. L'amplitude de la réponse fréquentielle du filtre continu est préservée. Mais l'expression (I.26) montre qu'il y a une distorsion non linéaire de l'échelle de fréquence. Pour résoudre ce problème, Golden et Kaiser §7§ suggèrent la prédistorsion des fréquences de coupures de $H(p)$ selon l'expression (I.26). Dans le cas général, il est nécessaire de définir une nouvelle fonction de transfert $H(p)$, qui peut être, ensuite, numérisée à l'aide de (I.28). Néanmoins, dans le cas où l'amplitude de la réponse fréquentielle du filtre prototype passe-bas $H(p)$ est de type constante par morceau, on procède ainsi :

(i) On transforme les fréquences des coupures de $H(p)$ selon (I.26).

(ii) On applique les transformations des bandes voulues au prototype passe-bas ainsi obtenu. Ces transformations ont pour effet de modifier les pôles et les zéros du filtre pour faire correspondre $H'(p)$ au gabarit voulu.

(iii) On échantillonne la F.T. ainsi obtenue à l'aide de (I.28). Le filtre numérique ainsi obtenu répondra aux spécifications imposées au départ.

I.4 Détermination et réalisation de la fonction de transfert échantillonnée du filtre numérique :

Il y a deux méthodes pour déterminer la F.T. du filtre numérique récursif qui satisfait aux spécifications imposées sur l'amplitude de la réponse fréquentielle. La première consiste à choisir la F.T. d'un filtre analogique H(p), puis de l'échantillonner. Il existe une documentation assez vaste concernant la F.T. analogique H(p) qui approche un gabarit de départ; Nous avons vu dans la section précédente les deux méthodes d'échantillonnage de H(p) : transformée en z ou transformée bi-linéaire. La méthode de la transformée en z est appelée méthode de "l'invariance impulsionnelle" par Rader et Gold §8§. En effet, la réponse impulsionnelle du filtre échantillonné est égale à la réponse impulsionnelle de H(p) aux instants d'échantillonnages. Les considérations suivantes illustrent ce fait. Soit h(t) la réponse impulsionnelle de H(p). Si l'on applique à ce système le train d'impulsions de (I.17), il en résulte le signal de sortie continu suivant :

$$y(t) = \sum_{n=0}^{\infty} f(nT)h(t-nT) \qquad (I.29)$$

Si, maintenant, on échantillonne y(t) avec la même période T en synchronisme avec l'échantillonnage du signal d'entrée, nous avons :

$$y(mT) = \sum_{n=0}^{\infty} f(nT)h\ ((m-n)T) \qquad (I.30)$$

Ceci étant un produit de convolution, en prenant la transformée en z de deux membres de I.30, on a : $Y(z) = F(z)\ H(z)$
où H(z) est la transformée en z de la suite /h(n)/ obtenue en échantillonnant h(t) avec la même période T.

Quand la F.T. H(p) est rationnelle en p, degré n, la transformée en z de H(p) est aussi une fonction rationnelle en z du même degré n. Quand H(p) est un rapport de deux polynômes rationnels en p (c'est le cas des F.T. des filtres analogiques), on décompose H(p) en fractions simples :

$$H(p) = \frac{N(p)}{D(p)} = \sum_{\substack{\text{pôles de} \\ H(p)}} \frac{1}{1 - e^{Tp_k}} \text{ (Résidus de H(p))} \qquad (I.31)$$

Cette décomposition donne lieu à une somme de termes tels que :

$$\frac{1}{(s+a)} \quad \text{ou} \quad \frac{1}{(s+a)^{k+1}} \quad \text{ou} \quad \frac{1}{s^{k+1}}$$

Ces termes ont des transformées en z bien connues (il est facile de les évaluer à partir de la définition de base (A.1)).

$$\left(\frac{1}{s+a}\right) = \frac{1}{1 - \exp(-aT)z^{-1}} \tag{I.32}$$

$$\left(\frac{1}{(s+a)^{k+1}}\right) = (-1)^k \frac{1}{k!} \frac{\partial^k}{\partial a^k} \left(\frac{1}{1-\exp(-aT)z^{-1}}\right) \tag{I.33}$$

$$\left(\frac{1}{s^{k+1}}\right) = (-1)^k \frac{1}{k!} \lim_{a \to 0} \frac{\partial^k}{\partial a^k} \left(\frac{1}{1-\exp(aT)\, z^{-1}}\right) \tag{I.34}$$

Il suffit ensuite de simplifier l'expression obtenue et de la mettre sous la forme suivante :

$$H(z) = \frac{\sum_{i=0}^{N} a(i) z^{-i}}{1 + \sum_{i=1}^{M} b(i) z^{-i}} \tag{I.35}$$

où les a(i) et les b(i) sont des réels.

Pour les filtres numériques à larges bandes où ω_c est voisin de $\frac{\omega_e}{2}$ ou, suivant le comportement de H(p) quand $p \to j\infty$, on pourrait envisager l'emploi de la transformée bi-linéaire, comme indiqué dans la section précédente. Dans ce cas, nous aboutissons aussi, après simplification, à une expression analogue à celle de (I.35). "L'implémentation" de cette F.T. échantillonnée dans une machine, est très directe. En effet, (I.35) donne lieu à l'équation (I.8) répétée ici §3§ :

$$y(nT) = (a_0x(nT) + a_1x(nT-T) + \ldots + a_Nx(nT-NT)) - b_1y(nT-T) - \ldots - b_My\ (nT-MT) \qquad (I.36)$$

La réalisation de cette équation nécessite la mémorisation de (N+1) coefficients a(i), (N+1) échantillons du signal d'entrée x(nT) M coefficients b(i), et M échantillons des sorties précédentes. Nous pouvons aussi faire un bilan des opérations arithmétiques nécessaires : il faudra (N+M+1) multiplications (si l'on suppose que a(i), b(i) $\neq$ 0 ou 1 pour tout (i)) et (N+M-1) additions au plus pour calculer la sortie y(nT). Si toutes ces opérations s'effectuent dans un temps t $\leq$ T, on dit que le filtre numérique opère en temps réel. Le schéma bloc correspondant à cette réalisation est donné fig. (I.4).

Nous verrons plus loin, (chapître IV) pourquoi cette forme de réalisation de H(z), dite forme "directe", est à déconseiller (précision nécessaire concernant les coefficients a(i) et b(i)).

Nous pouvons exprimer H(z) sous la forme suivante :

$$H(z) = \frac{1}{D(z)}\, N(z)$$

$$\text{où } N(z) = \sum_{i=0}^{N} a(i)\, z^{-i} \qquad (I.37)$$

$$\text{et } D(z) = 1 + \sum b(i)\, z^{-i}$$

et réaliser H(z) en deux parties, celle qui correspond à $\frac{1}{D(z)}$ d'abord, celle qui correspond à N(z) ensuite. Cette réalisation

est strictement équivalente à la précédente à l'exception de deux points suivants : (i) On économise les mémoires nécessaires au stockage des variables intermédiaires ; (ii) en ce qui concerne la propagation de l'erreur d'arrondi des produits, les deux formes sont très différentes l'une de l'autre comme nous le verrons dans le chapître V.

La réalisation du filtre dans cette forme est appelée "forme canonique". Nous pouvons écrire :

$$H(z) = H_1(z)\, H_2(z) = \frac{1}{D(z)}\, N(z)$$

soit $w(nT)$ la sortie du système $H_1(z) = \frac{1}{D(z)}$

ce qui nous donne les équations de récurrence :

$$w(nT) = x(nT) - b_1 w(nT-T) - \ldots b_M w(nT-MT) \quad \text{et} \qquad (I.38)$$

$$y(nT) = a_0 w(nT) + a_1 w(nT-T) + \ldots a_N w(nT-NT) \qquad (I.39)$$

Le schéma bloc correspondant à la réalisation de ces deux équations est donné fig. (I.6). Il suffit, dans cette forme, de mémoriser M variables w(n) et l'entrée présente x(nT) d'où une économie de mots-mémoires.

On peut aussi écrire $H(z)$ sous les deux autres formes suivantes §6§ :

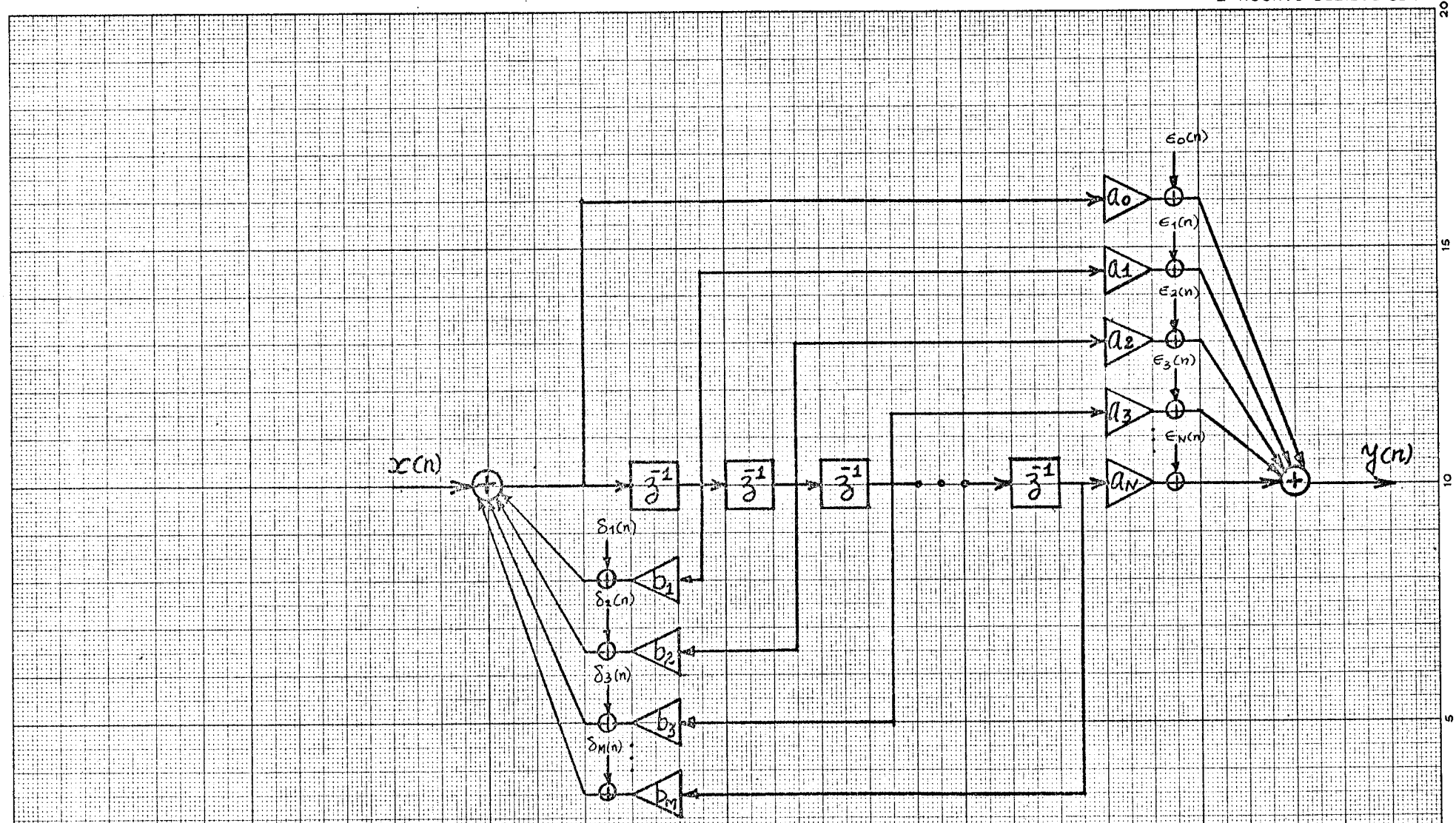
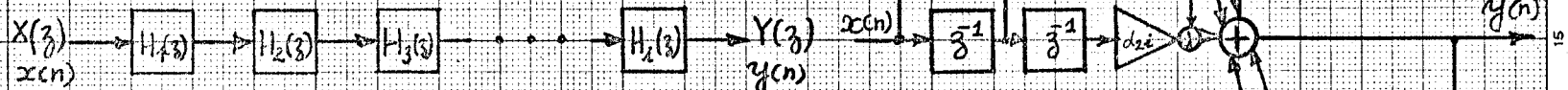


fig. I.6

La réalisation de filtre récursif dans la forme non-factorisée canonique.

La forme cascade



La forme parallele

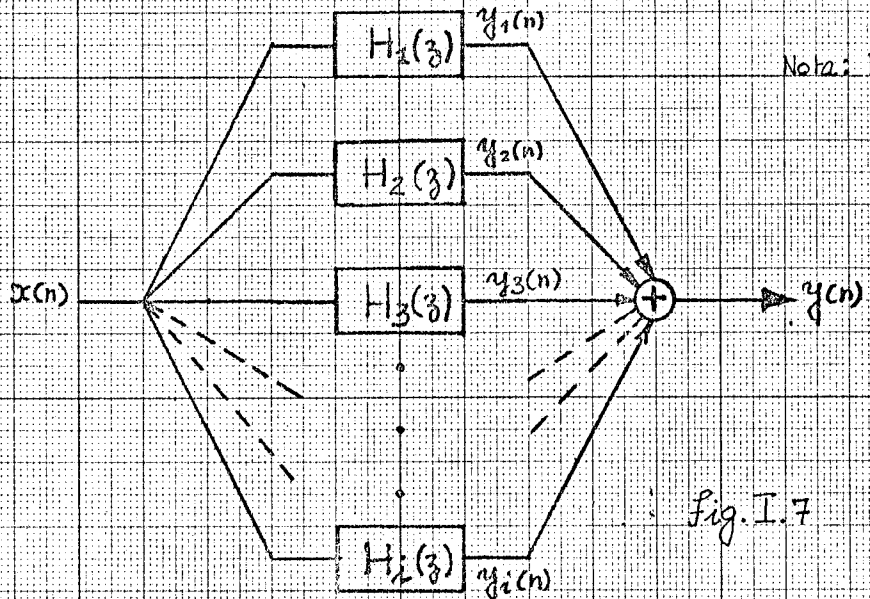


fig. I.7

$H_i(z)$: une cellule typique

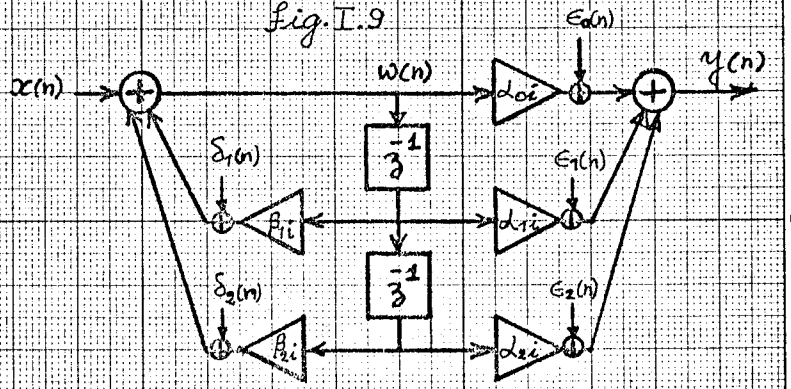
Realisation de $H_i(z)$

Sous la forme directe.

Nota: Pour la forme parallele $d_{2i} = 0$

fig. I.8

fig. I.9



la réclisation de $H_i(z)$ dans la forme canonique

$$H(z) = H_1(z) + H_2(z) + \ldots = \sum_i H_i(z) \qquad (I.40)$$

ou

$$H(z) = H'_1(z) \cdot H'_2(z) \cdot H'_3(z) \ldots = \prod_i H'_i(z) \qquad (I.41)$$

La forme de décomposition (I.40) est appelée forme parallèle. Cette décomposition donne lieu à la mise en parallèle de cellules de deuxième ordre comme le montre le schéma bloc fig. (I.7).

Chaque cellule $H_i(z)$ a pour fonction de transfert :

$$H_i(z) = \frac{A(0i) + A(1i)\, z^{-1}}{1 + \beta(1i)\, z^{-1} + \beta(2i)\, z^{-2}} \qquad (I.42)$$

où A(1i) et $\beta(2i)$ peuvent être nuls. Dans cette décomposition, nous supposons toujours que H(p) a seulement des pôles simples, réels ou complexes, ce qui est vrai en général.

La forme (I.41) est dite de forme cascade. Le schéma bloc correspondant est donné fig. (I.7). Une cellule $H_i(z)$ type a pour F.T. :

$$H_i(z) = \frac{\alpha(0i) + \alpha(1i)\, z^{-1} + \alpha(2i)\, z^{-2}}{1 + \beta(1i)\, z^{-1} + \beta(2i)\, z^{-2}} \qquad (I.43)$$

Ces deux décompositions peuvent être réalisées sous forme canonique, exactement comme dans le cas de la forme non décomposée, ce qui conduit aux schémas de bloc donnés fig. (I.8) et (I.9).

En fait, il existe d'autres formes de réalisation de ces filtres élémentaires. Rader et Gold §9§ ont proposé une forme particulière. Jackson §IO§ a proposé des formes transposées de I.42 et I.43. Nous nous bornerons ici à l'étude des formes I.42 et I.43 qui sont courantes et plus particulièrement la forme (I.43).

On peut trouver la relation qui relie les coefficients $\beta(1i)$ et $\beta(2i)$ et les pôles complexes conjugués correspondants, exprimés en coordonnées polaires.

On a : $1 + \beta(1i) z^{-1} + \beta(2i) z^{-2} = 0$

La résolution de cette équation quadratique fournit l'expression des pôles, soit $z_{1,2} = -\frac{\beta(1i)}{2} \pm \frac{\sqrt{\beta(1i) - 4\beta(2i)}}{2}$

Si r est le module et θ la phase correspondant du pôle, on a les relations suivantes :

$$\beta(2i) = r^2 \qquad \text{(I.44)}$$

et

$$\beta(1i) = -2r \cos\theta \qquad \text{(I.45)}$$

Les considérations de stabilité du système §5§ imposent les contraintes suivantes sur le module :

$$r < 1 \quad \text{d'où} \quad \beta(2i) < 1 \qquad \text{(I.46)}$$

et donc

$$|\beta(1i)| < 2 \qquad \text{(I.47)}$$

Cette transposition en numérique du filtre analogique ayant pour F.T. H(p) est évidemment très utile quand on veut se limiter aux filtres analogiques classiques (Butterworth, Chebycheff, Elliptique, etc ...) qui donnent des gabarits de type passe-bas, passe-bande, passe-haut et stop-bande.

Si l'on veut réaliser des filtres ayant des gabarits quelconques, il faut recourir à des méthodes beaucoup plus générales. Etant donné l'importance que prend la simulation sur ordinateur des systèmes physiques et biologiques, où ces types de gabarits complexes peuvent représenter des modèles de F.T., une méthode générale pour déterminer la nature et le nombre de cellules d'un filtre numérique quelconque, s'avère nécessaire. Les filtres que nous avons cherché à déterminer, doivent "s'implémenter" facilement sur des petites unités de traitement numérique. C'est dans cette optique que nous avons mis en oeuvre en centre de calcul, un programme utilisant l'algorithme puissant d'optimisation non linéaire, celui de Fletcher et Powell, pour la conception de tels filtres. La méthode ainsi que les résultats obtenus seront développés dans le chapître II.

C H A P I T R E II

CONCEPTION ASSISTEE PAR ORDINATEUR

DES FILTRES NUMERIQUES RECURSIFS

II.1. Généralités sur le problème de l'optimisation :

Dans le chapître précédent, nous avons vu qu'il y avait deux méthodes pour concevoir les filtres numériques récursifs, soit : (i) échantillonner la F.T. du filtre analogique qui correspond aux spécifications imposées sur le gabarit fréquentiel. On obtient directement les pôles et les zéros du système.

(ii) évaluer les coefficients a(i) et b(i) qui caractérisent la F.T. du filtre numérique à l'aide d'un programme d'optimisation. Partant d'un jeu initial les valeurs pour des a(i) et b(i) et faisant varier ceux-ci, le programme approche le gabarit selon un critère d'erreur choisi au préalable.

Nous allons étudier la deuxième méthode et présenter les résultats que nous avons obtenus lors de l'élaboration de quelques filtres numériques.

Avant de procéder à l'étude proprement dite de l'algorithme d'optimisation de Fletcher et Powell, nous allons d'abord aborder le problème général de l'optimisation et des généralités théoriques sur la méthode du gradient dite de la "plus grande pente" ("Steepest descent" en anglais).

Le problème général de l'optimisation repose sur deux considérations fondamentales : la première est la mesure du comportement du système pour un jeu donné des paramètres à optimiser.

La deuxième concerne le choix d'un critère d'erreur à minimiser. Précisons plus en détail ces deux considérations.

Puisque nous nous interessons à la réponse en amplitude du filtre, nous allons formuler ces considérations en ces termes. Pour ce faire, nous allons utiliser les notations suivantes :

$X_1, X_2, X_3, \ldots X_M$: sont les M paramètres ajustables du système à optimiser. Dans notre cas, ce sont les coefficients a(i) et b(i) de la F.T. discrète.

$y(f, \hat{X})$: est la réponse en amplitude du système à optimiser. Elle est une fonction de la fréquence et du vecteur $\hat{\hat{X}}$ représentant les M paramètres.

$y_d(f)$: est la réponse en amplitude désirée du système. Elle est aussi une fonction de la fréquence. Elle est définie sur l'intervalle $0, f_N$ où f_N = la fréquence de Nyquist = $\frac{1}{2} f_e$

Il est bien évident que la valeur de $y(f,\hat{X})$ devra être facilement estimée pour tout f et $\hat{X}$. Des considérations de temps de calcul imposent que cette estimation repose sur des calculs les plus simples possible, car ces calculs vont se répéter des centaines de fois (voire même des milliers) avant la détermination finale du vecteur $\hat{X}$.

On peut définir la fonction d'erreur $e(f,\hat{X})$ comme :

$$e(f,\hat{X}) = y(f,\hat{X}) - y_d(f) \qquad \text{(II.1)}$$

L'erreur est donc aussi une fonction de la fréquence f, et du vecteur des paramètres $\hat{X}$.

Par nécessité, la critère d'erreur doit être un nombre réel pour pouvoir ordonner et comparer les différentes réponses résultant des différents $\hat{X}$. Pour ce faire, on élimine f de (II.1) et l'on définit une fonction de coût $c(\hat{X})$. Communément, deux sortes de fonctions de coût sont utilisées. Elles sont définies comme :

$$c(\hat{X}) = \int_{\Delta f} e^2(f,\hat{X})\, w(f)\, df \qquad \text{(II.2)}$$

ou

$$c(\hat{X}) = \max_{f} \left| e(f,\hat{X}) \right| \qquad \text{(II.3)}$$

Dans la définition (II.2) on intègre le carré de l'erreur, pondéré (si on désire) par la fonction $w(f)$, définie sur un intervalle (ou bande) de fréquence Δf. Pour l'évaluation numérique, (II.2) peut se mettre sous la forme suivante :

$$c(\hat{X}) = \sum_{i=1}^{N} e^2(f_i,\hat{X}_i)\, w_i \qquad \text{(II.4)}$$

Avant d'aborder la formulation mathématique de la méthode de la plus grande pente, nous allons rappeler les problèmes posés par cette méthode et les solutions que nous avons retenues pour les résoudre.

(i) l'algorithme de Fletcher-Powell est un algorithme qui n'applique aucune contrainte sur les variables à optimiser. Dans le cas du filtrage numérique, la stabilité du filtre impose que le module des pôles soit < 1. Les considérations de déphasage minimale donnent lieu à la même contrainte pour les zéros du filtre. La méthode que nous avons utilisée pour incorporer ces contraintes dans le programme d'optimisation est détaillée dans le sous-chapître II.2.

(ii) le choix d'un critère d'erreur s'impose et ce choix est très important. En général, ce choix dépend de la nature du problème à résoudre. Pour la conception des filtres numériques le critère II.3 est plus efficace que celui de II.4 parce qu'il nous permet de borner l'erreur, c'est-à-dire la déviation entre le gabarit réalisé et le gabarit imposé, pour toutes les fréquences. Dans la conception des filtres passe-bande (ou stop-bande) où la bande passante (stoppante) est très étroite, le critère II.4 peut cacher des déviations importantes du fait qu'il est une moyenne pondérée. Néanmoins, nous avons choisi le critère II.4 pour la raison suivante :

- le sous programme de Fletcher-Powell fait appel un grand nombre de fois (plusieurs centaines) à un sous-programme d'évaluation analytique du vecteur gradient.

- compte tenu de (i) la technique d'application des contraintes sur les pôles et les zéros peut conduire à l'appel du sous-programme de Fletcher-Powell plusieurs fois.

- Donc, une réduction du temps de calcul dans le sous-programme d'évaluation du vecteur gradient entraîne un gain appréciable sur le temps global de calcul.

La méthode d'évaluation du vecteur gradient est indiquée dans l'annexe C.

(iii) Tout algorithme de minimisation basé sur le développement en série de Taylor de la fonction du coût $c(\hat{X})$ opère sur le principe de "l'excursion locale", c'est-à-dire que le minimum trouvé $\hat{X}min$ est tel que $c(\hat{X}min) \leq c(\hat{X}min + \Delta\hat{X})$ pour des $\Delta\hat{X}$ suffisamment petits au voisinage du point $\hat{X}min$ ($\hat{X}min$ et $\Delta\hat{X}$ sont des vecteurs de dimensions M dans notre cas). Donc, nous ne pourrons jamais affirmer que le minimum obtenu est le minimum minimorum. Dans notre cas, nous augmentons la probabilité pour que le minimum obtenu soit aussi le vrai minimum, en utilisant la même procédure d'optimisation deux ou trois fois avec des valeurs initiales $\hat{X}$ différentes. Nous verrons plus loin, comment le programme tient compte des contraintes sur les pôles et les zéros du filtre, les a(i) et les b(i) et de l'unicité du minimum.

Après ces remarques préliminaires, nous abordons la représentation mathématique de la méthode de la plus grande pente.

Le point de départ de cette méthode d'optimisation est le développement de la fonction de coût en une série de Taylor autour du point minimum $\hat{X}min$:

$$c(\hat{X}+\Delta\hat{X}) = c(\hat{X}) + \sum_{j=1}^{M} \frac{\partial c(\hat{X})}{\partial X_j} \Delta X_j + \frac{1}{2} \sum_{j=1}^{M} \sum_{k=1}^{M} \frac{\partial^2 c(\hat{X})}{\partial X_j \partial X_k} \Delta X_j \Delta X_k + \cdots \qquad \text{(II.5)}$$

expression que l'on peut réécrire sous le forme matricielle suivante (en négligeant les termes d'ordre supérieurs à deux) :

$$c + \Delta c = c + G^T \Delta\hat{X} + \frac{1}{2} \Delta\hat{X}^T H \Delta\hat{X} \qquad \text{(II.6)}$$

avec

$$\hat{X} = \begin{bmatrix} \Delta X_1 \\ \Delta X_2 \\ \vdots \\ \Delta X_M \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_M \end{bmatrix} \qquad \text{(II.7)}$$

$$G = \begin{bmatrix} \frac{\partial c(\hat{X})}{\partial X_1} \\ \vdots \\ \frac{\partial c(\hat{X})}{\partial X_M} \end{bmatrix} \qquad \text{(II.8)}$$

et

$$H = \begin{bmatrix} \frac{c(\hat{X})}{\partial X_1 \partial X_2} & \cdots & \frac{c(\hat{X})}{\partial X_1 \partial X_M} \\ \vdots & & \vdots \\ \frac{c(\hat{X})}{\partial X_M \partial X_1} & & \frac{c(\hat{X})}{\partial X_M \partial X_M} \end{bmatrix} \qquad \text{(II.9)}$$

et où $\hat{X}^T$ = la transposée du vecteur $\hat{X}$

Le vecteur gradient G est orienté dans la direction de la croissance maximum de $c(\hat{x})$. C'est dire que pour un $\Delta\hat{X}$ petit et une norme $\|\Delta\hat{X}\|$ constante, la plus grande augmentation de $c(\hat{X})$ est obtenue lorsque $\hat{X}$ a la direction du vecteur gradient G. On démontre ceci facilement à partir de l'inégalité Cauchy-Schwarz §6§ :

$$\left|\frac{(G,\Delta\hat{X})}{\|G\| \quad \|\hat{X}\|}\right| \leq 1 \qquad \text{(II.IO)}$$

où $(\hat{X},\hat{Y})$ est le produit scalaire des vecteurs $\hat{X}$ et $\hat{Y}$.

Dans (II.IO), l'égalité est vérifiée pour $G = a\,\Delta\hat{X}$ où a = constante. C'est-à-dire, quand $\hat{X}$ a la même direction que G. Si l'on se déplace dans la direction opposée au gradient G, on suit "la plus grande pente" conduisant au minimum ; d'où le nom de la méthode.

La matrice H est appelée la matrice "Hessienne". Elle est symétrique et représente le comportement quadratique de la fonction du coût $c(\hat{X})$. En particulier, à un point stationnaire de $c(\hat{X})$, le gradient G devient nul et l'équation (II.6) devient :

$$c + \quad c = c + \frac{1}{2}\,\Delta\hat{X}^T\, H\, \Delta\hat{X} \qquad \text{(II.11)}$$

La condition nécessaire et suffisante pour que le point stationnaire soit le minimum est que la forme quadratique $\Delta\hat{X}^T\, H\, \Delta\hat{X}$ soit définie positive pour tout $\Delta\hat{X}$ différent de zéro.

L'équation (II.5) peut se mettre sous la forme suivante :

$$c(\hat{X} + \Delta\hat{X}) = c(\hat{X}) + \sum_{i=1}^{M} g(i)x(i) + \frac{1}{2}\sum_{i=1}^{} \sum_{j=1}^{} h(ij)\ x(i)\ x(j) \qquad \text{(II.12)}$$

où $x(i) = \Delta\hat{X}(i)$

$$g(i) = \frac{\partial c(\hat{X})}{\partial X(i)}$$

et

$$h(ij) = h(ji) = \frac{\partial^2 c(\hat{X})}{\partial X(i)\partial X(j)}$$

On détermine les M conditions nécessaires pour le minimum en évaluant les dérivées partielles par rapport à $x(k)$ pour $k = 1, 2, \ldots M$ de l'équation (II.12). Ces dérivées doivent être nulles et par suite :

$$g(k) + \frac{1}{2}\sum_{i=1}^{M} h(ik)x(i) + \frac{1}{2}\sum_{j=1}^{M} h(kj)x(j) = 0 \qquad \text{(II.13)}$$

$$k = 1, 2 \ldots M$$

Puisque la matrice H est symétrique, nous avons $h(kj) = h(jk)$. (II.13) s'écrit sous la forme suivante :

$$\sum_{j=1}^{M} h(kj)x(j) = -g(k),\ k = 1, 2, \ldots M \qquad \text{(II.14)}$$

ou en notation matricielle :

$$H \Delta\hat{X} = -G \qquad \text{(II.15)}$$

Ainsi les variations à appliquer aux paramètres pour obtenir le minimum sont obtenues en résolvant le système (II.14).

Cependant il est à remarquer que cette résolution nécessite l'évaluation préalable des M dérivées partielles de second ordre. Cette approche est rarement utilisée à cause du temps de calcul prohibitif.

Néanmoins, les équations (II.14) et (II.15) représentent le point de départ de l'algorithme de Fletcher-Powell qui évitera le calcul direct des éléments de la matrice H. Cependant il nécessite le calcul du gradient de $c(\hat{X})$ pour $\hat{X}$ donné. La méthode d'évaluation du gradient, dans le cas des filtres numériques que nous avons choisis, est indiquée dans l'annexe C. L'étude de l'algorithme de Fletcher-Powell est faite dans l'annexe B.

II.2 Conception des filtres numériques par la méthode d'optimisation de Fletcher-Powell §13§.

Examinons plus en détail notre cas particulier :

La F.T. du filtre numérique donnée par la relation sera le point de départ. Mais celle-ci présente les inconvénients suivants :

(i) Si l'on veut respecter les contraintes sur les pôles (nécessaires à la stabilité), on est amené à mettre en facteur le dénominateur.

(ii) Les racines du dénominateur peuvent varier très sensiblement en fonction de la précision de calcul.

Il est donc préférable de choisir l'une des formes décomposées (I.40) ou (I.41). La relation (I.40) qui aboutit à la réalisation en forme parallèle nécessite une factorisation du numérateur si l'on veut appliquer les contraintes de déphasage minimal. C'est pourquoi nous avons choisi la forme cascade (I.41) dont une cellule type a pour F.T. la relation (I.43) répétée ici :

$$H(z^{-1}) = \prod_{i=0}^{K} H(i)(z^{-1})$$

$$H(i)(z^{-1}) = \frac{1 + a(1i)z^{-1} + a(2i)z^{-2}}{1 + b(1i)z^{-1} + b(2i)z^{-2}} \qquad \text{(II.I6)}$$

et $H(0)(z^{-1}) = A$ gain statique du filtre.

Le problème peut être énoncé ainsi :

soit $R^d(i)$ $(i = 1, 2, \ldots M)$ le module de la réponse fréquentielle désirée, spécifiés en M points, W(1), W(2), ... W(M)

où $W(i) = \frac{2f(i)}{fe}$, fe = la fréquence d'échantillonnage.

Soit $R(z(i))$, $i = 1, 2, \ldots M$ le module de la réponse fréquentielle évaluée aux fréquences W(i). Il est donné par :

$$R(z(i)) = \prod_{i=0}^{K} \left| H(i)(z(i)) \right|$$

$$= \prod_{i=1}^{K} \left| \frac{1+a(1i)z^{-1}(i)+a(2i)z^{-2}(i)}{1+b(1i)z^{-1}(i)+b(2i)z^{-2}(i)} \right| A \qquad \text{(II.17)}$$

où $z(i) = \exp(j\omega(i)T)$

où les a(1i), a(2i), b(1i), b(2i) représentent un choix particulier des coefficients.

Soit $c(\theta)$, la fonction de coût à minimiser, θ étant le vecteur des paramètres de dimension (4K+1), θ = (A,a(11), a(21), b(11), b(21), ... a(1K), a(2K), b(1K), b(2K))^T. Si nous prenons comme critère de coût la somme des carrés des erreurs, nous avons :

$$c(\theta) = \sum_{i=1}^{M} \left(R(z(i)) - R^d(i)\right)^2 \qquad \text{(II.18)}$$

Alors, le problème est de trouver un θ^* tel que :

$$c(\theta^*) \leq c(\theta) \quad \text{pour tout } \theta \text{ admissible.}$$

Pour ce faire, on a intérêt à réduire, si possible, la dimension du vecteur θ. En effet, en minimisant $c(\theta)$ par rapport à A, on peut définir un nouveau vecteur $\hat{\theta}$ de dimensions 4K.

On a :

$$\text{soit } \hat{\theta}^T = (a(11), a(21), b(11), b(21), \ldots a(1K), a(2K), b(1K), b(2K))$$

$$\text{soit } R(z(i)) = A \left| H(z(i), \hat{\theta}) \right|$$

$$\text{où } H(z(i), \hat{\theta}) = \left| \prod_{i=1}^{K} \right| H(i)(z(i), \hat{\theta})$$

$$\text{et } c(\hat{\theta}) = \sum_{i=1}^{M} \left(\left| AH(z(i), \hat{\theta}) \right| - R^d(i) \right)^2 \qquad \text{(II.19)}$$

En différenciant par rapport à A , on trouve le A qui minimise $c(\hat{\theta})$. Soit :

$$\frac{\partial c(\theta)}{\partial |A|} = \sum_{i=1}^{M} \left(2 \left| AH(z(i), \hat{\theta}) \right| - R^d(i)\right) \left| H(z(i), \hat{\theta}) \right| = 0$$

$$\text{d'où } |A^*| = \frac{\sum_{i=1}^{M} H(z(i), \hat{\theta})\, R^d(i)}{\sum_{i=1}^{M} \left| H(z(i), \hat{\theta}) \right|^2} \qquad \text{(II.20)}$$

Donc, on a la nouvelle fonction de coût :

$$\hat{c}(\hat{\theta}) = c(A^*, \hat{\theta}) \qquad \text{(II.21)}$$

Le signe de A^* est pris positif, ce signe est sans influence sur le gabarit de la réponse en amplitude.

Maintenant, on peut utiliser l'algorithme de Fletcher-Powell pour minimiser $\hat{c}(\hat{\theta})$. On a besoin pour cela du vecteur gradient $\frac{\partial \hat{c}}{\partial \hat{\theta}(n)}$ $n = 1, 2 \ldots 4K$. Ce dernier peut être évalué analytiquement par un programme assez court. La méthode est développée en annexe C.

Pour les filtres numériques, il y a deux contraintes :

(i) Les racines (ou pôles) du dénominateur doivent être situées à l'intérieur du cercle unité dans le plan z, pour que le système soit stable.

(ii) Pour éviter un déphasage excessif, la même contrainte s'applique aux zéros.

Ces contraintes sont assez faciles à introduire dans le programme. Pour cela, il suffit de remarquer les faits suivants : soit b, un zéro réel de D(z). Si l'on remplace b par 1/b, D(z) est multiplié par le facteur $(z-b) / (z-1/b)$.

Ce facteur a, pour $z = \exp(j\omega T)$, une amplitude constante égale à b. Donc, l'inversion des pôles (zéros) qui se trouvent à l'extérieur du cercle unité ne modifie pas la réponse en amplitude de D(z). Quand on sort du sous-programme d'optimisation, les pôles et les zéros sont dans un ordre quelconque dans le plan de z. On les teste et on inverse ceux qui sont à l'extérieur du cercle unité. Puis on redémarre l'optimisation avec ces nouveaux paramètres. L'organigramme (II.1) indique la manière dont se déroulent ces différents tests et la procédure pour terminer l'itération.

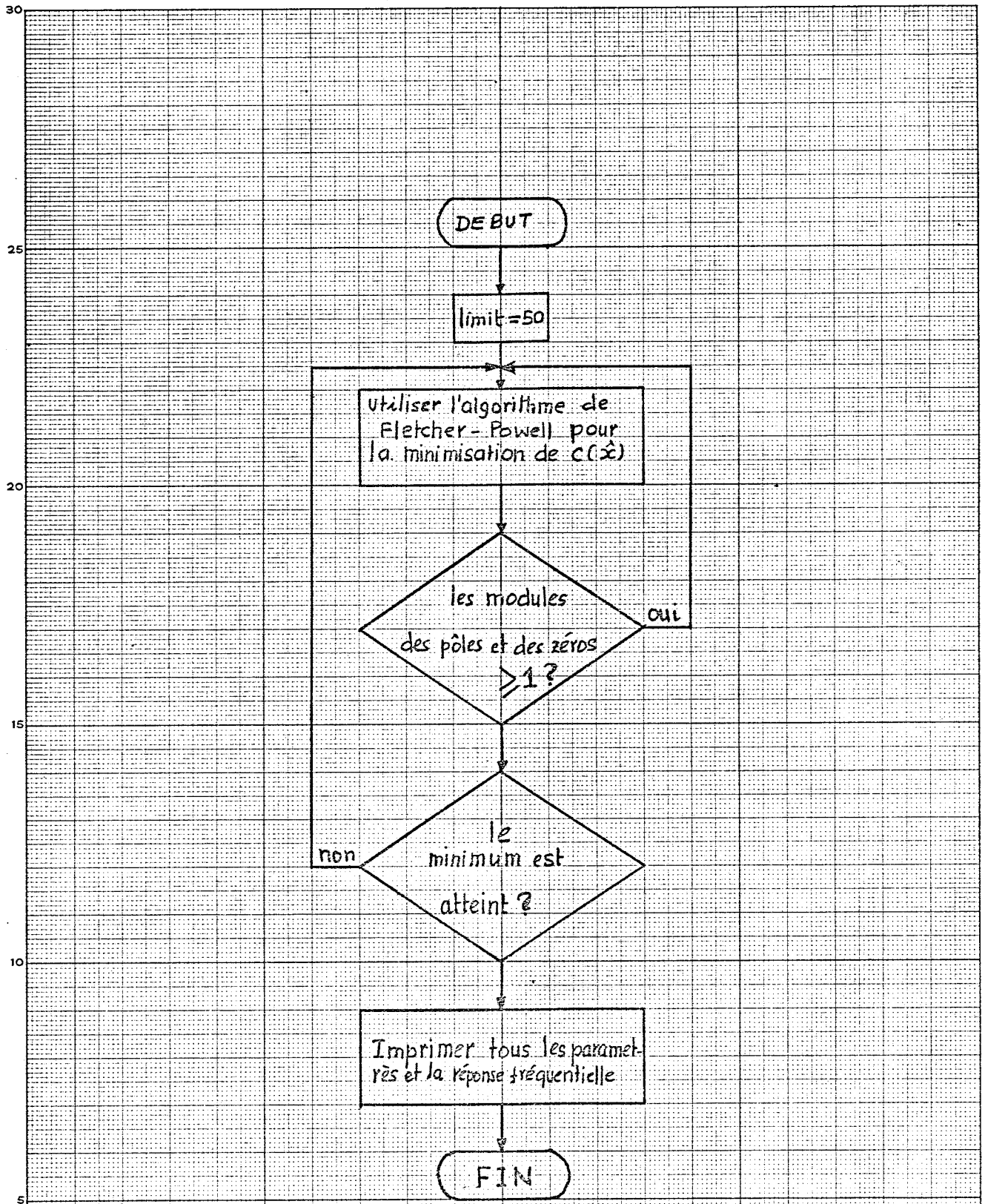

DEBUT
limit=50
Utiliser l'algorithme de Fletcher-Powell pour la minimisation de C(x̂)
les modules des pôles et des zéros ≥1 ?
oui
le minimum est atteint ?
non
Imprimer tous les paramètres et la réponse fréquentielle
FIN

fig. II.1

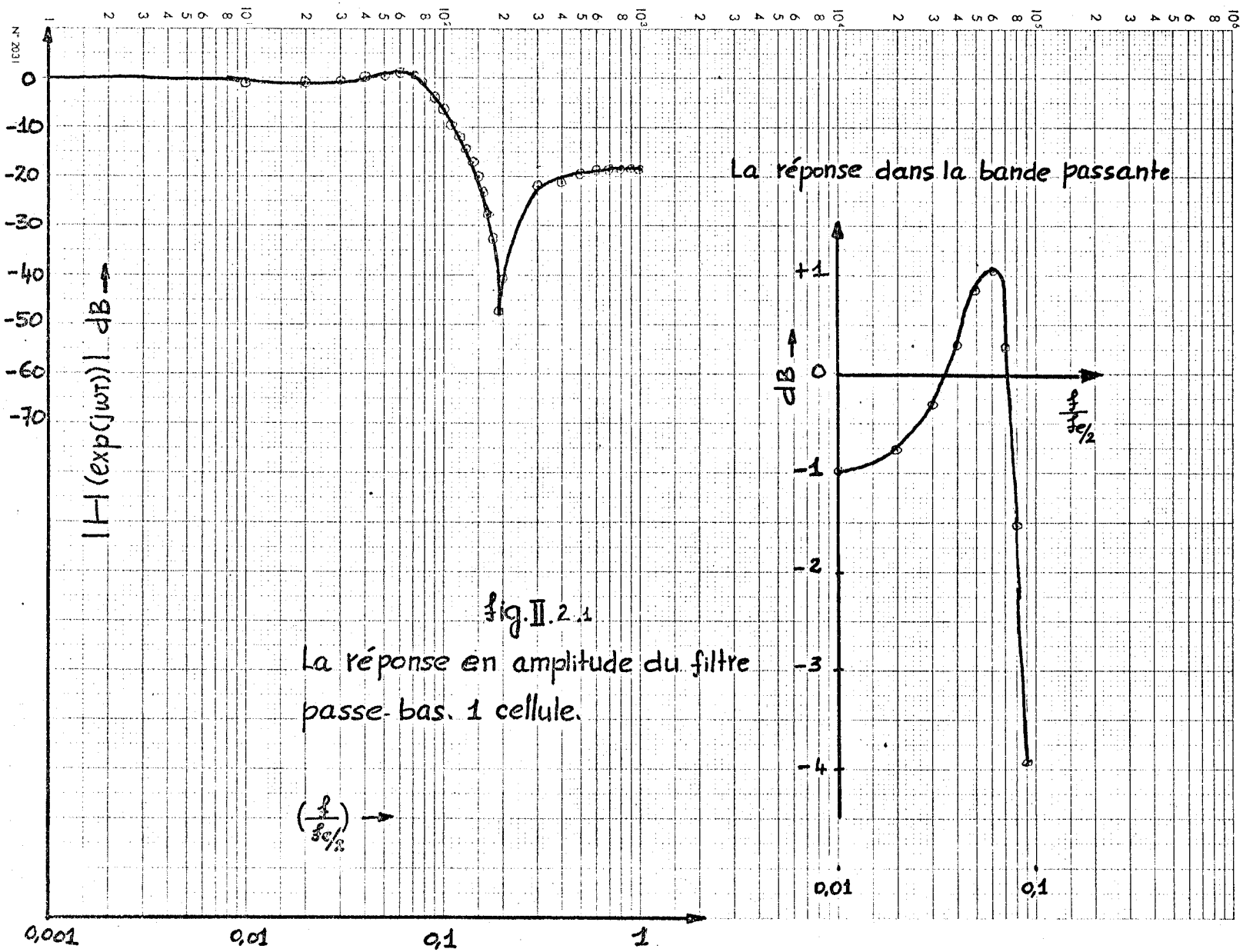


fig.II.2.1

La réponse en amplitude du filtre passe-bas. 1 cellule.

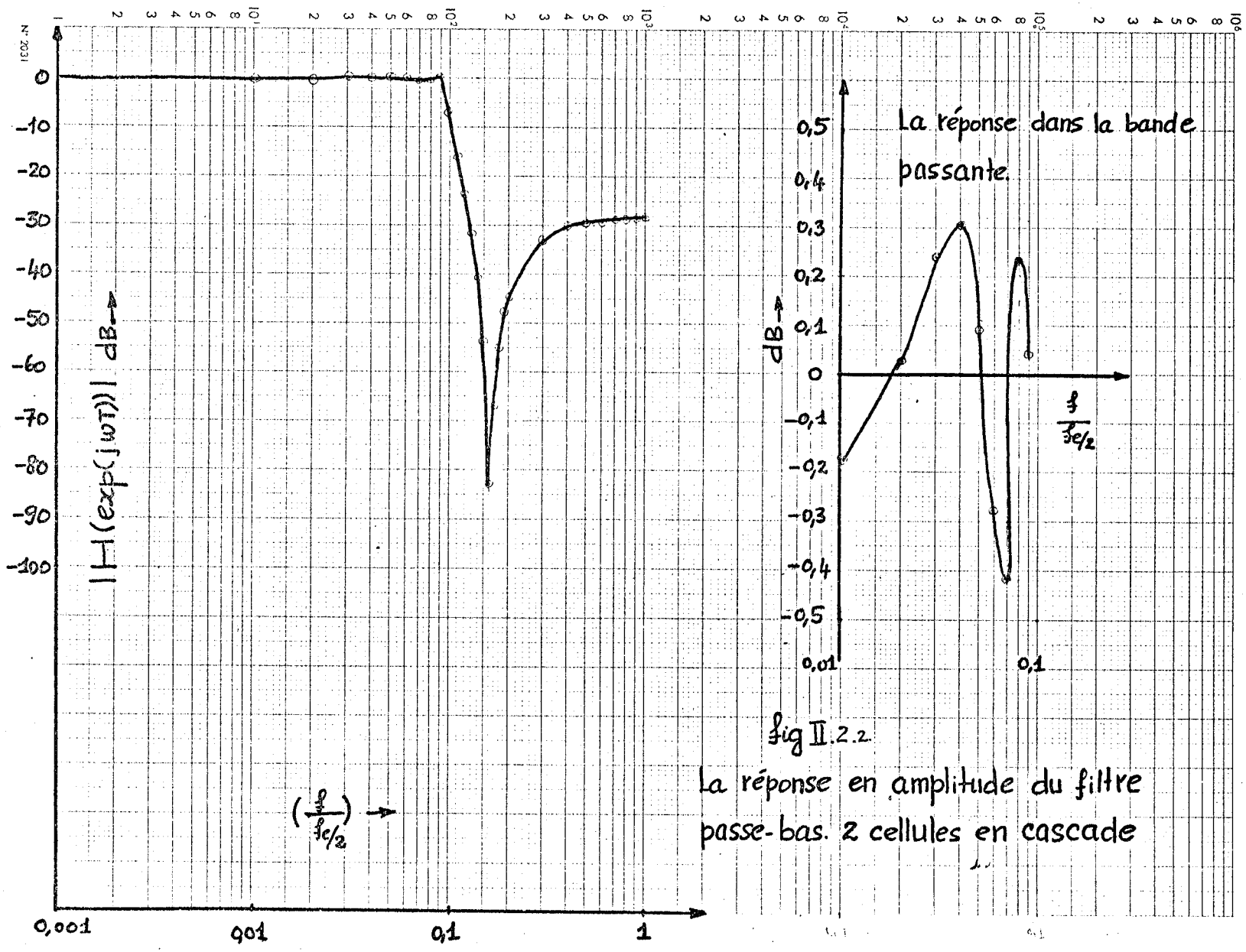
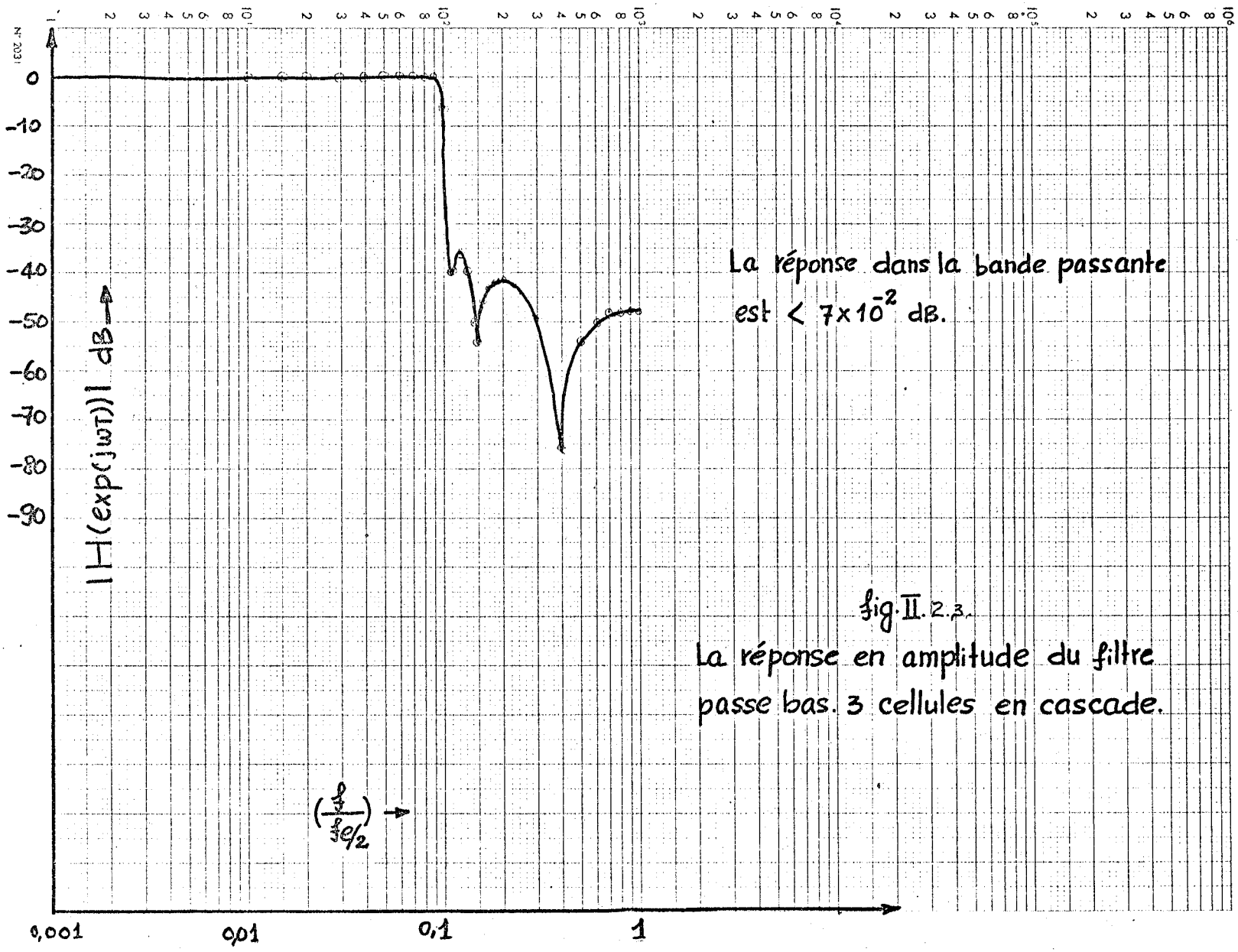


fig II.2.2.

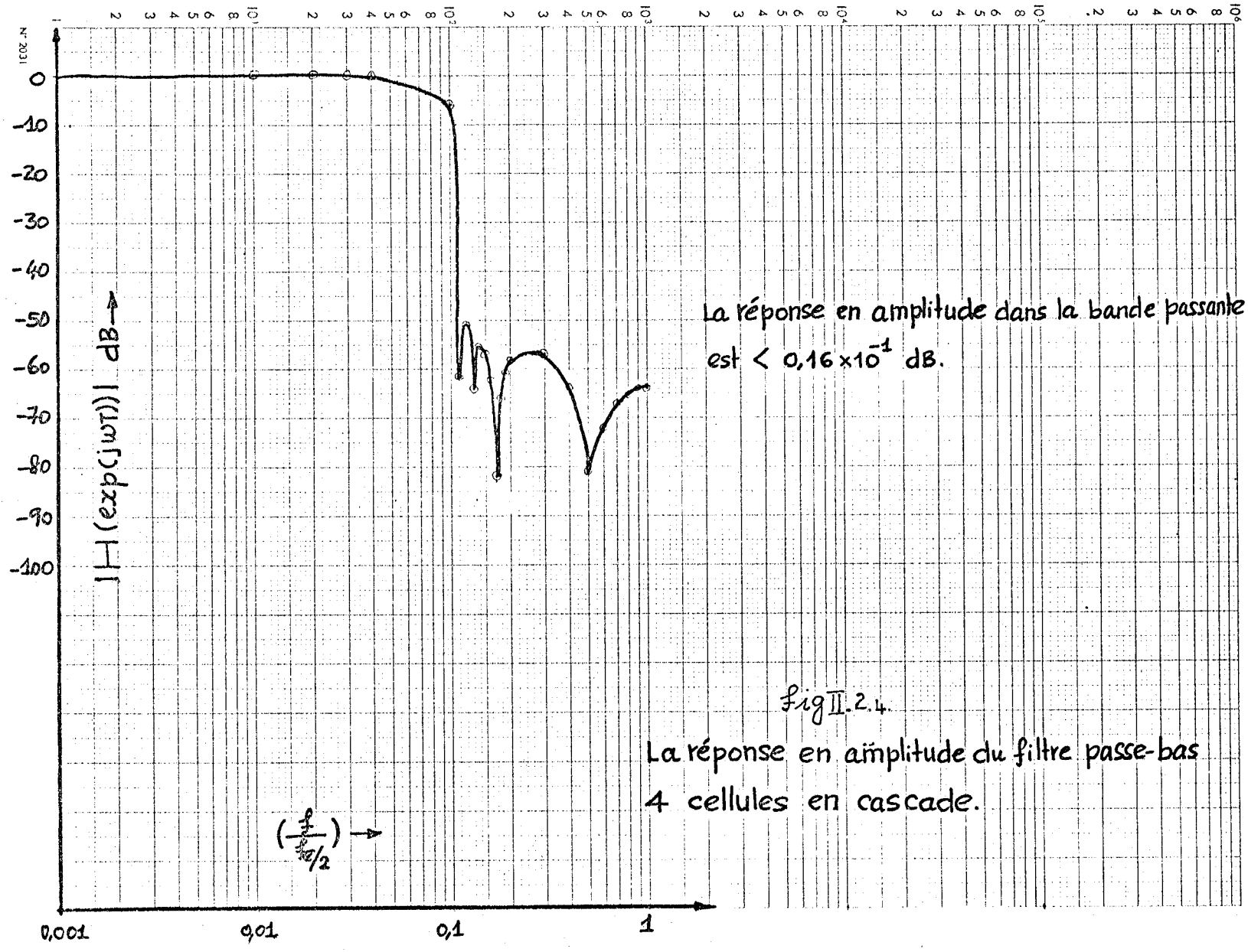
La réponse en amplitude du filtre passe-bas. 2 cellules en cascade



La réponse dans la bande passante est $< 7 \times 10^{-2}$ dB.

fig. II.2.3.

La réponse en amplitude du filtre passe bas. 3 cellules en cascade.



La réponse en amplitude dans la bande passante est $< 0{,}16 \times 10^{-1}$ dB.

fig II.2.4.

La réponse en amplitude du filtre passe-bas 4 cellules en cascade.

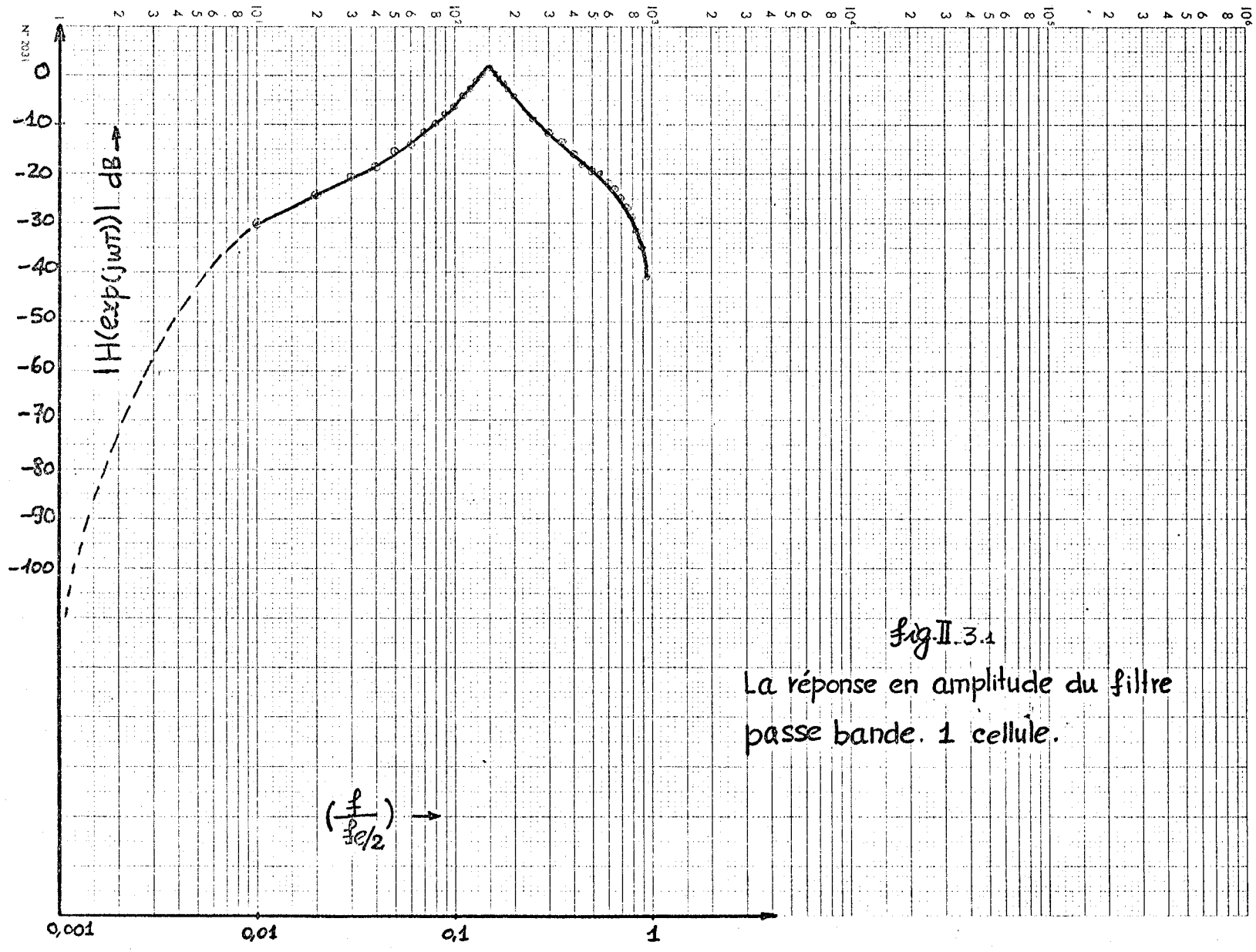
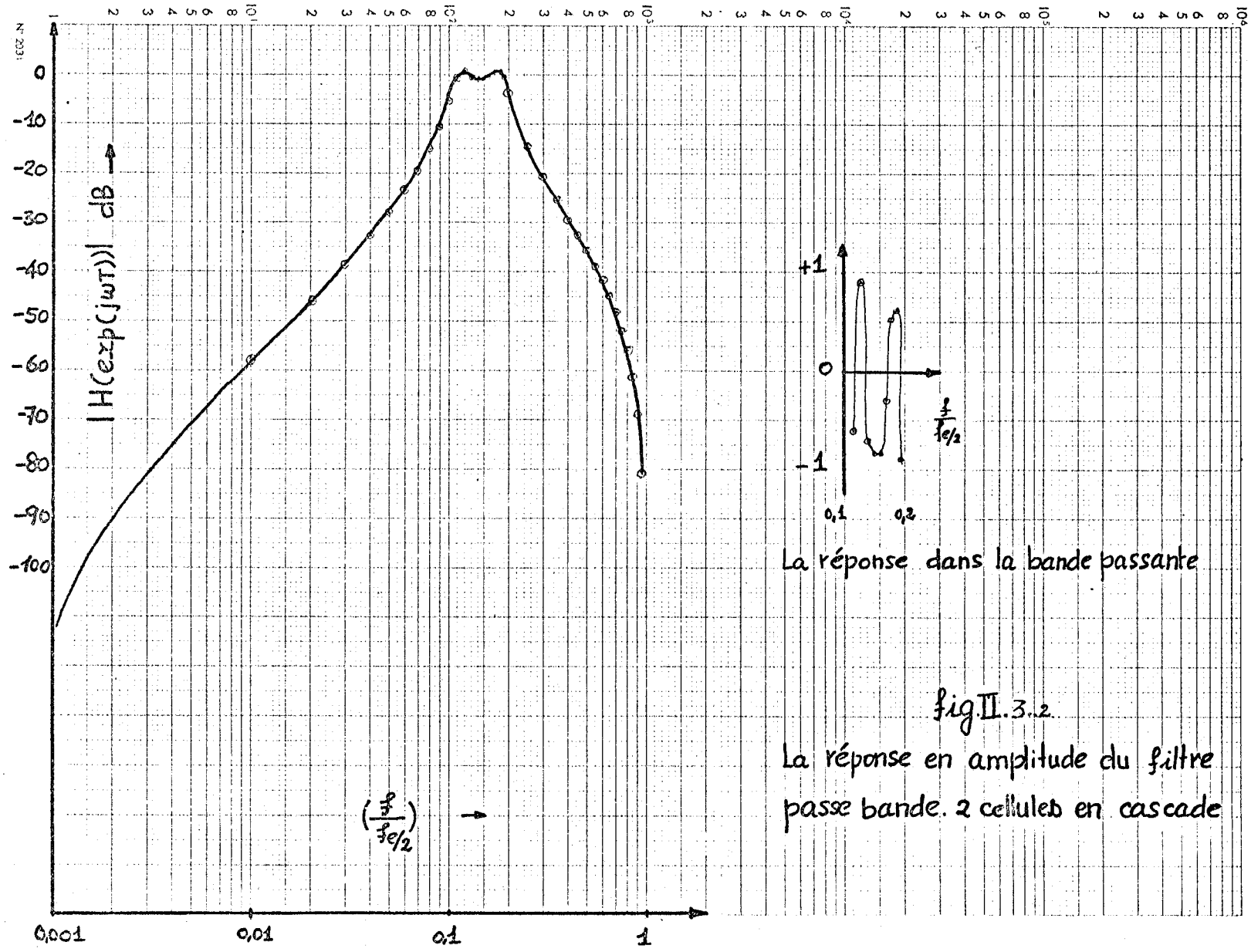


fig.II.3.1

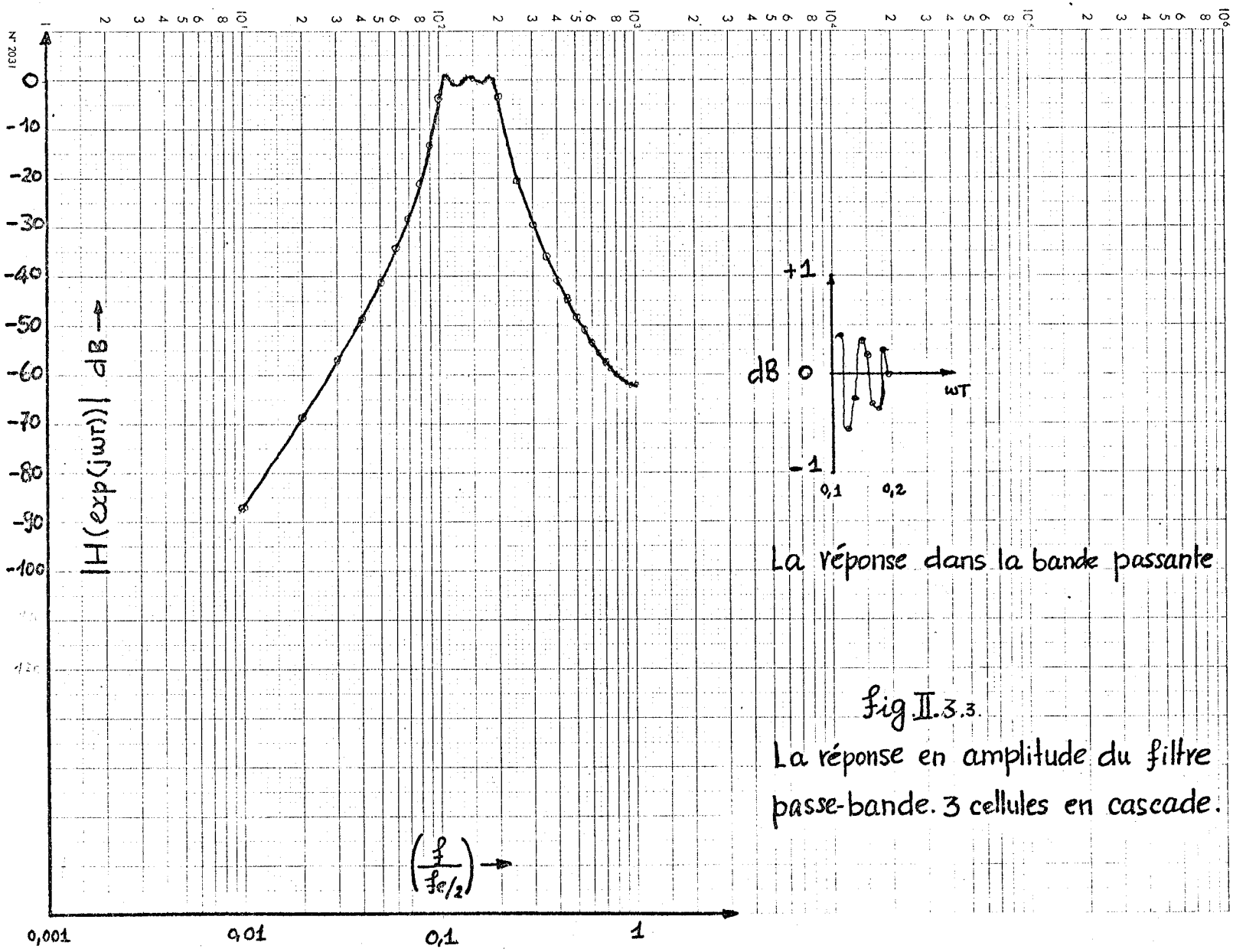
La réponse en amplitude du filtre passe bande. 1 cellule.



La réponse dans la bande passante

fig.II.3.2.

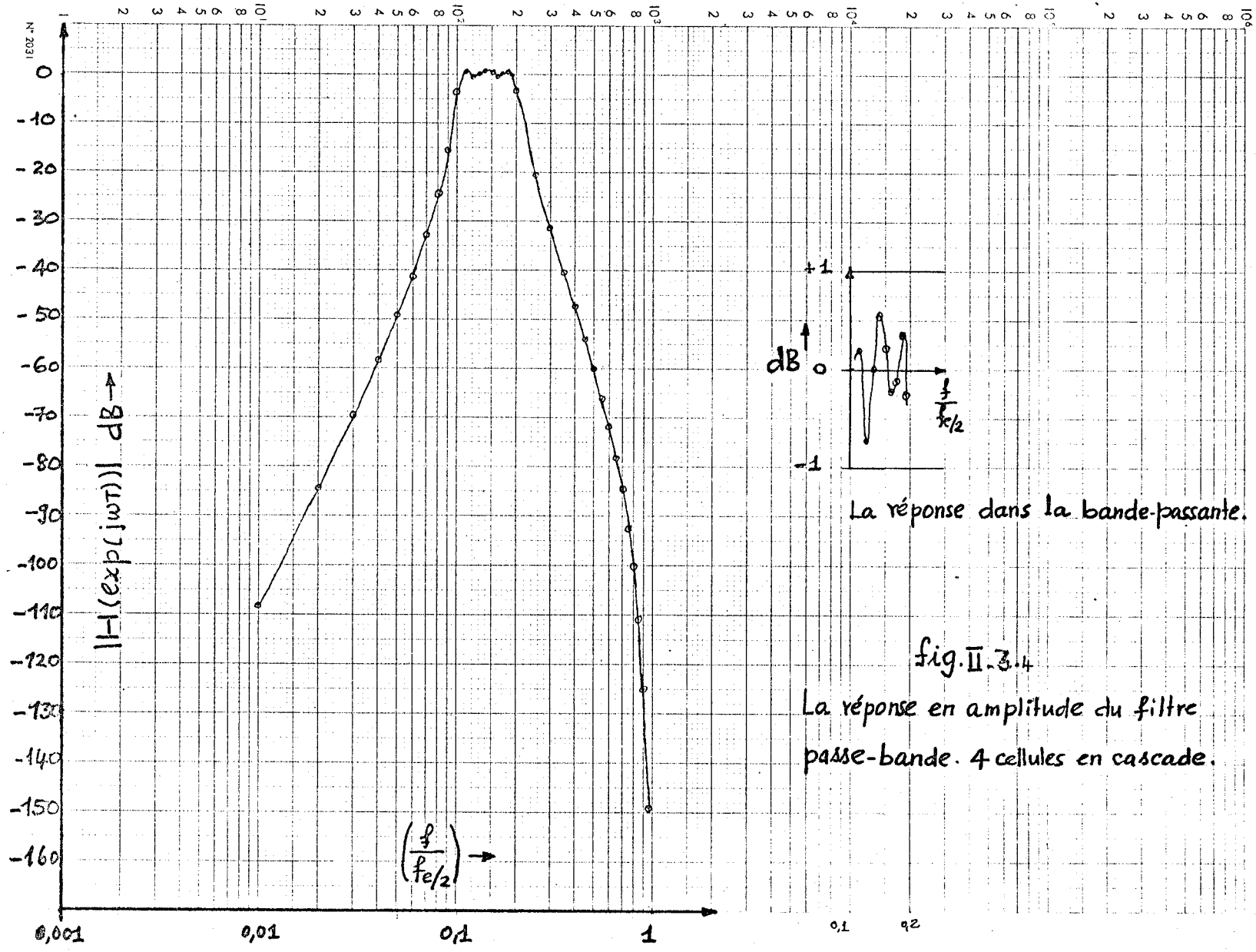
La réponse en amplitude du filtre passe bande. 2 cellules en cascade



La réponse dans la bande passante

fig II.3.3.

La réponse en amplitude du filtre passe-bande. 3 cellules en cascade.



La réponse dans la bande-passante.

fig. II-3.4

La réponse en amplitude du filtre passe-bande. 4 cellules en cascade.

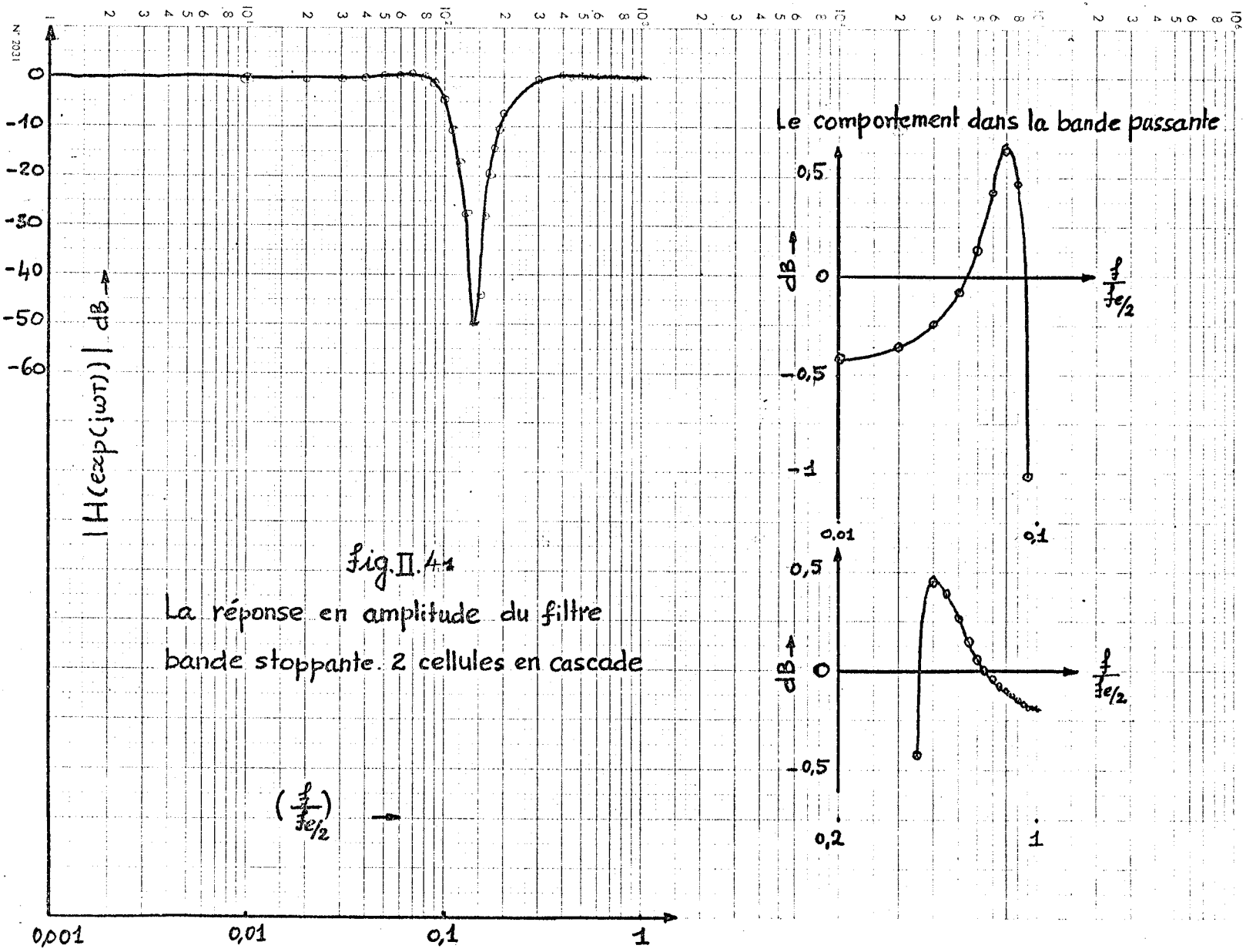


Fig. II.41

La réponse en amplitude du filtre bande stoppante. 2 cellules en cascade

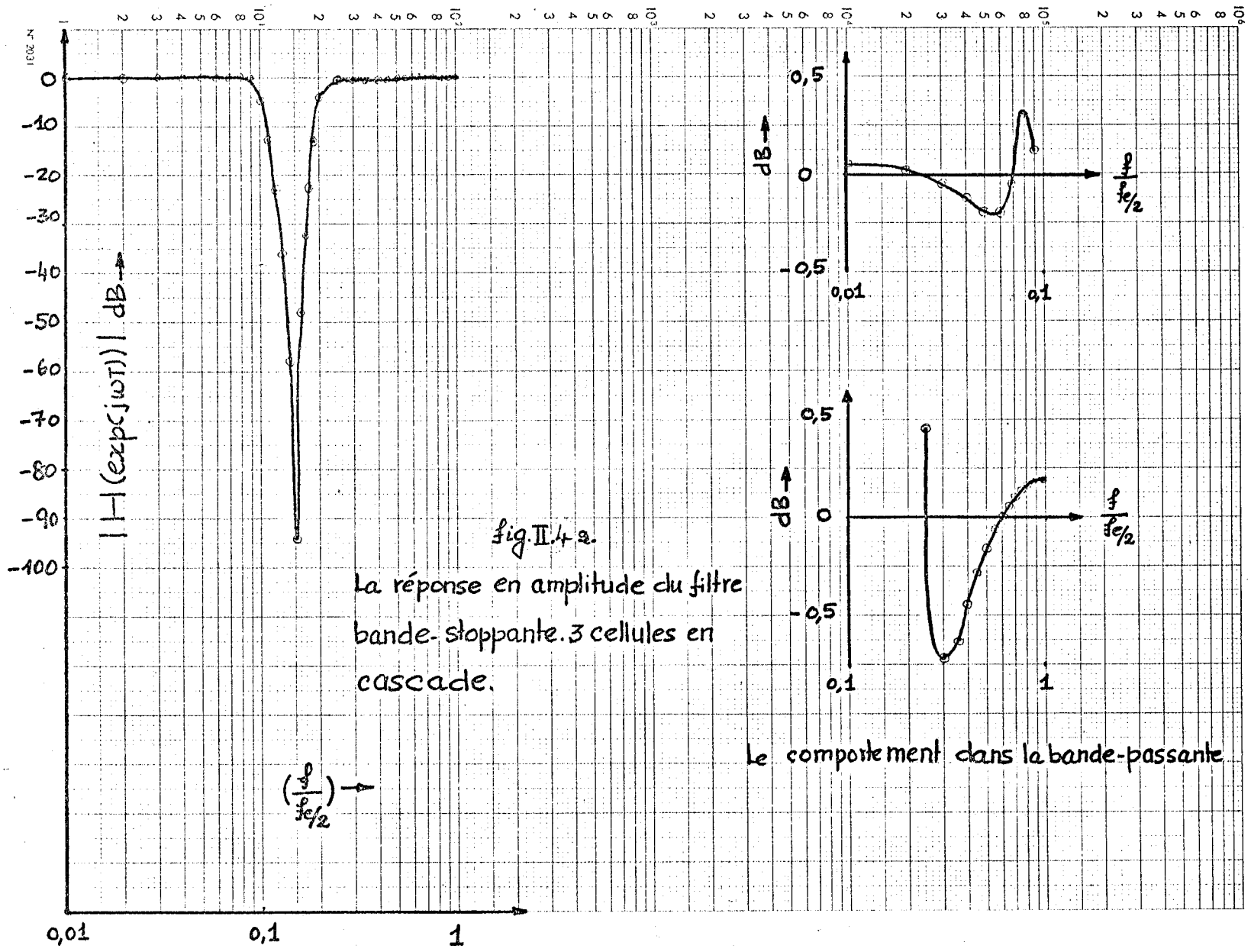


fig. II.4.2.

La réponse en amplitude du filtre bande-stoppante. 3 cellules en cascade.

Le comportement dans la bande-passante

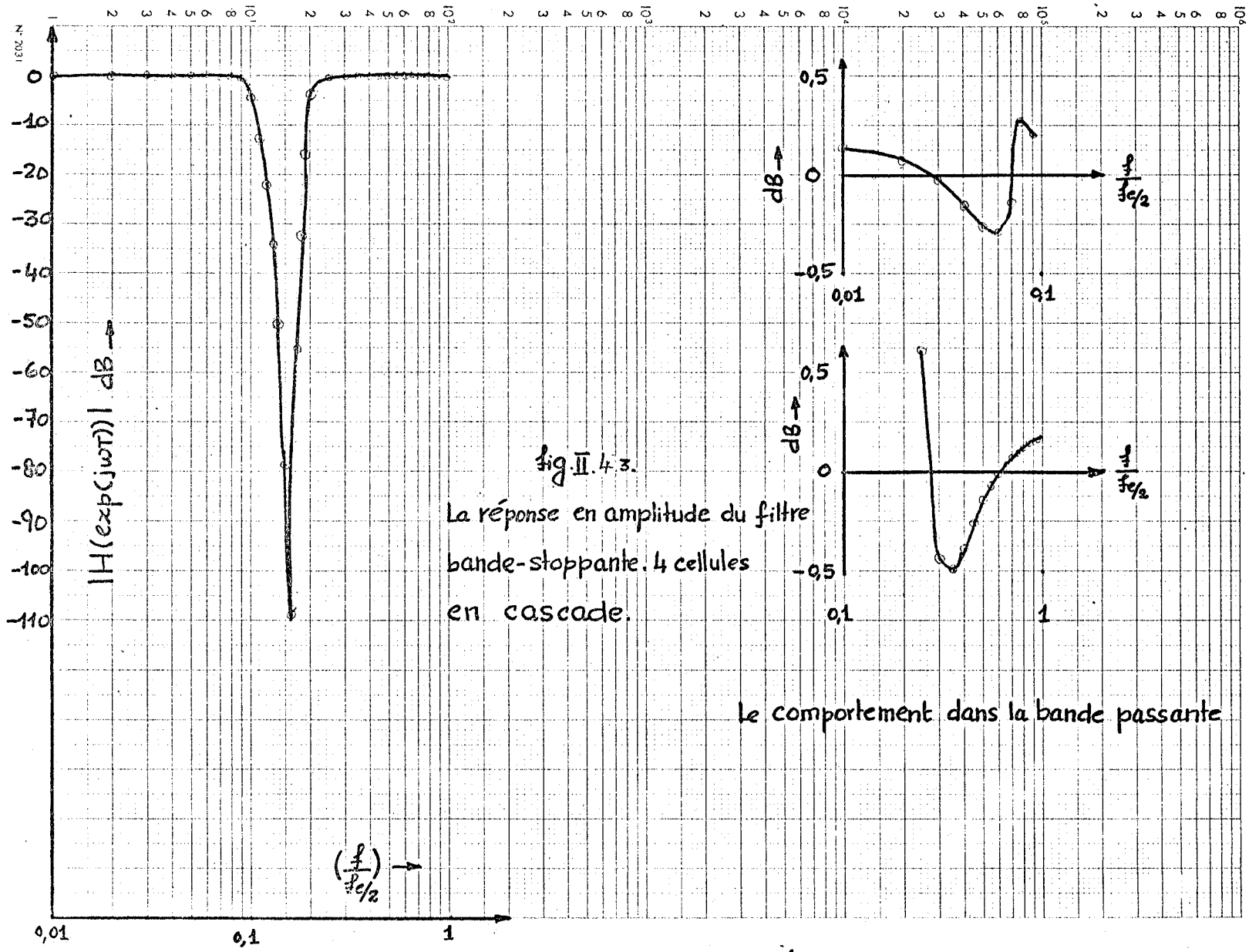


fig. II. 4.3.

La réponse en amplitude du filtre bande-stoppante. 4 cellules en cascade.

Le comportement dans la bande passante

Gain	0,1173397
$C(\hat{\theta})$	0,5673182
No. d'evaluations du gradients	137
No. des itérations	41
Nbr des Inversions des pôles et des zéros	1
Temps du calcul	0 min 46,78 sec en 360/50.

$R_i^d = 1,0$ pour $\frac{f}{f_{e/2}}$ entre 0 et 0,1 par pas de 0,01

$R_i^d = 0,5$ pour $f/(f_{e/2})$ égal à 0,1

$R_i^d = 0,0$ pour $f/(f_{e/2})$ entre 0,11 et 0,2 par pas de 0,01

$R_i^d = 0,0$ pour $f/(f_{e/2})$ entre 0,2 et 1 par pas de 0,1

Coefficients	le Gradient evalué au pt minimum	Partie Reel	Partie Imaginaire	Module	Argument
$a_1 = -1,6438232$	$0,588 \times 10^{-6}$	Zéro : 0,8219116	0,5696150	1.0	34,7233°
$a_2 = 1.0000000$	$0,728 \times 10^{-6}$				
$b_1 = -1,7935278$	$0,367 \times 10^{-5}$	Pôle : 0,8967639	0,1918108	0,917	12,07°
$b_2 = 0,8409769$	$0,630 \times 10^{-5}$				

les paramètres et quelques Renseignements utiles pour une cellule passe-bas.

TABLEAU II.1.1

Gain	c(θ̂)	Nbr. d'évaluations du gradient	Nbr. d'itérations	Nbr. des inversions des Pôles et des zéros	Temps de calcul
0,0360738	0,05140276	371	64	0	0 mn 06,175 sec

	coefficients	le gradient	Partie Réel	Partie Imaginaire	Module	Argument
1ère cellule	$a_1 = -1,744588$	$0,715 \times 10^{-8}$	0,8722944	0,4889811	0,9999999	29,2736°
	$a_2 = 0,9999999$	$-0,232 \times 10^{-7}$				
	$b_1 = -1,7565569$	$-0,271 \times 10^{-5}$	0,8782784	0,1267081	0,8873714	8,209°
	$b_2 = 0,7874279$	$-0,266 \times 10^{-5}$				
2ème cellule	$a_1 = -1,7445888$	$0,715 \times 10^{-8}$	0,8722944	0,4889810	0,9999999	29,2736°
	$a_2 = 0,9999999$	$-0,232 \times 10^{-7}$				
	$b_1 = -1,8660158$	$0,200 \times 10^{-5}$	0,9330078	0,27203963	0,97185867	16,252°
	$b_2 = 0,9445093$	$0,235 \times 10^{-5}$				

Les Paramètres d'un filtre passe-bas (2 cellules en cascade)

Tableau II.12.

Gain	$C(\hat{\theta})$	Nbr. d'evaluations du gradient	Nbr. d'Itérations	Nbr. d'Inversions des Pôles et des Zéros	Temps de calcul
0,005374483	0,001179117	1609	274	3	0min. 41,41 sec

	Coefficients	gradient	Partie Réelle	Partie Imaginaire	Module	Argument
1ère cellule	$a_1 = -1,7943826$	$0,381 \times 10^{-5}$	0,8971913	0,4416381	0,999998	26,208°
	$a_2 = 0,999996$	$0,258 \times 10^{-5}$				
	$b_1 = -1,8197763$	$0,393 \times 10^{-4}$	0,90988817	0,2398085	0,940959	14,765°
	$b_2 = 0,8854046$	$0,484 \times 10^{-4}$				
2ème cellule	$a_1 = -1,8787609$	$0,186 \times 10^{-4}$	0,93938049	0,34287761	0,9999998	20,052°
	$a_2 = 0,9999997$	$0,169 \times 10^{-4}$				
	$b_1 = -1,8916793$	$-0,427 \times 10^{-4}$	0,94583967	0,29702184	0,99138018	17,438°
	$b_2 = 0,9828346$	$-0,337 \times 10^{-4}$				
3ème cellule	$a_1 = -0,6578588$	$0,111 \times 10^{-6}$	0,32892294	0,94434342	0,999989	70,795°
	$a_2 = 0,9999790$	$0,297 \times 10^{-7}$				
	$b_1 = -1,7165032$	$0,155 \times 10^{-5}$	0,85825164	0,1006441	0,86413259	66,883°
	$b_2 = 0,7467251$	$-0,603 \times 10^{-7}$				

Les Paramètres d'un filtre passe bas (3 cellules en cascade)

Tableau II.1.3.

Gain	C($\hat{\theta}$)	Nbr. d'Evaluations du gradient	Nbr. d'Iterations	Nbr. d'Inversions des Pôles et des Zéros	Temps de calcul
0,0010833064	0,00002375666	1822	292	6	1 min. 09,32 sec.

	coeffients	gradient	Partie Reelle	Partie Imaginaire	Module	Argument
1ère cellule	$a_1 = 0,14420469$	$-0,241 \times 10^{-6}$				
	$a_2 = 0,9973670 6$	$0,392 \times 10^{-8}$	$-0,07210234$	$0,99607645$	$0,99868266$	$94,141°$
	$b_1 = -1,68263029$	$-0,137 \times 10^{-4}$				
	$b_2 = 0,71402535$	$-0,142 \times 10^{-4}$	$0,84131514$	$0,07883005$	$0,84500021$	$53,529°$
IIème cellule	$a_1 = -1,88110900$	$-0,218 \times 10^{-4}$				
	$a_2 = 1,0$	$-0,204 \times 10^{-4}$	$0,94055450$	$0,33964312$	$1,0$	$19,855°$
IIIème cellule	$b_1 = -1,84303175$	$-0,181 \times 10^{-4}$				
	$b_2 = 0,92005834$	$-0,251 \times 10^{-4}$	$0,92151587$	$0,26620824$	$0,95919671$	$16,112°$
	$a_1 = -1,71731640$	$-0,147 \times 10^{-5}$				
	$a_2 = 0,99995694$	$-0,166 \times 10^{-4}$	$0,85865820$	$0,51250662$	$0,99997847$	$30,831°$
IVème cellule	$b_1 = -1,76164106$	$-0,185 \times 10^{-4}$				
	$b_2 = 0,81665990$	$-0,588 \times 10^{-5}$	$0,88082053$	$0,20202746$	$0,90369237$	$12,918°$
	$a_1 = -1,84174038$	$-0,546 \times 10^{-5}$				
	$a_2 = 0,99999520$	$-0,873 \times 10^{-4}$	$0,92087049$	$0,38986246$	$0,99999760$	$22,946°$
	$b_1 = -1,90720932$	$-0,131 \times 10^{-5}$				
	$b_2 = 0,99999301$	$-0,838 \times 10^{-4}$	$0,95360466$	$0,30105009$	$0,99999650$	$17,520°$

Filtre Passe-bas

Tableau II.1.4.

Gain	$C(\hat{\theta})$	Nbr. d'evaluations du gradient	Nbr. d'Itérations	Nbr. d'Inversions des pôles et des zéros	Temps de Calcul
0,09109567 5	1,201203	192	65	2	1 min 22,44 sec (IBM 360/50)

coefficients	gradient	Partie Reelle	Partie Imaginaire	Module	Argument
$a_1 = -0,55866 \times 10^{-7}$	$0,1622 \times 10^{-8}$	$-0,9999999$			
$a_2 = -0,999999$	$0,1105 \times 10^{-7}$	$0,9999999$			
$b_1 = -1,6456318$	$-0,1452 \times 10^{-6}$				
$b_2 = 0,8432614$	$-0,116 \times 10^{-5}$	$0,8228159$	$0,4077198$		

$R_i^d = 0,0$ pour $0 \leq (f/f_N) < 0,1$ par pas de $0,01$

$R_i^d = 0,707$ pour $f/f_N = 0,1$

$R_i^d = 1,0$ pour $0,1 < (f/f_N) < 0,2$ par pas de $0,01$

$R_i^d = 0,707$ pour $f/f_N = 0,2$

$R_i^d = 0,0$ pour $0,2 < (f/f_N) \leq 1,0$ par pas de $0,05$

Filtre Passe-bande

Tableau II.2.1

Gain	C(ê)	Nbr. d'evaluations du gradient	Nbr. d'Itérations	Nbr. d'Inversions des pôles et des zéros	Temps de calcul
0,011416107	0,2733431	783	137	2	0min 17,32 sec

	coefficients	gradient	Partie Reelle	Partie Imaginaire	Module	Argument
1ère cellule	$a_1 = -0,56476 \times 10^{-5}$	$-0,873 \times 10^{-8}$	$-0,9999943$			
	$a_2 = -0,9999943$	$0,696 \times 10^{-8}$	$1,0$			
	$b_1 = -1,795954$	$-0,289 \times 10^{-6}$	$0,89797735$	$0,3356859$	$0,95867011$	$20,496°$
	$b_2 = 0,9190483$	$-0,4 \times 10^{-5}$				
2ème cellule	$a_1 = -0,54318 \times 10^{-5}$	$-0,890 \times 10^{-8}$	$-0,9999945$			
	$a_2 = -0,9999945$	$0,151 \times 10^{-8}$	$1,0$			
	$b_1 = -1,575544$	$0,112 \times 10^{-5}$	$0,78777238$	$0,5042865$	$0,93533556$	$32,620°$
	$b_2 = 0,8748902$	$0,646 \times 10^{-6}$				

Filtre Passe-bande.

Tableau II.2.2

Gain	C(θ)	Nbr. d'evaluations du gradient	Nbr. d'Itérations	Nbr. d'Inversions des pôles et des zéros	Temps de calcul
0,006661	0,0863028	806	148	1	0min 21,16 sec

	coefficients	gradient	Partie Reelle	Partie Imaginaire	Module	Argument
1ère cellule	$a_1 = -0,8808875$	$-0,132 \times 10^{-6}$	$-0,1191128$			
	$a_2 = -0,1191128$	$-0,178 \times 10^{-4}$	1,000000			
	$b_1 = -0,1845587 \times 10^{1}$	$0,117 \times 10^{-4}$	0,9227935	0,316998	0,97572521	0 18,958°
	$b_2 = 0,9520357$	$0,122 \times 10^{-4}$				
2ème cellule	$a_1 = -0,8808875$	$-0,132 \times 10^{-6}$	$-0,1191128$			
	$a_2 = -0,1191128$	$-0,178 \times 10^{-6}$	1,0			
	$b_1 = -1,68400511$	$-0,233 \times 10^{-4}$	0,84200255	0,3988334	0,9316847	25,345°
	$b_2 = 0,8680364$	$-0,285 \times 10^{-4}$				
3ème cellule	$a_1 = -0,8808875$	$-0,132 \times 10^{-6}$	$-0,1191128$			
	$a_2 = -0,1191128$	$-0,178 \times 10^{-6}$	1,0			
	$b_1 = -1,5789788$	$0,189 \times 10^{-4}$	0,7894893	0,5381748	0,95547146	34,281°
	$b_2 = 0,9129257$	$0,265 \times 10^{-4}$				

Filtre Passe-bande.

Tableau II.2.3.

Gain	C(θ̂)	Nombre d'evaluations du gradient	Nombre d'itérations	Nbr d'Inversions des pôles et des zéros	Temps de calcul
0,00043096998	0,064053321	807	151	1	0min 32,518cc

	coefficients	gradient	Partie Reelle	Partie Imaginaire	Module	Argument
1ère cellule	$a_1 = -0,0268678$	$0,2345 \times 10^{-3}$	$-0,9648207 \pm 0,99168860$			
	$a_2 = -0,9568017$	$0,241 \times 10^{-3}$				
	$b_1 = -1,8572975$	0,869				
	$b_2 = 0,96185811$	$-0,199$	$0,92864875$	$0,31538802$	$0,9807436$	$18,758°$
2ème cellule	$a_1 = -0,0268673$	$0,234 \times 10^{-3}$	$-0,9648204$			
	$a_2 = -0,9568006$	$0,241 \times 10^{-3}$	$0,9916877$			
	$b_1 = -1,6996958$	$0,141 \times 10^{1}$				
	$b_2 = 0,87325719$	$0,191 \times 10^{1}$	$0,8498479$	$0,38860731$	$0,93448229$	$24,573°$
3ème cellule	$a_1 = -0,0268652$	$0,234 \times 10^{-3}$	$-0,9648208$			
	$a_2 = -0,9567993$	$0,241 \times 10^{-3}$	$0,99168605$			
	$b_1 = -1,5721407$	$-0,947$				
	$b_2 = 0,90138653$	$-0,116 \times 10^{1}$	$0,78607036$	$0,53242831$	$0,94941378$	$34,110°$
4ème cellule	$a_1 = -0,0269167$	$0,233 \times 10^{-3}$	$-0,96480152$			
	$a_2 = -0,9568112$	$0,240 \times 10^{-3}$	$0,99171825$			
	$b_1 = -1,1681787$	$-0,591 \times 10^{-2}$				
	$b_2 = 0,475739$	$-0,654 \times 10^{-2}$	$0,58408937$	$0,3668498$	$0,68973849$	$32,131°$

Filtre Passe-bande

Tableau II.2.4.

Gain	$C(\hat{\theta})$	Nbr. d'évaluations du gradient	Nbr. d'Itérations	Nbr. d'Inversions des pôles et des zéros	Temps de Calcul
0,59970334	0,62530040	514	119	1	0min 14,37 sec

	coefficients	gradient	Partie Reelle	Partie Imaginaire	Module	Argument
1ère cellule	$a_1 = -1,45634721$	$-0,113 \times 10^{-5}$	0,728171108	0,68539536	0,99999998	43,266°
	$a_2 = 0,99999996$	$-0,601 \times 10^{-6}$				
	$b_1 = -1,77646449$	$0,167 \times 10^{-6}$	0,88823224	0,24610552	0,921696510	15,486°
	$b_2 = 0,84952445$	$-0,761 \times 10^{-6}$				
2ème cellule	$a_1 = -1,81957648$	$0,455 \times 10^{-6}$	0,909763243	0,41512733	0,99999993	24,527°
	$a_2 = 0,99999986$	$-0,328 \times 10^{-6}$				
	$b_1 = -0,71268802$	$0,670 \times 10^{-6}$	0,35633440	0,66569869	0,75506844	61,840°
	$b_2 = 0,57012859$	$0,834 \times 10^{-6}$				

Filtre Stop-bande

Tableau II.3.1

$R_i^d = 1,0$ pour $0 \leq (f/f_N) < 0,1$ par pas de 0,01

$R_i^d = 0,7$ pour $f/f_N = 0,1$

$R_i^d = 0,0$ pour $0,2 < (f/f_N) < 0,3$ par pas de 0,01

$R_i^d = 0,7$ pour $(f/f_N) = 0,3$

$R_i^d = 1,0$ pour $0,3 < (f/f_N) \leq 1$ par pas de 0,05

Gain	$c(\hat{\theta})$	Nbr. d'évaluations du gradient	Nbr. d'itérations	Nbr. d'inversions des pôles et des zéros	Temps de calcul
0,144391	0,14864359	193	55	0	0 min 3,83 sec

	Coefficients	gradient	Partie Reelle	Partie Imaginaire	Module	Argument
1ère cellule	$a_1 = -1,541863$	$0,268 \times 10^{-6}$				
	$a_2 = 1,0$	$0,304 \times 10^{-6}$	0,7709319	0,636917	1,0	39,562°
	$b_1 = -1,0965512$	$-0,282 \times 10^{-7}$				
	$b_2 = 0,999999$	$0,129 \times 10^{-6}$	0,54827564	0,83629765	0,9999999	56,751°
2ème cellule	$a_1 = -1,5418638$	$0,268 \times 10^{-6}$				
	$a_2 = 1,0$	$0,304 \times 10^{-6}$	0,7709319	0,636917	1,0	39,562°
	$b_1 = -0,4644911$	$0,226 \times 10^{-6}$				
	$b_2 = 0,3224912$	$-0,290 \times 10^{-7}$	0,23224558	0,51822124	0,5678831	65,86°
3ème cellule	$a_1 = -1,5418638$	$0,268 \times 10^{-6}$				
	$a_2 = 1,0$	$0,304 \times 10^{-6}$	0,7709319	0,636917	1,0	39,562°
	$b_1 = -1,7755760$	$0,179 \times 10^{-5}$				
	$b_2 = 0,83577303$	$0,196 \times 10^{-6}$	0,8877880	0,2181866	0,914206	13,807°

Filtre Stop-bande

Tableau II.3.2.

Gain	C(θ̂)	Nbr. d'evaluations du gradient	Nbr. d'Itérations	Nbr. d'Inversions des Pôles et des Zéros	Temps de calcul
0,33792759	0,32306254	141	54	0	0min. 7,39 sec

	Coefficient	gradient	Partie Reelle	Partie Imaginaire	Module	Argument
1ère cellule	$a_1 = -1,61605249$	$-0,720 \times 10^{-6}$				
	$a_2 = 0,9999998$	$-0,548 \times 10^{-6}$	0,80802624	0,58914636	0,99999992	36,096°
	$b_1 = -0,95101814$	$0,106 \times 10^{-6}$				
	$b_2 = 0,71580676$	$0,283 \times 10^{-6}$	0,47550907	0,69978417	0,84605364	55,803°
2ème cellule	$a_1 = -1,61605249$	$-0,720 \times 10^{-6}$				
	$a_2 = 0,99999980$	$-0,548 \times 10^{-6}$	0,80802624	0,58914636	0,99999992	36,096°
	$b_1 = -1,84104703$	$0,291 \times 10^{-5}$				
	$b_2 = 0,91838587$	$0,229 \times 10^{-5}$	0,92052351	0,26650014	0,95832451	16,146°
3ème cellule	$a_1 = -1,61605249$	$-0,720 \times 10^{-6}$				
	$a_2 = 0,9999998$	$-0,548 \times 10^{-6}$	0,80802624	0,58914636	0,99999992	36,096°
	$b_1 = -0,98381586$	$0,548 \times 10^{-6}$	0,17266094			
	$b_2 = 0,14005477$	$0,280 \times 10^{-6}$	0,81115492			
4ème cellule	$a_1 = -1,61605249$	$-0,720 \times 10^{-6}$				
	$a_2 = 0,9999998$	$-0,548 \times 10^{-6}$	0,80802624	0,58914636	0,99999992	36,096°
	$b_1 = -0,95101814$	$0,106 \times 10^{-6}$				
	$b_2 = 0,71580676$	$0,283 \times 10^{-6}$	0,47550907	0,69978417	0,84605364	55,803°

Tableau II.3.3

II.3 Résultats obtenus

Nous avons déterminé à l'aide de cet algorithme les paramètres des trois types de filtres suivants : passe-bas, passe-bande et stop-bande et de plus nous avons étudié les deux aspects suivants :

- souplesse d'utilisation de cet algorithme, c'est-à-dire l'encombrement en mémoire, le temps de calcul, etc ...
- performances de cet algorithme, c'est-à-dire les caractéristiques fréquentiels des filtres conçus à l'aide de celui-ci.

Nous avons imposé pour ces différents filtres des spécifications sévères (la bande de transition très courte étant d'un octave, voir les tableaux II.1 à II.3). Nous avons déterminé les paramètres de filtres comprenant de 1 à 4 cellules dans chaque catégorie qui répondent au mieux à ces spécifications. Les réponses en amplitude obtenues sont données fig. II.2 à II.4. Les paramètres correspondants à ces filtres ainsi que les coûts $\hat{c}(\hat{\theta})$ sont fournis par les tableaux II.1 à II.3.

L'examen de ces résultats nous conduit à formuler les conclusions suivantes :

1. Une cellule de 2ème ordre n'a pas en général un nombre suffisant de paramètres pour satisfaire aux spécifications imposées. Pour une cellule passe-bas nous observons que la pente dans la bande de transition est très raide, de l'ordre de -40dB par octave. Par contre, la réponse dans la bande de réjection est à peu près constante à -20dB. Ceci est dû à l'insuffisance, dans le plan z, des zéros dans le deuxième quadrant à l'intérieur du cercle unité. La seule paire

de zéros complexes conjugués du filtre est très proche de pôles complexes conjugués pour pouvoir satisfaire au mieux les spécifications de la bande de transition. Dans le cas du filtre passe-bande, 1 cellule, la bande passante est beaucoup trop étroite par rapport à la bande d'un octave fixée par les spécifications. La pente dans les bandes de transitions est de l'ordre de 12 dB par octave. Cette pente augmente très rapidement aux alentours des fréquences 0 et f_N à cause de la présence des zéros réels dans le plan z au voisinage de ± 1.

2. Le filtre passe-bande apparaît plus facile à concevoir que le filtre passe-bas : temps de calcul plus faible, meilleure précision. En effet, la présence de zéros dans les deux quadrants (θ élevé) rend l'approximation dans les bandes de réjections plus satisfaisantes pour les filtres de type passe-bande. En revanche, dans le cas des filtres passe-bas, la pente dans la bande de transition est beaucoup plus satisfaisante : elle est de l'ordre de -40 dB à $0{,}11\ f_N$ pour un filtre à 3 cellules. La fréquence de coupure étant à $0{,}1\ f_N$, ceci donne une pente de -40 dB par $\frac{I}{10}$ de décade.

3. En conclusion, pour les applications courantes, les pentes des bandes de transition (au moins 36 dB/octave pour les filtres passe-bande et supérieur à 36 dB/octave pour les filtres passe-bas) semblent être satisfaisantes avec des filtres à 3 cellules.

4. Les filtres à bande-stoppante que nous avons examinés approchent mal le gabarit désiré. Les temps de calcul sont nettement inférieurs aux deux autres types de filtres. Cette mauvaise approximation est dûe à l'insuffisance des paramètres, compte tenu de la sévérité des spécifications. Bien que les pentes des bandes de transitions augmentent avec le nombre de cellules (-4 dB à la fréquence de coupure et -80 dB à I/3 d'octave pour un filtre de 4 cellules), la largeur de la bande stoppante est trop étroite pour pouvoir mieux approcher le gabarit imposé avec 4 cellules.

5. Les erreurs dans les bandes passantes ainsi que dans les bandes stoppantes ne sont pas des ondulations à amplitude constante (equi-ripple).

C H A P I T R E III

CONCEPTION DES FILTRES NUMERIQUES NON RECURSIFS

III.1 Généralités sur le problème de l'approximation :

Les filtres non récursifs, nous l'avons déjà signalé au chapître I, ont une réponse impulsionnelle finie. De ce fait, ils sont bien adaptés au filtrage des blocs de données. Les approximations réalisables conduisent pour la F.T. à un polynome trigonométrique. Ceci est à opposer aux filtres récursifs dotés d'une F.T. qui est le rapport de deux polynômes trigonométriques (voir chapître I). De ce fait, les premiers ont besoin de beaucoup plus de paramètres pour répondre aux mêmes spécifications que les seconds. La réalisation d'un filtre non récursif doit tenir compte des remarques suivantes :

(i) le filtre idéal a une réponse impulsionnelle de durée infinie.
(ii) On ne peut l'approcher que par un filtre dont la durée de la réponse impulsionnelle est finie.
(iii) la troncature de la réponse impulsionnelle (en abrégé R.I.) donne lieu au phénomène bien connu §14§ des oscillations dans le domaine fréquentiel dit "phénomène de Gibbs".
(iv) l'amplitude maximum de ces oscillations ne diminue que lentement lorsque la durée de la R.I. augmente.

Pour réduire ce phénomène d'oscillation, on peut multiplier la R.I. tronquée du filtre idéal par une fonction du temps ("fenêtre") qui décroît lentement vers zéro (voir les fig. III.1 et III.2 , III.3).

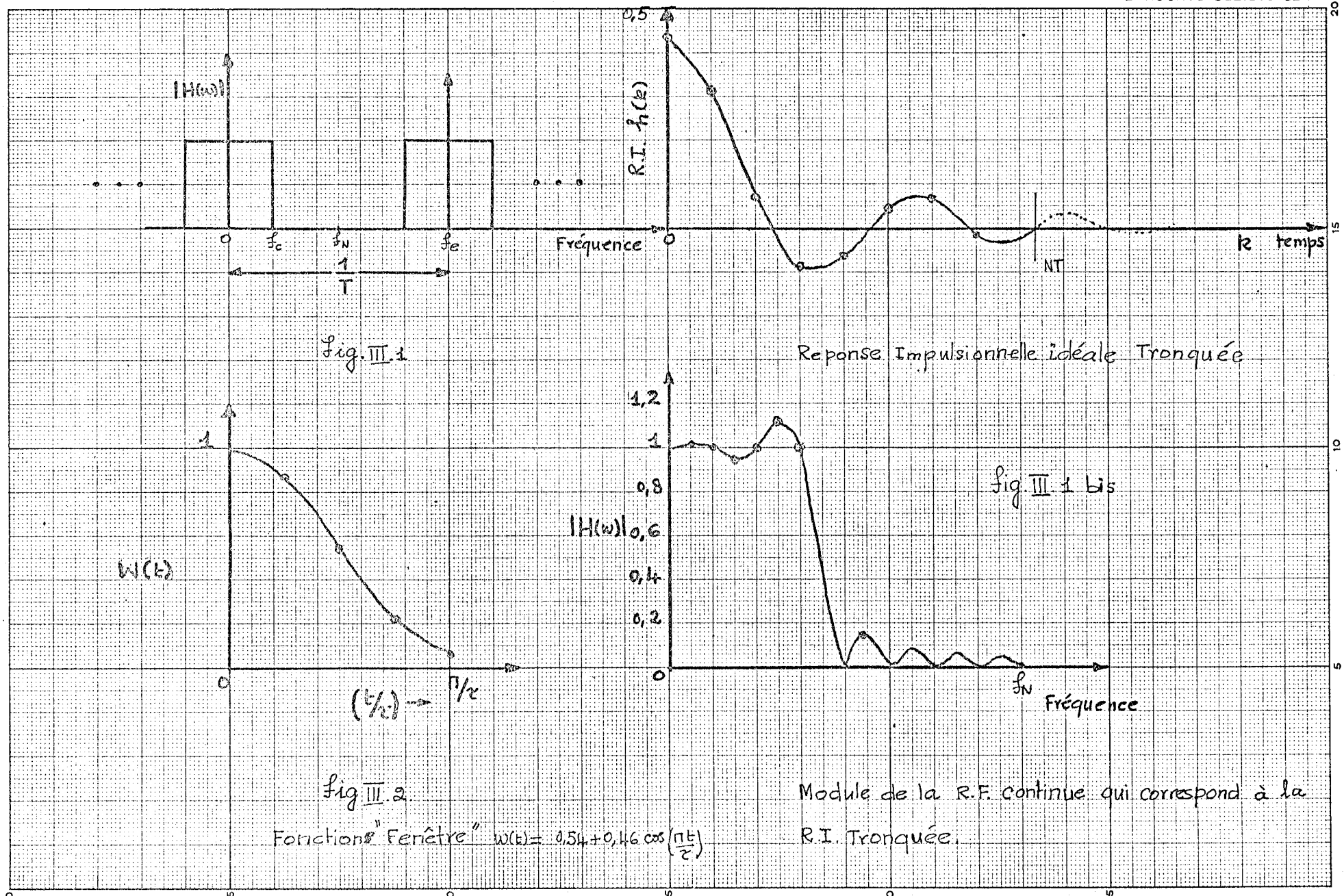


fig. III.1 Reponse Impulsionnelle idéale Tronquée

fig III.1 bis

fig III.2.

Fonction "Fenêtre" $w(k) = 0{,}54 + 0{,}46 \cos\left(\frac{\pi k}{2}\right)$

Module de la R.F. continue qui correspond à la R.I. Tronquée.

fig. III. 3.

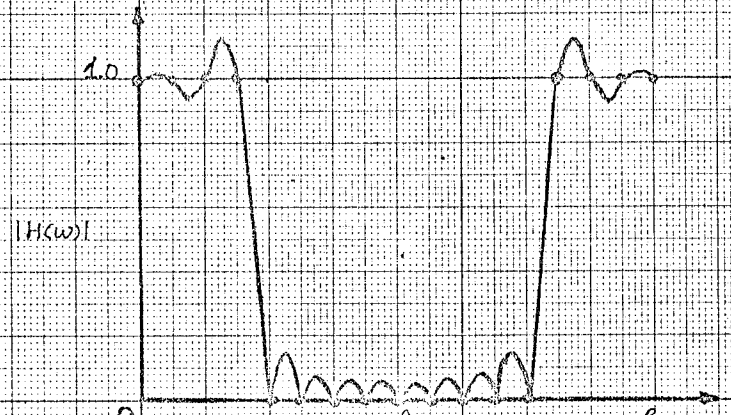


fig III. 4

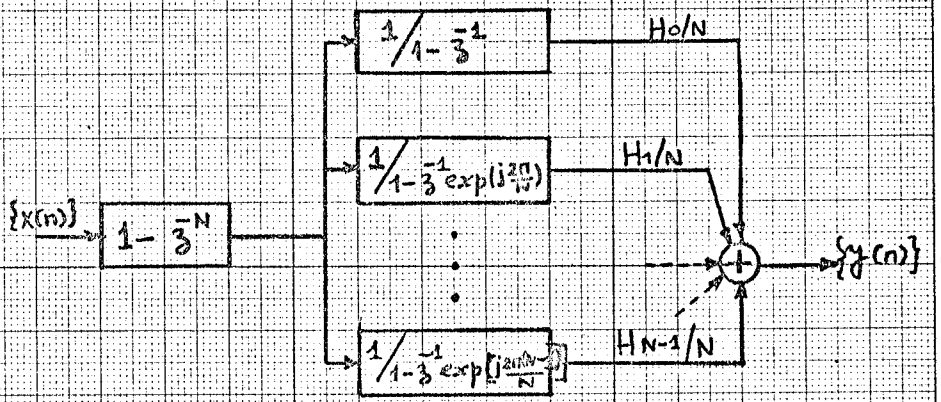


fig. III. 5

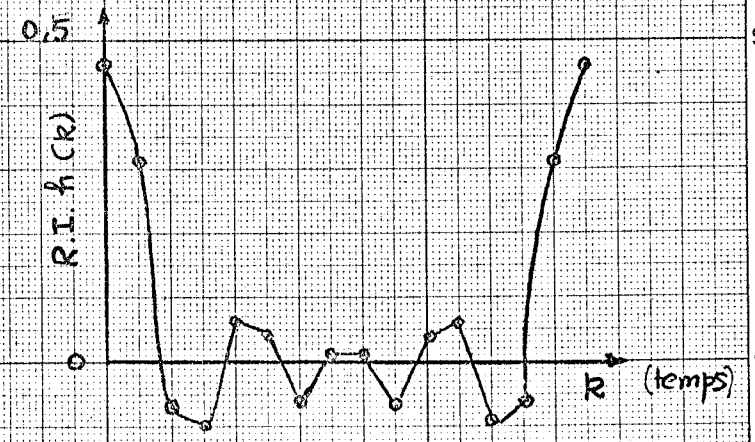


fig. III. 6

Soit f(nT) une telle fonction fenêtre (discrète) ayant pour transformée en z,F(z). Soit h(nT) la séquence de la R.I. tronquée du filtre idéal, ayant pour transformée en z,H(z). La transformée en z du produit x(nT) · f(nT) (n entier) est donnée par le théorème de la convolution complexe §3§ soit :

$$H'(z) = \frac{1}{2\pi j}\int_C H(z/\lambda)\ F(\lambda)\ \frac{d\lambda}{\lambda} \qquad \text{(III.1)}$$

Ainsi, un choix judicieux de la fonction fenêtre f(nT) peut atténuer les lobes secondaires dans les bandes passantes et stoppantes (voir fig. III.3). Kaiser §6§ a proposé une gamme de fenêtres qui donne une certaine flexibilité à la conception de tels filtres. En ajustant un paramètre, on peut diminuer les amplitudes des lobes secondaires au prix d'une augmentation de la bande de transition. Helms §15§ a aussi proposé les fenêtres de Dolph-Chebycheff.
La synthèse des filtres numériques non récursifs procède ainsi :

(i) Evaluation de la R.I. du filtre définie par N points, équi-espacés en temps (intervalle T) soit h(n).

(ii) Choix d'une fonction fenêtre définie également par N points, équi-espacés, (intervalle T) soit f(n).

(iii) Evaluation de la R.I. modifiée h'(n) = h(n) f(n)
n = 0,1 ... N-1

(iv) Réalisation du produit de convolution de la séquence d'entrée /x(n)/ par la séquence de la R.I. h'(n) :

$$y(m) = \sum_{j=0}^{N-1} h'(j)x(m-j), \quad \text{pour } m = 0,1 \ldots N-1 \qquad \text{(III.2)}$$

Ce produit de convolution peut être réalisé, soit par la convolution directe (voir le schéma de la fig. I.2) ; soit par l'algorithme de la F.F.T. §16§. La dernière méthode consiste à trouver la réponse fréquentielle du filtre H'(n), n = 0,1 ... N-1 (en prenant la transformée de Fourier discrète (D.F.T.) de la séquence h'(n)), et de la multiplier par la transformée de Fourier discrète de la séquence d'entrée /x(n)/, soit X(n). La transformée inverse du produit donne la séquence de sortie filtrée /y(n)/.
Mathématiquement, on a :

$$H'(n) = \frac{1}{N} \sum_{k=0}^{N-1} h'(k) W^{-kn} \qquad n=0,1 \ldots N-1 \qquad \text{(III.3)}$$

$$X(n) = \frac{1}{N} \sum_{k=0}^{N-1} x(k) W^{-kn}, \quad n=0,1 \ldots N-1 \qquad \text{(III.4)}$$

$$Y(n) = H'(n)X(n), \quad n=0,1 \ldots N-1$$

$$y(n) = \sum_{k=0}^{N-1} Y(k) W^{kn} \qquad n=0,1 \ldots N-1 \qquad \text{(III.5)}$$

où $W = \exp(j\frac{2\pi}{N})$

La conception des filtres numériques non récursifs par cette méthode est appelée "méthode de la fenêtre" (Windowing Technique).

Dans ce chapître, nous allons étudier une autre méthode de conception de ces filtres. Cette méthode est appelée "l'échantillonnage fréquentiel" §17§.

Soit H(k), k = 0,1 ... N-1, la séquence des valeurs de la R.F. du filtre à réaliser (spécifications). On a :

$$H(k) = \left| H(k) \right| \exp\,(j\theta(k)), \quad k = 0,1 \ \ldots\ N-1 \tag{III.6}$$

On peut trouver la R.I. finie, qui correspond à ces spécifications, en utilisant la transformée de Fourier inverse discrète (I.D.F.T.). Soit :

$$h(n) = \frac{1}{N} \sum_{k=0}^{N-1} H(k) W^{kn} \quad n=0,1 \ \ldots\ N-1 \tag{III.7}$$

La transformée en z de la séquence /h(n)/ est définie comme :

$$H(z) = \sum_{n=0}^{N-1} h(n)\ z^{-n} \tag{III.8}$$

Pour les valeurs de z sur le cercle unité, on a :

$$H(z) = \sum_{n=0}^{N-1} h(n) \exp\left(j \frac{2\pi kn}{N}\right) \tag{III.9}$$

$$= H(z) \Big|_{z=\exp(j2\pi k/N)} = H(k)$$

Ainsi, on voit que la réponse fréquentielle continue a les valeurs désirées (ou spécifiées) aux instants d'échantillonnages de la R.F. spécifiée. Mais, entre les points d'échantillonnages, la R.F. continue peut être différente de celle que l'on veut approcher (voir fig.III.4). On se propose d'étudier une façon de minimiser cet écart. Pour cela, on va d'abord développer une relation explicite qui relie les H(k) et la R.F. continue.
En substituant l'expression de h(n) dans (III.8), on a :

$$H(z)=\sum_{n=0}^{N-1}\left(\frac{1}{N}\sum_{k=0}^{N-1}H(k)W^{kn}\right)z^{-n} \qquad \text{(III.10)}$$

En permutant les sommations, on a :

$$H(z)=\frac{1}{N}\sum_{k=0}^{N-1}\sum_{n=0}^{N-1}H(k)\,W^{kn}\,z^{-n}$$

On peut effectuer la deuxième sommation. Ceci est une série géométrique de raison $W^k z^{-1}$. Nous avons donc :

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

$$=\frac{1-z^{-N}}{N}\sum_{k=0}^{N-1}\frac{H(k)}{(1-z^{-1}W^k)} \qquad \text{(III.11)}$$

Cette dernière équation est très importante. En effet, non seulement elle nous permet de calculer la R.F. continue en fonction des H(k), mais aussi, elle nous indique une autre façon de réaliser les filtres non récursifs. Le deuxième terme représente N cellules élémentaires des filtres récursifs, ayant toutes des pôles sur le cercle unité, à $z = \exp\left(j\frac{2\pi k}{N}\right)$, $k = 0,1 \ldots N-1$. Le premier terme est un filtre "peigne" qui a N zéros situés aux mêmes endroits que les pôles. Le schéma de réalisation d'un tel filtre est indiqué dans la fig.(III.5).

Nous allons nous intéresser à l'optimisation de la réponse du filtre réalisé. A cette fin, nous allons évaluer H(z) pour

$z = \exp(j\omega T)$. Nous pouvons mettre l'équation III.11 sous la forme suivante pour $z = \exp(j\omega T)$:

$$H(\exp(j\omega T)) = \frac{1-\exp(-jN\omega T)}{N} \sum_{k=0}^{N-1} \frac{H(k)}{1-\exp(-j\omega T)\exp(j\frac{2\pi k}{N})}$$

$$= \frac{1-\exp(-jN\omega T)}{N} \sum_{k=0}^{N-1} \frac{H(k)\exp(-j\frac{\pi k}{N})\exp(j\frac{\omega T}{2})}{\exp(j\frac{\omega T}{2})\exp(-j\frac{\pi k}{N}) - \exp(-j\frac{\omega T}{2})\exp(j\frac{\pi k}{N})}$$

$$= \frac{1-\exp(-jN\omega T}{N} \sum_{k=0}^{N-1} \frac{H(k)\exp(-j\frac{\pi k}{N})\exp(j\frac{\omega T}{2})}{2j\ \sin(\frac{\omega T}{2} - \frac{\pi k}{N})}$$

$$= (\exp(-\frac{jN\omega T}{2})(\exp(\frac{jN\omega T}{2}) - \exp(-\frac{jN\omega T}{2}))) / N$$

$$\sum_{k=0}^{N-1} \frac{H(k)\exp(-j\frac{\pi k}{N})\exp(j\frac{\omega T}{2})}{2j\ \sin(\frac{\omega T}{2} - \frac{\pi k}{N})}$$

$$= \frac{\exp(-j\frac{N\omega T}{2})\exp(j\frac{\omega T}{2})}{N} \sum_{k=0}^{N-1} \frac{H(k)\sin(\frac{N\omega T}{2})\exp(-j\frac{\pi k}{N})}{\sin((\omega T/2 - \pi k/N)}$$

$$= \frac{\exp(-j\frac{N\omega T}{2}(1-\frac{1}{N}))}{N} \sum_{k=0}^{N-1} \frac{H(k)\sin(\frac{N\omega T}{2})\exp(-j\frac{\pi k}{N})}{\sin(\omega T/2 - \pi k/N)} \qquad \text{(III.12)}$$

Cette équation III.12 nous permet de déduire que :

(i) La R.F. continue est une fonction linéaire des échantillons fréquentiels (E.F.)H(k).

(ii) Le déphasage est linéaire. Celui-ci est dû au premier terme.

(iii) Si, au départ, N est pair et H(k) est une séquence réelle et symétrique H(k) = H(N-k), la réponse fréquentielle interpolée n'est pas réelle (à un déphasage linéaire près). Une petite composante imaginaire s'introduit à cause du terme $\exp(-j\pi k/N)$ au numérateur.

(iv) Quand $H(\exp(j\omega T))$ est réel (à un déphasage linéaire près) on s'aperçoit qu'il est égal à une somme de termes élémentaires de la forme :

$$\sin\left(\frac{\omega NT}{2}\right) / \sin\left(\frac{\omega T}{2} - \theta\right)$$

Donc, par exemple, dans la conception des filtres passe-bas, on peut définir les H(k) dans la bande-passante, ayant tous une amplitude égale à 1.0 et tous égaux à 0.0 dans la bande de rejection. En définissant une bande de transition avec un, deux ou trois H(k) on peut intuitivement espérer annuler les ondulations dûes aux termes de la bande-passante par les ondulations dues aux termes de la bande de transition.

La petite composante imaginaire peut être négligée pour N petit. Pour N grand, on peut rendre $H(\exp(j\omega T)$ purement réel (sauf pour le déphasage linéaire) en faisant la substitution suivante :

$$H(k) = G(k) \exp\left(j \frac{\pi k}{N}\right) \qquad \text{(III.13)}$$

D'après (III.12), il est évident que la sommation devient purement réelle. On définit $G(k) = -G(N-k)$ et $G(\frac{N}{2}) = 0$.

On a, pour la réponse impulsionnelle h(n) :

$$h(n) = \frac{1}{N}\sum_{k=0}^{N-1} G(k)\exp(j\frac{\pi k}{N})\exp(j\frac{2\pi kn}{N}),\ n=0,1 \ldots N-1 \qquad \text{(III.14)}$$

qui peut s'écrire :

$$h(n) = \frac{1}{N}\sum_{k=1}^{N/2-1} G(k)\exp(j\frac{\pi k}{N})\exp(j\frac{2\pi kn}{N}) + \frac{1}{N}\sum_{k=N/2+1}^{N-1} G(k)\exp(j\frac{\pi k}{N})\exp(j\frac{2\pi kn}{N})$$

$$+ \frac{G(0)}{N} + G(\frac{N}{2})$$

En posant $k' = N-k$ dans le deuxième terme, on a :

$$h(n) = \frac{1}{N}\sum_{k=1}^{N/2-1} G(k)\exp(j\frac{\pi k}{N})$$

$$+ \sum_{k'=1}^{N/2-1} G(k')\exp(j\frac{\pi(N-k')}{N})\exp(j\frac{2\pi n(N-k')}{N}) + \frac{G(0)}{N}$$

$$= \frac{G(0)}{N} + \frac{1}{N}\sum_{k=1}^{N/2-1} 2G(k)\cos(\frac{\pi k}{N} + 2\frac{\pi kn}{N}) \qquad \text{(III.15)}$$

Donc, la réponse impulsionnelle est purement réelle et sa propriété de symétrie est différente de celle qui correspond aux H(k). En effet,

$$h(n) = h(N-1-n) \quad \text{pour} \quad n = 0,1 \ldots \frac{N}{2} - 1$$

Une telle séquence est montrée sur la fig. (III.6).

III.2 Spécification du filtre à réaliser et optimisation :

Les spécifications d'un filtre passe-bas sont données fig. III.7. Les paramètres NP, NBP et NBT représent respectivement le nombre d'échantillons fréquentiels, le nombre de points dans la bande-passante et le nombre de points de transition. Ce dernier nombre est arbitraire. On va ajuster ces paramètres de la bande de transition pour réduire le maximum des lobes secondaires. Le programme d'optimisation a deux autres paramètres d'entrée ; le type du filtre (soit pass-bas, soit bande-passante), et la fréquence d'échantillonnage. Sur la fig.(III.7) nous avons fixé (2NBP-1) échantillon à 1.0 et (2 NBT) échantillons à optimiser, les autres étant fixés à zéro.

A titre d'exemple, prenons les deux points de transition T_1 et T_2, le programme d'optimisation pourra être résumé ainsi :

Considérations préliminaires :

(1) Pour NP donné, il faut déterminer le nombre de points à évaluer sur la R.F. interpolée. Nous avons utilisé un rapport de 16:1.

(2) Pour l'interpolation, on utilise l'algorithme de F.F.T. On trouve d'abord la réponse impulsionnelle qui correspond aux spécifications. Ensuite, on ajoute 15NP zéros symétriquement, comme indiqué sur les fig.(III.8) et (III.9). La D.F.T. de cette nouvelle séquence de R.I. donne la R.F. interpolée. Les G(k) étant réels et symétriques, les h(n) le seront aussi, ainsi que la R.F. interpolée (aux erreurs de calcul près).

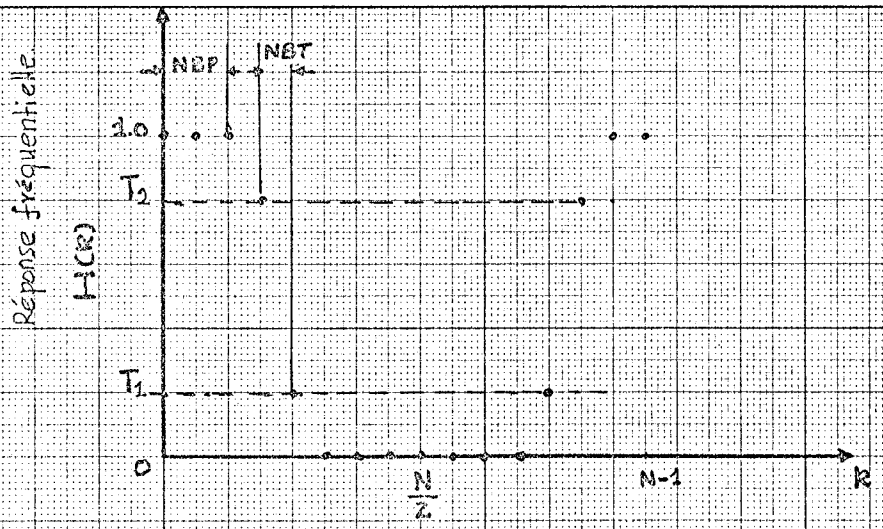


fig. III. 7

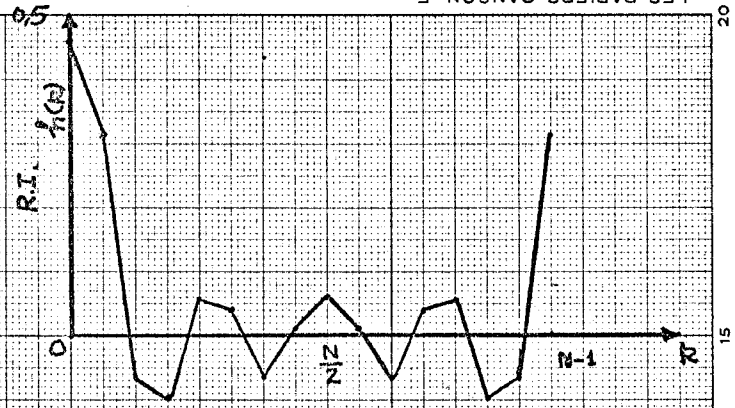


fig. III. 8

Les spécifications.

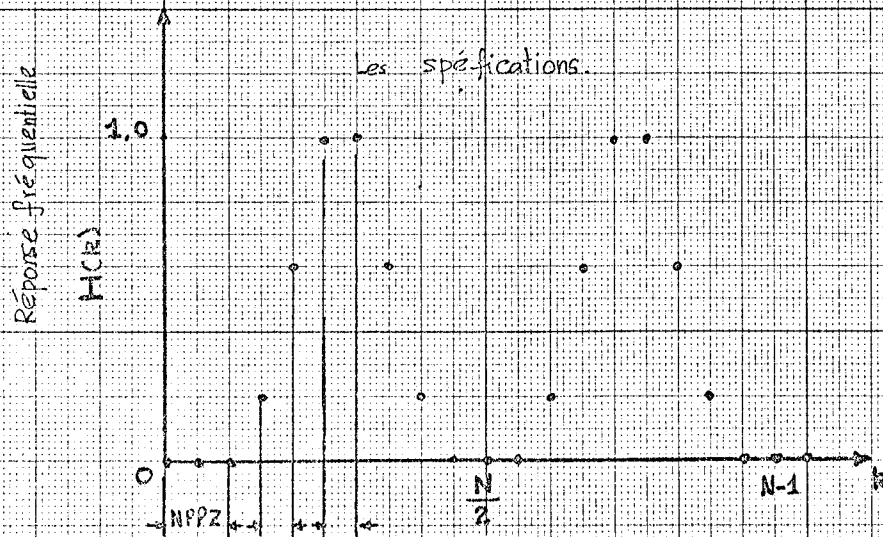


fig III. 7bis

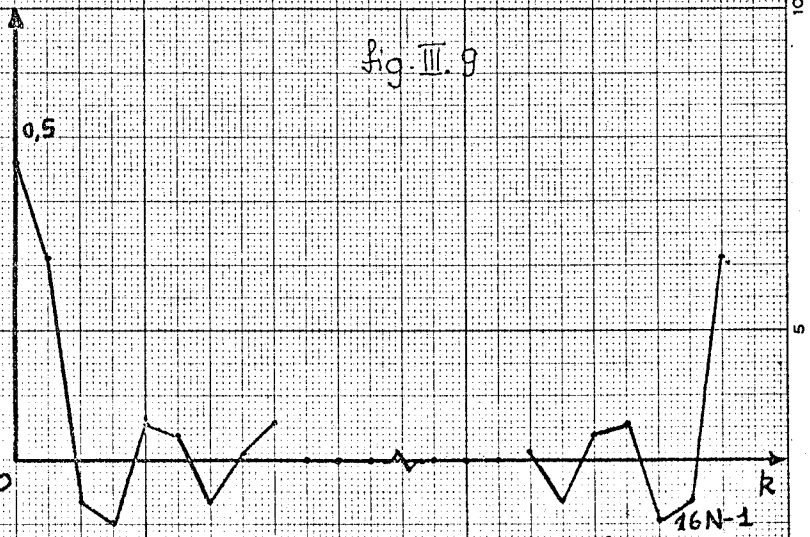


fig. III. 9

Introduction de zéros pour obtenir la réponse fréquentielle interpolée.

Stratégie de minimisation :

(1) On commence par une recherche unidimensionnelle (dans le cas de deux échantillons de transition T_1 et T_2, on fixe $T_2 = 1$ pour commencer). On fait varier T_1 dans l'intervalle (0, 1), et on trouve la valeur de T_1 qui minimise le maximum de lobes secondaires. Soit le point A sur la fig. (III.10).

(2) Pour définir un autre point sur la droite (à deux dimensions), on perturbe T_2 de sa valeur présente à une valeur légèrement inférieure. On répète la recherche unidimensionnelle en faisant varier T_1 comme dans l'étape (1). Soit B le point ainsi déterminé. Les deux points A et B déterminent la droite de recherche dans l'espace à deux dimensions, T_1, T_2.

(3) La recherche de minimum le long de cette droite donne le point C.

(4) Pour obtenir une nouvelle direction de recherche, on fixe T_2 à cette dernière valeur (au point C), et on fait varier T_1 entre (0, 1) pour minimiser le maximum de lobes secondaires (M.L.S.) de $|H(\exp(j\omega T)|$. Soit D, un tel point, comme en (2). On perturbe T_2 de sa valeur en C et la recherche de T_1 qui minimise le M.L.S. nous conduit au point E. Les deux points D et E déterminent la nouvelle direction de recherche.

(5) La nouvelle recherche le long de la droite DE nous conduit au point F. Si la différence entre les deux M.L.S. en C et en F est inférieure à une valeur prédéterminée, la recherche est terminée et le point F donne la solution (dans l'espace à deux dimensions). Sinon, on réitère l'étape (4) jusqu'à l'obtention d'un minimax.

fig III.10

Pratiquement, nous avons constaté que l'algorithme converge toujours en trois itérations pour deux points de transitions. Une difficulté qui a été ressentie pendant la mise au point du programme a été la détermination du pas de recherche entre (0, 1). Si le pas n'est pas assez fin, la direction de recherche peut s'éloigner du point optimum. On a utilisé, pour commencer, un pas de 1/50 qui a été ensuite modifié du cours de l'optimisation selon les critères suivants :

(1) Diminution du pas pour les directions successives de recherches lorsqu'on s'approche du point optimum de manière à augmenter la précision de la détermination de la direction de ce point.

(2) Test de chaque direction de cherche. Notons ici le cas particulier où la vraie pente est voisine de 90°. La pente évaluée sera de 90° à cause de l'imprécision des calculs. Il est donc nécessaire de tester cette condition et de reprendre l'optimisation de la manière suivante :
   (i) on augmente la perturbation de T_2
   (ii) on diminue le pas de recherche

Il faut remarquer à ce propos, que l'on ne peut obtenir l'optimum qu'à un epsilon près, à cause de la quantification de l'espace de recherche. Néanmoins, le résultat sous-optimum obtenu est très intéressant et peut suffire dans beaucoup d'applications.

La recherche s'arrête dès que la différence entre deux M.L.S. successifs ne dépasse pas 0, 1 dB. Cependant il a été nécessaire de prendre pour l'arrêt, les deux autres critères suivants :

(1) On arrête si le nombre d'itérations dépasse trois.

(2) On arrête si le M.L.S. commence à croître (par rapport à sa valeur précédente), au lieu de décroître.

Pour cette méthode de recherche directe, le temps de calcul est excessif. Un filtre passe-bande de 32 points avec deux points de transition a coûté 4 minutes et 32,5 secondes (sur IBM 360/91). A cause de cela nous n'avons pas pu étudier une gamme étendue de filtres, ni augmenter les points de transition. En effet, le temps de calcul semble croître d'une façon exponentielle avec le nombre de points de transition. Ainsi nous nous sommes contentés de vérifier l'algorithme et de donner des résultats pour 1 et 2 points de transition, pour des filtres passe-bas et passe-bande spécifiés sur 16 et 32 points.

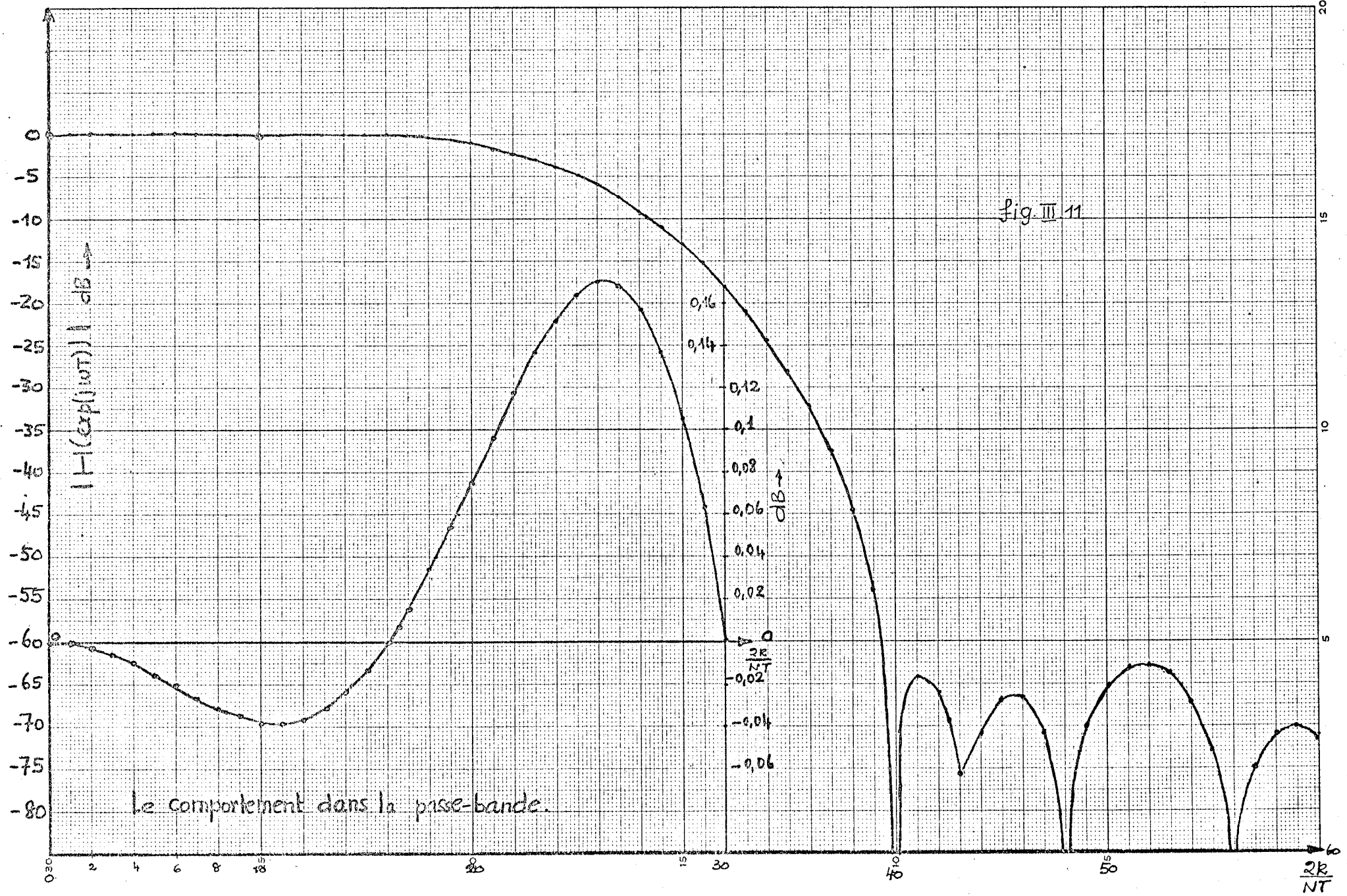
III.3 Résultats obtenus :

Nous avons conçu deux types de filtres, soit passe-bas et passe-bande. Les spécifications pour ces filtres sont données dans le tableau (III.1). Pour les filtres passe-bande, nous avons introduit un paramètre supplémentaire soit NPPNZ, qui indique les nombres de points (E.F.) précédant le premier point non nul. Les filtres conçus sont définis sur 16 et 32 points avec un ou deux points de transition. Les résultats obtenus sont donnés dans le tableau III.1. Les courbes de R.F. sont tracées sur les fig. (III.11 à III.13). On remarque bien que la minimisation ait été faite sur les valeurs absolues des lobes secondaires, que les ondulations dans la bande passante n'excèdent guère 0,5 dB. On remarque également que les ondulations dans les deux bandes ne sont pas de type "à amplitude constante" (équi-ripple).

Nous n'avons pas pu à cause des temps de calcul excessifs, étudier les variations des valeurs des points de transition en fonction de la largeur de la bande passante et du nombre de points N.P.

Type de filtre	NP	NPPZ	NBP	NBT	T_1	T_2	Min max dB
Passe-bas	16	0	3	2	0,1253375	0,6239880	– 62,92
Passe-bas	32	0	8	2	0,1271487	0,6193245	– 60,72
Passe-bande	32	4	4	2	0,6470053	0,4765139	– 73,87
Passe-bas	16	0	4	1	0,3799999	–	– 35,67
Passe-bas	32	0	9	1	0,3599999	–	– 40,61
Passe-bande	32	4	6	1	0,2999999	–	– 50,21

Tableau III.1



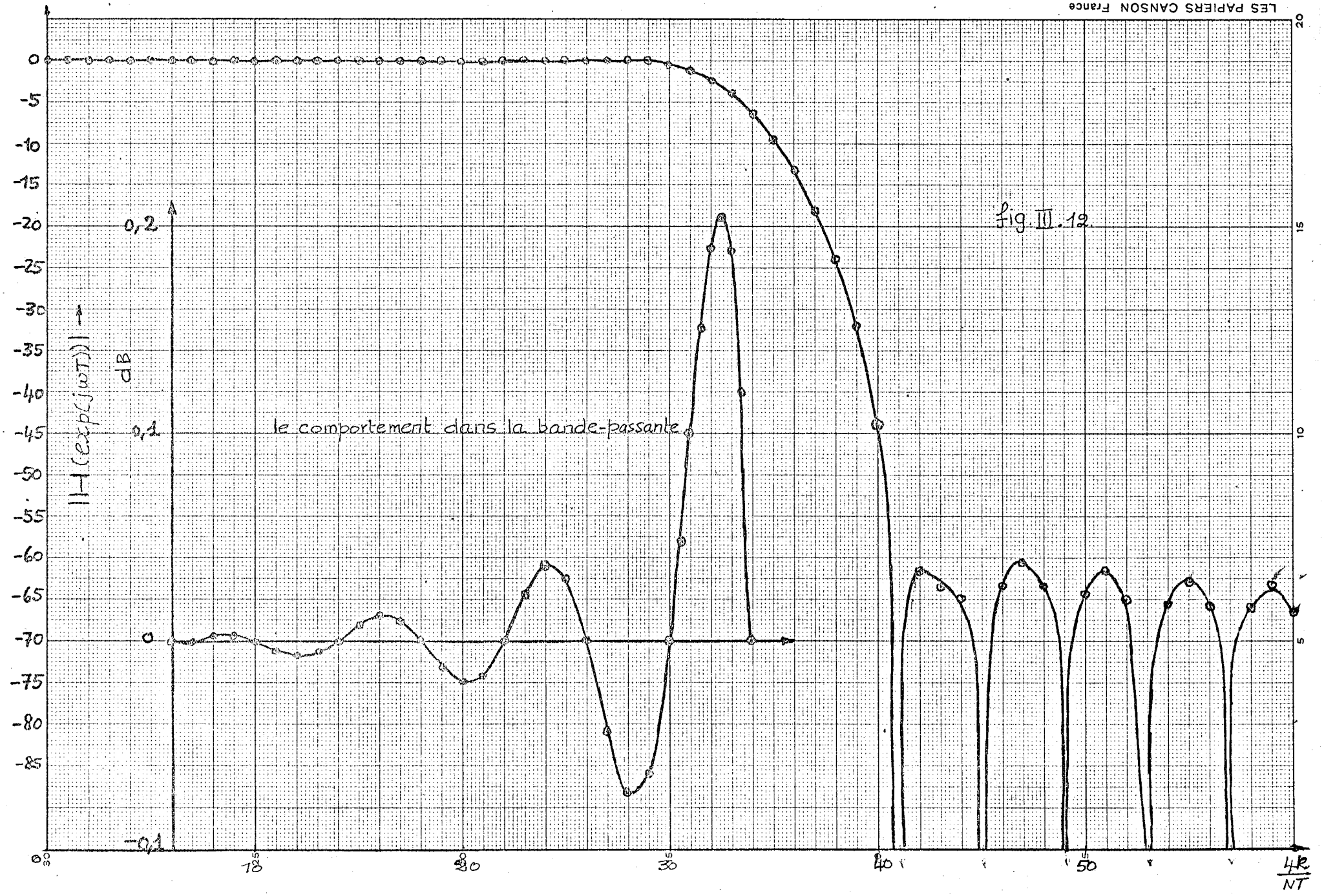
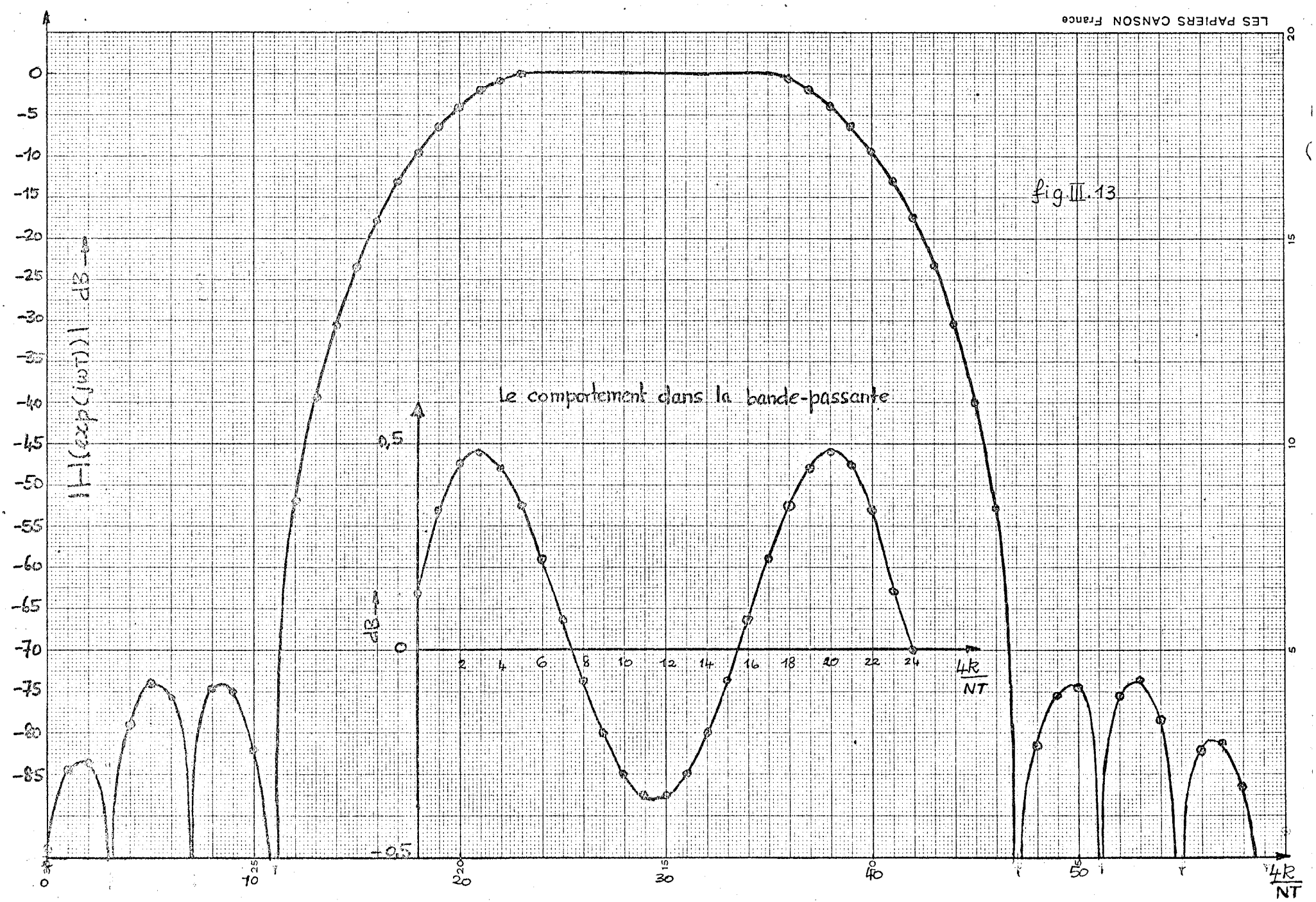

LES PAPIERS CANSON France
0
-5
-10
-15
-20
-25
-30
-35
-40
-45
-50
-55
-60
-65
-70
-75
-80
-85
0,2
0,1
0
-0,1
dB
|H(exp(jωT))| →
le comportement dans la bande-passante
20
15
10
5
0
30
18
20
30
40
50
4k/NT

Fig. III.12.


LES PAPIERS CANSON France
fig.III.13
Le comportement dans la bande-passante
|H(exp(jωT))| dB
dB
0
-5
-10
-15
-20
-25
-30
-35
-40
-45
-50
-55
-60
-65
-70
-75
-80
-85
0,5
-0,5
2
4
6
8
10
12
14
16
18
20
22
24
4R/NT
0
10
20
30
40
50
4R/NT

III.4 CONCLUSION

Une méthode d'approximation des filtres non récursifs a été développée. Les résultats obtenus montrent que cette méthode donne une bonne approximation du gabarit recherché. Mais malheureusement, le temps de calcul est excessif. L'extension de cette méthode à l'approximation de filtres différentiateurs ne semble poser aucun problème. Une fois que les points de transition sont évalués, le filtre peut être réalisé suivant une des trois méthodes indiquées dans la section (III.1). L'analyse d'erreur, pour la réalisation dite de "convolution directe", sera développée dans le chapître (V). La méthode de "l'échantillonnage fréquentiel" est une combinaison de filtres récursifs élémentaires en parallèle et d'un filtre non récursif en cascade avec celle-ci. Les méthodes développées dans le chapître (V) s'appliquent donc également pour ce cas. Si on le réalise en utilisant l'algorithme de F.F.T., l'analyse d'erreur porte essentiellement sur ce dernier.

C H A P I T R E I V

SYNTHESE DES TRAVAUX ACTUELS

SUR LES ERREURS DE QUANTIFICATION

Dans ce chapître, nous allons mettre en évidence les difficultés que pose la réalisation pratique des filtres numériques. Après avoir déterminé les coefficients du filtre qui satisfont (selon un critère d'optimalité préalablement choisi) aux spécifications sur le gabarit fréquentiel, il reste à choisir le type de réalisation. Nous allons montrer tout d'abord que la réalisation d'un filtre numérique sous sa forme non décomposée n'est pas souhaitable. Après avoir éliminé cette forme, il reste quatre configurations courantes. En fait, compte tenu des configurations transposées, il en existe beaucoup d'autres §10§, §32§.

Nous nous proposons d'étudier en détail les propriétés des quatre formes courantes de réalisation d'un filtre numérique récursif :

(i) cascade directe

(ii) cascade canonique

(iii) parallèle directe

(iv) parallèle canonique, dont les schémas bloc dont donnés fig. (I.7), (I.8), (I.9).

IV.1 Etude de la forme non décomposée.

Soit la fonction de transfert d'un filtre à réaliser :

$$H(z^{-1}) = \frac{\sum_{n=0}^{N} a(n) z^{-n}}{1 + \sum_{n=1}^{N} b(n) z^{-n}} \qquad \text{(IV.1)}$$

Cette fonction de transfert donne lieu à l'équation de récurrence suivante :

$$y(n) = \sum_{k=0}^{N} a(k)\, x((h-k)T) - \sum_{k=1}^{N} b(k)\, y((n-k)T) \qquad \text{(IV.1 bis)}$$

Cette équation peut être réalisée, directement, de deux façons différentes comme indiqué sur les deux schémas bloc. Le schéma (I.4) correspond à la réalisation dite "directe", et le schéma (I.6) à celle dite "canonique". Dans la première réalisation, les coefficients a(i)(ou les zéros) précèdent les coefficients b(i) (ou les pôles), et inversement dans la deuxième.

Kaiser §6§ a étudié ces formes de réalisation du point de vue de la précision nécessaire pour représenter les coefficients, afin d'assurer la stabilité. Il a déduit une borne supérieur pour le nombre des digits nécessaires à la représentation des coefficients qui assurent la stabilité absolue. Il est utile de rappeler, ici, quelques résultats de ses travaux pour avoir une idée des problèmes rencontrés dans la réalisation des filtres numériques d'ordre élevé

dans la forme non factorisée, surtout lorsque la fréquence d'échantillonnage devient très élevée.

Le dénominateur de IV.1 peut être mis en facteur comme suivant :

$$D(z^{-1}) = \prod_{n=1}^{N} (1 - z^{-1}/z(n)) \qquad \text{(IV.2)}$$

où les z(n) sont les pôles du système qu'on a supposé être simples. Si l'on suppose que l'on approche un filtre linéaire analogique ayant la fonction de transfert H(p) par la technique de la transformée en z, le pôle z(n) correspond au pôle exp(p(n)T) dans le plan de Laplace. Donc, IV.2 se met sous la forme :

$$D(z^{-1}) = \prod_{n=1}^{N} (1 - z^{-1}\exp(p(n)T) \qquad \text{(IV.3)}$$

D'autre part, la contrainte de Nyquist impose que la partie imaginaire de $(p(i)T) < \Pi$. Donc, en normalisant par rapport à la moitié de la fréquence d'échantillonnage, on a :

$$\mu(n) = \frac{p(n)T}{\Pi} = \frac{p(n)}{(\omega e/2)} \qquad \text{(IV.4)}$$

et $|\mu(n)| < 1$ pour tout n

Quand $T \to 0$, c'est à dire quand $\omega e \to \infty$, les $\mu(n) \to 0$
Donc, on a pour T voisin de 0 :

$$(1 - \exp(p(n)T)z^{-1}) \longrightarrow (1 - (1+\mu(n)\Pi)z^{-1}) \qquad \text{(IV.5)}$$

Donc, dans le plan z^{-1}, les pôles sont

$$z(n) \simeq \frac{1}{1+\mu(n)\pi} \simeq 1-\mu(n)\pi \qquad \text{(IV.6)}$$

où $|\mu(n)| \ll 1$

L'équation IV.6 montre que lorsque ωe s'approche de l'infini, les pôles du système tendent à se concentrer autour du point $z^{-1} = 1$ dans le plan z^{-1}. Donc, quand ωe est très elevée, les imprécisions dans la représentation des b(n) peuvent faire bouger les pôles à l'intérieur du cercle unité et ainsi causer l'instabilité du système. On peut estimer les variations des b(n) pour qu'un pôle se situe à $z^{-1}=1$; soit la limite de la stabilité. D'après IV.3 et IV.6, pour $z^{-1}=1$, on a :

$$D(z^{-1})\Big|_{z^{-1}=1} = \prod_{n=1}^{N} p(n)T \qquad \text{(IV.7)}$$

D'après IV.1, on a :

$$D(z^{-1})\Big|_{z^{-1}=1} = 1 + \sum_{n=1}^{N} b(n) = D_0 \qquad \text{(IV.8)}$$

Si l'un des b(n) subit un changement de valeur égal à D_0, $D(z^{-1})$ aura effectivement un zéro à $z^{-1}= 1$

Par exemple, prenons le cas d'un système de premier ordre :

$$D(z^{-1}) = 1 - bz^{-1} = 0$$

$$D(z^{-1})\Big|_{z^{-1}=1} = 1 - b = D_0$$

La distance entre le point courant sur le cercle unité à $z^{-1} = 1$, et le pôle à 1/b est égale à (1/b-1).

(Voir fig.IV.1)

Si b change de valeur, par exemple $b' = b+\Delta b$ tel que $\Delta b = D_0$, soit $\Delta b = 1-b$, la même distance devient égale à

$$1/b'-1 = \frac{1}{b+1-b} - 1 = 0$$

Donc, le système a effectivement un pôle à $z^{-1} = 1$. Tout changement supplémentaire de b tel que la valeur $D(z^{-1})\Big|_{z^{-1}=1}$ change de signe entraînerait l'instabilité.

Donc, l'équation IV.7 représente, pour T voisin de zéro, grosso modo, une borne minimum pour la précision des b(n). Cette équation nous permet de déduire les deux conclusions importantes :

(i) Pour T voisin de zéro, la précision nécessaire pour représenter les coefficients dépend de ω_e et de l'ordre N du polynôme du dénominateur. Pour T constant, voisin de zéro, doubler l'ordre du dénominateur entraîne le doublement des nombres de digits nécessaires (approximativement, compte tenu du fait que les pôles tendent à se concentrer au voisinage du point $z^{-1}=1$) pour représenter les b(n).

(ii) Pour N, ordre du polynôme $D(z^{-1})$, constant, doubler la fréquence d'échantillonnage nécessite $n \log_{10} 2$ digits supplémentaires pour représenter les b(n).

C'est pour ces raisons qu'il est nécessaire de décomposer l'expression $H(z^{-1})$ du filtre d'ordre N en cellules élémentaires. La forme cascade correspond à la décomposition suivante :

$$H(z^{-1}) = A \prod_{i=1}^{K} \frac{1+ \alpha(1i)z^{-1}+ \alpha(2i)z^{-2}}{1+ \beta(1i)z^{-1}+ \beta(2i)z^{-2}} \qquad \text{(IV.9)}$$

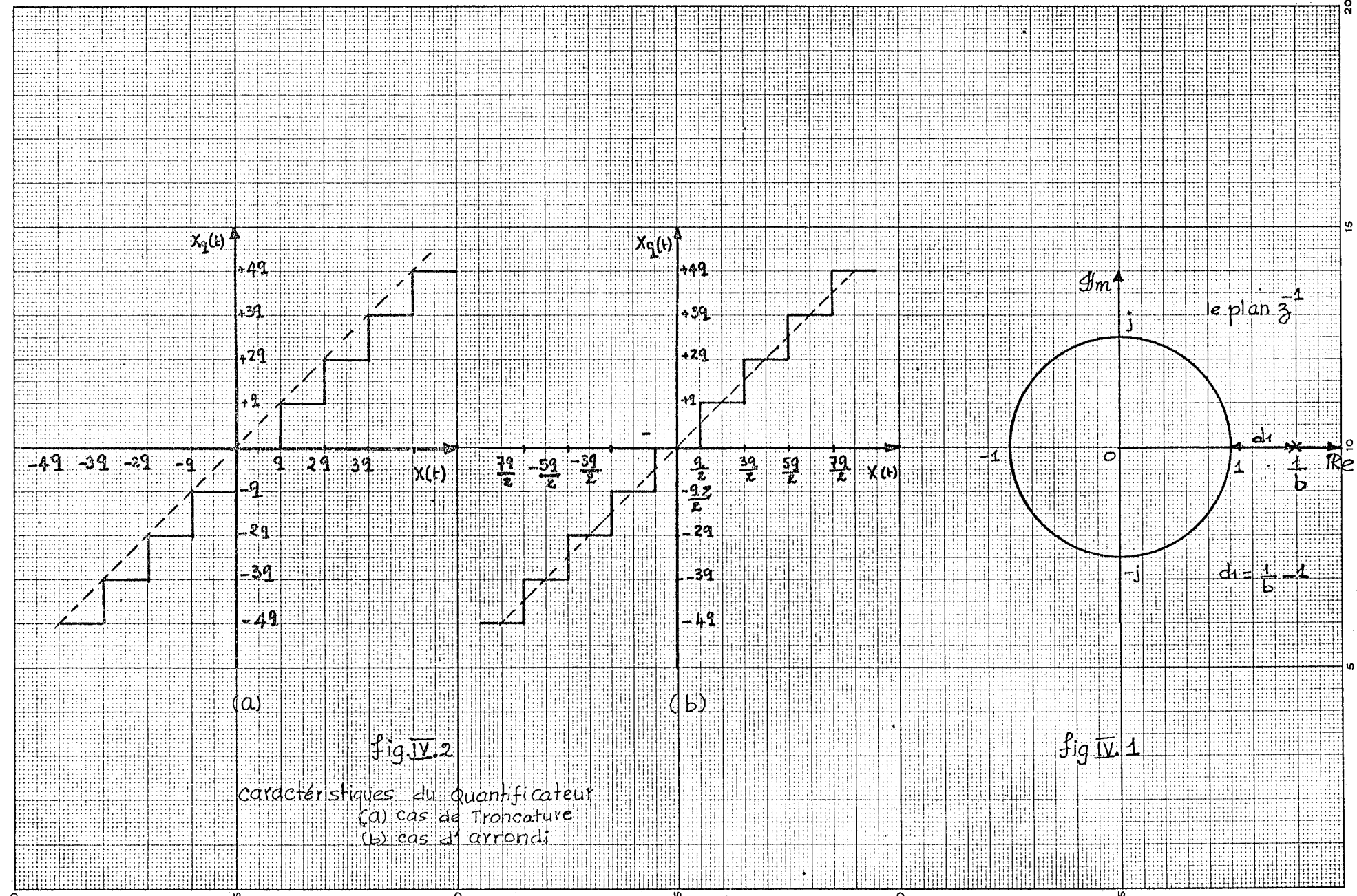


fig IV.2

caractéristiques du Quantificateur
(a) cas de Troncature
(b) cas d'arrondi

fig IV.1

La forme parallèle correspond à la décomposition suivante :

$$H(z^{-1}) = \lambda_{0+} \sum_{i=1}^{K} \frac{\lambda_{(0i)} + \lambda_{(1i)} z^{-1}}{1 + \beta_{(1i)} z^{-1} + \beta_{(2i)} z^{-2}} \qquad \text{(IV.10)}$$

Chacune de ces cellules peut être réalisée en forme directe ou en forme canonique.

IV.2 Erreur dûe à la quantification du signal d'entrée :

Tout traitement numérique d'un signal continu est précédé par la quantification de celui-ci. Cette opération est non linéaire. Pour faciliter l'analyse du processus de quantification, Bennett §18§ a proposé un modèle statistique et a étudié la validité de ce modèle. Deux types de quantification sont employés couramment et sont schématisés fig.IV.2. Mathématiquement, on peut définir ces deux opérations par les deux équations suivantes :

soit $x(t)$ le signal d'entrée continu, $x_q(t)$ le signal d'entrée quantifié et q le pas de quantification.
Le premier type de quantification assigne des valeurs discrètes telles que :

$$x_q(t) = nq \quad \text{tel que } nq \leq x(t) < (n+1)q \qquad \text{(IV.11)}$$

où n est un entier positif, et le deuxième tel que :

$$x_q(t) = nq \quad \text{tel que } (n-\tfrac{1}{2})q \leq x(t) < (n+\tfrac{1}{2})q \qquad \text{(IV.12)}$$

Puisque la valeur exacte du signal d'entrée est perdue à cause de cette transformation non linéaire, il est important de connaître la nature et l'ampleur de l'erreur introduite. Bennett §18§ pour le deuxième type de quantification, a étudié cette erreur, (équation IV.12) en l'assimilant à un bruit blanc additif uniformément réparti entre $-q/2$ et $+q/2$. Avec cette hypothèse, il est facile d'évaluer la

variance de l'erreur définie comme $\epsilon_q = x(t) - x_q(t)$

Elle est égale à

$$\sigma^2_{\epsilon_q} = \frac{1}{q} \int_{-q/2}^{q/2} x^2 \, dx = q^2/12 \qquad \text{(IV.13)}$$

La valeur moyenne est $m_q = \frac{1}{q} \int_{-q/2}^{q/2} x\,dx = 0.$ (IV.13 (bis))

Cette hypothèse reste valable pour une grande classe de signaux x(t), allant des signaux à bande étroite aux signaux à large bande, dans les conditions suivantes :

(i) Le pas de quantification est assez fin ; autrement dit le nombre des niveaux de quantification est assez important, compte tenu de la dynamique du signal (supérieur à 16).

(ii) L'amplitude du signal est telle qu'il traverse plusieurs niveaux entre les instants d'échantillonnage. Widrow §19§ qui a aussi étudié ce problème, a démontré, d'autre part, que, bien que la transformation réalisée par l'opération de quantification est non linéaire, son effet sur la fonction de densité de probabilité (et la fonction caractéristique) du signal d'entrée est linéaire.

IV.3 Erreur dûe à la quantification des coefficients.

Il a été mentionné, en début de ce chapître, l'importance que peut prendre cette erreur qui peut conduire à l'instabilité du filtre réalisé.

Le problème peut être résumé ainsi :
on a évalué les valeurs des coefficients a(n) et b(n) du filtre à réaliser soit par une méthode directe d'optimisation, soit par l'une des transformations déjà mentionnées. Ces valeurs, étant des nombres réels appartenant à un intervalle continu, on les suppose connues (sous forme de mots) de N bits ou N est un entier positif en général très grand. On ne peut représenter les coefficients qu'à l'aide d'un nombre fini de bits très inférieur à N. Il en résulte une modification de la réponse fréquentielle désirée. Il s'agit d'évaluer cette distorsion, et, en se donnant un critère d'erreurs acceptable, il s'agit de minimiser le nombre de bits nécessaires pour les représentations de ces coefficients. Posé ainsi, le problème est très difficile à résoudre. Ce problème appartient au domaine de la programmation non-linéaire dans un espace ponctuel("Non linear Integer Programming"). Néanmoins il est très important de pouvoir réduire le nombre de bits nécessaires pour représenter les coefficients, tout en restant le plus près possible du gabarit de départ. En effet, la complexité du filtre numérique étant essentiellement fonction de l'opérateur qui réalise les multiplications, une réduction des longueurs de mots des coefficients réduit sa complexité ; ou bien, pour la même complexité, cette réduction nous permettra d'augmenter les longueurs de mots des variables intermédiaires et de sortie, réduisant ainsi le niveau de bruit dû à l'arrondi des produits, comme nous le verrons dans le chapître V.

Après cette courte introduction montrant l'importance et la complexité de ce problème, nous allons résumer les travaux réalisés dans ce domaine et indiquer dans quelle optique nous avons traité ce problème.

Kaiser §6§ qui a étudié ce problème, suggère une étude basée sur le lieu d'Evans. On peut considérer l'effet de la quantification des coefficients comme de petits déplacements des pôles et des zéros du filtre (dans le plan z). Ce qui conduit à utiliser les techniques du lieu d'Evans pour l'évaluation de la

sensibilité de la structure du filtre pour des petites perturbations des coefficients. Cette approche nous paraît peu intéressante car elle nécessite une étude détaillée pour chaque filtre à réaliser.

Knowles et Olcayto §23§ l'ont étudié par une approche statistique. Dans cette approche on choisit comme critère d'erreur la valeur quadratique moyenne de la différence entre le gabarit du filtre réalisé et le gabarit du filtre optimum. Ce critère n'est évidemment pas satisfaisant quand on s'interesse à l'écart pour une fréquence particulière. C'est cependant cette approche que nous avons adoptée, car elle est de mise en oeuvre simple dans un petit calculateur.

IV.4 Erreur d'arrondi des produits.

Cette catégorie d'erreur est dûe à l'arrondi ou à la troncature des produits. Nous considérons ici le cas de l'arithmétique virgule fixe en complément à deux. Quand un nombre représenté sur n bits (bit de signe exclu), est multiplié par un autre nombre de n bits, le résultat est sur 2n bits (bit de signe exclu). Dans le cas de troncature, les n bits de poids faible ne sont pas pris en compte. Dans le cas de l'arrondi, le (n+1) ième bit est testé. S'il vaut zéro, on élimine les n bits de poids faible, comme dans le cas précédent. Si, par contre, il vaut 1, on ajoute 1 au n ième bit et on ne prend en compte que les n bits de poids fort. Ceci s'applique également pour les nombres négatifs à ceci près : on a le choix de faire l'arrondi du produit soit sur le module, soit sur sa représentation complément à deux, les deux étant équivalents sauf dans le cas où seul le (n+1) ième bit vaut un, tous les autres du (n+2) ième jusqu'au 2n ième étant nuls. Par la suite, nous considérerons ce cas particulier comme ayant une probabilité de réalisation quasi nulle.

L'effet de ces erreurs, dans le cas des filtres numériques non récursifs, est relativement facile à analyser, moyennant un certain nombre d'hypothèses simplificatrices. En effet, d'une part ces filtres sont toujours stables, et, d'autre part, l'erreur ne dépend pas de la forme de réalisation. En plus, si le signal d'entrée est stationnaire, les statistiques de l'erreur le sont aussi.

Avec les hypothèses suivantes, on peut assez facilement évaluer la valeur moyenne $\bar{e}$ et la variance $\sigma_{\bar{e}}^2$ à la sortie du filtre :

(i) L'excursion en amplitude du signal d'entrée est suffisamment grande et son spectre est suffisamment "riche".

(ii) Les sources d'erreur ne sont pas corrélées entre elles.

(iii) Chaque source d'erreur est une source de bruit blanc, c'est-à-dire :

$$E(e_i(nT), e_i(kT)) = 0 \qquad \text{pour tout } n \neq k$$

(iv) Il n'y a pas de débordement des registres quand on effectue l'opération de l'addition.

Dans ces conditions, on a :

espérance mathématique de l'erreur à la sortie :

$$\bar{e} = \sum_{i=0}^{M} \bar{e}(i) \qquad \text{(IV.16)}$$

variance de l'erreur :

$$\sigma_{\bar{e}}^2 = \sum_{i=0}^{M} a^2(i)\, \sigma_{e(i)}^2 \qquad \text{(IV.17)}$$

où $\sigma_{e(i)}^2$ est la variance de chaque source d'erreur (voir fig. I.4).

Cependant l'analyse se complique lorsqu'on cherche à connaître la nature de la distribution des erreurs e(i). Celle-ci dépend des coefficients multiplicateurs. Dans l'annexe D, nous avons déterminé les expressions de la valeur moyenne et de la variance de l'erreur e(i) dans l'hypothèse d'une distribution discrète et uniforme. Une analyse détaillée des filtres non récursifs est faite dans le chapître V .

L'évaluation de la valeur moyenne et de la variance de l'erreur de sortie dans les filtres numériques récursifs est beaucoup plus compliquée. En effet, les erreurs se propageant à travers la boucle de retour dûe aux coefficients b(i) pour rentrer dans le système. Donc, la valeur moyenne $\bar{e}$ et la variance $\sigma_{\bar{e}}^2$ sont des fonctions du temps avant d'atteindre le régime permanent. Leurs valeurs dépendent de la fonction de transfert du filtre et de la forme de réalisation.

Les flèches représentées fig. (I.4) et (I.6) indiquent les sources d'erreurs d'arrondi pour les deux formes des réalisations non décomposées. On voit que, dans le cas de la forme directe, les erreurs instantanées sont à l'entrée d'un système tous pôles, alors que dans la forme canonique les erreurs sont injectées dans un système ayant des pôles et des zéros. On peut en déduire qu'en général, la variance de l'erreur de sortie de la forme canonique est inférieure à celle de la forme directe. Dans le chapître V nous établissons les conditions pour lesquelles ceci est vrai.

Rader et Gold §24§ ont étudié l'effet du bruit d'arrondi dans les filtres numériques avec l'hypothèse que les sources de bruit sont non corrélées. Ils ont conclu que l'erreur d'arrondi et l'erreur de conversion analogique numérique (CAN) sont équivalentes, la seule différence étant topologique. C'est à dire que l'erreur de CAN s'ajoute au signal d'entrée du système alors que les erreurs d'arrondi entrent dans le système à divers points. Donc, ils en déduisent que l'on peut étudier ces dernières selon le modèle s'appliquant aux premiers et qui a été étudié par plusieurs personnes, en particulier Bennett §18§ et Widrow §19§.

Nous en retiendrons que la source de bruit qui s'introduit à chaque point de multiplication a les caractéristiques suivantes :

(i) bruit blanc uniformément réparti sur l'intervalle $(-q/2, +q/2)$ q : pas de quantification

(ii) variance de chaque source est égale à $q^2/12$

Sous cette hypothèse qui néglige la dépendance entre l'erreur d'arrondi et le coefficient multiplicateur particulier, la variance de l'erreur totale à la sortie peut être évaluée assez facilement. En effet, pour la forme canonique, les erreurs instantanées dûes aux coefficients a(i) s'ajoutent à la sortie, tandis que les erreurs dûes aux coefficients b(i) s'ajoutent à l'entrée du filtre. Si l'on suppose qu'il y a m coefficients a(i) et n coefficients b(i), la variance de l'erreur de sortie est :

$$\sigma_e^2 = \frac{nq^2}{12} \frac{1}{2\pi j} \int_C H(z)\, H(z^{-1})\, z^{-1} dz + \frac{mq^2}{12} \qquad \text{(IV.18)}$$

où H(z) est la fonction de transfert du filtre discret §24§, le contour C d'intégration étant le cercle unité §3§. Cette intégrale peut être évaluée par la méthode des résidus. Jury et al ont proposé une méthode d'évaluation numérique §25§.

Pour la forme directe, toutes les erreurs instantanées s'ajoutent au point (voir fig. I.4) d'entrée d'un système tous pôles. Donc, pour cette forme, si l'on a (m+n) coefficients au total, la variance de l'erreur totale à la sortie est :

$$= \frac{(m+n)q^2}{12} \frac{1}{2\pi j} \int_C \frac{1}{D(z)D(z^{-1})} \frac{dz}{z} \qquad \text{(IV.19)}$$

où D(z) est le dénominateur de la fonction de transfert H(z) du filtre numérique.

Les expressions équivalentes à IV.18 et IV.19 dans le domaine temporel sont données par : §3§ , §24§

Pour la forme canonique :

$$\sigma_e^2 = \frac{nq^2}{12} \sum_{i=0}^{\infty} h^2(iT) + \frac{mq^2}{12} \qquad \text{(IV.20)}$$

Pour la forme directe :

$$\sigma_e^2 = \frac{(m+n)q^2}{12} \sum_{i=0}^{\infty} h_d^2(iT) \qquad \text{(IV.21)}$$

où /h(iT)/est la réponse impulsionnelle du filtre et /h_d(iT)/ est la réponse impulsionnelle du système ayant pour fonction de transfert 1/D(z).

Pour un filtre de deuxième ordre ayant deux pôles complexes conjugués dans le plan de z, Gold et Rader §24§ ont évalué l'équation IV.18 (forme canonique) comme étant :

$$\sigma_e^2 = \frac{2q^2}{12} \frac{1}{1-r^2} \left(1 - \frac{r^2 \sin^2\theta\,(1+r^2)}{1+r^4-2r^2\cos(2\theta)}\right) + \frac{mq^2}{12} \qquad \text{(IV.22)}$$

où les pôles du filtre sont situés à $z = r \exp(\pm j\theta)$

Pour la forme directe IV.19 a été évaluée comme :

$$\sigma_e^2 = \frac{(m+2)q^2}{12} \frac{1+r^2}{1-r^2} \frac{1}{1+r^4-2r^2\cos 2\theta} \qquad \text{(IV.23)}$$

Ces deux équations démontrent bien que la variance de l'erreur peut être très différente dans les deux cas. On remarque également que la variance de l'erreur est indépendante de la variance du signal d'entrée.

Après avoir décomposé la fonction de transfert H(z) en filtres de premier ou/et de deuxième ordre soit en cascade soit en parallèle, il reste le choix important de la forme de réalisation de ces cellules élémentaires. Pour cela, les équations IV.18 et IV.23 s'appliquent. Cependant les équations IV.18 à IV.21 sont difficiles à évaluer et les équations IV.22 et IV.23 ne sont applicables qu'aux pôles complexes. Donc, le but essentiel du chapître V est de développer des expressions simples à évaluer, reliées directement aux coefficients et s'appliquant aux pôles complexes et aux pôles réels. Ainsi, ces expressions simples nous permettront de comparer les deux formes de réalisation des filtres numériques et de choisir celle pour laquelle la variance de l'erreur totale à la sortie est minimale.

IV.5 Contrainte dûe aux débordements de registres

La fig. I.7 montre les différentes réalisations d'un filtre élémentaire de deuxième ordre. Si l'on prend, par exemple, la réalisation en cascade canonique, il faut s'assurer que les variables w(n) et y(n) ne "débordent" pas des registres. Si l'on ne prend pas garde à ceci, on introduit une distorsion importante dans le système.

La condition sur w(n) pour éviter le débordement dépend du dénominateur de la FT. $H(z^{-1})$. Si l'on trouve la condition pour que $|w(n)| < 1$, il est facile, ensuite, de trouver la condition pour que $|y(n)| < 1$, parce que la FT entre w(n) et y(n) est le numérateur de la FT $H(z^{-1})$ et sa réponse impulsionnelle est finie. Donc nous allons dans ce qui suit, considérer un système purement récursif (c'est-à-dire $N(z^{-1}) = 1$).

Le modèle du bruit d'arrondi, comme nous l'avons signalé dans la section précédente, repose sur l'hypothèse qu'il n'y a pas de débordement lorsqu'on effectue la sommation indiquée dans l'équation IV.1 bis. Dans le cas de l'arithmétique en virgule flottante, il y a un cadrage (scaling) avant chaque addition, et, de ce fait, on introduit une source de bruit supplémentaire. Mais en arithmétique fixe, compte tenu de la dynamique des registres, on devra atténuer l'entrée pour faire face à ce problème de débordement. Mais, puisque la variance de l'erreur d'arrondi est indépendante de la variance du signal d'entrée, l'atténuation de la dynamique de celui-ci entraîne une dégradation du rapport $(S/_B)$. Donc, il faut l'atténuer au minimum. Nous allons examiner en détail la nature de cette contrainte.

La sortie à la n ième itération w(n) est donnée par le produit de convolution de la réponse impulsionnelle h(n) et de l'entrée x(n) :

$$w(n) = \sum_{k=0}^{n} h(k) \times (n-k) \qquad \text{(IV.24)}$$

Si l'entrée est bornée en valeur absolue par un nombre positif B, on a :

$$|x(n)| \leqslant B \quad \text{quelque soit} \quad n \geqslant 0 \qquad \text{(IV.25)}$$

En substituant en IV.24, on a :

$$|w(n)| \leqslant B \sum_{k=0}^{n} |h(k)| \qquad \text{(IV.26)}$$

Pour qu'il n'y ait pas débordement, la condition nécessaire et suffisante est :

$$|w(n)| < 1 \quad \text{quelque soit} \quad n \geqslant 0 \qquad \text{(IV.27)}$$

compte tenu de IV.26, cette condition se traduit par la contrainte suivante sur le signal d'entrée :

$$B \leqslant \frac{1}{\sum_{k=0}^{\infty} |h(k)|} \qquad \text{(IV.28)}$$

L'équation IV.28 nous donne donc une borne supérieure du signal d'entrée qui nous garantit l'absence de débordement des registres. Mais nous pouvons remarquer les faits suivants :

1/- Cette borne est"pessimiste". Pour une grande classe de signaux d'entrée, l'égalité dans la relation (IV.26) ne sera jamais atteinte, car c'est seulement lorsqu'on a $x(n) = \pm B$ pour tout $n \geqslant 0$ et que signe $(x(n-k))$ = signe $(h(k))$ pour tout $k \in [0,\infty]$ que l'égalité sera atteinte. Donc, la condition IV.28 est très restrictive.

2/- La borne B d'après IV.28 n'est pas toujours facile à évaluer, sauf pour les cas les plus simples.
Pour le système de 1er ordre, il est assez facile d'évaluer IV.28. En effet, la fonction de transfert d'un tel système est :

$$H(z^{-1}) = \frac{1}{1-b\ z^{-1}}$$

où b est une constante < 1.
La réponse impulsionnelle de ce système est :

$$h(n) = b^n \quad \text{pour} \quad n \in [0,\infty], \quad n \text{ entier}$$

$$\text{d'où } B \leq \frac{1}{\frac{1}{1-b}} \leq (1-b) \qquad \text{(IV.29)}$$

Il faut remarquer que, pour b voisin de 1, le facteur (1-b) est très petit. Ceci implique la nécessité d'une grande dynamique des registres. Mais, en pratique, nous n'avons pas besoin d'appliquer la contrainte IV.28. Jackson §10§ a étudié ce problème et a développé des expressions qui nous permettent de choisir une stratégie de cadrage pour chaque classe de signaux d'entrée rencontrés dans la pratique. Dans l'annexe E, nous évaluons la contrainte IV.28 pour les systèmes de II^{ème} ordre pour bien montrer la difficulté d'une analyse exacte de ce problème. Pour les filtres que nous avons conçus par la méthode d'optimisation du chapître II, le gain du filtre normalise le module de la réponse fréquentielle à 1 dans la bande passante. De ce fait, il y a déjà une atténuation du signal d'entrée. Donc, intuitivement, on peut espérer que cette diminution est suffisante pour éviter les débordements. Pour les filtres que nous avons simulés, la contrainte de IV.28 a été évaluée. Ceci ne dépasse guère un facteur $\frac{1}{8}$, c'est à dire qu'il faut prévoir 3 bits supplémentaires pour représenter les états intermédiaires du filtre.

Oppenheim §26§ a proposé une arithmétique semi-flottante, en utilisant la normalisation de toutes les variables du filtre. L'arithmétique des additions et des multiplications est en virgule fixe. Le facteur de cadrage, dû à la normalisation, est une puissance de deux et la sortie à chaque instant est recadrée (décalée) par ce même facteur. De ce fait, la dynamique est considérablement augmentée et les performances de ce mode d'arithmétique se situent entre l'arithmétique fixe et l'arithmétique flottante.

IV.6 Cycles limites dans les filtres numériques récursifs

Les cycles limites dans les filtres numériques récursifs sont dûs à l'arrondi des produits intermédiaires §27§ §28§. Les signaux de type échelon unité ou impulsion de dirac ainsi que toute remise à zéro du signal d'entrée après l'instant nT provoque le phénomène des cycles limites. Dans le premier cas, au lieu de s'approcher assymptotiquement de la valeur constante déterminée par la F.T. du filtre, la sortie oscille autour de celle-ci. Dans les deux derniers cas la sortie, au lieu de tendre assymptotiquement vers zéro, oscille autour du zéro.

Ces cycles limites accompagnent toutes les réalisations des filtres numériques récursifs. L'amplitude de ces oscillations dépend des coefficients du filtre, de la forme de la réalisation et du nombre de bits utilisés pour la quantification des produits. Dans ces conditions on ne peut que se contenter de donner des bornes supérieures de l'amplitude des oscillations en fonction de ces trois paramètres. Puisque ces phénomènes sont indépendants de la base du système numérique employé (représentation des nombres réels dans la machine) nous avons conduit l'étude pour la base décimale.

Système de Ier ordre :

Soit le filtre du premier ordre dont l'équation de reccurence est :

$$y(n) = -b\, y(n-1) + x(n) \qquad \text{(IV.30)}$$

Sa F.T. est : $h(z) = \frac{z}{z+b}$ (IV.31)

Si $|b| < 1$, le filtre est strictement stable. Mais, si dans la réalisation du filtre, on arrondit le produit $[b\, y(n-1)]$ sa R.I. peut ne pas tendre asymptotiquement vers zéro.

Soit $x(n) = 10\,\delta(n)$ et supposons que l'arrondi soit fait à l'entier le plus proche (q=1). Le tableau (IV.1) indique les valeurs successives de sortie pour les différentes valeurs de b. Nous voyons que les cycles limites alternant en signe et ayant une amplitude K se produisent pour $0,5 \leq b < 1$. Nous voyons aussi que dans chaque cas, les cycles limites ont lieu lorsque la valeur effective de b devient 1, c'est à dire quand le filtre se comporte comme s'il avait un pôle à z = -1. Pour b négatif, le pôle effectif sera a z = +1, mais le cycle limite est unipolaire. Pour mieux comprendre ce phénomène, évaluons : la sortie réelle, la sortie arrondie et la valeur apparente du coefficient b pour b = -1/2. Quand l'entrée est $10\,\delta(n)$, on a, pour $n \geq 1$, l'équation de sortie :

$$y(n) = -\left[by(n-1)\right]_A \qquad \text{(IV.32)}$$

où $[x]_A$ = valeur de x arrondie à l'entier le plus proche.
Le tableau (IV.2) montre les valeurs successives de y(n), n = 0, 1... 8, ainsi que les valeurs apparentes du coefficient b à chaque itération. Nous voyons bien que la sortie, au lieu de tendre vers zéro, reste constante à 1. Soit K, la valeur absolue du cycle limite. La relation (IV.32) nous donne :

n	b=0,1 $y(n)$	b=0,2 $y(n)$	b=0,3 $y(n)$	b=0,4 $y(n)$	b=0,5 $y(n)$	b=0,6 $y(n)$	b=0,7 $y(n)$	b=0,8 $y(n)$	b=0,9 $y(n)$
0	10	10	10	10	10	10	10	10	10
1	-1	-2	-3	-4	-5	-6	-7	-8	-9
2	0	0	1	2	3	4	5	6	8
3			0	-1	-2	-2	-4	-5	-7
4				0	1	1	3	4	6
5					-1	-1	-2	-3	-5
6					1	1	1	2	5
7					-1	-1	-1	-2	-5
8					1	1	1	2	5
9					-1	-1	-1	-2	-5

Tableau IV.1

$p = -0,5$

n	$y(n)$	b effectif	Position du pôle
0	10	$-\frac{1}{2}$	$\frac{1}{2}$
1	5	$-\frac{1}{2}$	$\frac{1}{2}$
2	3	$-\frac{3}{5}$	$\frac{3}{5}$
3	2	$-\frac{2}{3}$	$\frac{2}{3}$
4	1	-1	1

Tableau IV.2

$$K = -\left[b\ K\right]_A \qquad \text{(IV.33)}$$

et pour $b < 0$

$$K = \left[|b|\ K\right]_A \qquad \text{(IV. 33 bis)}$$

Mais, nous avons aussi :

$$\left[|b|\ K\right]_A \leq |b|\ K + \frac{q}{2} \qquad \text{(IV.34)}$$

où q = pas de quantification = 1 dans notre cas. Ces deux rélations nous donnent :

$$K \leq |b|\ K + \frac{q}{2} \qquad \text{(IV.35)}$$

ou :

$$K(1-|b|) \leq \frac{q}{2} \qquad \text{(IV.36)}$$

ou encore :

$$K \leq \frac{q/2}{1-|b|} \qquad \text{(IV.37)}$$

Pour $b > 0$, K alterne en signe. Donc, nous avons :

$$K = -\left[b(-K)\right]_A = \left[bK\right]_A \qquad \text{(IV.38)}$$

(IV.34) et (IV.38) nous donnent :

$$K \leq bK + q/2$$

et finalement :

$$K \leq \frac{q/2}{1-b} \qquad \text{(IV.37 bis)}$$

Donc, dans tous les cas, nous avons bien la relation (IV.37).
L'amplitude, (crête-à-crête) des oscillations est :

$$L_0 = K - (-K) = 2K \frac{q}{1-|b|} \qquad |b| < 1 \qquad \text{(IV.40)}$$

Si la quantification après l'arrondi est faite sur n bits, signe exclu, on a : $q = 2^{-n}$ et :

$$K \leq \frac{2^{-n}}{2(1-|b|)} \qquad \text{(IV.41)}$$

Quand $x(n) = C$, (valeur constante) quel que soit $n \geq 0$, nous avons :

$$y(n) = -b\, y(n-1) + ax(n) \qquad \text{(IV.42)}$$

Soit $\left[ax(n)\right]_A = C_1$ C_1 : valeur constante
qui conduit à :

$$K = \left[|b|\, K\right]_A + C_1$$

D'après (IV.34) :

$$K \leq |b| K + q/2 + C_1 \qquad \text{(IV.43)}$$

et finalement :

$$K \leq \frac{q/2 + C_1}{1 - |b|} \qquad \text{(IV.44)}$$

Puisque :

$$\left[|b|\ K\right]_A \geq |b| K - q/2 \qquad \text{(IV.45)}$$

la borne supérieure est :

$$K \geq \frac{C_1 - q/2}{(1-|b|)} \qquad \text{(IV.46)}$$

(IV.44) et (IV.46) nous permettent d'écrire :

$$\frac{C_1 - q/2}{1 - |b|} \leq K \leq \frac{C_1 + q/2}{1 - |b|} \qquad \text{(IV.46 bis)}$$

L'amplitude crête-à-crête des oscillations est :

$$L_1 = \frac{q}{1 - |b|} = L_0 \qquad \text{(IV.47)}$$

Ceci montre que l'amplitude crête-à-crête reste la même dans les deux cas.

Pour les filtres numériques de deuxième ordre, Gowdy §28§ développe des expressions permettant de déterminer les bornes supérieures des oscillations et ceci pour les deux formes de réalisation (canonique et directe).

C H A P I T R E V

ANALYSE DES ERREURS D'ARRONDI

DANS LES FILTRES NUMERIQUES

Dans le chapître précédent, il a été indiqué que le signal d'erreur à la sortie du filtre dû à la troncature (ou l'arrondi) des produits intermédiaires dépend de la forme de réalisation du filtre. Nous avons vu aussi les difficultés que pose la réalisation non factorisée d'un filtre d'ordre élevé, et la nécessité de réaliser des filtres élémentaires de deuxième ou premier ordre en cascade ou en parallèle.

Dans ce chapître, nous allons étudier et déduire des expressions de la valeur moyenne et de la variance de l'erreur de sortie pour les deux formes d'"implémentation" (la forme directe et la forme canonique) des filtres numériques élémentaires. Ces expressions sont importantes puisqu'elles nous permettent de choisir entre les formes différentes de réalisations d'un filtre pour que l'erreur d'arrondi soit le minimum.

Les résultats théoriques développés dans ce chapître sont comparés aux résultats expérimentaux obtenus par simulation. Le chapître VI décrit cette simulation sur petit calculateur.

Nous allons étudier d'abord les filtres non récursifs pour démontrer la dépendance de l'erreur d'arrondi et des coefficients du filtre.

V.1 Analyse des erreurs pour les filtres non récursifs

Il est relativement facile d'étudier l'effet des erreurs d'arrondi dans les filtres non récursifs. Ceci est dû à l'indépendance du signal de sortie y(nT) par rapport aux sorties précédentes y(nT-T), etc.
Donc, l'erreur qui résulte de l'arrondi d'un produit à la $n^{ième}$ itération est indépendante des sorties précédentes. Si la suite d'entrée est une suite temporelle stationnaire, l'erreur totale à la sortie est aussi une suite stationnaire dont la valeur moyenne est égale à la somme des valeurs moyennes des erreurs d'arrondi de chaque produit et la variance est égale à la somme des variances de chaque terme d'erreur.

Il est utile d'insister sur les hypothèses qu'on va faire pour modeler le processus de l'erreur d'arrondi. On considère que :

(i) les sources d'erreur sont mutuellement non corrélées entre elles. Sous cette hypothèse, la variance de l'erreur à la sortie est tout simplement la somme des variances individuelles de chaque source.

(ii) chaque source d'erreur est une source de bruit blanc, indépendant du signal d'entrée. Cette hypothèse sera valable pour une grande classe de signaux allant des signaux à bande étroite aux signaux à large bande. Evidemment, cette hypothèse n'est guère valable pour des signaux constants (un échelon unité), ou une impulsion de dirac.

(iii) il n'y a pas de débordement des registres dans les additions.

Même avec ces hypothèses qui simplifient l'analyse, une difficulté majeure subsiste qui est dûe à la dépendance de l'erreur et de la valeur des coefficients du filtre. Il est impossible

de tenir compte de cette dépendance dans le modèle statistique que nous avons étudié. Néanmoins, le modèle reste valable pour une large classe de coefficients, quelques unes des valeurs particulières étant écartées. Compte tenu du fait que, dans la réalisation d'un filtre d'ordre élevé, il y aura plusieurs coefficients non identiques et prenant toutes les valeurs possibles dans un intervalle bien défini, le modèle que nous avons développé sera très utile pour une évaluation raisonnablement exacte du niveau de bruit d'arrondi dans les filtres numériques. A partir de ces remarques, nous allons aborder l'étude des filtres non récursifs.

Nota : Nous appellerons "bruit d'arrondi" indifféremment les bruits dûs à la troncature et à l'arrondi, sauf là où il pourrait y avoir une confusion.

V.1bis Système du premier ordre non récursif

Soit la fonction du transfert en z^{-1} du filtre à réaliser :

$$H(z^{-1}) = 1 + a\, z^{-1} \qquad (V.1)$$

L'équation de récurrence qui réalise ce filtre est donnée par :

$$y(n) = x(n) + a\, x(n-1) \qquad (V.2)$$

où y(n) est la sortie à l'instant nT et x(n) et x(n-1) sont les valeurs d'entrée aux instants nT et (nT-T).
Compte tenu de l'erreur d'arrondi, l'équation devient :

$$y'(n) = x(n) + ax(n-1) - \epsilon(n) \qquad (V.3)$$

où l'on suppose qu'il n'y a pas d'erreur dans la représentation du coefficient a. Si x(n) et a sont quantifiés sur n bits (bit de signe exclu), le produit (a x(n)) aura 2n bits (bit de signe exclu). On est amené à tronquer (arrondir) les n bits de poids faible, ce qui constitue l'erreur de troncature (arrondi) $\epsilon(n)$.
Si l'on définit l'erreur totale comme étant

$$e(n) = y(n) - y'(n) \qquad (V.4)$$

on a

$$e(n) = \epsilon(n) \qquad (V.5)$$

En prenant l'espérance mathématique de deux membres de l'équation (V.5) on a

$$\bar{e}(n) = \overline{\epsilon}(n) = \frac{2^{-n}}{2} = \frac{q}{2} \quad \text{(d'après l'annexe D)}$$

où q est le pas de quantification.
La valeur quadratique moyenne est :

$$\overline{e^2}(n) = \overline{\epsilon^2}(n) = \frac{2^{-2n}}{3} = \frac{q^2}{3} \quad \text{(d'après l'annexe D)}$$

Donc, la variance de l'erreur σ_e^2 est égale à :

$$\sigma_e^2 = \overline{e^2(n)} - (\bar{e}(n))^2$$

$$= \overline{2^{2n}}/12 = q^2/12 \qquad (V.6)$$

Les courbes (V.1) et (V.2) montrent les variations de $\bar{e}(n)$ et σ_e^2 en fonction du coefficient a. On voit que, pour certaines valeurs du coefficient a, ce modèle n'est pas valable. En effet, si a vaut 0,5 la multiplication par a correspond à un décalage à droite du nombre à multiplier, et l'erreur est soit zéro, soit $2^{(n-1)}$ (ou $\frac{q}{2}$). Donc $\bar{e} = \frac{q}{4}$ et $\sigma_e^2 = \frac{q^2}{16}$. Les histogrammes (V.3) (V.4) de l'erreur qui correspondent à ces valeurs de coefficients confirment bien que la distribution de l'erreur n'est pas uniforme comme nous l'avons supposée, mais, concentrée autour des valeurs particulières.

Cas de Troncature :

Variance du signal d'entrée réduite

---- Théorique

⊙ Expérimentale

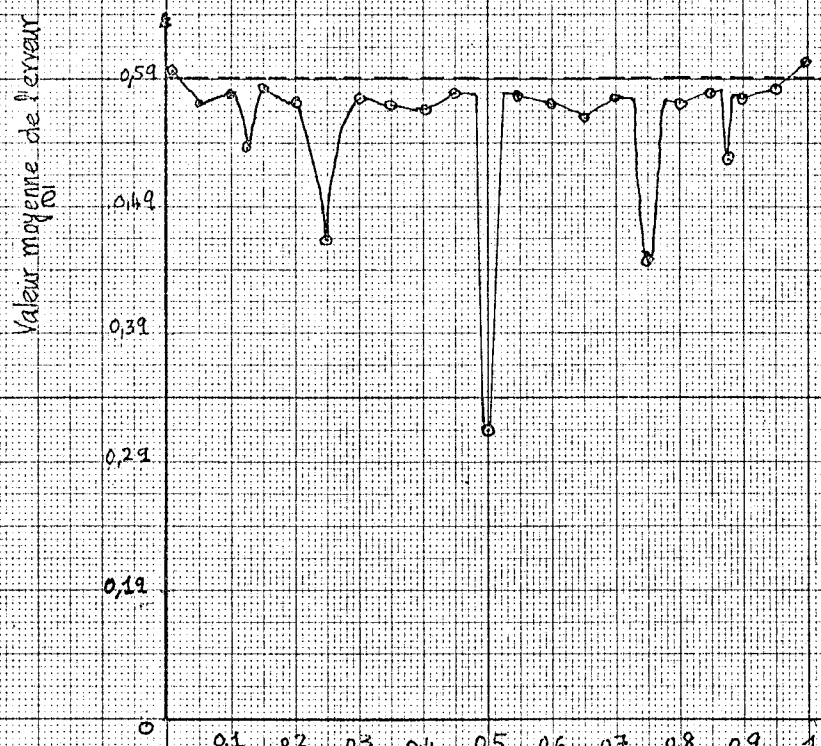


fig V 5

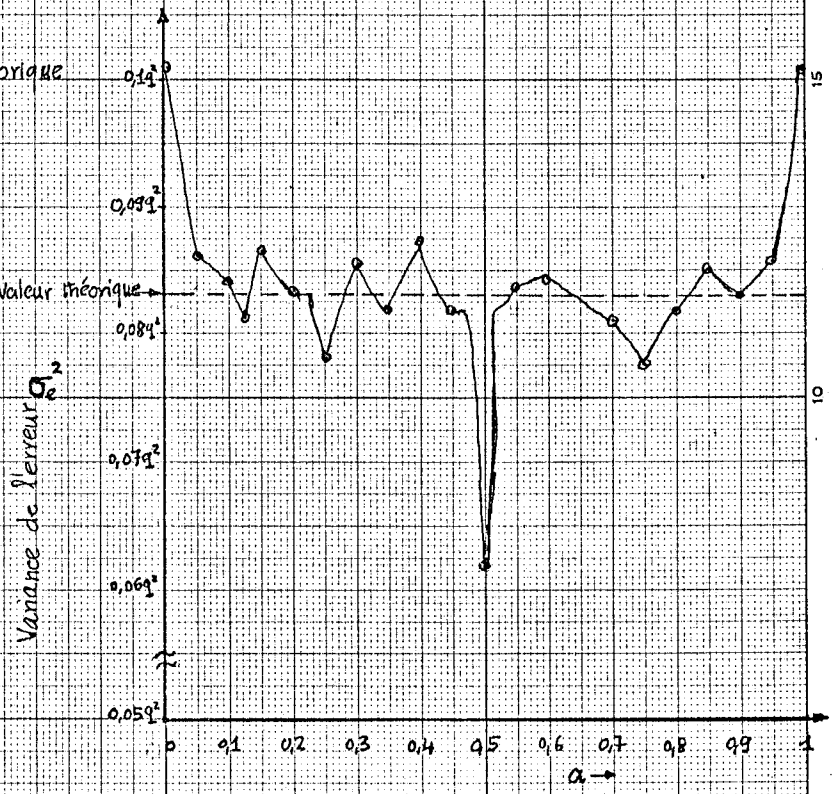


fig. V 6

Variance du signal d'entrée $\sigma_s'^2 = \frac{1}{(16)^2}\sigma_s^2$.

Il est à remarquer que : la variance de l'erreur ne dépend pas de la variance du signal d'entrée. Donc, si la variance du signal d'entrée diminue, le rapport signal sur bruit ($S/_B$) diminue aussi. Les courbes (V.5) et (V.6) vérifient bien cette affirmation.

V.2 Système de 1er ordre récursif :

Soit le filtre à réaliser :

$$H(z^{-1}) = \frac{1}{1+b\ z^{-1}}$$

Donc l'équation de récurrence qui réalise ce filtre, quand aucune erreur ne se produit est :

$$y(n) = -(b\ y(n-1)) + x(n) \qquad (V.7)$$

Dans la réalisation pratique de cette équation de récurrence, une source d'erreur s'introduit comme indiqué sur le schéma bloc fig. V.11 Cette erreur à la n^{ième} itération est $\delta(n)$. Avec les mêmes hypothèses que précédemment la sortie réelle est :

$$y'(n) = -(b\ y'(n-1) - \delta(n)) + x(n) \qquad (V.8)$$

Dans le cas de la troncature, nous avons démontré par ailleurs (annexe D) que $\delta(n)$est toujours positive.

Si l'on définit l'erreur totale à la n^{ième} itération comme étant e(n) = y(n) - y'(n), nous avons, en substituant dans (V.8) et soustrayant (8) de (7)

$$e(n) = -b\ e(n-1) - \delta(n) \qquad (V.9)$$

Cette équation est importante. Elle montre que nous pouvons prendre pour modèle du processus d'erreur une source de bruit blanc à l'entrée du système. L'erreur à la sortie obéit au même type d'équation de récurrence que le signal d'entrée. Le processus d'erreur /e(n)/ est une séquence stationnaire (après la période transitoire) si x(n) l'est et si en plus $|b| < 1$. Prenons l'espérance mathématique de deux membres de l'équation (V.9).

$$\overline{e(n)} = -b\ \overline{e(n-1)} - \overline{\delta}(n) \qquad \text{(V.10)}$$

Dans le cas de la troncature $\overline{\delta(n)} = q/2 = \overline{\delta}$

dans le cas de l'arrondi $\overline{\delta(n)} = 0$

En prenant la transformée en z de (V.10)

$$E(z^{-1}) = -b\ z^{-1}\ E(z) - \overline{\delta}(z)\ \frac{z}{z-1}$$

d'où

$$E(z^{-1}) = -z\ \overline{\delta}(z)\ \Big/ (z-1)(1+b\ z^{-1}) \qquad \text{(V.11)}$$

En utilisant le théorème de la valeur finale, nous pouvons calculer $\overline{e}(n)$ quand $n \to \infty$

$$\overline{e} = \lim_{z \to 1} \frac{(z-1)\ (-z\ \overline{\delta}(z))}{(z-1)(1+b\ z^{-1})}$$

$$\overline{e} = \frac{-\ \overline{\delta}(z)}{1+b} = \frac{-q/2}{1+b} \qquad \text{(V.12)}$$

On voit que la valeur moyenne de l'erreur dépend du coefficient b.

Evaluons le moment d'ordre deux de l'erreur :

$$\overline{e^2(n)} = b^2\ \overline{e^2(n-1)} + \overline{\delta^2(n)} + 2b\ \overline{\delta(n)\ e(n-1)}$$

Puisque $\delta(n)$ est l'erreur injectée dans le système à la $n^{ième}$ itération, et le processus étant blanc, elle n'est pas correlée avec l'erreur totale à la $(n-1)^{ième}$ itération.

Donc,

$$\overline{e^2(n)} = b^2 \overline{e^2(n-1)} + \overline{\delta^2(n)} + 2b\ \overline{\delta(n)}\ \overline{e(n-1)}$$

Dans le cas de l'arrondi,

$$\overline{\delta(n)} = 0$$

Donc,

$$\overline{e^2(n)} = b^2 \overline{e^2(n)} + \overline{\delta^2(n)}$$

$$\sigma_{eA}^2 = b^2 \sigma_{eA}^2 + \sigma_\delta^2 \qquad (\text{quand } n \to \infty)$$

où σ_{eA}^2 = la variance de l'erreur (arrondi)

$$\sigma_{eA}^2 = \frac{\sigma_\delta^2}{1-b^2} \tag{V.13}$$

Dans le cas de la troncature, on a :

$$\overline{e^2(n)} = m_2 = b^2 m_2 + \overline{\delta^2(n)} + 2b\ \overline{e(n-1)\ \delta(n)}$$

ou m_2 représente le moment d'ordre 2 de la variable e(n) soit encore

$$e^2(n) = m_2 = b^2 m_2 + \overline{\delta^2(n)} + 2b\ \overline{e(n-1)}\ \overline{\delta(n)}$$

en substituant les valeurs théoriques de e(n-1) et $\overline{\delta(n)}$ dans la relation précédente,
Nous avons :

$$m_2' = \frac{q^2}{3(1-b^2)} - \frac{q^2 b}{2(1+b)\ (1-b^2)} \tag{V.14}$$

$$\sigma^2_{eT} = m_2 - (\overline{e(n)})^2$$

où σ^2_{eT} est la variance de l'erreur dûe aux troncatures

$$= \frac{q^2}{3(1-b^2)} - \frac{q^2}{4(1+b^2)} \left[2b/_{1-b} + 1\right]$$

$$= \frac{q^2}{3(1-b^2)} - \frac{q^2}{4(1+b^2)} \left(\frac{1+b}{1-b}\right)$$

$$= \frac{q^2}{3(1-b^2)} - \frac{q^2}{4(1-b^2)} = \frac{q^2}{12} \frac{1}{(1-b^2)} \qquad (V.15)$$

Sur les graphiques (V.7) et (V.8) sont tracées les courbes de $\overline{e}$ et σ^2_e en fonction de $\frac{1}{1-b}$ et $\frac{1}{1-b^2}$. On s'aperçoit que l'hypothèse du bruit de troncature uniformément répartie (valeur moyenne q/2 et variance $q^2/_{12}$) est vérifiée pour une classe assez large de coefficients b comprise entre $2^{-(n-1)}$ et $1-2^{-(n-1)}$ (les valeurs particulières de b telles que 1/2, 1/4, 3/4 ont une valeur moyenne et une variance inférieures à la prédiction théorique). Pour les valeurs extrêmes de b nous avons des valeurs nettement supérieures à la prédiction théorique. Cela peut être expliqué assez facilement. Prenons par exemple $b = -1/2$; la multiplication par 1/2 est effectivement un décalage à droite d'une position. Donc, l'erreur est soit 0, soit q/2. Dans ce cas, l'erreur est une variable aléatoire discrète qui ne prend que deux valeurs soit 0, soit q/2 avec une probabilité 1/2. C'est ce que montre l'histogramme de l'erreur Fig.V.4 et V.9 .

En évaluant la valeur moyenne et la variance d'une telle variable, nous avons :

$$\overline{X} = \sum_k p_k X_k = 0 + \frac{1}{2} \frac{q}{2} = q/4$$

$$X^2 = \sum_k p_k X_k = 0 + \frac{1}{2} \frac{q^2}{4} = q^2/8$$

$$\sigma^2_X = q^2/8 - q^2/16 = q^2/16$$

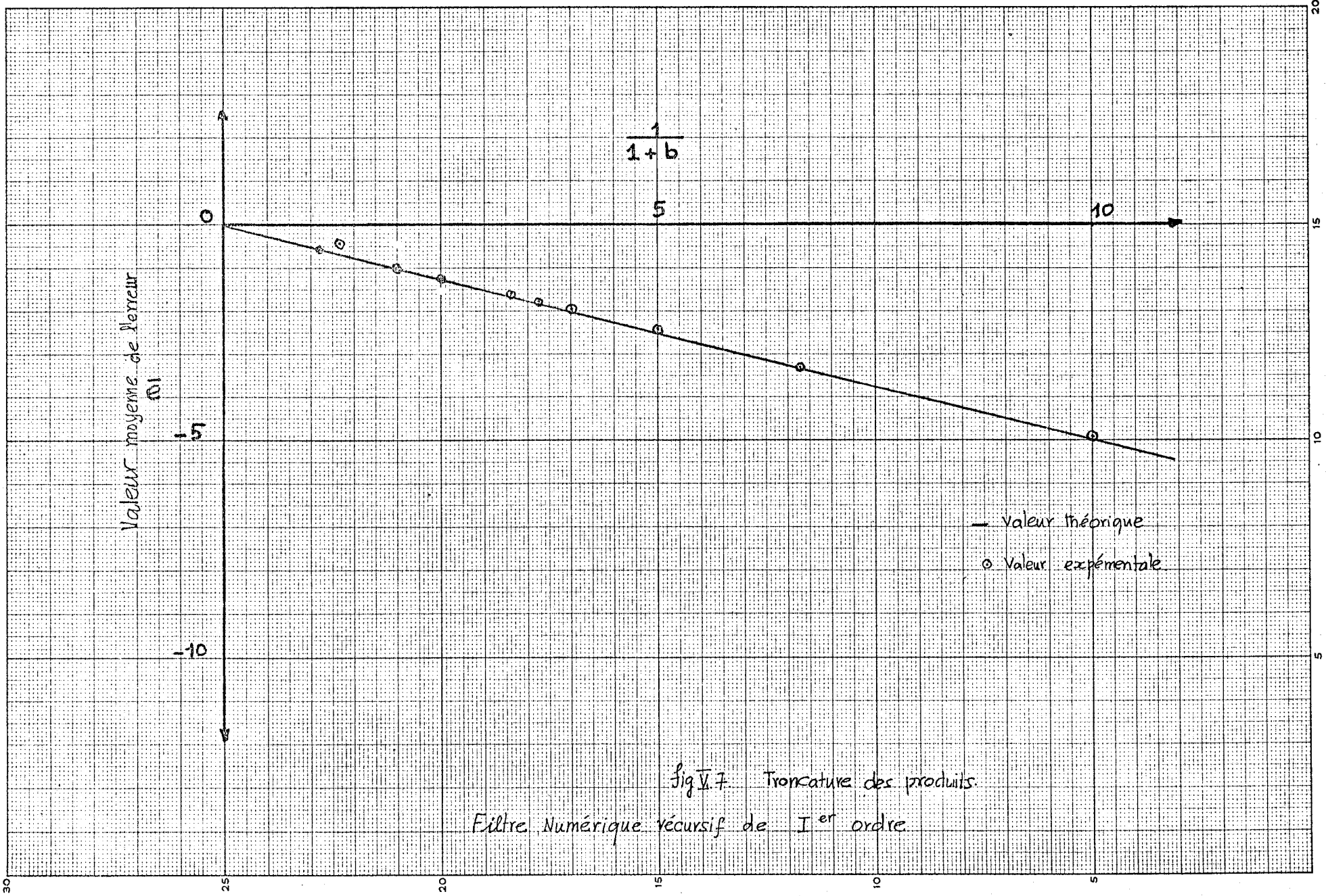
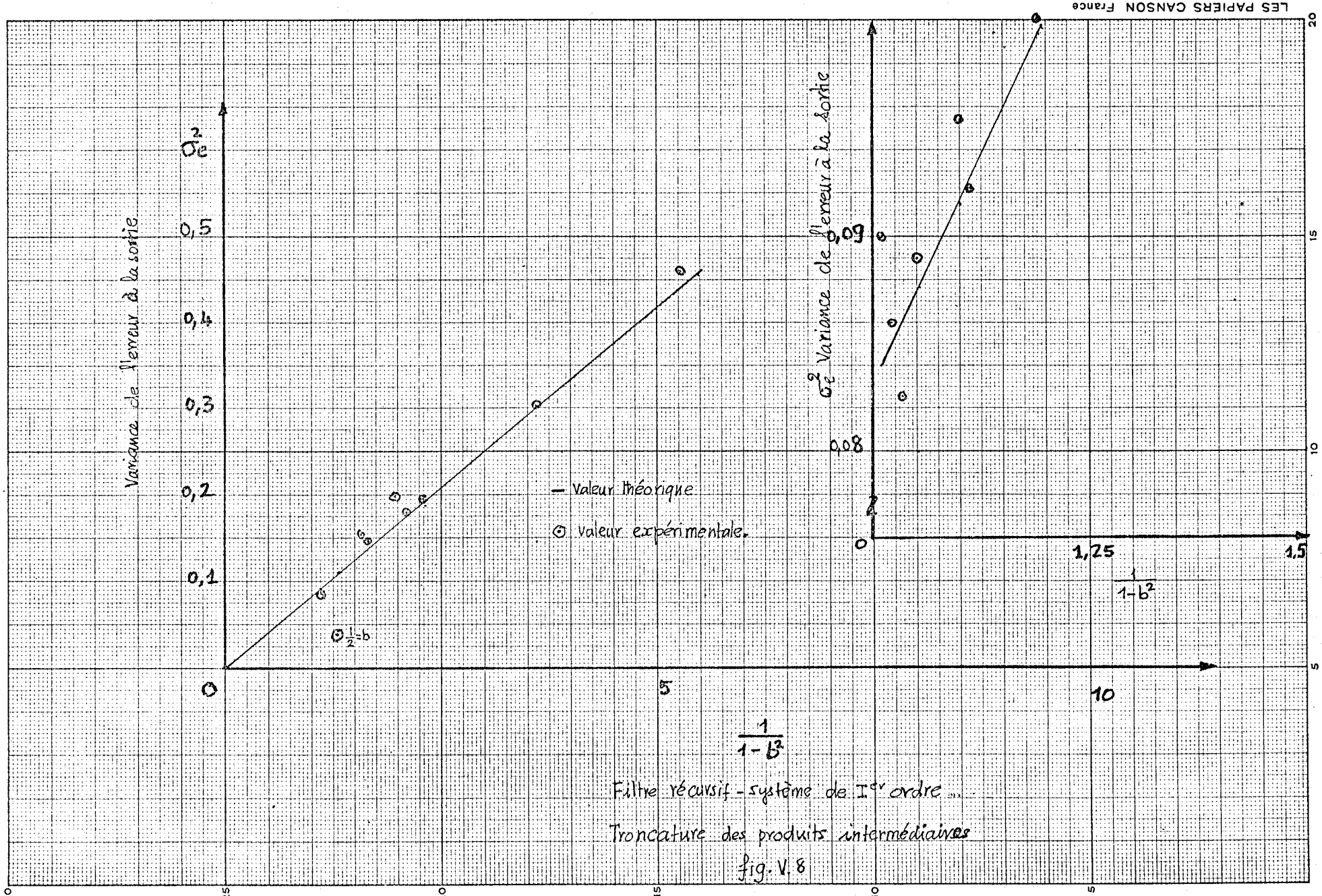


fig V.7 Troncature des produits.

Filtre Numérique récursif de I^er ordre



Filtre récursif - système de Ier ordre

Troncature des produits intermédiaires

fig. V. 8

Fig V.8 bis. Arrondi des produits.

Avec cette variable à l'entrée, à la sortie du système nous avons :

$$\bar{e} = \frac{1}{1-\frac{1}{2}} (q/4) = q/2 = 0,5$$

$$= q^2/16 \frac{1}{1-(\frac{1}{2})^2} = q^2/12 = 0,0833$$

C'est ce qui est vérifié expérimentalement.

Considérons maintenant le même système du 1er ordre d'un autre point de vue qui, pensons-nous, doit simplifier beaucoup les calculs. Quand l'entrée d'un système linéaire est un bruit blanc stationnaire avec une valeur moyenne nulle et une variance σ_e^2 (voir le schéma de bloc V.10) on peut très facilement évaluer la variance de la sortie d'un tel système en utilisant le produit de convolution :

$$s_e(n) = \sum_{k=o}^{n} h(k)\, e(n-k)$$

où h(k) est la réponse impulsionnelle du système

$$\overline{s_e^2(n)} = \sum_{k=o}^{n} \sum_{j=o}^{n} h(k)h(j)\overline{e(n-k)e(n-j)}$$

$$= \sum_{k=o}^{n} h^2(k)\, \sigma_e^2 \qquad (V.16)$$

puisque /e(n)/ est une séquence de bruit blanc stationnaire avec une variance σ_e^2 nous avons :

$$\overline{e(n-k)e(n-j)} = \sigma_e^2\, \delta(k-j)$$

Donc, pour un système de 1er ordre, quand $n \to \infty$

$$\overline{s_e^2(n)} = \sigma_e^2 \sum_{k=o}^{\infty} h^2(k)$$

La valeur moyenne de $/\delta_e(n)/$ étant nulle

$$\sigma^2_{\delta e} = \sigma^2_e \sum_{K=0}^{\infty} h^2(k)$$

$$= \sigma^2_e \sum_{K=0}^{\infty} b^{2K} , \quad |b| < 1$$

$$= \frac{\sigma^2_e}{1 - b^2}$$

Soit σ^2_x la variance du signal d'entrée d'un tel système, sa valeur moyenne étant nulle.
La variance du signal de sortie y est :

$$\sigma^2_y = \sigma^2_x / 1-b^2$$

On peut définir un rapport signal sur bruit comme :

$$\frac{\sigma^2_y}{\sigma^2_{\delta e}} = \sigma^2_e \cdot 12 \cdot 2^{2(n-1)}$$

Cela nous montre que, pour la même variance du signal d'entrée, diminuer d'un bit la longueur de mot après la multiplication diminue d'un facteur quatre le rapport signal sur bruit.

V.3 Etude d'un système de IIème ordre purement récursif (ou auto-régressif) :

Soit le filtre à réaliser :

$$H(z^{-1}) = \frac{1}{1 + b_1 z^{-1} + b_2 z^{-2}}$$

Les hypothèses sont :

(i) la séquence à l'entrée de ce filtre est une séquence de bruit blanc stationnaire, valeur moyenne nulle et variance σ^2_x. On la note : /x(n)/

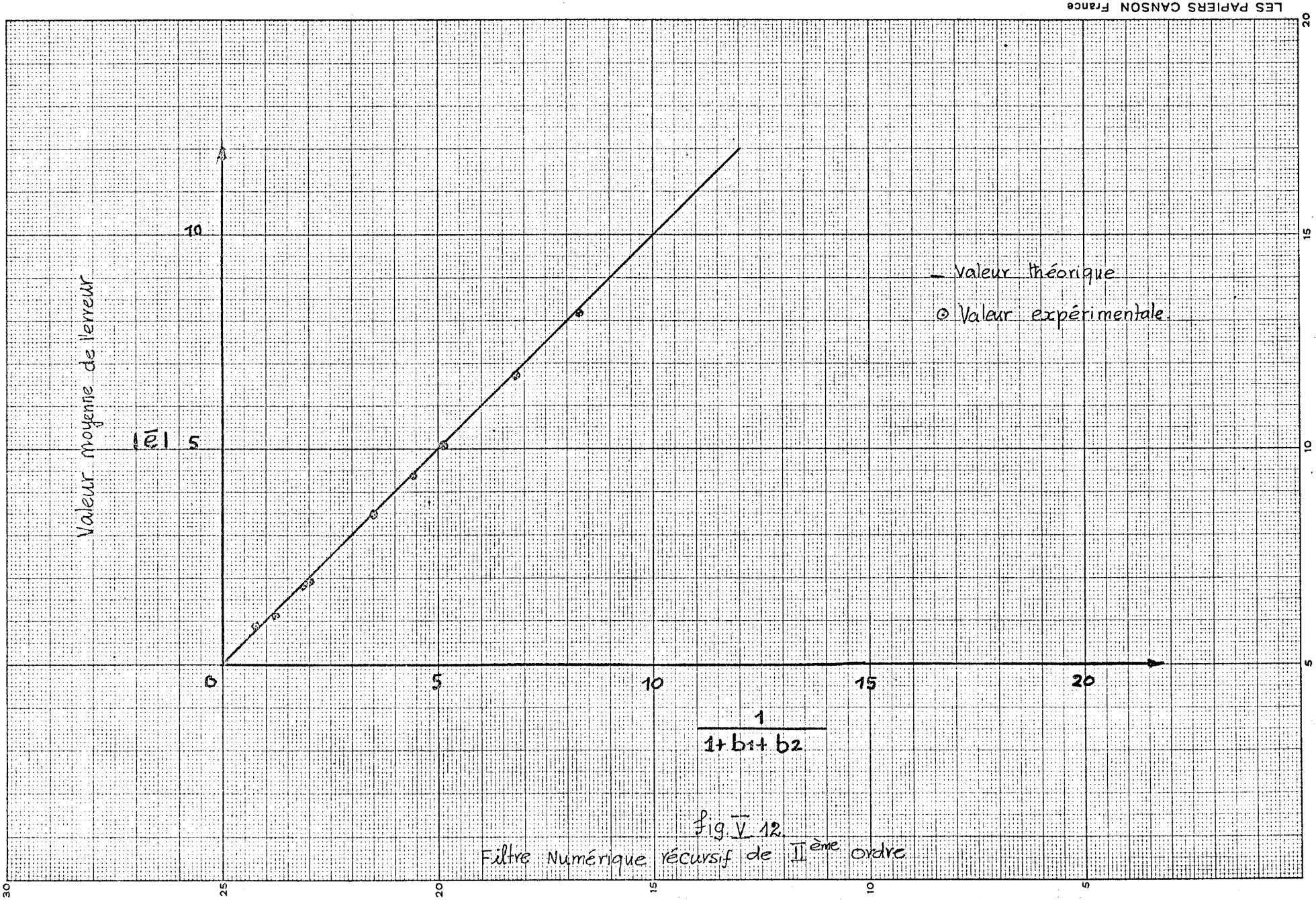


Fig V 12.

Filtre Numérique récursif de II ème ordre

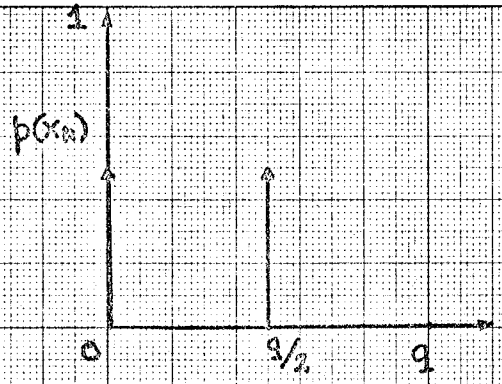


fig. V.9.

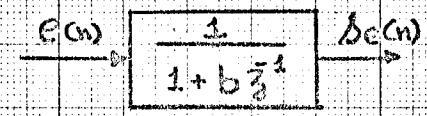


fig. V.10.

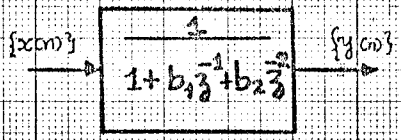


fig. V.13.

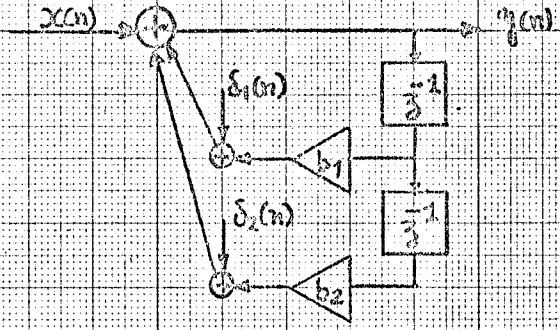
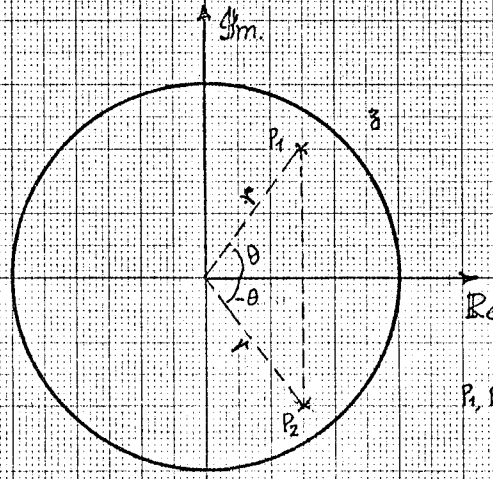


fig. V.11



P1, P2 : Pôles complexes conjugués

fig. V.14.

(ii) Les erreurs d'arrondi ou de troncature après la multiplication sont considérées comme des sources de bruit mutuellement non corrélées entre elles et avec le signal d'entrée.

(iii) chaque source de bruit a une valeur moyenne nulle (cas de l'arrondi) ou q/2 (cas de troncature) et une variance de $\frac{1}{12} q^2$ où q est le pas de quantification.

Le schéma fonctionnel qui montre l'injection du bruit dans le système est donné fig. V.11.

On voit que les deux sources d'erreur s'introduisent à l'entrée du système. Puisqu'elles sont, par hypothèse, non corrélées entre elles, et, statistiquement indépendantes d'une itération à l'autre, on peut évaluer la variance du bruit à la sortie en utilisant l'équation V.16, soit :

$$\overline{s_e^2} = \sigma_e^2 \sum_{k=0}^{\infty} h^2(k)$$

où h(k) est la réponse impulsionnelle du système et σ_e^2 est la variance de l'erreur à l'entrée soit :

$$\sigma_e^2 = \frac{q^2}{12} + \frac{q^2}{12}$$

selon l'hypothèse de l'indépendance statistique de ces deux sources d'erreur. Puisque la somme des carrés de la réponse impulsionnelle n'est pas toujours facile à évaluer, nous allons développer des expressions pour la valeur moyenne et pour la variance de l'erreur à la sortie en fonction des coefficients bi.
Soit, y(n) la sortie idéale (sans erreur) à la $n^{ième}$ itération du système quand on a, à l'entrée, /x(n)/, nous avons :

$$y(n) = x(n) - b_1\, y(n-1) - b_2\, y(n-2) \qquad (V.17)$$

Soit, $y'(n)$ la sortie réelle et $\delta_1(n)$, $\delta_2(n)$ les erreurs de troncature après les multiplications suivantes :

$$b_1\ y(n-1) \quad \text{et} \quad b_2\ y(n-2)$$

Si nous supposons qu'il n'y a pas d'erreur dans les additions (c'est à dire, qu'il n'y a pas de débordement des registres), nous avons :

$$y'(n) = x(n) - b_1\ y'(n-1) + \delta_1(n) - b_2\ y'(n-2) + \delta_2(n) \qquad (V.18)$$

Soit l'erreur totale $e(n)$ à la $n^{\text{ième}}$ itération, définie comme

$$e(n) = y(n) - y'(n)$$

Donc, (17) - (18), nous donne :

$$e(n) = -b_1\ y(n-1)+b_1\ y'(n-1)-b_2(n-2)+ b_2\ y'(n-2)$$

$$-\delta_1(n) - \delta_2(n)$$

Soit, $e(n-i)$ et $e(n-2)$ l'erreur totale à la $(n-1)^{\text{ième}}$ et $(n-2)^{\text{ième}}$ itérations respectivement définies comme :

$e(n-1) = y(n-1)-y'(n-1)$; $e(n-2) = y(n-2)- y'(n-2)$
alors,

$$e(n) = -b_1\ e(n-1)- b_2\ e(n-2)- \delta_1(n)- \delta_2(n) \qquad (V.19)$$

En prenant l'espérance mathématique de deux membres de cette équation, nous avons :

$$\overline{e}(n) = -b_1\ \overline{e(n-1)}-b_2\overline{e(n-2)} - \overline{\delta_1(n)}- \overline{\delta_2(n)} \qquad (V.19\ bis)$$

En prenant la transformée en z de V.19 bis, nous avons :

$$\overline{e}(z^{-1}) = \frac{-z\ 2\ \overline{\delta}}{(z-1)(1+b_1 z^{-1} + b_2 z^{-2})}$$

où $\bar{\delta} = q/_2$ pour troncature

$= 0$ pour arrondi

Le théorème de valeur finale nous donne :

$$\bar{e} = \frac{-2\,\bar{\delta}}{1 + b_1 + b_2} \qquad (V.20)$$

Le graphique V.12 montre les courbes théoriques et expérimentales de la valeur moyenne de l'erreur en fonction de b_1 et b_2. Nous pouvons constater un très bon accord entre résultats théoriques et résultats expérimentaux.

Pour la variance de l'erreur, considérons le système suivant schématisé figure V.13.

$/x_n/$ est une séquence aléatoire, stationnaire de valeur moyenne nulle et de variance σ_x^2. En plus, $/x_n/$ est une séquence de bruit blanc. La sortie y(n) est donnée par :

$$y(n) = -b_1\, y(n-1) - b_2\, y(n-2) + x(n) \qquad (V.21)$$

Multiplions les deux membres de l'équation V.21 par $y(n-k)$ et prenons l'espérance mathématique :

$$\overline{y(n)\,y(n-k)} = b_1\,\overline{y(n-1)\,y(-k)} - b_2\,\overline{y(n-2)\,y(n-k)} + \overline{x(n)\,y(n-k)},\quad k>0$$

que nous pouvons écrire sous la forme :

$$\gamma(k) = -b_1\gamma(k-1) - b_2\,\gamma(k-2), \qquad k>0 \qquad (V.22)$$

avec $\gamma(k)$ représentant auto-covariance (au retard k) de la séquence de sortie $/y(n)/$,
et où $\overline{x(n)\,y(n-k)} = 0$ pour $k>0$
puisque $x(n)$ et $y(n-k)$ ne sont pas corrélés et que $\overline{x(n)} = 0$ (par hypothèse).

Pour k = 0 , nous avons :

$$\overline{x(n)\ y(n)} = \sigma_x^2$$

puisque

$$\overline{x(n)\ y(n-1)} = 0 \quad \text{et} \quad \overline{x(n)\ y(n-2)} = 0$$

L'équation (V.22) nous montre que l'auto-covariance de la sortie d'un système purement récursif obéit à la même équation de récurrence que la sortie /y(n)/.
En divisant (V.22) par $\gamma(0)$, nous avons :

$$\rho(k) = -b_1 \rho(k-1) - b_2 \rho(k-2) \qquad (V.23)$$

où $\rho(k)$ est le coefficient d'auto-corrélation (au retard k) de la sortie /y(n)/.
L'équation (V.23) nous permet d'exprimer les coefficients d'auto-corrélation $\rho(1)$ et $\rho(2)$ en fonction des coefficients b_1 et b_2 du système.
Sachant que :

$$\rho(0) = 1 \quad \text{et} \quad \rho(1) = \rho(-1)$$

nous avons :

$$\rho(1) = -b_1 \rho(0) - b_2 \rho(1)$$

d'où

$$\rho(1) = \frac{-b_1}{1 + b_2} \qquad (V.24)$$

et

$$\rho(2) = \frac{-b_1(-b_1)}{1 + b_2} - b_2 \rho(0)$$

$$= \frac{b_1^2}{1 + b_2} - b_2 \qquad (V.25)$$

En posant k = 0 dans (V.22)nous avons :

$$\gamma(0) = -b_1 \gamma(-1) - b_2 \gamma(-2) + \sigma_x^2$$

$$\gamma(0) = -b_1 \gamma(1) - b_2 \gamma(2) + \sigma_x^2$$

où $\gamma(0) = \sigma_y^2$: la variance de la sortie /y(n)/

en normalisant par rapport à $\gamma(0)$,

$$1 = -b_1 \rho(1) - b_2 \rho(2) + \frac{\sigma_x^2}{\sigma_y^2}$$

en substituant pour $\rho(1)$ et $\rho(2)$ les expressions (V.24) et (V.25) nous avons :

$$(1+b_1)\left[\frac{-b_1}{1+b_2}\right] + b_2\left[\frac{b_1^2}{1+b_2} - b_2\right] = \frac{\sigma_x^2}{\sigma_y^2}$$

ou

$$\frac{(1-b_2)}{(1+b_2)}\left[(1+b_2)^2 - b_1^2\right] = \frac{\sigma_x^2}{\sigma_y^2}$$

d'où

$$\sigma_y^2 = \frac{(1+b_2)\,\sigma_x^2}{(1-b_2)\left[(1+b_2)^2 - b_1^2\right]} \qquad \text{(V.26)}$$

Dans notre cas, où l'on a, à l'entrée, une somme de 2 variables aléatoires, indépendantes, ayant chacune une variance de $\sigma_e^2 = q^2/12$ la variance de l'erreur à la sortie est :

$$\sigma_{se}^2 = \frac{(1+b_2)\;2\,(q^2/12)}{(1-b_2)\left[(1+b_2)^2 - b_1^2\right]} \qquad \text{(V.27)}$$

L'équation (V.27) qui permet le calcul de la variance de l'erreur de sortie est beaucoup plus générale que l'expression (IV.23) de Rader et Gold, car elle est applicable quelque soit la nature des pôles réels ou complexes.

Remarquons aussi que la mathode que nous venons de développer est beaucoup plus simple que celle de Gowdy, J.N. §28§ dont la formulation mathématique est complexe. Montrons comment nous pouvons retrouver dans le cas des pôles complexes conjugués l'expression (IV.23) de Rader et Gold.
Le dénominateur du système est :

$$D(z^{-1}) = 1 + b_1 z^{-1} + b_2 z^{-2}$$

les pôles sont les racines de $1 + b_1 z^{1} + b_2 z^{-2} = 0$

c'est à dire de $z^2 + b_1 z + b_2 = 0$ (V.28)

Dans le plan complexe de z, les pôles complexes conjugués sont :

$r \exp(\pm j\theta)$ ou r est le module, et θ l'angle du pôle normalisé, (période d'échantillonnage = 1 seconde), comme l'indique la fig. V.14.

En identifiant les racines de l'équation(V.28)et les pôles dans la représentation polaire, nous avons les deux relations (I.44) et (I.47) répétées ici :

$$b_1 = - 2r \cos \quad \text{(V.29)}$$

$$b_2 = r^2 \quad \text{(V.30)}$$

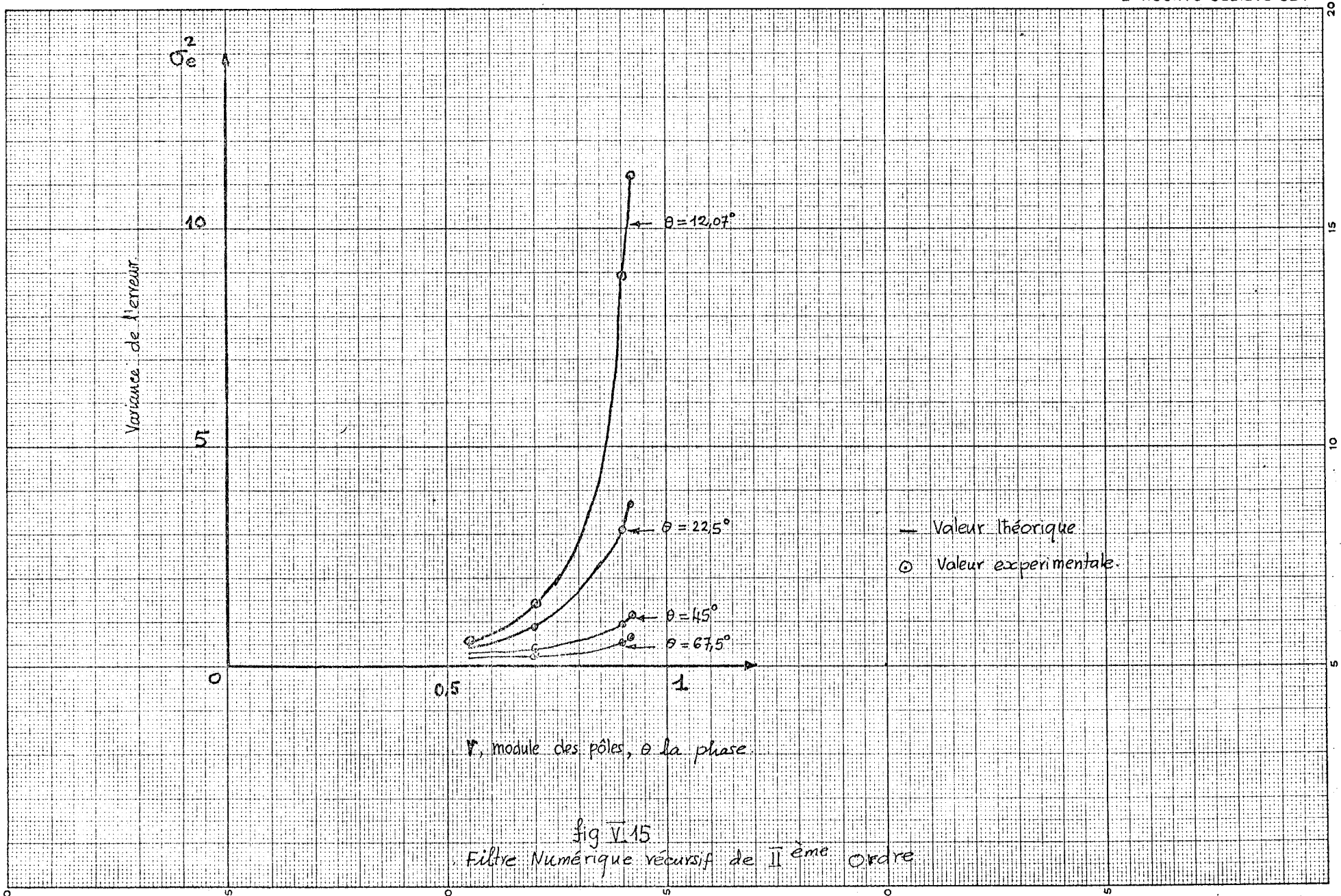


fig V.15

Filtre Numérique récursif de II ème ordre

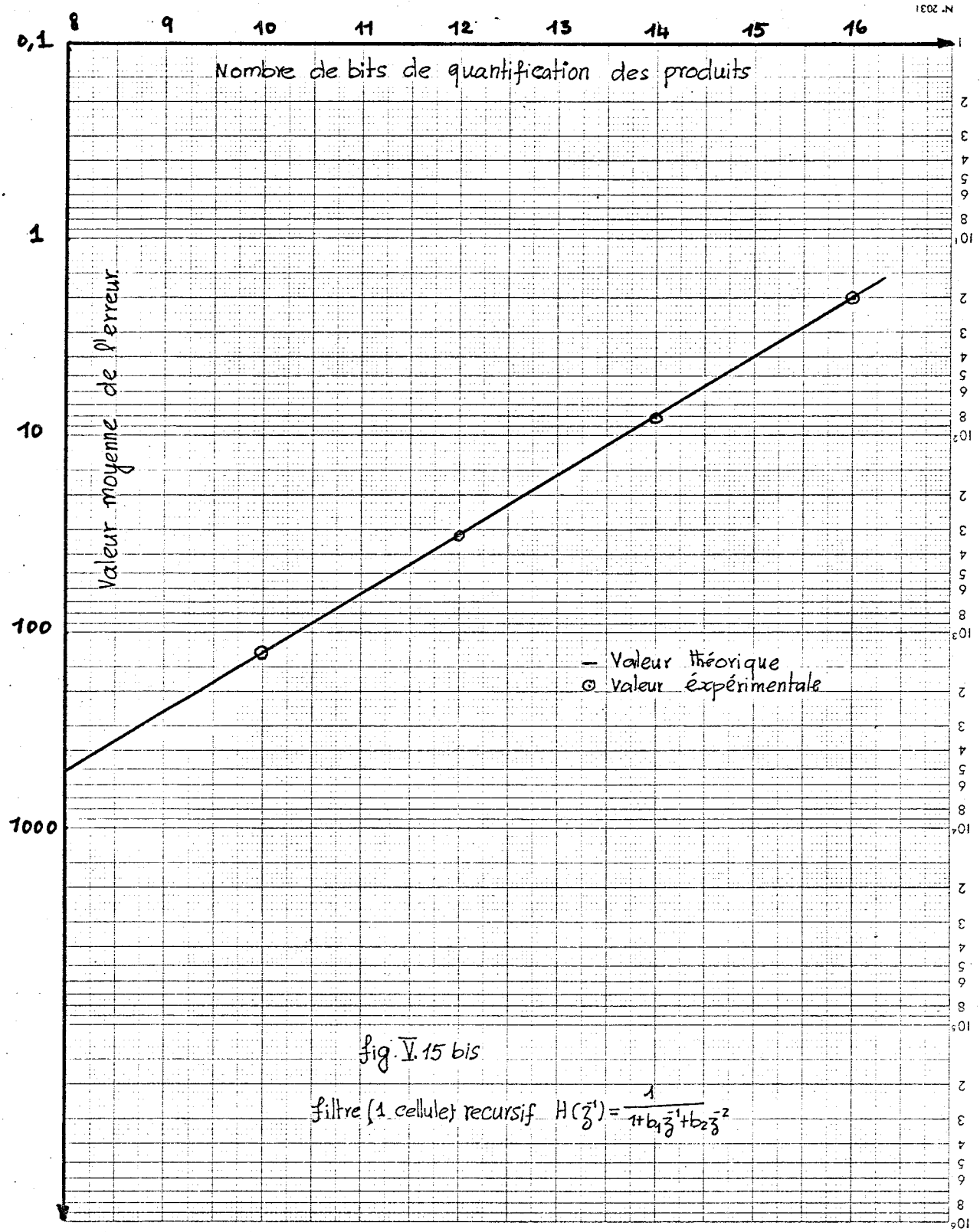


fig. V.15 bis

filtre (1 cellule) recursif $H(z^{-1}) = \frac{1}{1+b_1 z^{-1}+b_2 z^{-2}}$

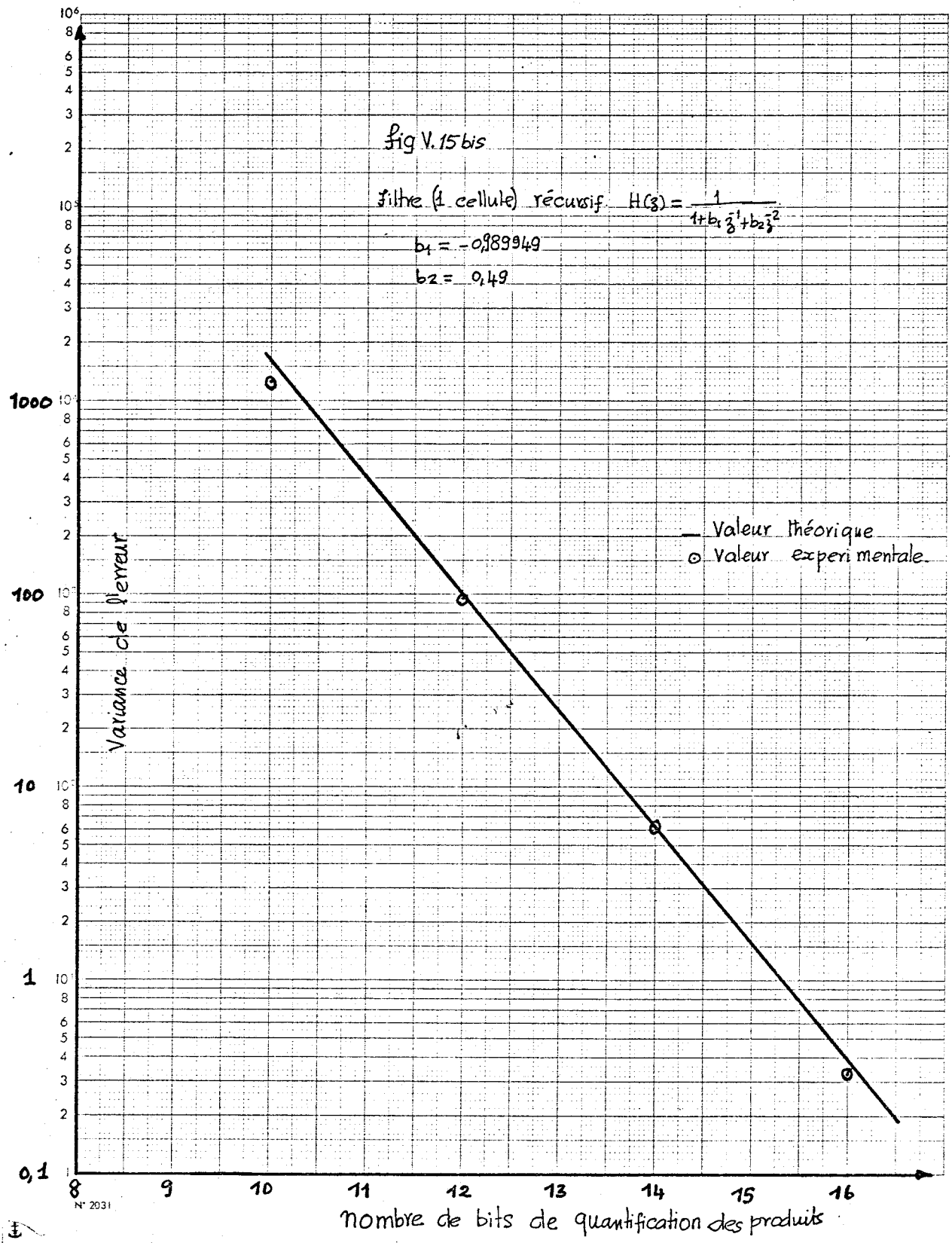


Fig V.15 bis
Filtre (1 cellule) récursif $H(z) = \frac{1}{1+b_1 z^{-1}+b_2 z^{-2}}$
$b_1 = -0,989949$
$b_2 = 0,49$
Valeur théorique
Valeur expérimentale
Variance de l'erreur
1000
100
10
1
0,1
8
9
10
11
12
13
14
15
16
nombre de bits de quantification des produits
N° 2031

En substituant dans la relation (V.27), nous avons :

$$\sigma_{se}^2 = \frac{2\,(q^2/12)\,(1+r^2)}{(1-r^2)\left[(1+r^2)^2 - 4r^2\cos^2\theta\right]}$$

D'autre part nous avons aussi :

$$2\cos^2\theta = \cos 2\theta + 1$$

donc :

$$\sigma_{se}^2 = \frac{2\,(1+r^2)\,(q^2/12)}{(1-r^2)\left[1 + r^4 + 2r^2 - 2r^2(\cos 2\theta + 1)\right]}$$

$$= \frac{(1+r^2)}{(1-r^2)} \quad \frac{2(q^2/12)}{\left[1 + r^4 - 2\,r^2\cos 2\theta\right]} \qquad \text{(V.31)}$$

L'expression (V.31) est bien identique à l'expression (IV.23). Les courbes tracées sur le graphique V.15 (pour les différentes valeurs de b_1 et b_2) montrent la concordance des valeurs théoriques prévues par les équations (V.27) et (V.31) et des valeurs expérimentales de la variance de l'erreur totale.

V.4 Réalisation d'une cellule sous forme directe

Soit la cellule élémentaire

$$H(z^{-1}) = \frac{1 + a_1 z^{-1} + a_2 z^{-2}}{1 + b_1 z^{-1} + b_2 z^{-2}}$$

qui, réalisée en forme directe a pour schéma bloc la fig. (I.8)

En utilisant les résultats que nous avons obtenus dans les cas précédents, nous pouvons facilement évaluer l'expression de la variance de l'erreur à la sortie. En effet, il est évident que les erreurs instantanées $\epsilon_1(n)$, $\epsilon_2(n)$ $\delta_1(n)$, et $\delta_2(n)$ s'ajoutent à l'entrée d'un système de deuxième ordre purement récursif. Donc, le schéma équivalent pour les termes d'erreur, sera celui de la fig. V.13.

Ce qui fait que :

$$e(n) = y(n) - y'(n) = -b_1 e(n-1) - b_2 e(n-2) + \epsilon_1(n) + \epsilon_2(n) - \delta_1(n) - \delta_2(n)$$

ce qui donne, en prenant la moyenne d'ensemble pour chaque n et ensuite la transformée en z :

$$E(z^{-1}) = \frac{z}{(z-1)} \frac{(\bar{\epsilon}_1 + \bar{\epsilon}_2 - \bar{\delta}_1 - \bar{\delta}_2)}{(1 + b_1 z^{-1} + b_2 z^{-2})} \qquad (V.32)$$

De ce fait, la valeur moyenne de l'erreur en régime permanent est obtenue en appliquant le théorème de la valeur finale,

soit :

$$\bar{e} = \lim_{z \to 1} \frac{(z-1)\, z\, (\bar{\epsilon}_1 + \bar{\epsilon}_2 - \bar{\delta}_1 - \bar{\delta}_2)}{(z-1)(1 + b_1 z^{-1} + b_2 z^{-2})}$$

$$\bar{e} = \frac{A_0 - B_0}{1 + b_1 + b_2} \qquad (V.33)$$

où

$$A_0 = \bar{\epsilon}_1 + \bar{\epsilon}_2 = q/2 + q/2$$

$$B_0 = \bar{\delta}_1 + \bar{\delta}_2 = q/2 + q/2$$

Dans le cas où les coefficients a_1, a_2, b_1 et b_2 sont tous différents de zéro ou 1, la valeur moyenne de l'erreur $\bar{e}$ est nulle.

Evaluons la variance de l'erreur à la sortie du système :

selon l'hypothèse statistique faite, les erreurs sont indépendantes. Ce qui fait que la variance de l'erreur totale est la somme des contributions de chaque terme d'erreur à l'entrée. Donc :

$$\sigma_{se}^2 = \frac{4q^2}{12} \frac{(1+b_2)}{(1-b_2)\left[(1+b_2)^2 - b_1^2\right]} \qquad (V.34)$$

ou plus généralement :

$$\sigma_{se}^2 = \frac{(n_a+n_b)\,q^2}{12} \frac{(1+b_2)}{(1-b_2)\left[(1+b_2)^2 - b_1^2\right]} \qquad (V.35)$$

où a : nombre des coefficients $a_i \neq 0$ ou 1
b : nombre des coefficients $b_i \neq 0$ ou 1

V.5 Réalisation d'une cellule sous forme canonique

Soit le même filtre que dans le cas précédent. Nous pouvons le réaliser selon le schéma I.9 qui correspond aux deux équations aux différences simultanées suivantes :

$$w(n) = x(n) - b_1\, w(n-1) - b_2\, w(n-2) \qquad (A)$$

$$y(n) = w(n) + a_1\, w(n-1) + a_2\, w(n-2) \qquad (B)$$

Autrement dit, nous le réalisons de telle façon que les pôles précédent les zéros, contrairement au cas précédent, Soit :

$$H(z^{-1}) = \frac{N(z^{-1})}{D(z^{-1})} = \frac{1 + a_1 z^{-1} + a_2 z^{-2}}{1 + b_1 z^{-1} + b_2 z^{-2}}$$

La réalisation sous forme canonique correspond à la factorisation de $H(z^{-1})$ en deux cellules en cascade et le schéma correspondant est donné dans la fig. I.9

Il est évident, en regardant le schéma, que les erreurs instantanées de troncature, après les multiplications par les b_i, s'ajoutent à l'entrée du système, alors que les erreurs instantanées dûes aux termes a_i s'ajoutent à la sortie du système. Déduisons l'expression de la valeur moyenne de l'erreur à la sortie :
la sortie réelle de la première cellule qui correspond à un système purement récursif est $w'(n)$.

$$w'(n) = x(n) - b_1 w'(n-1) - b_2 w'(n-2) + \delta_1(n) + \delta_2(n)$$

En procédant d'une façon analogue au système de II^{ème} ordre purement récursif, nous avons :

$$e(n) = w(n) - w'(n) = b_1 e(n-1) - b_2 e(n-2) - \delta_1(n) - \delta_2(n) \qquad (V.36)$$

et enfin la sortie réelle est :

$$y'(n) = w'(n) + a_1 w'(n-1) - \epsilon_1(n) + a_2 w'(n-2) - \epsilon_2(n)$$

$$y(n) = w(n) + a_1 w(n-1) + a_2 w(n-2)$$

d'où

$$y(n) - y'(n) = e(n) + a_1 e(n-1) + a_2 e(n-2) + \epsilon_1(n) + \epsilon_2(n)$$

Si l'on définit l'erreur totale à la sortie comme

$$y(n) - y'(n) = e_T(n)$$

nous avons :

$$e_T(n) = e(n) + a_1\, e(n-1) + a_2\, e(n-2) + \epsilon_1(n) + \epsilon_2(n) \qquad (V.37)$$

En prenant la moyenne d'ensemble, puis la transformée en z des équations (V.36) (V.37), nous obtenons :

$$E(z^{-1}) = \frac{-z}{(z-1)} \quad \frac{(\bar{\delta}_1 + \bar{\delta}_2)}{(1 + b_1 z^{-1} + b_2 z^{-2})}$$

$$E_T(z^{-1}) = E(z^{-1})\,(1 + a_1 z^{-1} + a_2 z^{-2}) + \frac{z}{(z-1)}\,(\bar{\epsilon}_1 + \bar{\epsilon}_2)$$

En appliquant le théorème de la valeur finale, nous avons :

$$\bar{e}_T = -B_0 \left[\frac{1 + a_1 + a_2}{1 + b_1 + b_2} \right] + A_0$$

où B_0 et A_0 sont définis comme précédemment.

Cette équation nous confirme bien que les erreurs dûes aux termes b_i voient la fonction de transfert du système alors que les erreurs dûes aux a_i s'ajoutent tout simplement à la sortie.

Evaluons la variance du signal de sortie du filtre :

Il s'agit d'évaluer d'une façon simple la variance de la sortie $/y(n)/$ d'un système linéaire discret ayant pour fonction de transfert en z^{-1},

$$H(z^{-1}) = \frac{1 + a_1 z^{-1} + a_2 z^{-2}}{1 + b_1 z^{-1} + b_2 z^{-2}}$$

excité à l'entrée par une séquence de bruit blanc (valeur moyenne nulle et variance $= \sigma_x^2$) stationnaire.

Soit l'équation aux différences reliant la sortie et l'entrée d'un tel système :

$$y(n) = b_1 y(n-1)+b_2 y(n-2)+a_1 x(n-1)+a_2 x(n-2)+x(n)$$

En multipliant les deux nombres de l'équation par $y(n-k)$ et en prenant l'espérance mathématique de deux membres, nous obtenons :

$$\overline{y(n)y(n-k)} = b_1\overline{y(n-1)y(n-k)} + b_2\overline{y(n-2)y(n-k)} + a_1\overline{x(n-1)y(n-k)} + a_2\overline{x(n-2)y(n-k)} + \overline{x(n)y(n-k)}$$

soit encore :

$$\gamma(k) = b_1\,\gamma(k-1)+b_2\,\gamma(k-2) + a_1\,\gamma_{xy}(k-1) + a_2\,\gamma_{xy}(k-2) + \gamma_{xy}(k)$$

pour $k \leq 2$ (V.38)

où $\gamma(k)$ est l'auto-covariance (au retard k) de la sortie /y(n)/
$\gamma_{xy}(k)$ est la cross-covariance entre la sortie et l'entrée au retard k.

Il faut noter que $\gamma_{xy}(k) = 0$ pour $k > 0$

car $$\gamma_{xy}(k) = \overline{x(n)\ y(n-k)}$$

et l'entrée présente $x(n)$ est complètement indépendante des sorties antérieures.

Nous avons aussi : $\overline{x(n)\ y(n)} = \sigma_x^2$

En faisant $k = 0$ dans (V.38) nous avons :

$$\gamma(0) = b_1\,\gamma(-1)+b_2\,\gamma(-2)+a_1\,\gamma_{xy}(-1)+a_2\,\gamma_{xy}(-2)+\gamma_{xy}(0)$$

qui devient, à cause de la symétrie de la fonction d'auto-covariance

$$\gamma(0) = b_1 \gamma(1)+b_2 \gamma(2)+a_1 \gamma_{xy(-1)}+a_2 \gamma_{xy(-2)}+ \gamma_{xy(0)} \qquad (V.39)$$

Evaluons chaque terme de cette équation afin de déterminer $\gamma(0)$

$$\gamma_{xy(0)} = \overline{y(n)x(n)} = \sigma_x^2 \qquad (V.40)$$

$$\gamma_{xy(-1)} = \overline{y(n)x(n-1)}$$

$$y(n)-b_1y(n-1)-b_2y(n-2) = x(n)+a_1x(n-1)+a_2x(n-2)$$

En multipliant les deux membres de cette équation par x(n-1) et en prenant l'espérance mathématique, nous avons :

$$\overline{y(n)x(n-1)} = b_1\overline{y(n-1)x(n-1)} + b_2\overline{y(n-2)x(n-1)} + \overline{x(n)\ x(n-1)} + a_1\overline{x(n-1)x(n-1)} + a_2\overline{x(n-2)x(n-1)} \qquad (V.41)$$

$$\gamma_{xy(-1)} = b_1 \sigma_x^2 + 0 + 0 + a_1 \sigma_x^2 + 0$$

d'où

$$\gamma_{xy(-1)} = \sigma_x^2 (a_1 + b_1) \qquad (V.42)$$

Calculons $\gamma_{xy(-2)}$. Par définition :

$$\gamma_{xy(-2)} = x(n-2)\ y(n)$$

En multipliant (V.40) par x(n-2) et en prenant l'espérance mathématique nous obtenons :

$$\overline{y(n)x(n-2)} = b_1\overline{y(n-1)x(n-2)} + b_2\overline{y(n-2)x(n-2)} + \overline{x(n)x(n-2)} + a_1\overline{x(n-1)x(n-2)} + a_2\overline{x(n-2)x(n-2)}$$

d'où

$$\gamma_{xy(-2)} = b_1\overline{y(n-1)x(n-2)} + b_2\sigma_x^2 + 0 + 0 + a_2\sigma_x^2$$

$$\overline{y(n-1)x(n-2)} = b_1\overline{y(n-2)x(n-2)} + b_2\overline{y(n-3)x(n-2)} + \overline{x(n-1)x(n-2)} + a_1\overline{x(n-2)x(n-2)} + a_2\overline{x(n-3)x(n-2)}$$

$$= b_1\sigma_x^2 + 0 + 0 + a_1\sigma_x^2 + 0$$

et finalement :

$$\gamma_{xy(-2)} = \sigma_x^2 (a_2 + a_1b_1 + b_1^2 + b_2) \qquad (V.43)$$

L'équation (V.38) nous donne

pour $k = 1$:

$$\gamma_{(1)} = b_1\gamma_{(0)} + b_2\gamma_{(1)} + a_1\gamma_{xy(0)} + a_2\gamma_{xy(-1)} + \gamma_{xy(1)}$$

En substituant les valeurs de $\gamma_{xy(0)}$ et $\gamma_{xy(-1)}$ nous avons :

$$\gamma_{(1)} = b_1\gamma_{(0)} + b_2\gamma_{(1)} + a_1\sigma_x^2 + a_2\sigma_x^2 (a_1+b_1) + \gamma_{xy(1)}$$

$$\gamma_{xy(1)} = \overline{x(n)y(n-1)} = 0$$

$$\gamma_{(1)} = \frac{b_1\gamma_{(0)}}{1-b_2} + \frac{\sigma_x^2}{1-b_2} (a_1 + a_1a_2 + b_1a_2) \qquad (V.44)$$

pour $k = 2$,

$$\gamma_{(2)} = b_1\gamma_{(1)} + b_2\gamma_{(0)} + a_1\gamma_{xy(1)} + a_2\gamma_{xy(0)} + \gamma_{xy(2)}$$

$$\gamma_{(2)} = b_1\gamma_{(1)} + b_2\gamma_{(0)} + a_2\sigma_x^2 ,$$

puisque $\gamma_{xy(1)} = \gamma_{xy(2)} = 0$. (V.45)

En substituant les valeurs de $\gamma_{(1)}$, $\gamma_{(2)}$, $\gamma_{xy(-1)}$ et $\gamma_{xy(-2)}$ données par les équations (44), (45), (40), (42) et (43) respectvement dans (39), nous avons :

$$\gamma(0) = \frac{b_1^2\,\gamma(0)}{1-b_2} + \frac{b_1\,\sigma_x^2}{1-b_2}\,(a_i + a_1a_2 + b_1a_2)$$
$$+ b_2b_1\left[\frac{b_1\,\gamma(0)}{1-b_2} + \frac{\sigma_x^2}{1-b_2}\,(a_1+a_1a_2+b_1a_2)\right]$$
$$+ b_2^2\,\gamma(0) + b_2a_2\,\sigma_x^2 + a_1\,\sigma_x^2\,(a_1+b_1)$$
$$+ a_2(a_2+a_1b_1+b_1^2+b_2)\,\sigma_x^2 + \sigma_x^2$$

$$\gamma(0) - \frac{b_1^2\,\gamma(0)}{1-b_2} - \frac{b_2b_1^2\gamma(0)}{1-b_2} - b_2^2\gamma(0) = \sigma_x^2\left[\frac{b_1}{1-b_2}(a_1+a_1a_2+b_1a_2)\right.$$
$$+\frac{b_2b_1}{1-b_2}(a_1+a_1a_2+b_1a_2)$$
$$+ a_2b_2 + a_1(a_1+b_1)$$
$$\left.+ a_2(a_2+a_1b_1+b_1^2+b_2)+1\right]$$

Le 1er membre s'écrit :

$$\frac{\gamma(0)}{1-b_2}\left[(1+b_2)\left\{(1-b_2)^2 - b_1^2\right\}\right]$$

Le deuxième membre s'écrit :

$$\frac{\sigma_x^2}{1-b_2}\left[(1-b_2)(1+a_1^2+a_2^2) + 2a_2(b_1^2-b_2^2)+2a_1b_1(1+a_2)+2a_2b_2\right]$$

La variance de la sortie est donc :

$$\sigma_y^2 = \gamma(0) = \sigma_x^2\,\frac{\left[(1-b_2)(1+a_1^2+a_2^2)+2a_2(b_1^2-b_2^2)+2a_1b_1(1+a_2)+2a_2b_2\right]}{(1+b_2)\left[(1-b_2)^2-b_1^2\right]}$$

(V.46)

Dans notre cas, $\sigma_x^2\; n_b\; q^2/12$ où n_b = nombre des $b_i \neq 0$ ou 1

Donc, la variance de l'erreur à la sortie est :

$$\sigma_{se}^2 = n_b\, q^2/12 \frac{\text{Numérateur de (46)}}{\text{Dénominateur de (46)}} + n_a \frac{q^2}{12}$$

où n_a = nombre des $a_i \neq 0$ ou 1

En comparant les variances de l'erreur dans les deux cas (forme directe V.34 et forme canonique V.46) nous déduisons la condition pour que σ_{se}^2 de la forme canonique soit inférieure à σ_{se}^2 de la forme directe.

En changeant les signes de b_1 et b_2, pour être en accord avec l'écriture de l'équation de récurrence (A) nous avons :

$$\frac{n_b q^2}{12} \frac{(1+b_2)(1+a_1^2+a_2^2)+2a_2(b_1^2-b_2^2)-2a_1b_1(1+a_2)-2a_2b_2}{(1-b_2)\left[(1+b_2)^2-b_1^2\right]} + \frac{n_a q^2}{12}$$

$$< \frac{(n_a+n_b)q^2}{12} \frac{1+b_2}{(1-b_2)\left[(1+b_2)^2-b_1^2\right]}$$

Posons :

$$N1 = (1+b_2)(1+a_1^2+a_2^2)+2a_2(b_1^2-b_2^2)-2a_1b_1(1+a_2)-2a_2b_2$$

$$N2 = (1+b_2) \quad \text{et} \quad D1 = (1-b_2)\left[(1+b_2)^2-b_1^2\right]$$

Finalement nous avons la condition suivante :

$$n_b\, N1 + n_a \leq (n_a + n_b)\, N2 \tag{V.47}$$

L'équation (47) fournit un critère de choix de la forme de réalisation (soit canonique, soit directe) compte tenu des coefficients.

De même que précédemment, remarquons que les équations (V.46) et (V.47) sont très générales en ce sens qu'elles tiennent compte de toutes les configurations possibles de pôles et de zéro par opposition avec l'expression IV.22 de Rader et Gold qui ne s'applique qu'au cas de pôles complexes conjugués.

Notons également que la méthode développée ici est beaucoup plus simple et générale que celle de Gowdy J.N.§28§. En effet, la dernière méthode utilise une formulation mathématique complexe et ne s'applique que lorsque l'on a le numérateur de $H(z^{-1})$ égal à $a_0 + a_1 z^{-1}$.

De l'équation (V.46) nous pouvons très facilement déduire la variance de la sortie pour d'autres formes de numérateur.

Exemple 1 : Système réalisé en cascade des cellules élémentaires, mais sans normalisation du coefficient a_0 à 1. La cellule type a pour le numérateur et le dénominateur de la F.T. :

$$N_i(z^{-1}) = a_0+a_1 z^{-1}+a_2 z^{-2}$$

$$D_i(z^{-1}) = 1+b_1 z^{-1}+b_2 z^{-2}$$

$$\sigma_y^2 = \frac{\sigma_x^2 \quad (1-b_2)(a_0^2+a_1^2+a_2^2)+ 2a_2(b_1^2-b_2^2)+2a_1b_1(a_0+a_2)+2a_2b_2}{D1} \qquad (V.48)$$

Exemple 2 : Système réalisé en cellules en parallèle. La cellule type a pour F.T. :

$$H_i(z^{-1}) = \frac{N_i(z^{-1})}{D_i(z^{-1})}$$

où $$N_i(z^{-1}) = a_0 + a_1 z^{-1}$$

et $$D_i(z^{-1}) = 1+b_1 z^{-1}+b_2 z^{-2}$$

$$\sigma_{ij}^2 = \sigma_x^2 \frac{(1-b_2)(a_0^2+a_1^2)+2a_1b_1a_0}{D1} \qquad (V.49)$$

Avec ces expressions pour la variance du signal à la sortie, nous pouvons évaluer le rapport signal sur bruit (S.B.) pour différentes réalisations du filtre numérique et comparer les performances pour choisir celle qui convient le mieux, compte tenu des spécifications.

De l'équation(V.49)nous pouvons facilement déduire l'expression donnée par Rader et Gold §24§ dans le cas particulier d'un zéro réel, et d'un pôle complexe conjugué. Ce cas correspond aux coefficients suivants :

$$a_0 = 1,\ a_1 = -r\cos\theta$$

$$b_1 = 2r\cos\theta \quad \text{et} \quad b_2 = -r^2$$

En substituant dans (V.49)

$$\sigma_y^2 = \sigma_x^2 \frac{(1+r^2)(1+r^2\cos^2\theta) - 2r\cos\theta(2r\cos\theta)}{(1-r^2)\left[(1+r^2)^2 - 4r^2\cos^2\theta\right]}$$

$$= \sigma_x^2 \frac{1+r^2\cos^2\theta + r^2 + r^4\cos^2\theta - 4r^2\cos^2\theta}{(1-r^2)\,D_1}$$

où $D_1 = 1+2r^2+r^4-4r^2\cos^2\theta$

le numérateur est égal à : $\sigma_x^2\left[1 + r^2(1-\sin^2\theta) + r^2 + r^4(1-\sin^2\theta) - 4r^2\cos^2\theta\right]$

$$= \sigma_x^2\left[1+2r^2+r^4 - 4r^2\cos^2\theta - r^2\sin^2\theta - r^4\sin^2\theta\right]$$

$$= \sigma_x^2\left[D_1 - r^2\sin^2\theta(1+r^2)\right]$$

d'où

$$\sigma_y^2 = \frac{\sigma_x^2}{(1-r^2)}\left[1 - \frac{r^2\sin^2\theta\,(1+r^2)}{D_1}\right]$$

θ des Pôles	b_1	$\frac{1}{1+b_1+b_2}$	$\frac{1+a_1+a_2}{1+b_1+b_2}$	Forme directe				Forme canonique			
				Valeur moy. Théorique	Valeur moy. expérimentale	Variance Théorique	Variance expérimentale	Valeur Moy. Théorique	Valeur Moy. expérimentale	Variance Théorique	Variance expérimentale
12,07°	−1,793526	21,07	3,09	0	0,0219	22,4	22,698	−2,09	−2,098	0,447	0,458
22,5°	−1,694482	6,826	1,0036	0	0,0211	7,43	7,428	−0,0036	−0,0053	0,337	0,328
45°	−1,296900	1,838	0,269	0	0,00054	2,26	2,223	+0,731	+0,729	0,504	0,5008
67,5°	−0,701877	0,878	0,1675	0	0,00219	1,33	1,307	+0,832	+0,867	0,964	0,99

$$N(z_0^{-1}) = 1 + a_1 z_0^{-1} + a_2 z_{02}^{-2}$$

$a_1 = -1,662981$ (constant)

$a_2 = 0,81$ constant.

$b_2 = 0,8409766$ (constant)

Tableau V.1

et comme D_1 est égal à $1 + r^4 - 2r^2 \cos 2\theta$,

nous avons :

$$\sigma_y^2 = \frac{\sigma_x^2}{(1-r^2)} \left[1 - \frac{r^2 \sin^2\theta \,(1+r^2)}{1+r^4-2r^2\cos 2\theta} \right] \qquad (V.50)$$

Ceci est l'expression donnée par Rader et Gold.

Afin de vérifier expérimentalement ces expressions, nous avons simulé 4 filtres ayant tous les mêmes zéros complexes à 0,9 exp (±j 22.5°). Nous avons fait varier l'angle des pôles tout en gardant le module constant à 0,917, entre 12° et 67,5°. Dans les deux cas (forme directe et forme canonique), les résultats expérimentaux s'accordent bien avec les prévisions théoriques (voir le tableau V.1).

V.6. 1 Analyse des erreurs dûes à la quantification des coefficients :

Comme nous l'avons vu dans le chapître précédent, une autre conséquence des registres de longueur (nombre de bits) limitée, est l'impossibilité de représenter d'une façon exacte les coefficients du filtre. En effet, les procédures de conception des filtres conduisent, en général, à des coefficients nécessitant une précision très grande (voir infinie). La réalisation pratique de ces coefficients impose l'arrondi ou la troncature de ceux-ci, pour tenir compte des longueurs des registres utilisés. Une approche pour réduire l'erreur résultant de cette troncature consiste à considérer des formes de réalisation qui sont le moins sensible aux petits changements des paramètres. La difficulté principale dans l'évaluation de cette sensibilité est la définition d'une mesure convenable de celle-ci.

Par contraste avec le cas de la quantification (l'arrondi) des produits, l'effet de l'arrondi en ce qui concerne les coefficients, n'est pas du ressort de la statistique. Il est important de bien le mettre en évidence. En effet, les changements survenus sur les coefficients entraînent des mouvements des pôles et des zéros du filtre dans le plan complexe.

Prenons l'exemple d'une paire de pôles complexes conjugués. D'après les équations (I.44) à (I.47), le module r et l'angle θ du pôle sont liés aux coefficients b_1 et b_2 du filtre par les relations suivantes :

$$b_2 = r^2$$

$$|b_1| = 2r\cos\theta$$

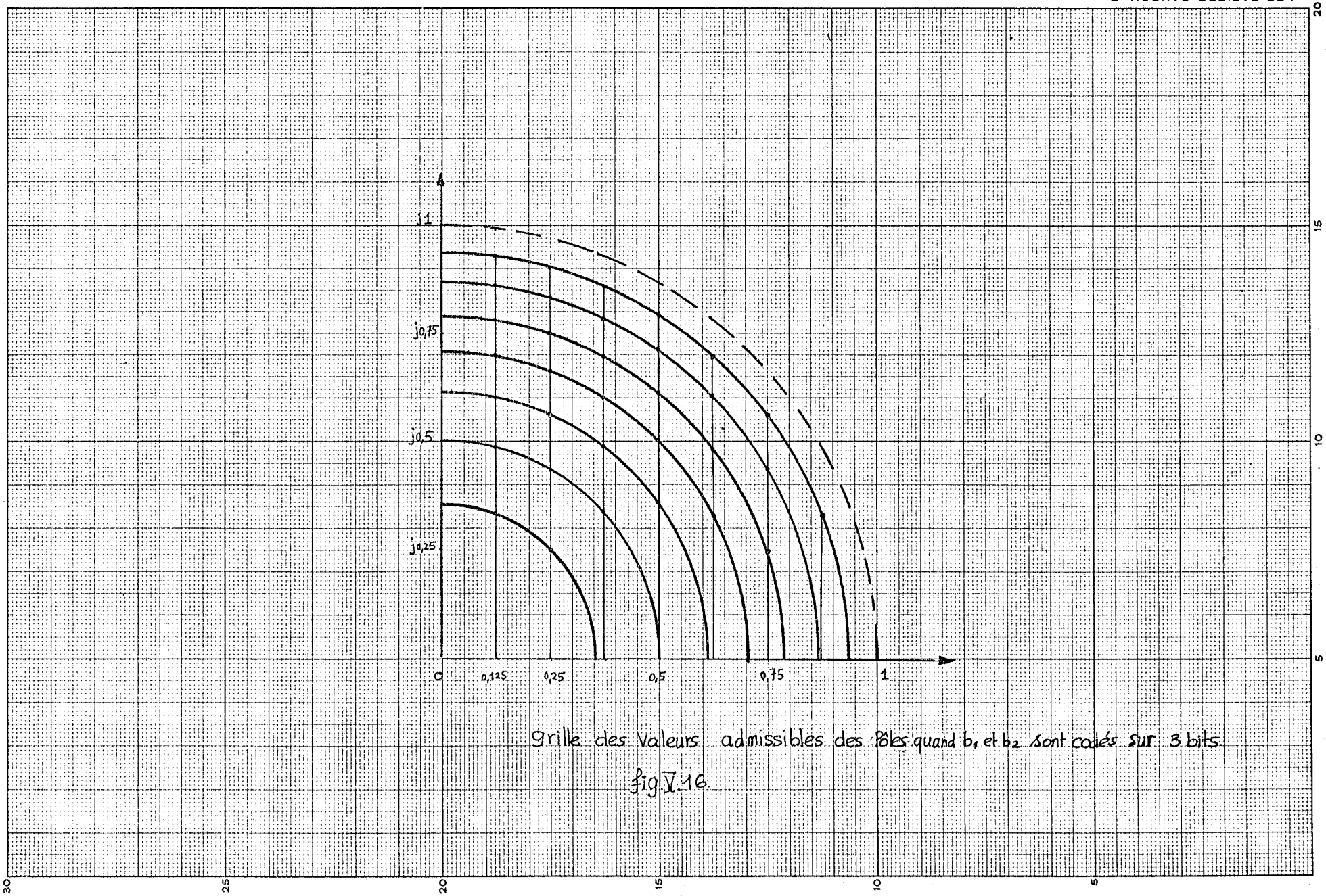
On peut considérer que la quantification des coefficients b_1 et b_2 se traduit pas une quantification du plan z (l'intérieur du cercle unité). La figure (V.16) montre une telle quantification quand b_1 et b_2 sont codés par trois bits. La quantification de b_2 se traduit (à cause de la relation I.44) par des cercles concentriques de module constant et celle de b_1 à des droites verticales. Dans notre cas, le pas de quantification q étant $\frac{1}{8}$, nous avons des cercles concentriques de rayon R tel que :

$$R^2 = \frac{1}{8}, \frac{2}{8}, \ldots \frac{7}{8}$$

(constants), ou bien encore

$$R = \sqrt{\frac{1}{8}}, \frac{1}{2}\sqrt{\frac{3}{8}} \ldots \sqrt{\frac{7}{8}},$$

comme indiqué sur la figure, tandis que b_1 ne pourra prendre que les valeurs $\frac{1}{8}, \frac{2}{8}, \ldots \frac{7}{8}$, (pas de $\frac{1}{8}$). Les pôles ne peuvent donc se placer qu'à l'une des intersections. Remarquons les points importants suivants :



grille des valeurs admissibles des pôles quand b_1 et b_2 sont codés sur 3 bits.

fig V.16.

(i) Pour faire une étude statistique de l'erreur, il faut faire varier aléatoirement les coefficients du filtre idéal (précision infinie) à l'intérieur de la surface élémentaire de quantification (ou la grille) définie par les deux droites et les deux cercles successifs, comme indiqué sur la figure (V.16). Supposons que les b_1 et b_2 sont quantifiés sur n bits et que l'on veuille évaluer la variance de l'erreur dûe à l'arrondi de ceux-ci en comparant sa sortie à celle d'un filtre idéal dont les coefficients sont quantifiés sur n' bits, $n' >> n$.

Pour pouvoir faire une analyse statistique, il faudra perturber les coefficients du filtre idéal en injectant un bruit blanc sur les bits de poids faible et cela plusieurs fois de façon à avoir plusieurs réalisations du "bruit" d'arrondi des coefficients. La signification physique d'une telle évaluation sera la suivante : on examine tous les filtres idéaux dont les coefficients se situent à l'intérieur de la grille de quantification définie auparavant. La mesure de la sensibilité est définie comme l'espérance mathématique de la différence au carré entre le signal de sortie de ces différentes réalisations du filtre idéal et celui du filtre réalisé, ayant ses pôles à une extrémité de la grille.

(ii) Pour les filtres ayant des pôles aux angles petits (filtres passe-bas), la définition de ceux-ci se dégrade beaucoup. Pour θ grand ($\geqslant 45°$), celle-ci s'améliore.

Dans l'analyse que nous avons faite, nous avons traité l'erreur de quantification des paramètres du filtre comme des sources de "bruit". Toutefois, dans la vérification expérimentale, nous avons maintenu les coefficients du filtre idéal (réalisé en virgule flottante) constants, sans perturbation d'une itération à l'autre, pour les raisons suivantes :

(i) l'utilisateur s'interesse, en général, à l'évaluation de l'erreur dûe à la quantification des coefficients d'un filtre réalisé (en virgule fixe, à précision limitée) particulier. Il semble donc utile de pouvoir donner une borne supérieure pour la mesure de sensibilité plutôt qu'une mesure moyenne.

(ii) sans perturbation, on n'observera qu'une seule réalisation du bruit et l'on obtiendra une mesure de l'erreur qui sera une borne supérieure.

Après ces remarques, nous allons énoncer les hypothèses d'un modèle statistique d'étude de cette erreur. Le schéma de la figure (VI.2) montre la réalisation expérimentale pour la vérification de ce modèle.
Pour bien comprendre la signification de l'analyse que nous avons faite, prenons une cellule élémentaire dont la fonction de transfert est donnée par la relation (I.43). Soit :

$$H(z^{-1}) = \frac{1+a_1 z^{-1}+a_2 z^{-2}}{1+b_1 z^{-1}+b_2 z^{-2}} \qquad (V.51)$$

où α_{oi} est pris +1, sans perte de généralité.
Les hypothèses sont les suivantes :

(1) Le filtre réalisé H'(z), dont les coefficients sont tronqués, est toujours stable.

(2) Les autres types d'erreurs de quantification sont : soit rendues négligeables, soit éliminées.
En effet, le même signal d'entrée quantifié, est utilisé par les deux filtres (l'idéal et la réalisation avec quantification des coefficients). Les produits intermédiaires sont tous quantifiés sur n bits, et le niveau du bruit dû à l'arrondi des produits intermédiaires est constant. La variance du signal d'erreur est maintenue de telle sorte qu'il n'y ait pas de débordement dans les additions. Il n'y a donc pas d'erreur dûe au cadrage.

Soit $H'(z^{-1})$, la F.T. du filtre réalisé :

$$H'(z^{-1}) = \frac{1+(a_1+\Delta a_1)z^{-1}+(a_2+\Delta a_2)z^{-2}}{1+(b_1+\Delta b_1)z^{-1}+(b_2+\Delta b_2)z^{-2}} \qquad (V.52)$$

où Δa_1, Δa_2, Δb_1, Δb_2, sont les erreurs de troncature des coefficients exacts a_1, a_2, b_1 et b_2.

La sortie exacte $y(n)$ est donnée par :

$$y(n) = -b_1y(n-1)-b_2y(n-2)+x(n)+a_1x(n-1)+a_2x(n-2) \qquad (V.53)$$

La sortie erronée /y'(n)/ est donnée par :

$$y'(n) = -(b_1+\Delta b_1)y'(n-1)-(b_2+\Delta b_2)y'(n-2)+(a_1+\Delta a_1)x(n-1)$$
$$+(a_2+\Delta a_2)x(n-2)+x(n) \qquad (V.54)$$

Soit $e(n)$ l'erreur à la $n^{ième}$ itération définie comme :

$$e(n) = y(n) - y'(n) \qquad (V.55)$$

D'où, (V.53) — (V.54) nous donne :

$$e(n) = -b_1\, e(n-1)-b_2\, e(n-2)+\Delta b_1y'(n-1)+\Delta b_2y'(n-2) -$$
$$\Delta a_1x(n-1)-\Delta a_2x(n-2) \qquad (V.56)$$

En substituant pour; $y'(n-1) = y(n-1)-e(n-1)$ et

$$y'(n-2) = y(n-2)-e(n-2)$$

nous avons :

$$e(n) = -b_1e(n-1)-b_2e(n-2)+\Delta b_1\left[y(n-1)-e(n-1)\right] +$$
$$\Delta b_2\left[y(n-2)-e(n-2)\right] -\Delta a_1x(n-1)-\Delta a_2x(n-2) \qquad (V.57)$$

En négligeant, dans une première approximation, les produits de deuxième ordre tel que $\Delta b_1 e(n-1)$, etc ...,il ressort :

$$e(n) = -be(n-1)-b_2e(n-2)+\Delta b_1 y(n-1)+\Delta b_2 y(n-2) - \Delta a_1 x(n-1)- \Delta a_2 x(n-2) \qquad (V.58)$$

L'équation (V.58) nous permet de déduire que la valeur moyenne de l'erreur $e(n)$, quand $n \to \infty$ est nulle, si la séquence d'entrée $/x(n)/$ a la valeur moyenne nulle.

En supposant que $/x(n)/$ a pour transformée en z, $X(z)$, la transformée en z des deux membres de (V.58) donne :

$$E(z^{-1}) = -b_1 z^{-1} E(z^{-1})-b_2 z^{-2} E(z^{-1})+ \mathcal{Z}\left[\sum_{k=1}^{2} \Delta b_k\, y(n-k)\right] - \mathcal{Z}\left[\sum_{k=1}^{2} \Delta a_k\, x(n-k)\right] \qquad (V.59)$$

ou :

$$E(z^{-1})\ 1+b_1 z^{-1} +b_2 z^{-2} = \mathcal{Z}\left[\sum_{k=1}^{2} \Delta b_k\, y(n-k)\right] - \mathcal{Z}\left[\sum_{k=1}^{2} \Delta a_k\, x(n-k)\right] \qquad (V.60)$$

C'est ici qu'intervient le choix d'une hypothèse sur la nature des termes Δa_k et Δb_k.

Si nous considérons ces termes comme des variables aléatoires, prenant toutes les valeurs possibles à l'intérieur d'une grille (fig.V.16)nous aboutissons à une expression pour la variance de l'erreur qui ne peut être qu'une valeur moyenne. Avec cette hypothèse, les Δa_k et les Δb_k sont donc des variables aléatoires, uniformément réparties dans les intervalles indiqués ci-dessous et statistiquement indépendantes, c'est à dire que

$$\overline{a_k \cdot a_n} = \delta(k-n) \quad \text{où} \quad \delta : \delta \text{ Kronecker}$$

$$\overline{b_k \cdot b_n} = \delta(k-n) \tag{V.61}$$

et $$\overline{a_k \cdot b_k} = 0 \ \forall \ k.$$

Le cas d'arrondi :

$$|\Delta a_k| \leq \frac{q}{2} \quad \forall \text{ k entier}$$

$$|\Delta b_k| \leq \frac{q}{2} \quad \forall \text{ k entier}$$

$$\overline{\Delta a_k^2} = \frac{q^2}{12}, \quad \overline{\Delta a_k} = 0 \tag{V.62}$$

$$\overline{\Delta b_k^2} = \frac{q^2}{12}, \quad \overline{\Delta b_k} = 0$$

le cas de troncature :

$$\Delta a_k \leq q$$

$$\forall \text{ k entier}$$

$$\Delta b_k \leq q$$

$$\overline{\Delta a_k^2} = \frac{q^2}{3}, \quad \overline{\Delta a_k} = \frac{q}{2} \tag{V.62}$$

$$\overline{\Delta b_k^2} = \frac{q^2}{3}, \quad \overline{\Delta b_k} = \frac{q}{2}$$

où $q = 2^{-n}$, n étant le nombre de bits utilisés pour représenter les coefficients arrondis (ou tronqués) du filtre réalisé en virgule fixe (signe exclu). Sous cette hypothèse, nous n'obtiendrons donc pas la valeur exacte de la variance de l'erreur qui correspond aux $H'(z-1)$ et $H(z^{-1})$ donnés.

Par contre, si nous faisons l'hypothèse que les Δa_k et les Δb_k pour $H'(z^{-1})$ et $H(z^{-1})$ sont gardés constants, nous aboutissons à l'expression pour la variance de l'erreur qui correspond à une seule réalisation des variables aléatoires Δa_k et Δb_k.

Nous allons développer une analyse avec la première hypothèse. De l'expression résultante nous déduirons l'expression relative à la deuxième hypothèse.

En utilisant le théorème de la convolution réelle, nous avons :

$$E(z^{-1})(1+b_1 z^{-1} + b_2 z^{-2}) = \Delta b(z^{-1}) Y(z^{-1}) - \Delta a(z^{-1}) X(z^{-1}) \qquad (V.62)$$

où $\Delta b(z^{-1}) = \sum_{k=1}^{2} \overline{\Delta b}_k z^{-k}$

et $\Delta a(z^{-1}) = \sum_{k=1}^{2} \overline{\Delta a}_k z^{-k}$ (V.62 bis)

En posant $D(z^{-1}) = 1+b_1 z^{-1} + b_2 z^{-2}$ et remplaçant $Y(z^{-1})$ par $H(z^{-1}) X(z^{-1})$, nous avons :

$$E(z^{-1}) = X(z^{-1}) \left[\frac{H(z^{-1}) \Delta b(z^{-1})}{D(z^{-1})} - \frac{\Delta a(z^{-1})}{D(z^{-1})} \right] \qquad (V.63)$$

D'autre part,

$$H(z^{-1}) = \frac{N(z^{-1})}{D(z^{-1})} = \frac{1+a_1(z^{-1})+a_2(z^{-2})}{D(z^{-1})}$$

Définissons la mesure de la sensibilité comme :

$$S = \frac{1}{2\pi j} \int_C \overline{\frac{E(z)\, E(z^{-1}) dz}{X(z)\, X(z^{-1})\, z}} \qquad (V.64)$$

D'après (V.55) nous avons :

$$E(z^{-1}) = Y(z^{-1}) - Y'(z^{-1}) = X(z^{-1}) H(z'(z^{-1}) - X(z^{-1}) H'(z^{-1}) \qquad (V.65)$$

d'où

$$\frac{E(z^{-1})}{X(z^{-1})} = H(z^{-1}) - H'(z^{-1}) \qquad (V.66)$$

alors

$$S = \frac{1}{2\pi j} \int_C \overline{\left[H(z^{-1}) - H'(z^{-1})\right] \left[H(z) - H'(z^{-1})\right] \frac{dz}{z}} \qquad (V.67)$$

Pour $z = \exp(j\omega T)$ où T = la période d'échantillonnage, on a §1§ :

$$S = \frac{1}{\omega_e} \int_0^{\omega_e} \overline{\left| H^*(j\omega) - H^{*\prime}(j\omega) \right|^2} \, d\omega \qquad (V.68)$$

avec ω_e : la fréquence d'échantillonnage en radians.
Cette relation nous montre que le critère d'erreur S (où mesure de sensibilité) est la moyenne quadratique de la différence entre la F.T. du filtre idéal et du filtre réalisé avec des coefficients tronqués (ou arrondis). S est aussi la valeur moyenne quadratique de la différence entre les deux réponses impulsionnelles. Il faut signaler ici que ce critère n'est pas toujours satisfaisant, et que parfois un critère de type "écart maximum" peut être mieux adapté pour certaines catégories de filtres.
D'après (V.62), (V.64) et (V.68), finalement :

$$S = \frac{1}{2\pi j} \int_C \overline{\left[\frac{N(z^{-1})\Delta b(z^{-1})}{D^2(z^{-1})}\right]\left[\frac{N(z)\Delta b(z)}{D^2(z)}\right]} \frac{dz}{z} - \frac{1}{2\pi j} \int_C \overline{\left[\frac{\Delta a(z^{-1})}{D(z^{-1})}\right]\left[\frac{\Delta a(z)}{D(z)}\right]} \frac{dz}{z} \qquad (V.69)$$

Avec la première hypothèse, nous pouvons, en utilisant la relation (V.62 bis), évaluer l'espérance mathématique du produit indiqué dans (V.69) en fonction de Δa_k^2 et de Δb_k^2 §1§ ce qui nous donne :

$$S = \left(\sum \overline{\Delta b_k^2}\right) \frac{1}{2\pi j} \int_C \frac{N(z^{-1})N(z)}{D^2(z^{-1})D^2(z)} \frac{dz}{z} - \left(\sum \overline{\Delta a_k^2}\right) \frac{1}{2\pi j} \int_C \frac{1}{D(z^{-1})D(z)} \frac{dz}{z} \qquad (V.70)$$

La relation (V.70) nous permet d'observer deux faits importants :

(1) Soit $h_1(n)$ la R.I. du système ayant pour la F.T. $N(z^{-1}) / D^2(z^{-1})$ et $h_2(n)$, celle du système ayant pour F.T. $1/D\,(z^{-1})$.
Les deux intégrales de la relation (V.70) représentent donc les deux sommations suivantes §1§ :

$$\sum_{n=0}^{\infty} h_1^2(n) = \frac{1}{2\pi j} \int_C \frac{N(z^{-1})N(z)}{D^2(z^{-1})D^2(z)} \frac{dz}{z}$$

$$\sum_{n=0}^{\infty} h_2^2(n) = \frac{1}{2\pi j} \int_C \frac{1}{D(z^{-1})\, D(z)} \frac{dz}{z} \qquad (V.71)$$

(V.71) nous permet d'évaluer ces deux intégrales, pour un $H(z^{-1})$ donné, par un programme très simple (réalisé en virgule flottante). Il suffit, en effet, de simuler les deux F.T. et d'envoyer à l'entrée de chaque système la séquence (1, 0, 0, 0 ...). En sommant les sorties respectives, on évalue :

$$\sum_{n=0}^{N} h_1^2(n) \text{ et } \sum_{n=0}^{N} h_2^2(n), \text{ ou } N \text{ doit être assez grand}$$

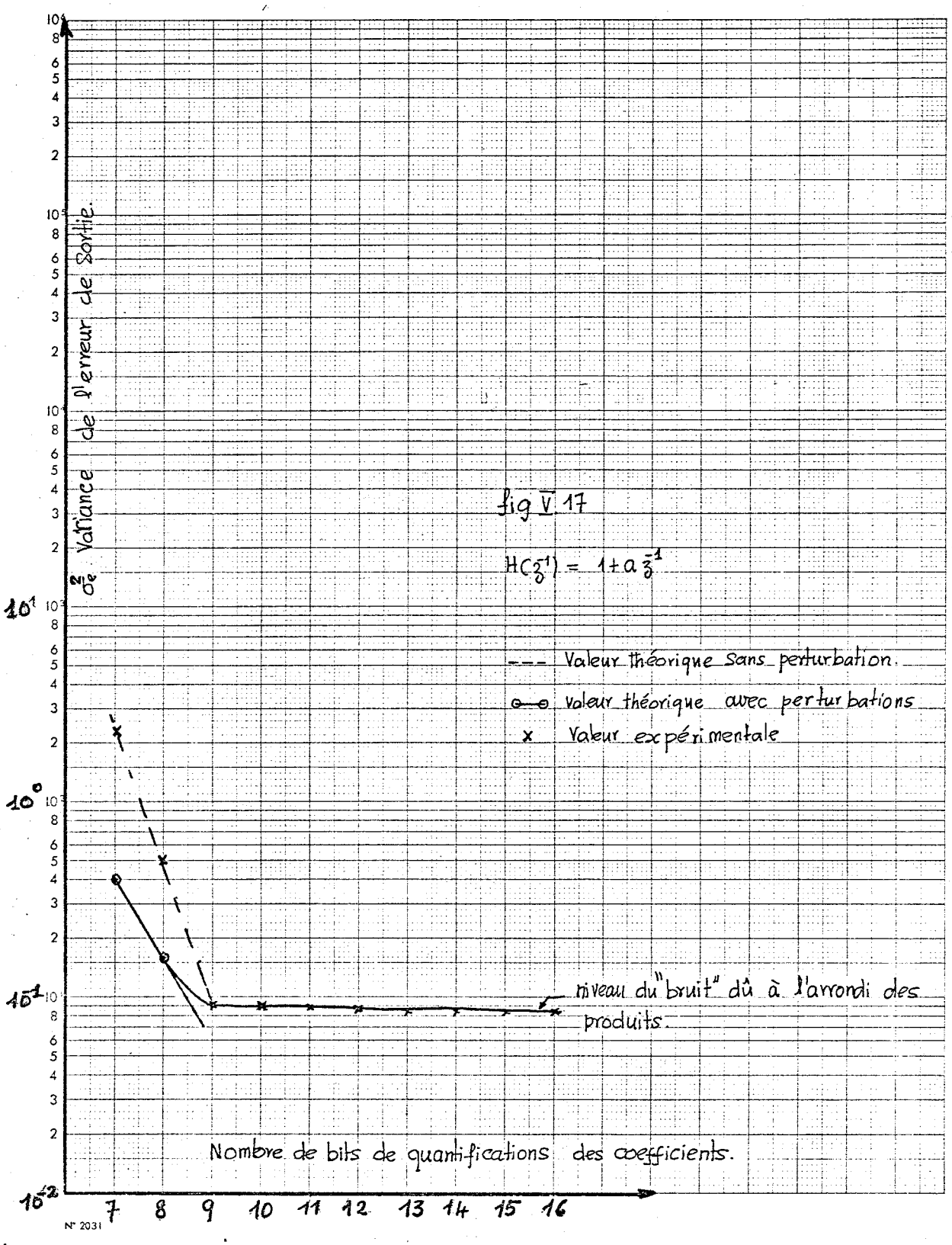
pour que les résultats coïncident avec ceux donnés par (V.71). En pratique, N dépend du module des pôles. Plus celui-ci se rapproche de 1, plus N doit être grand. Pour un filtre ayant des pôles dont le module est 0,917, N = 1024 fournit de bons résultats.

(2) Dans le cas où on perturbe de manière aléatoire et indépendante d'une itération à l'autre les coefficients idéaux de telle sorte qu'ils prennent des valeurs à l'intérieur de la grille des valeurs admises pour les filtres à coefficients tronqués (ou arrondis), on peut évaluer facilement les $\overline{\Delta a_k^2}$ et $\overline{\Delta b_k^2}$ (voir les relations V.62)

La relation (V.70) est donc facile à évaluer.

(3) Quand on apporte pas de perturbation comme dans le cas précédent, les erreurs de troncature (ou arrondis) restent alors constantes, pendant l'expérience. On ne peut donner qu'une borne supérieure pour les quantités $\overline{\Delta a_k^2}$ et $\overline{\Delta b_k^2}$. Cette borne supérieure est égale à :

$$\begin{aligned} &\text{(i) Le cas de l'arrondi :} && \overline{\Delta a_k^2}\ ,\ \overline{\Delta b_k^2} \leq \frac{q^2}{4} \\ &\text{(ii) Le cas de troncature :} && \overline{\Delta a_k^2}\ ,\ \overline{\Delta b_k^2} \leq q^2 \end{aligned} \qquad (V.73)$$


σ_e^2 Variance de l'erreur de sortie.
10^1
10^0
10^{-1}
10^{-2}
fig V.17
$H(z^{-1}) = 1 + a z^{-1}$
Valeur théorique sans perturbation.
Valeur théorique avec perturbations
Valeur expérimentale
niveau du "bruit" dû à l'arrondi des produits.
Nombre de bits de quantifications des coefficients.
7 8 9 10 11 12 13 14 15 16
N° 2031

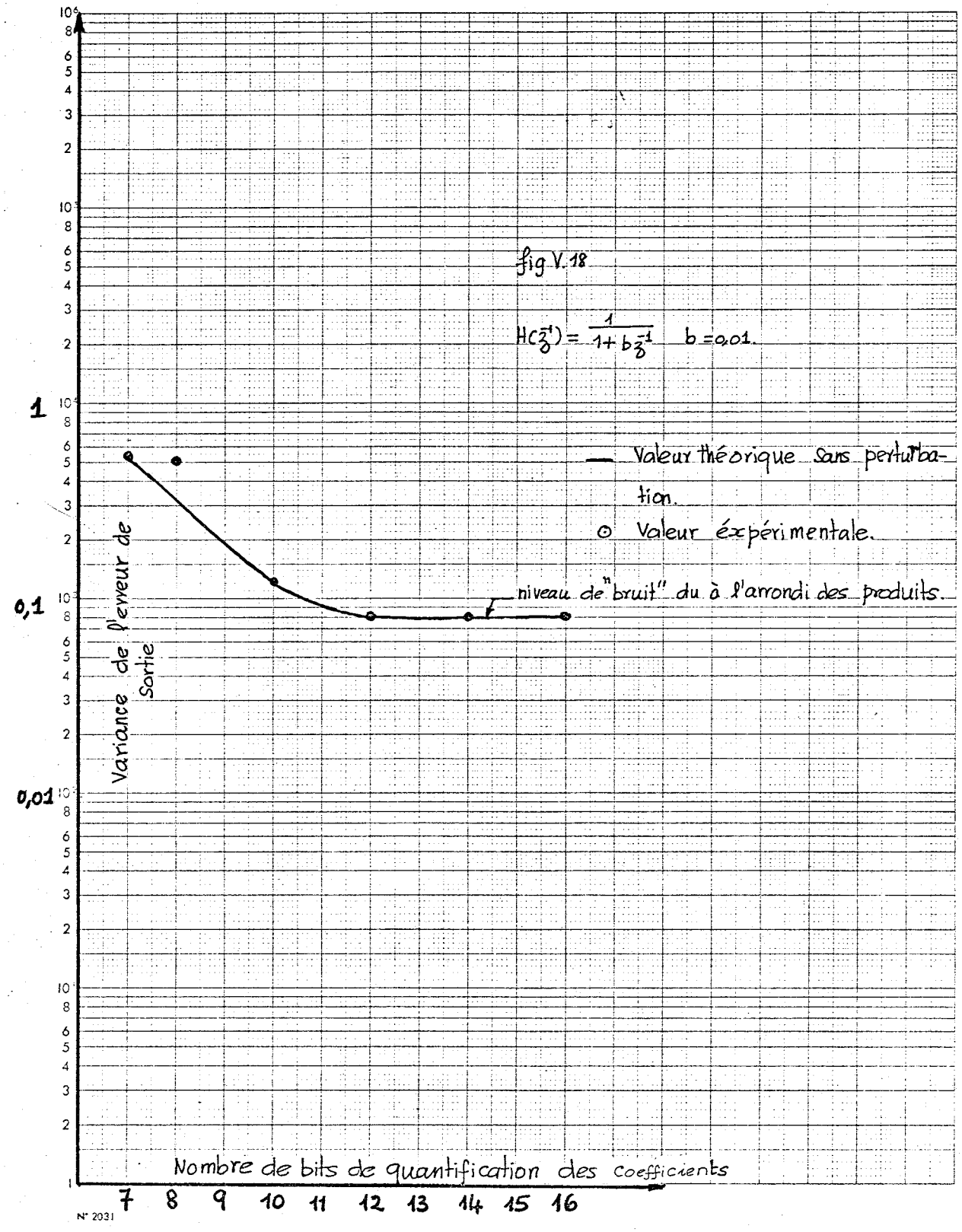


fig V.18
$H(z^{-1}) = \frac{1}{1+bz^{-1}}$ b = 0,01
— Valeur théorique sans perturbation.
⊙ Valeur expérimentale.
niveau de "bruit" du à l'arrondi des produits.
Variance de l'erreur de Sortie
1
0,1
0,01
Nombre de bits de quantification des coefficients
7 8 9 10 11 12 13 14 15 16
N° 2031

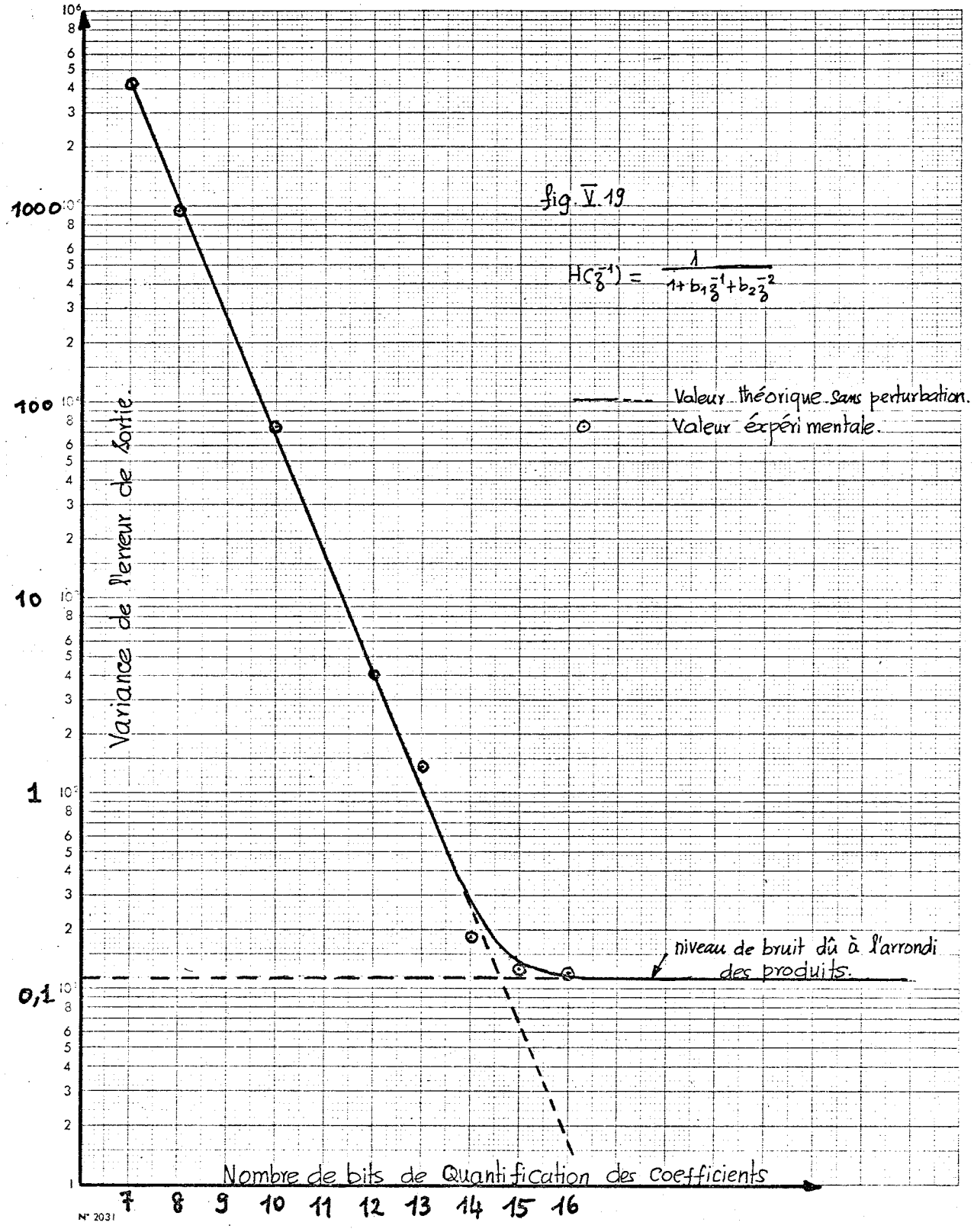


fig. V.19
$H(z_\delta^{-1}) = \frac{1}{1+b_1 z_\delta^{-1}+b_2 z_\delta^{-2}}$
Valeur théorique sans perturbation.
Valeur expérimentale.
Variance de l'erreur de sortie.
1000
100
10
1
0,1
niveau de bruit dû à l'arrondi des produits.
Nombre de bits de Quantification des coefficients
7 8 9 10 11 12 13 14 15 16
N° 2031

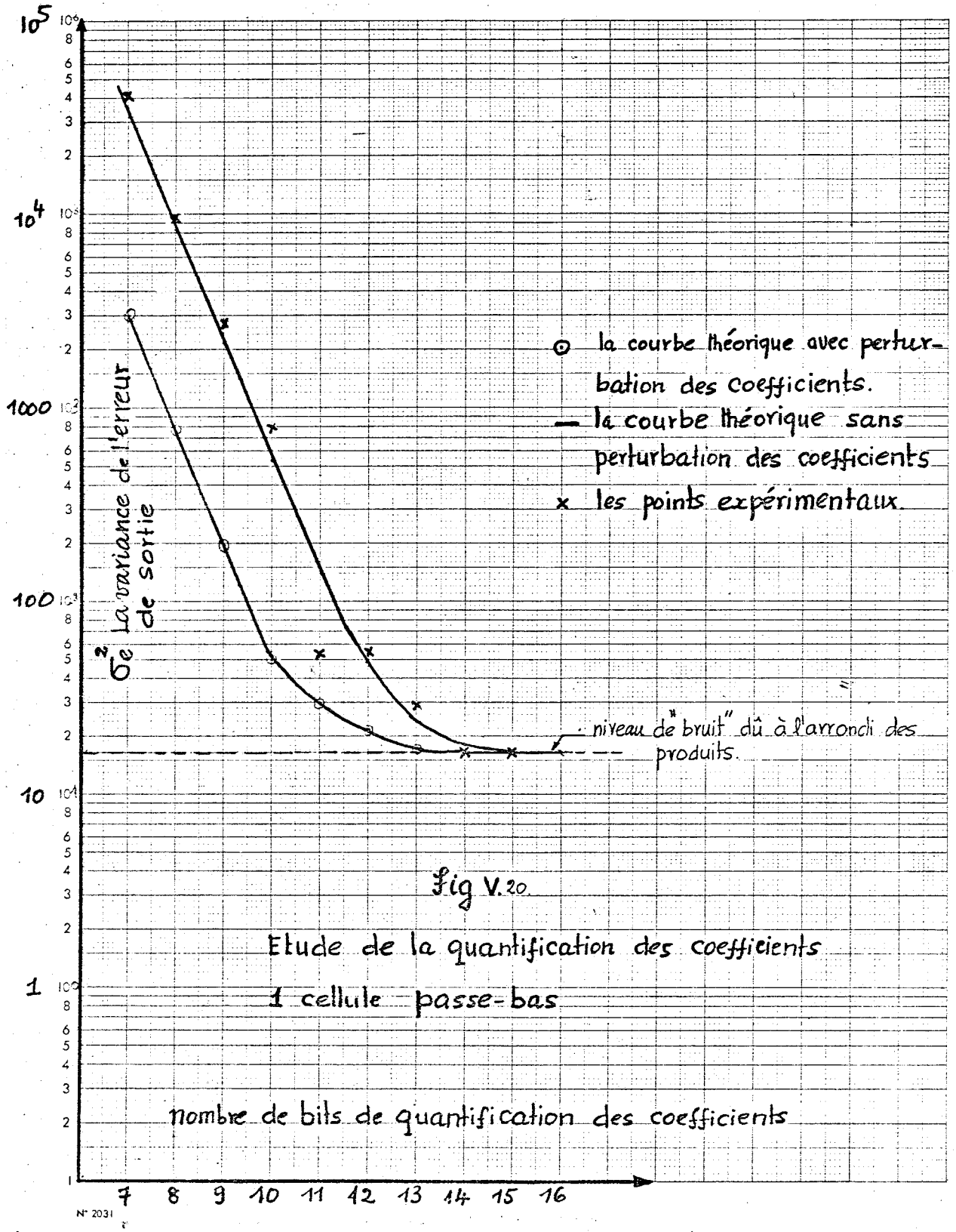


Fig V.20.

Etude de la quantification des coefficients

1 cellule passe-bas

N° 2031

Ainsi, il devient impossible de prédire analytiquement la valeur exacte de S. Cela dépend de la configuration des bits résultant pour les Δa_k et les Δb_k .

(4) On peut modeler le processus de l'erreur dûe aux troncatures (arrondis) des coefficients par deux F.T. parasites disposés en parallèle sur la F.T. idéale. Ce modèle se déduit des relations (V.63) et (V.66).

Les résultats expérimentaux présentés figures (V.17) (V.18), (V.19), (V.20) montrent que les prédictions théoriques s'accordent bien avec ces résultats. Dans le cas d'une cellule pass-bas, on voit que, en pratique, il n'y a aucune différence entre la quantification sur 11 bits et 12 bits (voir la fig.V.20), bien que la prédiction théorique donne des valeurs différentes pour S. Celle-ci est dûe au fait que, la théorie ne tient pas compte des configurations particulières des bits de quantification des coefficients. En effet, le 12ème bit relatif aux coefficients b_1 et b_2 est zéro. Donc, il n'y a aucune différence entre la quantification sur 11 bits et celle sur 12 bits. On montre aussi sur ces figures les courbes qui correspondent à la perturbations des coefficients (modèle statistique).

Pour les filtres d'ordre élevé, réalisés avec des cellules élémentaires en parallèle ou en cascade, l'évaluation théorique de S devient assez compliquée. Seule une simulation expérimentale, qui prend soin d'éliminer l'erreur dûe aux arrondis des produits, peut aider à la détermination de la dégradation de la réponse impulsionnelle.

Pour la réalisation en forme canonique, un calcul analogue à celui qui a été fait dans cette section, montre qu'il n'y a pas de différence entre les deux formes, du point de vue de l'erreur de quantification des coefficients. En effet, les deux équations correspondant à la forme canonique sont :

$$w(n) = x(n) - b_1 w(n-1) - b_2 w(n-2)$$
$$y(n) = w(n) + a_1 w(n-1) + a_2 w(n-2) \qquad (V.74)$$

En suivant une analyse semblable au cas précédent, on a :

$$w'(n) = x(n) - (b_1 + \Delta b_1) w'(n-1) - (b_2 + \Delta b_2) w'(n-2)$$
$$y'(n) = w'(n) + (a_1 + \Delta a_1) w'(n-1) + (a_2 + \Delta a_2) w'(n-2) \qquad (V.75)$$

et :

$$e(n) = w(n) - w'(n) = -b_1 e(n-1) - b_2 e(n-2) + \Delta b_1 w(n-1) + \Delta b_2 w(n-2) \qquad (V.76)$$

$$E(z^{-1}) = \Delta b(z^{-1})\, W(z^{-1}) \Big/ 1 + b_1 z^{-1} + b_2 z^{-2} \qquad (V.77)$$

Dans (V.76), nous avons négligé les produits tels que :
$\Delta b_1 \cdot e(n-1)$, etc.

L'erreur totale : $e_T(n) = y(n) - y'(n)$ (V.78)

qui s'écrit :

$$e_T(n) = e(n) + a_1 e(n-1) + a_2 e(n-2) - \Delta a_1 w(n-1) - \Delta a_2 w(n-2) \qquad (V.79)$$

d'où :

$$E_T(z^{-1}) = E(z^{-1})\,(1 + a_1 z^{-1} + a_2 z^{-2}) - \Delta a(z^{-1})\, W(z^{-1}) \qquad (V.80)$$

D'après (V.77)

$$E_T(z^{-1}) = \frac{\Delta b(z^{-1}) W(z^{-1}) (1 + a_1 z^{-1} + a_2 z^{-2})}{1 + b_1 z^{-1} + b_2 z^{-2}} - \Delta a(z^{-1}) W(z^{-1}) \qquad (V.81)$$

$$E_T(z^{-1}) = W(z^{-1}) \left[\Delta b(z^{-1}) H(z^{-1}) - \Delta a(z^{-1}) \right] \qquad (V.82)$$

D'autre part :

$$W(z^{-1}) = X(z^{-1}) \frac{1}{D(z^{-1})} \tag{V.83}$$

Donc nous avons :

$$E_T(z^{-1}) = X(z^{-1}) \left[\frac{\Delta b(z^{-1}) H(z^{-1})}{D(z^{-1})} - \frac{\Delta a(z^{-1})}{D(z^{-1})} \right] \tag{V.84}$$

(V.84)est identique à (V.63)

V.6. 2 Sensibilité des pôles vis à vis des perturbations des coefficients

Il est possible de faire une analyse théorique afin d'évaluer la précision nécessaire pour représenter les coefficients tenant compte des déplacements admissibles des pôles. Etudions, sur le cas particulier d'une cellule du 2ème ordre, la relation qui existe entre le déplacement de ces pôles et l'erreur sur les coefficients.

Pour une cellule de deuxième ordre, les pôles sont les racines de $D(z^{-1}) = 0$. Donc, on a, dans le cas des pôles complexes conjugués (dans le plan z):

$$z = -\frac{b_1}{2} \pm j \frac{(4b_2 - b_1^2)^{1/2}}{2} = \alpha + j\beta \tag{V.85}$$

Si b_1 est perturbé de Δb_1 et b_2, de Δb_2, les perturbations (ou déplacements) correspondantes des parties réelles et imaginaires des pôles sont facilement évaluables.

On a, pour la partie réelle : $\Delta\alpha = -\frac{\Delta b_1}{2}$ (V.86)

Pour la partie imaginaire, si Δb_1, Δb_2 et $\Delta\beta$ sont petits, on a approximativement :

$$\Delta\beta = \frac{\partial\beta}{\partial b_1}\Delta b_1 + \frac{\partial\beta}{\partial b_2}\Delta b_2 \qquad \text{(V.87)}$$

$$\left.\begin{aligned} \frac{\partial\beta}{\partial b_1} &= -b_1\,\frac{(4b_2-b_1^2)^{-1/2}}{2} \\ \frac{\partial\beta}{\partial b_2} &= (4b_2-b_1^2)^{-1/2} \\ (4b_2-b_1^2)^{-1/2} &= \frac{1}{2\beta} \end{aligned}\right\} \qquad \text{(V.88)}$$

compte-tenu de (V.88), on a finalement :

$$\Delta\beta = \frac{2\Delta b_2 - b_1 \Delta b1}{4\beta} \qquad \text{(V.89)}$$

Les relations (V.89) et (V.86) nous permettent, en imposant les valeurs de $\Delta\alpha$ et $\Delta\beta$ admissibles, d'évaluer les Δb_1, et Δb_2 correspondantes. D'autre part, la relation (V.89) explique la sensibilité plus élevée des pôles pour les perturbations Δb_1 et Δb_2, quand θ est petit. Quand β est voisin de un, θ est grand et $\Delta\beta$ est petit. Cette relation nous permet aussi d'assurer la stabilité absolue du filtre lors d'un choix des longueurs de mots pour représenter b_1 et b_2.

En conclusion, les équations (V.86) et (V.89) nous permettent de choisir le nombre n de bits de quantification des coefficients pour assurer la stabilité de chaque cellule.

V.7 Généralisation, au cas de plusieurs cellules, du modèle statistique du bruit d'arrondi des produits

Les expressions développées dans les sections V.1 à V.5 peuvent être facilement étendues pour le cas de plusieurs cellules, soit en cascade, soit en parallèle. Le schéma bloc de la figure V.21 indique la propagation du bruit d'arrondi dans le cas de la réalisation sous forme directe (cascade). Dans le cas de la forme cascade directe, le bruit généré par la première cellule se propage à travers les zéros et les pôles de la deuxième cellule, etc. Nous pouvons exprimer la variance du bruit à la sortie de la $i^{ème}$ cellule, en fonction de la sortie de la cellule précédente comme suit :

$$\sigma^2_{iD} = \sigma^2_{(i-1)D}\, H_c + (n_a + n_b)\, \frac{q^2}{12}\, H_D \qquad (V.90)$$

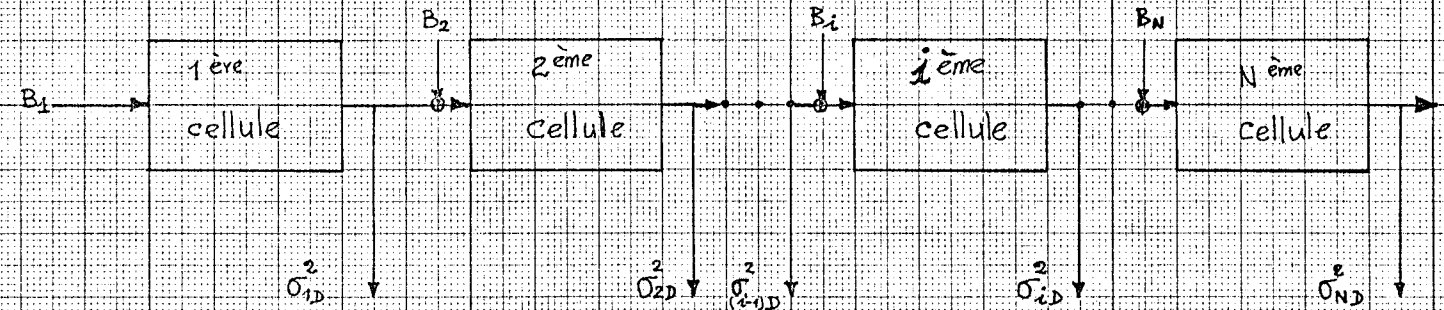
où : n_a = les nombres des a_i différents de zéro ou 1

n_b = les nombres des b_i différents de zéro ou 1

$H_C = \sum_{k=o}^{\infty} h^2(k)$, $/h(k)/$ = réponse impulsionnelle de la ième section

$H_D = \sum_{k=o}^{\infty} h^2_D(k)$, $/h_D(k)/$ = réponse impulsionnelle de la ième section avec $N_i(z^{-1}) = 1$

Les expressions développées dans les sections précédentes, peuvent être utilisées pour évaluer H_C et H_D. Les résultats expérimentaux vérifiant (V.90) sont donnés tableau V.2 dans le cas particulier de 2 cellules en cascade.



$$B_1 = (n_{1a}+n_{1b})\frac{q^2}{12}$$

$$B_2 = (n_{2a}+n_{2b})\frac{q^2}{12}$$

$$\vdots$$

$$B_i = (n_{ia}+n_{ib})\frac{q^2}{12}$$

$$\sigma^2_{1D} = B_1 H_{1D}$$

$$\sigma^2_{2D} = \sigma^2_{1D} H_{2C} + B_2 H_{2D}$$

$$\vdots$$

$$\sigma^2_{iD} = \sigma^2_{(i-1)D} H_{iC} + B_i H_{iD}$$

$$H_{iC} = \sum_{k=0}^{\infty} h^2_{(k)} \quad \text{où } \{h_{(k)}\} = \text{La R.I. de la } i^{\text{ème}} \text{ cellule}$$

$$H_{iD} = \sum_{k=0}^{\infty} h^2_{D(k)} \quad \text{où } \{h_{D(k)}\} = \text{La R.I. de la } i^{\text{ème}} \text{ cellule avec } N(z^{-1}) = 1$$

n_{ia} = Le nombre de coefficients $a_i \neq 0, ou 1$

n_{ib} = Le nombre de coefficients $b_i \neq 0, ou 1$

Fig. V.21

Propagation du bruit d'arrondi à travers N cellules réalisées sous forme directe- cascade.

Valeur Moyenne théorique de l'erreur	Variance théorique de l'erreur	Valeur Moy. de l'erreur (exp.)	Variance de l'erreur (exp.)
−68,898	113,8	−67,676	117,98

Tableau V.2

2 cellules réalisées sous forme circle-cascade

Pour la réalisation en forme canonique, en utilisant les mêmes notations, on a :

$$\sigma^2_{ic} = \left(\sigma^2_{(i-1)c} + n_b \frac{q^2}{12} \right) H_c + n_a \frac{q^2}{12} \qquad (V.91)$$

En ce qui concerne la forme parallèle, l'extension est immédiate, puisque la variance à la sortie du filtre est une somme des variances à la sortie de chaque cellule. Les expressions développées dans les sections précédentes peuvent être appliquées pour le choix de la réalisation et pour l'évaluation des performances de celle-ci.

C H A P I T R E V I

SIMULATIONS ET VERIFICATIONS EXPERIMENTALES

VI.1 Présentation des matériels et des programmes élaborés

Dans ce chapître, nous allons décrire le système et le matériel utilisé ainsi que les programmes de simulations mis au point pour les vérifications des expressions théoriques développées dans les chapîtres précédents, en ce qui concerne les erreurs dans les filtres numériques récursifs.

Les programmes de simulation des filtres ont été écrits en langage machine, et réalisés sur le petit calculateur Multi 8/ M-301 d'Intertechnique. Ce calculateur est doté d'une logique microprogrammée et d'une mémoire permanente. La mémoire vive (à tores) a un cycle de base de 1,1μSec et a une taille maximale de 32 K octets. Il ne s'agit pas, ici, de donner une description détaillée de cette machine ; on veut signaler simplement quelques caractéristiques utiles de celle-ci. En particulier, elle dispose des instructions qui permettent d'avoir une précision variable (la longueur des opérandes définie sur 1 à 4 octets) pour la plupart des opérations logiques et arithmétiques. La multiplication et la division constituent deux exceptions notoires dont les opérandes ont des longueurs fixes à 2 octets.

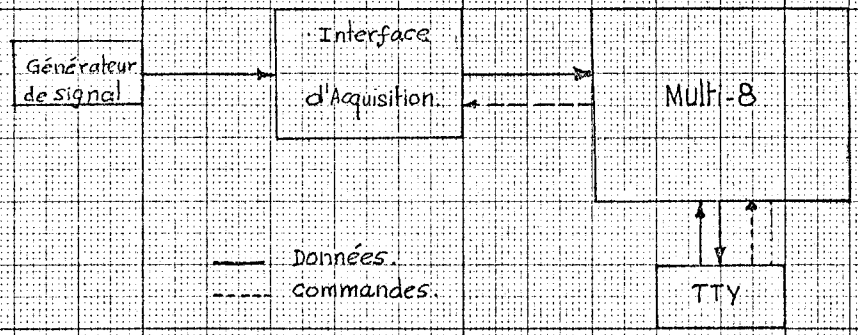
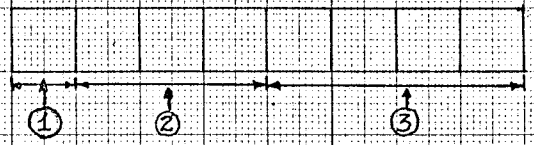


fig VI.1

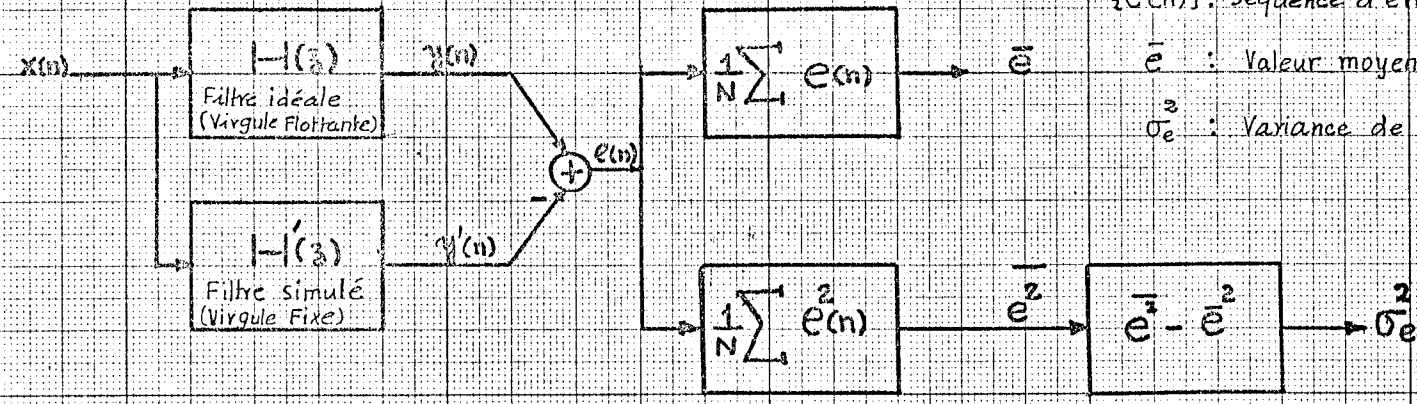
Mot de commande de l'inteface d'acquisition.



① : bit de commande du mode d'acquisition
0: Acquisition sur une seule voie
1: Acquisition cyclique de 8 voies

② : Adresse de voie 8 voies maximum

③ : commade de la fréquence d'échantillonnage
16 fréquences possibles.



{x(n)} : Séquence d'entrée

{y(n)} : Séquence de sortie du filtre idéal

{y'(n)} : Séquence de sortie du filtre simulé

{e(n)} : Séquence d'erreurs

$\bar{e}$: Valeur moyenne de l'erreur

σ_e^2 : Variance de l'erreur (estimation)

fig. VI. 21

Evaluation de Valeur moyenne et de Variance de l'erreur

Nous avons conçu un ensemble câblé d'acquisition de données. Cet ensemble comporte un convertisseur analogique-numérique (C.A.N.), un multiplexeur de 8 voies, et un sous-ensemble de logique et de commande. En particulier, celui-ci nous permet de commander par programme l'acquisition ainsi que toutes les conditions associées à celle-ci, c'est à dire le mode d'acquisition (cyclique ou permanent), l'adresse de voie et la fréquence d'échantillonnage choisie (l'une des 16 fréquences de 0,1 Hz à 10 KHz). Le C.A.N. numérise sur 10 bits (9 bits + bit de signe) le signal d'entrée continu entre $\pm$ 5 V. Ceci est résumé fig.(VI.1).

La méthode expérimentale que nous avons mise en oeuvre pour les vérifications des expressions développées au chapître V est résumée fig.VI.2. La F.T. du filtre prise ici comme référence est réalisée en arithmétique virgule flottante (mantisse de 24 bits et exposant de 8 bits). Les vérifications expérimentales des expressions théoriques développées au chapître V nécessitent une simulation des filtres numériques en arithmétique virgule fixe dans les conditions suivantes :

(i) Possibilité de faire varier le nombre de bits des résultats des produits intermédiaires

(ii) Les mêmes besoins s'appliquent à l'étude de l'erreur de quantification des paramètres du filtre

(iii) Pouvoir faire varier la variance du signal d'entrée afin d'étudier la relation de celle-ci avec les erreurs d'arrondi des produits.

Pour satisfaire ces conditions, nous avons mis au point trois programmes distincts. Le premier nous permet de simuler les filtres réalisés, soit sous forme directe -cascade, soit sous forme canonique -cascade, et d'étudier le processus d'erreur dû

à l'arrondi des produits. Le deuxième concerne la simulation du filtre afin d'étudier le processus d'erreur dû à la quantification des paramètres. Le troisième nous permet de générer un bruit blanc, soit uniformément réparti soit gaussien. Ce programme nous permet de modifier facilement la variance du bruit généré.

Dans tous les cas, un programme moniteur permet de choisir et définir les conditions d'étude. Celle-ci portent sur l'entrée, la sortie et les conditions de calcul. Le signal d'entrée peut être acquis directement à travers le C.A.N. ou simulé par programmation, où par lecture d'un ruban perforé (signal pré-enregistré).

La sortie des résultats peut être effectuée soit sur imprimante soit sur ruban perforé. Pour l'étude statistique des erreurs, le moniteur permet le choix entre l'évaluation de la valeur moyenne et de la variance de l'erreur à la sortie, et de l'histogramme de l'erreur. On peut également imprimer ou faire perforer l'erreur absolue à chaque itération.

Les conditions de calcul portent sur la définition du nombre de points à traiter, du nombre de cellules élémentaires en cascade, ainsi que du mode de quantification des produits (pour le premier programme), soit l'arrondi, soit la troncature. Une commande spéciale permet de rentrer les coefficients pour les filtres réalisés en arithmétique virgule fixe et pour les filtres idéaux réalisés en arithmétique virgule flottante. Dans le cas de la vérification du modèle de l'erreur de quantification des produits, les coefficients pour les filtres simulés en arithmétique fixe sont tous quantifiés sur 16 bits. L'organigramme qui résume le moniteur est donné fig. VI.6.

La sortie du filtre en arithmétique fixe est convertie en format virgule flottant ; les calculs de l'erreur et l'évaluation de l'histogramme se déroulent en arithmétique flottante. Les contraintes de dynamique pour chaque filtre (réalisé en virgule fixe) sont évaluées au préalable pour s'assurer de l'absence de débordement.

Pour la vérification du modèle de l'erreur dûe à la quantification des coefficients, on fixe la quantification des produits intermédiaires à 16 bits (constants). Donc, le niveau du "bruit" dû à celle-ci est connu. Une commande spéciale permet de choisir le nombre de bits pour quantifier les coefficients. Dans ce cas, aussi, la variance du signal d'entrée est réduite de telle sorte que les contraintes de dynamiques soient satisfaites. Dans tous les cas, le même signal d'entrée quantifié est injecté à chaque itération, dans les deux filtres (idéal et celui réalisé en virgule fixe), éliminant ainsi l'erreur de quantification du signal d'entrée.

Le signal d'entrée est obtenu soit à partir d'un générateur de bruit, soit par programmation. Le premier fournit un bruit blanc filtré à partir d'une séquence PRBS (Pseudo raudom binary séquence) générée par un registre à décalage bouclé. On peut choisir la bande passante du bruit, ainsi que sa variance. Le temps de calcul (c'est à dire l'acquisition, l'évaluation de deux sorties filtrées, de l'erreur, de l'erreur au carré ainsi que de l'histogramme de l'erreur) étant de l'ordre de 30 m sec. nous avons donc une période d'échantillonnage de 50 m sec. Nous avons estimé la fonction d'auto-corrélation du signal de sortie quantifié du générateur. La fig. VI.3 montre une estimation basée sur 6 évaluations de la fonction d'auto-corrélation, de celui-ci. On voit que le signal est de type "large bande", mais son spectre n'est pas parfaitement plat dans la bande passante. Par contre, un test statistique (Kolmogorov-smirnoff) a confirmé l'hypothèse que le signal quantifié est gaussien centré avec une probabilité d'avoir raison de 88%.

Nous avons aussi mis au point un programme de génération du bruit blanc, soit uniformément réparti, soit gaussien. Ce programme génère les nombres aléatoires gaussiens en sommant 16 nombres aléatoires uniformément répartis §29§. Ce programme est assez rapide et génère un tel nombre toutes les 7 mSec environ. Les propriétés statistiques sont meilleures que dans le cas précédent. La fonction d'auto-corrélation estimée est donnée sur la fig. VI.4. Elle correspond au cas des nombres définis sur 16 bits. La fig. VI.5 montre

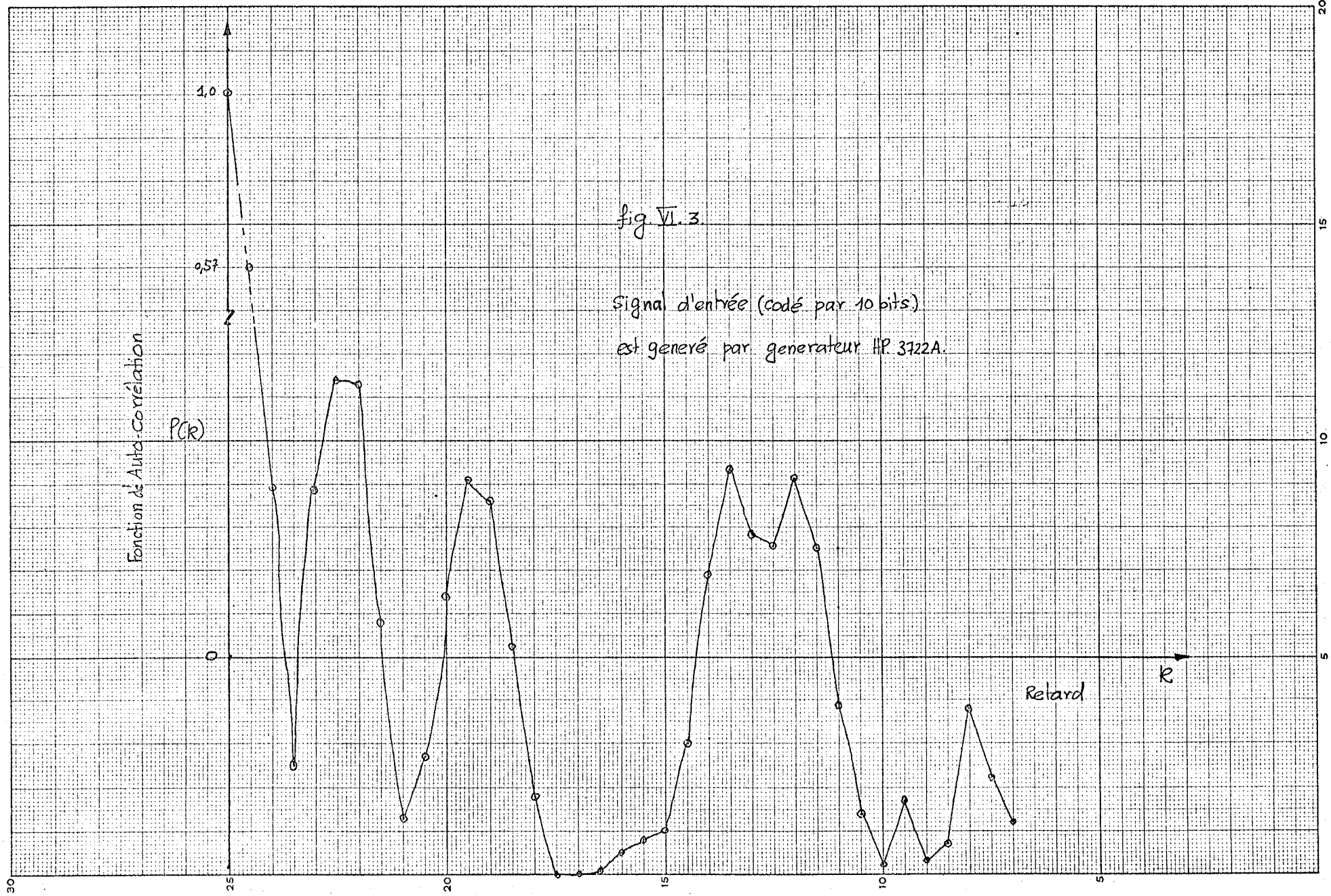


fig. VI.3.

Signal d'entrée (codé par 10 bits)

est generé par generateur HP 3722A.

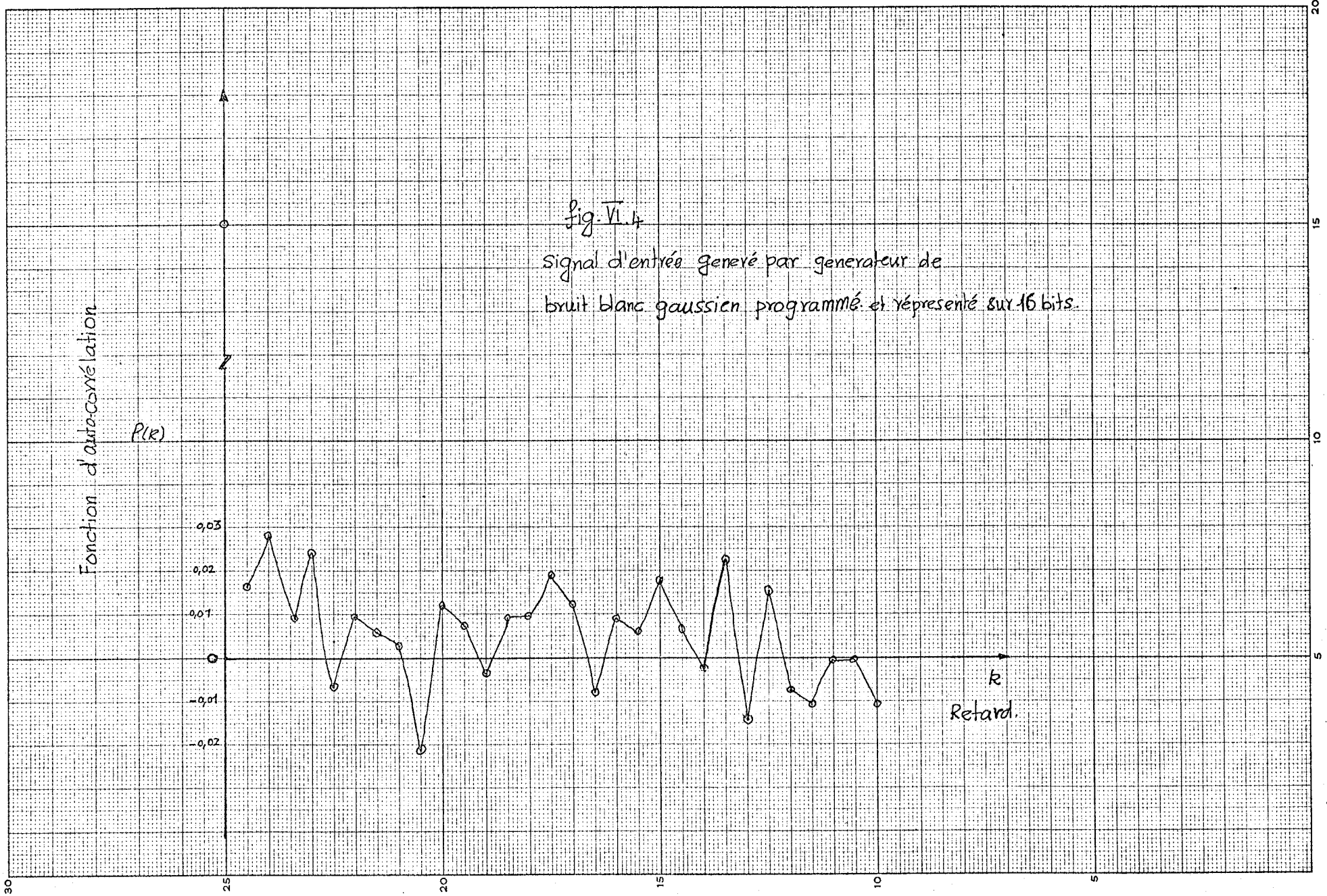


fig. VI.4

Signal d'entrée generé par generateur de bruit blanc gaussien programmé et répresenté sur 16 bits.

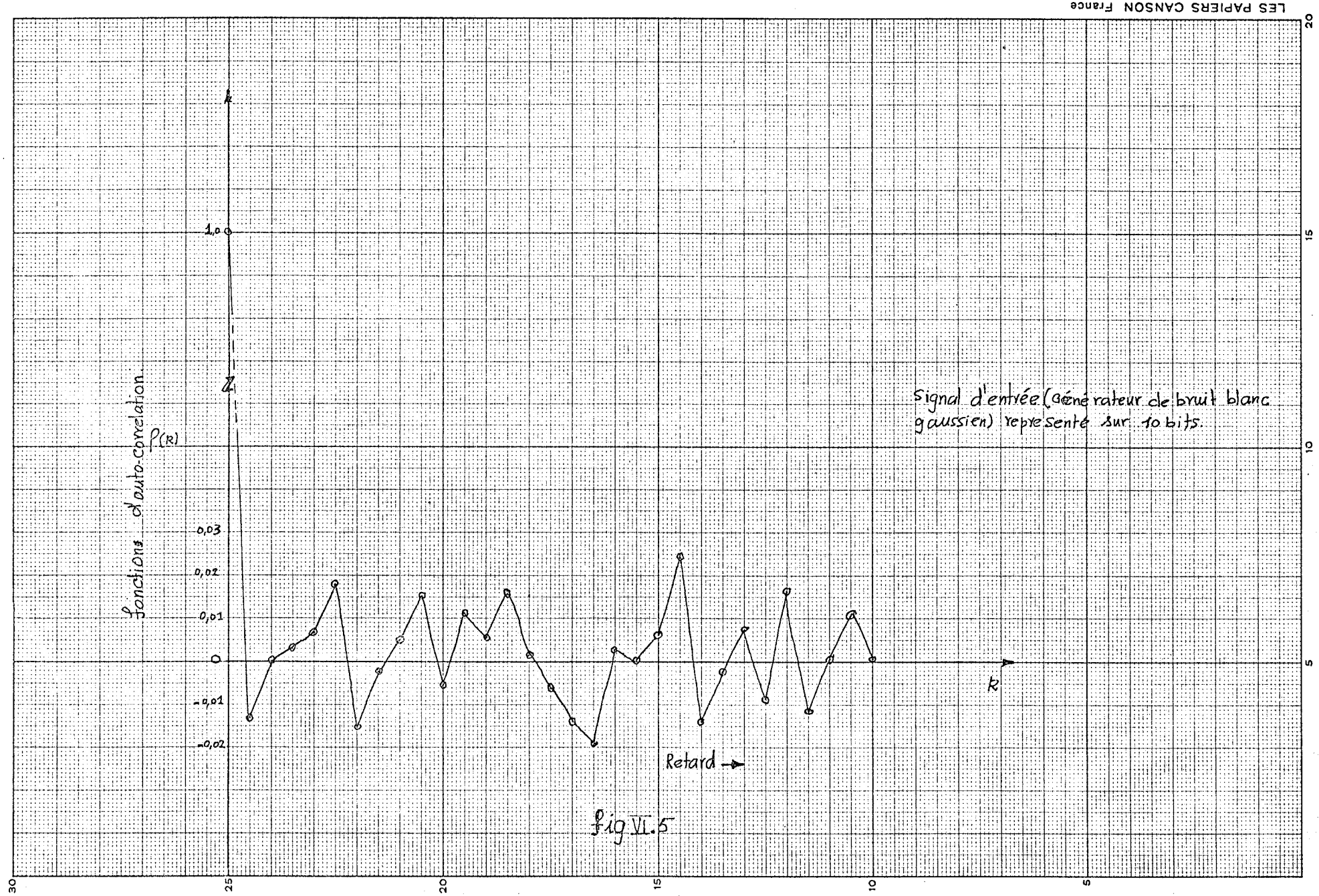


fig VI.5

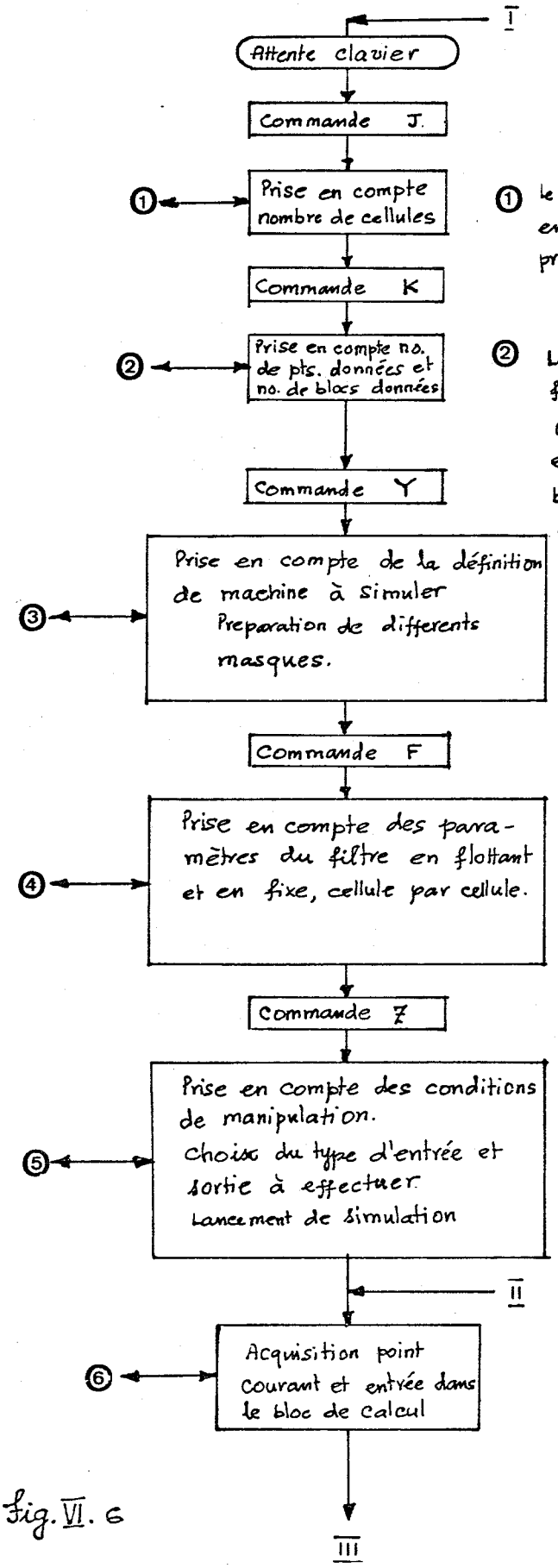


Fig. VI. 6

① Le nombre de cellules élémentaires en cascade. 8 cellules maximum de prévues.

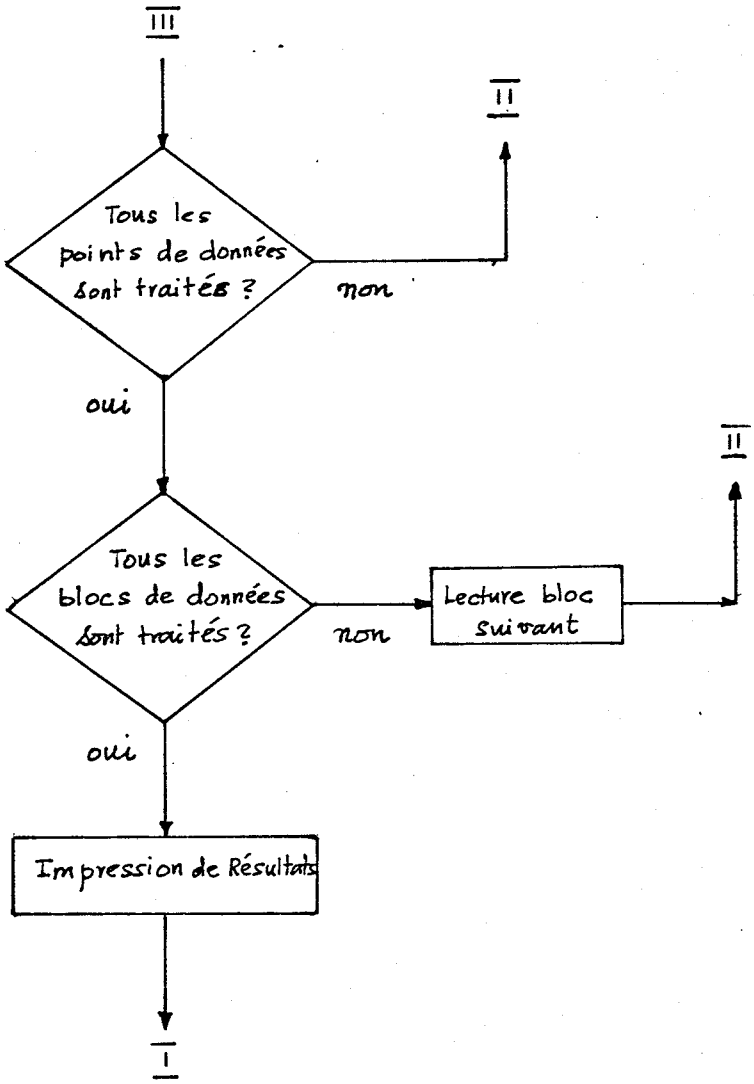
② Le nombre de points d'entrée à filtrer. Ce nombre est défini sur 4 octets en fixe. Si le signal est préenregistré sur ruban, lecture par bloc de 1024 points.

③ Définition de la longuer des mots pour representer les coefficients et les produits partiels. En fonction de celle-ci preparation de differents masques pour le cadrage necessaire à la simulation

④ Frappe de chiffres décimales. conversion decimal-hexadecimal. Normalisation en format flottant etc... Possibilité de correction des erreurs de frappe.

⑤ Le choix de l'arrondi ou troncature des produits. Définition de l'histogramme. Initialisations diverses

⑥ Le bloc de calcul comporte la simulation du filtre réalisé en arithmétique fixe et celui réalisé en arithmetique flottante. Evaluation de l'erreur, de l'erreur quadratique moyenne et elaboration de l'histogramme.


III
II
Tous les points de données sont traités ?
non
oui
II
Tous les blocs de données sont traités ?
Lecture bloc suivant
non
oui
Impression de Résultats
I

la même fonction évaluée pour les nombres définis sur 10 bits. Nous voyons que, dans les deux cas, l'hypothèse de l'indépendance statistique des nombres est mieux vérifiée que dans le cas précédent. Nous avons utilisé ce générateur pour vérifier l'indépendance de la variance de l'erreur dûe à la quantification des produits vis à vis de la variance du signal d'entrée.

Nous avons préféré utiliser ce dernier programme pour la génération du signal d'entrée à cause de sa souplesse d'emploi et de sa rapidité (gain de temps de 10 mSec par point).

VI.2 Propriété des estimateurs statistiques utilisés.

Les estimateurs utilisés pour estimer la valeur moyenne théorique et la variance théorique sont définis comme suit :

Soit $\bar{e}$ l'estimateur de la valeur moyenne de l'erreur μ et S^2, l'estimateur de la variance de l'erreur σ^2 définit comme :

$$\bar{e} = \frac{1}{N} \sum_{i=1}^{N} e_i \qquad \text{(VI.1)}$$

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{N-1} (e_i - \bar{e})^2$$

ou

$$S^2 = \sum_{i=1}^{N-1} \frac{e_i^2}{N-1} - (\bar{e})^2 \qquad \text{(VI.2)}$$

Puisque les caractéristiques statistiques de l'erreur dans le cas des filtres récursifs ne sont qu'asymptotiquement stationnaires, N devrait être très grand pour que les contributions des termes transitoires initiaux soient négligeables. Nous allons développer, ici,

des expressions nous permettant d'évaluer les erreurs d'estimation pour un système de premier ordre.

Avec les notations du chapître V, nous avons, pour les erreurs successives à chaque itération :

$$\begin{aligned} e(0) &= \delta(0) \\ e(1) &= b\delta(0) + \delta(1) \\ e(2) &= b^2\delta(0) + b\,\delta(1) + \delta(2) \\ &\vdots \\ e(n) &= b^n\delta(0) + b^{n-1}\delta(1) + \ldots + b\delta(n-1) + \delta(n) \end{aligned} \qquad \text{(VI.3)}$$

Donc, $\bar{e}$,l'estimateur de la valeur moyenne de l'erreur (pour ces n termes) est :

$$\bar{e} = \sum_{i=1}^{n} e(i)/n \qquad \text{(VI.4)}$$

ou :

$$\bar{e} = \frac{\delta(0)}{n}\sum_{i=0}^{n} b^i + \frac{\delta(1)}{n}\sum_{i=0}^{n-1} b^i + \ldots + \frac{\delta(n)}{n}\sum_{i=0}^{n-n} b^i \qquad \text{(VI.4bis)}$$

Chacune de ces sommations étant une série géométrique, en les sommant, nous avons :

$$n\bar{e} = \delta(0)\frac{1-b^{n+1}}{1-b} + \delta(1)\frac{1-b^n}{1-b} + \ldots + \delta(n)\frac{1-b}{1-b} \qquad \text{(VI.5)}$$

ou :

$$n\bar{e} = \sum_{i=0}^{n}\frac{\delta(i)}{1-b} - \sum_{i=1}^{n+1} b^i\,\delta\,(n+1-i) \qquad \text{(VI.6)}$$

En prenant l'espérance mathématique de deux membres, on a :

$$n\bar{e} = \frac{n\bar{\delta}}{1-b} - \frac{\bar{\delta}(b-b^{n+2})}{1-b} \qquad \text{(VI.7)}$$

où :

$$\bar{\delta} = E\left[\delta(i)\right] = \frac{1}{2}\ 2^{-n}$$

Finalement, on a :

$$\bar{e} = \frac{\bar{\delta}}{1-b} - \frac{\bar{\delta}}{n}\ \frac{(b-b^{n+2})}{1-b} \qquad \text{(VI.8)}$$

ou encore :

$$\bar{e} = \mu - \epsilon \qquad \text{(VI.9)}$$

avec :

$$\epsilon = \frac{\bar{\delta}\,(b-b^{n+2}}{n\ (1-b)}$$

On voit que la précision de l'estimateur $\bar{e}$ dépend de n, nombre de termes d'erreur utilisés, ainsi que du coefficient b. Puisque $|b|<1$, pour les filtres du Ier ordre, on aura $\bar{e} \simeq \mu$ quand n est grand. En tout cas, l'estimateur $\bar{e}$ est sans biais (asymptotiquement). A titre d'exemple, pour b = 0,9 nous avons une précision de 1% avec n = 1000 termes. Quand b s'approche de 1, il faudra beaucoup plus de termes pour la même précision.

En ce qui concerne l'estimateur S^2, en suivant une procédure analogue à celle utilisée par Davenport et al §30§, on peut évaluer sa variance σ_s^2. Le schéma bloc de la figure (VI.2) montre les opérations effectuées pour évaluer S^2.
Soit σ_s^2 la variance de l'estimateur S^2 et $\langle M \rangle$ sa valeur moyenne, soit σ^2, la vraie variance à estimer. Davenport §30§ donne l'expression de σ_s^2 comme :

$$\sigma_s^2 = \int_{-\infty}^{\infty} R_h(x) \left[R_z(x) - \langle z \rangle^2 \right] dx \qquad \text{(VI.10)}$$

ou $R_h(x)$ est la fonction d'auto-corrélation de la réponse impulsionnelle du moyenneur.

$R_z(x)$ est la fonction d'auto-corrélation de la sortie de l'opérateur "élévateur au carré".

$\langle z \rangle$ est la valeur moyenne de z, la sortie.

L'espérance mathématique de l'erreur relative est donnée par :

$$\frac{\sigma_s}{\langle M \rangle} = \frac{\int_{-\infty}^{\infty} R_h(x) \left[R_z(x) - \langle z \rangle^2 \right] dx}{\langle z \rangle \int_{-\infty}^{\infty} h(x,T)\, dx} \qquad \text{(VI.11)}$$

ou : h(x, T)est la réponse impulsionnelle du moyenneur.

Pour le cas d'un filtre passe-bas de Ier ordre, on a :

$$R_y(k) = R_y(0) b^k \quad \text{et} \quad R_y(0) = \frac{1}{1-b^2}$$

$$= R_y(0) \exp(k \log_e b) \qquad \text{(VI.12)}$$

et :

$$R_z(k) = R^2_y(0) \left[1 + 2\, e^{2k \log_e b} \right] \qquad \text{(VI.13)}$$

Finalement,

$$\frac{R_z(k) - \langle z \rangle^2}{\langle z \rangle^2} = 2 \exp(2k \log_e b)$$

La réponse impulsionnelle du moyenneur dans notre cas est tout simplement égale à 1 entre (0 et N-1). Donc, on a la somme des réponses :

$$\sum_{k=0}^{N-1} 1 = N \qquad \text{(VI.14)}$$

Nombre de termes	coefficient	Variance de l'estimateur $\sigma_s/\langle M\rangle$ %
10000	0,637628	2,596
	0,670320	2,696
	0,704688	2,818
	0,740818	2,977
	0,778800	3,188
	0,818731	3,482
	0,860708	3,927
	0,904837	4,696
	0,951229	6,480
1000	0,637628	8,207
	0,670320	8,519
	0,704688	8,909
	0,740818	9,409
	0,778800	10,074
	0,818731	11,003
	0,860708	12,405
	0,904837	14,821
	0,951229	20,404

Tableau VI.1

et la fonction d'auto-corrélation :

$$R_h(k) = \begin{cases} N - k & 0 \leq |k| < N \\ 0 & |k| \geq N \end{cases} \qquad \text{(VI.15)}$$

Nous pouvons donc évaluer l'erreur relative dans l'estimation comme :

$$\frac{\sigma_s}{\langle M \rangle} = \frac{2}{N} \sum_{k=o}^{N} (N-K)\, e^{-2k \log_e b} \qquad \text{(VI.16)}$$

en posant : $\log_a b = \alpha$ nous avons :

$$\frac{\sigma_s}{\langle M \rangle} = \frac{2}{N} \sum_{k=o}^{N} (N-K) \exp(-2k\alpha) \qquad \text{(VI.17)}$$

Nous voyons donc que l'erreur de l'estimation dépend du coefficient b et du nombre d'échantillons N. Nous avons évalué $\sigma_s / \langle M \rangle$ pour différentes valeurs de et de N. Le tableau (VI.1) montre les résultats obtenus. La valeur du coefficient b correspondant qui représente la bande passante du filtre, est aussi indiquée. On voit que la sommation de 1000 points nous donne une précision relative de l'estimation allant de 20% à 8%, en fonction de α. L'estimation basée sur 10,000 points nous donne une précision comprise entre 6,5 % et 2,6 %.

Les résultats pour les systèmes de IIème ordre sont faciles à évaluer, en suivant la même méthode. Mais, leur vérification pratique par calcul est assez onéreuse en temps machine.

C O N C L U S I O N S

Dans ce travail, nous avons étudié et mis en oeuvre deux techniques très différentes de filtrage numérique. A cet effet, nous avons élaboré deux programmes de conception de filtres, l'un pour les filtres numériques dits récursifs, l'autre pour les filtres numériques dits non-récursifs.

Le programme de génération des coefficients de filtres numériques récursifs a un encombrement mémoire de 50 K octets. Ce programme peut donc être exploité dans des petits calculateurs dotés d'une taille mémoire de l'ordre de 64 K octets. Ainsi l'utilisateur peut disposer d'un outil très souple pour la conception de filtres numériques récursifs quelconques ayant des gabarits complexes. La convergence de l'algorithme n'a posé aucun problème dans les cas que nous avons examinés. Pour obtenir une approche satisfaisante des gabarits que nous avons choisis, trois cellules paraissent suffisantes en général. Le temps de calcul pour concevoir un tel filtre (pass-bas) est de l'ordre de 40 secondes en IBM 360/91.

Des études complémentaires futures devraient permettre de généraliser ce programme en tenant compte en particulier des spécifications sur la réponse en phase et de critères d'erreur plus sévères (II.3) que celui utilisé.

Le programme d'optimisation que nous avons conçu et utilisé dans le cas du filtrage utilisant l'algorithme de F F T donne des résultats satisfaisants. La réduction d'amplitude des lobes secondaires est très importante pour le cas de 2 coefficients de transition, soit environ 60 dB. Mais l'utilisation du programme d'optimisation

qui détermine les valeurs de ces coefficients est onéreux en temps de calcul. Son exploitation sera justifiée, par contre, pour l'élaboration de tables à usage universel.

Nous avons également étudié les problèmes que pose la réalisation pratique des filtres récursifs sur petits calculateurs. Ces problèmes sont liés à la représentation des données et des paramètres par mots de longueur limitée. L'analyse de l'erreur de l'arrondi des produits a été faite en détail, en se donnant un modèle statistique. Ces expressions ont été vérifiées expérimentalement dans le cas des signaux d'entrée à large bande. Ces expressions conviennent donc parfaitement pour évaluer la longueur des mots nécessaire à la représentation de toutes les variables du filtre dans les deux formes d'"implémentation", directe et canonique. Ces expressions sont faciles à évaluer et apportent une aide appréciable dans l'élaboration d'un système de filtrage numérique.

Pour certaines classes de signaux particuliers tels : échelon unité, impulsion de dirac, sinusoïde pure, ces expressions peuvent être mises en défaut. Dans le cas de l'échelon unité ou de l'impulsion de dirac il y a apparition de cycles limites. Dans le cas de la sinusoïde pure les erreurs d'arrondi deviennent fortement correlées.

L'étude de l'erreur dûe à la quantification des coefficients est moins aisée en raison de la difficulté de choix d'un critère d'erreur. Le modèle que nous avons développé peut donner une idée de la sensibilité de la structure du filtre à étudier au sens de l'écart quadratique moyen entre la réponse impulsionnelle du filtre et la réponse impulsionnelle du filtre idéal. Une meilleure approche de ce problème serait d'incorporer les contraintes de la quantification des paramètres du filtre dans le programme de conception lui-même et, ainsi, de rechercher un jeu de paramètres optimum dans un espace de recherche quantifié §22§. Une autre

approche serait la transposition en numérique des structures des filtres analogiques ayant une sensibilité moindre vis à vis des variations de paramètres §20§ §21§.

Le problème de la contrainte dûeaux débordements de registres pour les filtres réalisés en arithmétique fixe, a été abordé. Des expressions qui permettent de borner l'amplitude du signal d'entrée, ont été développées.

Ainsi, avant de procéder à l'"implémentation" des filtres numériques récursifs, il convient donc de résoudre les problèmes suivants :

1 Evaluation de la longueur de mots nécessaire à la représentation des paramètres (coefficients) du filtre à l'aide de la relation (V.70), compte tenu d'une sensibilité imposée. Par suite des difficultés signalées, chapître V, concernant des configurations particulières des bits représentant les coefficients, il peut être utile de confirmer les résultats de l'évaluation par une simulation à l'aide d'un petit calculateur. En effet, il pourrait en résulter, pour la même sensibilité imposée, une réduction supplémentaire du nombre de bits nécessaire à la représentation des paramètres.

2 Le choix de la longueur des mots concernant les paramètres étant fait, il faut procéder à l'évaluation de la valeur moyenne et de la variance du bruit d'arrondi des produits pour les différentes formes de réalisation à l'aide des relations (V.35), (V.46), (V.47). Il convient de choisir la forme qui induit le plus faible bruit, compte tenu du rapport signal sur bruit requis.

3 Evaluation d'un facteur de cadrage (scaling) nécessaire à l'entrée de chaque cellule pour éviter le débordement de registres. Cette évaluation devra être faite pour chaque cellule. On dispose ensuite les cellules par ordre croissant de ce facteur, pour éviter une atténuation trop forte du signal d'entrée. Les expressions (IV.29) et (E.19) permettent cette évaluation.

Il est utile de signaler ici, pour des filtres utilisant des opérateurs arithmétiques et des registres de longueur fixe, qu'une relation existe en réalité entre le bruit dû à l'arrondi d produit et celui dû aux quantifications des coefficients.

Une augmentation de la précision des coefficients se traduit par un bruit d'arrondi plus grand. Parallèlement, une structure conduisant à une diminution de la précision requise pour représenter les coefficients, autorise une augmentation semblable de la longueur des mots représentant les données et ainsi, permet une réduction du bruit d'arrondi des produits.

REFERENCES BIBLIOGRAPHIQUES

§1§ Ragazzini J.R. et Franklin G.F. - Les systèmes asservis échantillonnés - Dunod, Paris 1962

§2§ Tou J.T. Digital and Sampled Data Control Systems, Mc Grawhill 1959

§3§ Jury E.I. - Theory and Application of the Z-Transform Method - John Wiley & Sons, New-York 1964

§4§ Wittlesey J.R.B. - "A Rapid Method for Digital Filtering" - Communicationsof ACM, volume 7, N° 9, September 1964

§5§ Box and Jenkins - Time Series Analysis - Forecasting and Control - Holden-Day 1970

§6§ Kuo F.F. and Kaiser J.F. - System Analysis by Digital Computer - John Wiley, New-York 1966

§7§ Golden R.M. and Kaiser J.F. - "Design of Wideband Sampled-Data Filters" - The Bell System Technical Journal, July 1964

§8§ Rader C.M. and Gold B - "Digital Filter Design Techniques in the Frequency Domain" - Proceeding of the IEEE, vol 55, N° 2 February 1967

§9§ Rader C.M. and Gold B - "Effects of Parameter Quantization on the Poles of a Digital Filter" - Proceedings of the IEEE, May 1967

§10§ Jackson L.B. - "On the Interaction of Roundoff Noise and Dynamic Tange in Digital Filters" - The Bell System Technical Journal, February 1970

§11§ Deczky A.G. - "Synthesis of Recursive Digital Filters using the minimum p Error Criterion" - IEEE Transactions on Audio & Electro Acoustics, Vol AU-20, October 1972

§12§ Fletcher R. and Powell M.J.D. - "A Rapidly Convergent Descent Method for minimization" - Computer Journal, Vol 6, N° 2 1963

§13§ Steiglitz K. - "Computer-Aided Design of Recursive Digital Filters" - IEEE Trans. on Audio & Electro-acoustics, Vol AU 18, June 1970

§14§ Lanczos C. - Linear Differential Operators - D.Van Nostrand and Co.

§15§ Helms H.D. - "Non Recursive Digital Filters : Design Methods for Achieving Specifications on Frequency Response" - IEEE Trans. Audio and Electro-acoustics, Vol AU-16, September 1968

§16§ Gentleman W.M. and Sande G. - " Fast Fourier Transforms for Fun and Profit " -Proceedings, Fall Joint Computer Conférence, 1966

§17§ Rabiner L.R., Gold B. and Mc Gonegal C.A. - "An Approach to the Approximation Problem for Non Recursive Digital Filters" - IEEE Trans. Audio & Electro-acoustics Vol. AU-18, June 1970

§18§ Bennett W.R. - "Spectra of Quantized Signals" - Bell System Technical Journal - Vol 27, July 1948

§19§ Widrow B. - "Statistical Analysis of Amplitude-Quantized Sampled-Data Systems" - AIEE Trans. on Application and Industry, Vol. 59, January 1961

§20§ Fetweiss A. - "Pseudopassivity, Sensitivity and Stability of Wave Digital Filters" - IEEE Trans. Circuit Theory, Vol. CT-19 N° 16, November 1972

§21§ Crochiere R.E. - "Digital Ladder Structures and Coefficient Sensitivity" - IEEE Transactions on Audio & Electro Acoustics vol. AU.20, October 1972

§22§ Avenhaus E. - "On the Design of Digital Filters with Coefficients of Limited Word Length" - IEEE, Trans. Audio and Electro-acoustics, Vol. AU.20, August 1972

§23§ Knowles J.B. and Olcayto E.M. - "Coefficient Accuracy and Digital Filter Response" - IEEE, Trans. Circuit Théory, Vol. CT-15, N° 1, March 1968

§24§ Gold B. and Rader C.M. - "Effects of Quantization Noise in Digital Filters"- Proceédings, Spring Joint Computer Conférence, 1966

§25§ Aström K.J., Jury E.I. and Agniel R.G. - "A Numerical Method for the Evaluation of Complex Integrals" - IEEE, Trans. Automatic Control, August 1970

§26§ Oppenheim A.V. - "Realization of Digital Filters using Block-Floating-Point Arithmetic" - IEEE, Trans. Audio and Electro-acoustics, Vol. AU-18, June 1970

§27§ Parker S.R. and Hess S.F. - "Limit-cycle Oscillations in Digital Filters" - IEEE, Trans. Circuit Theory, Vol.CT-18, November 1971

§28§ Gowdy J.N. - "Effects of Chopping and Rounding Errors in Digital Filters" - Thèse, University of Missouri, Columbia, 1971

§29§ Perry J.L., Schafer R.W. and Rabiner L.R. - "A Digital Hardware Realisation of Random Number Generator" - IEEE Transactions on Audio and Electro-acoustics, Vol. 20, October 1972

§30§ Davenport W.B., Johnson R.A. and Middleton D. - "Statistical Errors in Measurements on Random Time Functions" - Journal of Applied Physics, Vol. 23, N° 4, April 1952

§31§ Salisch A. - Rapport Interne, L.E.T.I.

§32§ Parker S.R. and Hess S. - "Canonic Realisations of Second Order Digital Filters Due to Finite Precision Arithmetic" - IEEE, Transactions on Circuit Theory, July 1972

A N N E X E

A N N E X E A

Définition et propriétés de la transformée en "z" §1§

La transformée en z, F(z), d'une fonction échantillonnée f représentée par une séquence de nombre $/f(n)/$ est définie par la série suivante, quand elle converge :

$$F(z) = \mathcal{Z}\,(f(n)) = \sum_{n=0}^{\infty} f(n)\, z^{-n} \qquad (A.1)$$

où z est une variable complexe. La série converge pour tout z tel que $|z| > R$ où R est un nombre positif. La région de convergence est donc l'extérieure d'un cercle de rayon R dans le plan complexe.

Cette transformée est un opérateur linéaire :

$$\mathcal{Z}\left[c_1 f_1(t) + c_2 f_2(t)\right] = c_1 \mathcal{Z}\left[f_1(t)\right] + c_2 \mathcal{Z}\left[f_2(t)\right] \qquad (A.2)$$

La transformée en z d'une séquence retardée est :

$$\mathcal{Z}\left[f(n-k)\right] = z^{-k}\, \mathcal{Z}\left[f(n)\right] \qquad (A.3)$$

Le produit de deux transformées H(z) et G(z) peut être représenté comme une convolution discrète dans le domaine temporel :

$$F(z) = H(z)\, G(z) \qquad (A.4)$$

$$f(n) = \sum_{j=0}^{\infty} h(j)\, g(n-j) \qquad (A.5)$$

où $F(z) = \mathcal{Z}\left[f(n)\right]$, $G(z) = \mathcal{Z}\left[g(n)\right]$, et $H(z) = \mathcal{Z}\left[h(n)\right]$

Théorème de la valeur finale :

$$\lim_{n\to\infty} f(n) = \lim_{z\to 1} (z-1)\,F(z) \tag{A.6}$$

Si la transformée F(z) d'une séquence $/f(n)/$ est connue, on peut obtenir la séquence $/f(n)/$ aux instants d'échantillonnage par la relation inverse suivante :

$$f(n) = \frac{1}{2\pi j} \int_C F(z)\, z^{n-1}\, dz \tag{A.7}$$

où C est un contour fermé à l'extérieur duquel et sur lequel F(z) est analytique. L'équation (A.7) s'applique à toute F(z) qui a une région de convergence. En pratique, quand on a affaire aux fonctions rationnelles en z^{-1} ayant des singularités isolées et limitées en nombre, on peut toujours trouver une suite $/f(n)/$ soit par division terme à terme, soit par décomposition de F(z) en fractions simples ayant une transformée inverse connue. Exemple :

$$F(z) = \frac{\sum_{i=0}^{L-1} a(i)\, z^{-i}}{1 + \sum_{i=1}^{L} b(i)\, z^{-i}} = z\, \frac{\sum_{i=0}^{L-1} a(i)\, z^{L-1-i}}{z^{L} + \sum_{i=1}^{L} b(i)\, z^{L-i}} \tag{A.8}$$

Si l'on suppose que les pôles $p(i)$ sont simples, en faisant la décomposition en fractions simples, on a :

$$F(z) = z \sum_{i=1}^{L} \frac{A(i)}{z - p(i)} = \sum_{i=1}^{L} \frac{A(i)}{1 - p(i)\, z^{-1}} \tag{A.9}$$

Chacune de ces L termes a pour transformée inverse $\sum_{n=0}^{\infty} A(i)\, p(i)^n$ nous avons alors :

$$f(n) = \sum_{i=1}^{L} A(i)\, p(i)^n \tag{A.10}$$

Si les $p(i)$ sont complexes, ils apparaissent en paires de pôles complexes conjugués.

A N N E X E B

Algorithme de Fletcher-Powell

Cette méthode a été développée à partir des fonctions quadratiques. Nous allons décrire cette méthode à partir de la forme quadratique à n dimensions et illustrer celle-ci à l'aide de la figure (B .1) dans l'espace à deux dimensions. Soit la fonction à minimaliser :

$$C = C_o + \sum_{i=1}^{n} a(i)x(i) + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} h(ij)x(i)x(j) \qquad (B.1)$$

qui s'écrit matriciellement sous la forme :

$$C = C_o + \hat{a}^T \hat{x} + \frac{1}{2}\hat{x}^T H \hat{x} \qquad (B.2)$$

où $\hat{a}^T$ = le vecteur ligne $(a_1, a_2, \ldots\ldots a_n)$, constant.

$\hat{x}$ = le vecteur colonne des paramètres $(x_1, x_2 \ldots x_n)$

H = la matrice "Hessienne".

Le vecteur gradient est, en prenant les dérivées partielles de (B.2) par rapport à $\hat{x}$:

$$\hat{g} = \hat{a} + H\hat{x} \qquad (B.3)$$

Au minimum $\hat{x}_{min}$ nous avons :

$$0 = \hat{a} + H\hat{x}_{min} \qquad (B.4)$$

Donc, pour le déplacement $(\hat{x} - \hat{x}_{min}) = \hat{s}$ on a :

$$\hat{s} = (\hat{x} - \hat{x}_{min}) = -H^{-1}\hat{g} \qquad (B.5)$$

Fig. B.1

Cette équation peut se déduire également de la relation (II.15) Dans la méthode décrite, la matrice H^{-1} n'est pas évaluée directement. Pour cela, on part d'une matrice G définie positive et symétrique. Cette matrice est améliorée après la $i^{ème}$ itération en utilisant l'information obtenue quand on se déplace dans la direction $\hat{s}^i$ selon (B.5) vers le minimum. On a :

$$\hat{s}^i = -G^i \hat{g}^i, \quad i = 0, 1, \ldots n-1 \qquad (B.6)$$

Initialement, pour i = 0, la matrice G est égale à la matrice unité. La relation (B.6) définit alors la direction opposée au gradient. Dans la fig. B.1, on part de $\hat{x}_0$. La direction $\hat{s}^0$ est représentée.

On effectue alors, dans la direction $\hat{s}^i$ une recherche unidimensionnelle, afin de rendre C minimale le long de la droite $\hat{x}^i + \lambda \hat{s}^i$. La méthode de la recherche d'un minimum local n'est pas essentielle à l'algorithme. On peut utiliser une méthode de recherche directe ou une méthode d'interpolation quadratique ou cubique. Mais, il est essentiel de minimiser localement C pour chaque itération. Celui-ci occupe un temps de calcul non négligeable §12§.

Soit α^i la valeur de λ qui rend C minimale. On définit un accroissement (ou incrément) des éléments du vecteur $\hat{x}$.

$$\hat{\sigma}^i = \alpha^i \hat{s}^i = \hat{x}^{i+1} - \hat{x}^i = \Delta \hat{x}^i \qquad (B.7)$$

ou

$$\hat{x}^{i+1} = \hat{x}^i + \hat{\sigma}^i \qquad (B.7 \text{ bis})$$

Dans la figure B.1, la recherche le long de $\hat{s}0$ nous conduit au point $\hat{x}^1$. Donc, nous aurons

$$\hat{\sigma}^0 = \hat{x}^1 - \hat{x}^0$$

On calcule ensuite C ($\hat{X}^{i+1}$) et $\hat{g}^{i+1}$. A ce stade, on dispose des éléments qui permettent d'améliorer la matrice G^1

de la relation (B.6). L'amélioration consiste à déterminer le vecteur $\hat{\sigma}^{i+1}$ de telle sorte qu'il soit un vecteur propre de G^{i+2} H. Ceci nous assure que, lorsque la procédure converge vers le minimum, G tend vers H^{-1} et les $\hat{\sigma}^{i}$, i = 0, 1 ... n-1 constituent une base complète de l'espace de recherche. Voyons comment se construit le vecteur $\hat{\sigma}^{i+1}$. A cet effet, définissons la différence suivante :

$$\hat{y}^{i} = \hat{g}^{i+1} - \hat{g}^{i} \qquad (B.8)$$

D'après (B.3), nous avons :

$$\hat{y}^{i} = H(\hat{x}^{i+1} - \hat{x}^{i}) = H\,\hat{\sigma}^{i} \qquad (B.9)$$

D'autre part, α^{i} étant évalué de telle sorte qu'il minimise localement C, on a : $\hat{\sigma}^{iT}\ \hat{g}^{i+1} = 0$ (B.10)
C'est à dire que le pas d'accroissement $\hat{\sigma}^{i}$ dans la direction de recherche $\hat{s}^{i}$ est orthogonal au gradient évalué au point $\hat{x}^{i+1}$.

Dans la figure (B..1) le vecteur $\hat{g}^{1}$ est perpendiculaire au vecteur $\hat{\sigma}^{o}$
Maintenant, construisons le vecteur $\hat{\sigma}^{i+1}$ perpendiculaire à $\hat{y}^{i}$, ce qui implique :

$$\hat{y}^{iT}\ \hat{\sigma}^{i+1} = 0 \qquad (B.11)$$

et compte tenu de (B.9), on a :

$$\hat{\sigma}^{iT} H\,\hat{\sigma}^{i+1} = 0 \qquad (B.12)$$

C'est à dire que les vecteurs $\hat{\sigma}^{i}$ et $\hat{\sigma}^{i+1}$ sont conjugués par rapport à la matrice définie positive H.

Pour construire le vecteur $\hat{\sigma}^{i+1}$, il faut déterminer la nouvelle direction de recherche. Cela veut dire, compte tenu de la relation (B.6), qu'il s'agit de déterminer la matrice G^i. Pour ce faire, on modifie la matrice G^0 de départ comme suit :

$$G^{i+1} = G^i + A^i + B^i \tag{B.13}$$

où

$$A^i = \frac{\hat{\sigma}^i \hat{\sigma}^{iT}}{\hat{\sigma}^{iT} \hat{y}^i} \tag{B.14}$$

et

$$B^i = - \frac{G^i \hat{y}^i \hat{y}^{iT} G^i}{\hat{y}^{iT} G^i \hat{y}^i} \tag{B.15}$$

En effet, en multipliant les deux membres de (B.13) par $\hat{y}^i$ nous avons :

$$G^{i+1} \hat{y}^i = \hat{\sigma}^i \tag{B.16}$$

qui s'écrit, compte tenu de (B.9)

$$G^{i+1} H \hat{\sigma}^i = \hat{\sigma}^i \quad , \quad 0 \leq i \leq n-1 \tag{B.17}$$

C'est à dire que le $\hat{\sigma}^i$ est un vecteur propre de G^{i+1} H, de valeur propre unité. A la $n^{ième}$ itération, on aura G^n H = I (I est la matrice unité) et, le minimum est atteint.

Dans la figure (B.1) on obtient la direction $\hat{s}^1$ à partir de $\hat{y}^0$, $\hat{\sigma}^0$, et G^0. A noter que $\hat{s}^1$ est perpendiculaire à $\hat{y}^0$. Une nouvelle recherche et minimisation le long de $\hat{s}^1$ nous conduit au point $\hat{x}^2$, etc ...

Résumons donc l'algorithme :

Soit le point courant $\hat{x}^i$ et le gradient correspondant $\hat{g}^i$ et la matrice correspondante G^i. L'itération procède ainsi :

1) évaluer $\hat{s}^i = -G^i \hat{g}^i$

2) trouver α^i le long de $\hat{x}^i + \lambda \hat{s}^i$ telle que $\mathbf{c}(\hat{x}^i + \alpha^i \hat{s}^i)$ soit un minimum local par rapport à λ ($\alpha^i > 0$)

3) évaluer le pas d'accroissement des paramètres avec :

$$\hat{\sigma}^i = \alpha^i \hat{s}^i$$

4) modifier les paramètres de manière que : $\hat{x}^{i+1} = \hat{x}^i + \hat{\sigma}^i$

5) évaluer $\mathbf{c}(\hat{x}^{i+1})$ et $\hat{g}^{i+1}$
(vérifier que $(\hat{\sigma}^i, \hat{g}^{i+1}) = 0$)

6) définir $\hat{y}^i = \hat{g}^{i+1} - \hat{g}^i$

7) évaluer $G^{i+1} = G^i + A^i + B^i$ (selon (B.14) et (B.15))

8) tester si le minimum est atteint. S'il l'est, on arrête le calcul, sinon, $i = i+1$ et retour en (1).

Le test pour le minimum est un point crucial. Compte tenu de la précision du calcul, on est amené à faire les tests suivants pour s'assurer que le minimum est atteint :

(1) Test portant sur toutes les composantes du gradient. Il faut qu'elles soient inférieures ou égales à ϵ, précision requise.

(2) Tester le pas d'accroissement $\hat{\sigma}^i$ dans la direction $\hat{s}^i$ qui doit être $\leq \epsilon$

(3) Il faut s'assurer que l'on a itéré au moins n fois (n étant le nombre de paramètres à optimiser) avant de s'arrêter.

(4) Il faut tester également la matrice G^i et vérifier qu'elle est toujours définie positive pour chaque i. Ceci nous assure que $\mathbf{C}(\hat{\mathbf{X}})$ est réduit à chaque itération. Ainsi la procédure ne divergera pas. Si, pour un i donné la matrice G^i n'est pas définie positive, on la remplace par la matrice unité et la procédure est poursuivie.

La précision de calcul ainsi que le nombre d'itérations sont des paramètres d'entrée pour le sous-programme.

Il faut noter aussi que l'algorithme n'impose aucune contrainte sur les paramètres $\hat{\mathbf{X}}$. Ainsi, si l'on veut utiliser cet algorithme dans le cas d'un problème avec des contraintes (comme le nôtre) il faut trouver une stratégie qui tienne compte du problème particulier à résoudre.

A N N E X E C

Evaluation du gradient

La fonction de coût $C(\hat{\theta})$ est une fonction de 4 K variables, K étant le nombre de cellules élémentaires en cascade. D'après (II.21) nous avons :

$$\hat{C}(\hat{\theta}) = C(A^*, \hat{\theta}) \tag{C.1}$$

Le gradient de $\hat{C}$ par rapport à $\hat{\theta}$ est défini comme :

$$\frac{\partial \hat{C}}{\partial \hat{\theta}(n)} = \frac{\partial C(A^*, \hat{\theta})}{\partial \hat{\theta}(n)} + \frac{\partial C(A^*, \hat{\theta})}{\partial A^*} \frac{\partial A^*}{\partial \hat{\theta}(n)} \tag{C.2}$$

$$n = 1, 2, \ldots 4K$$

Le terme $\frac{\partial C(A^*, \hat{\theta})}{\partial A^*}$ est nul, puisque A^* rend $C(A, \hat{\theta})$ minimum (voir la relation (II.20)

Nous avons donc, compte tenu de la relation (II.18) :

$$\frac{\partial C(\hat{\theta})}{\partial \hat{\theta}(n)} = 2A^* \sum_{i=1}^{M} \left(A^* \left|H(z(i), \hat{\theta})\right| - R^d_{(i)}\right) \frac{\partial \left|H(z(i), \hat{\theta})\right|}{\partial \hat{\theta}(n)} \tag{C.3}$$

D'autre part :

$$\left|H(z(i), \hat{\theta})\right| = \sqrt{H(z(i), \hat{\theta})\ \overline{H(z(i), \hat{\theta})}} \tag{C.4}$$

avec $\overline{H(z(i), \hat{\theta})}$: complexe conjugué de $H(z(i), \hat{\theta})$.

Nous avons donc :

$$\frac{\partial \left|H(z(i), \hat{\theta})\right|}{\partial \hat{\theta}(n)} = \frac{\left[\overline{H(z(i), \hat{\theta})} \frac{\partial}{\partial \hat{\theta}(n)} H(z(i), \hat{\theta}) + H(z(i), \hat{\theta}) \frac{\partial}{\partial \hat{\theta}(n)} \overline{H(z(i), \hat{\theta})}\right]}{2\sqrt{H(z(i), \hat{\theta})\ \overline{H(z(i), \hat{\theta})}}} \tag{C.5}$$

Le terme entre crochet est de la forme $\bar{A}B + A\bar{B}$
celui-ci se simplifie donc:

$$\overline{H(z(i),\hat{\theta})}\frac{\partial}{\partial\hat{\theta}(n)}H(z(i),\hat{\theta})+H(z(i),\hat{\theta})\frac{\partial}{\partial\hat{\theta}(n)}\overline{H(z(i),\hat{\theta})}=2\,Re\;\overline{H(z(i),\hat{\theta})}\frac{\partial}{\partial\hat{\theta}(n)}H(z(i),\hat{\theta}) \qquad (C.6)$$

La relation (C.5.) compte tenu des relations (C.4) et (C.6) se simplifie comme :

$$\frac{\partial|H(z(i),\hat{\theta})|}{\partial\hat{\theta}(n)}=\frac{Re\left[\overline{H(z(i),\hat{\theta})}\frac{\partial}{\partial\hat{\theta}(n)}H(z(i),\hat{\theta})\right]}{|H(z(i),\hat{\theta})|} \qquad (C.7)$$

$$n = 1, 2, \ldots 4K$$

Chaque terme du premier membre de la relation (C.7) peut être évalué à partir de la définition de base (II.16). répétée ici :

$$H(z^{-1}(i)) = A\prod_{k=1}^{K}\frac{1+a(1k)z^{-1}(i)+a(2k)z^{-2}(i)}{1+b(1k)z^{-1}(i)+b(2k)z^{-2}(i)} \qquad (C.8)$$

Pour simplifier l'écriture, posons $H(z(i),\hat{\theta})$ égal à $H(i)$.
Evaluons donc les dérivées partielles suivantes à l'aide des relations (C.7) et (C.8)

$$(1)\quad \frac{\partial|H(i)|}{\partial a(1j)}=\frac{Re\left[\overline{H(i)}\frac{\partial H(i)}{\partial a(1j)}\right]}{|H(i)|}$$

$$=\frac{1}{|H(i)|}Re\left[\overline{H(i)}\frac{\partial}{\partial a(1j)}\left[\prod_{k=1}^{K}\frac{1+a(1k)z^{-1}(i)+a(2k)z^{-2}(i)}{1+b(1k)z^{-1}(i)+b(2k)z^{-2}(i)}\right]\right]$$

$$=\frac{1}{|H(i)|}Re\left[\overline{H(i)}\left(\prod_{\substack{k=1\\k\neq j}}^{K}\frac{1+a(1k)z^{-1}(i)+a(2k)z^{-2}(i)}{1+b(1k)z^{-1}(i)+b(2k)z^{-2}(i)}\right)\left(\frac{z^{-1}(i)}{1+b(1j)z^{-1}(i)+b(2j)z^{-2}(i)}\right)\right]$$

$$\frac{\partial|H(i)|}{\partial a(1j)} = \frac{1}{|H(i)|} Re\left[\overline{H(i)}\, H(i) \quad \frac{z^{-1}(i)}{1+a(1j)z^{-1}(i)+a(2j)z^{-2}(i)}\right]$$

et finalement

$$\frac{\partial|H(i)|}{\partial a(1j)} = |H(i)|\, Re\left[\frac{z^{-1}(i)}{1+a(1j)z^{-1}(i)+a(2j)z^{-2}(i)}\right]$$

$$j = 1, 2, \ldots K, \quad i = 1, 2 \ldots M. \qquad (C.9)$$

(2) D'une manière identique, nous aurons :

$$\frac{\partial|H(i)|}{\partial a(2j)} = |H(i)|\, Re\left[\frac{z^{-2}(i)}{1+a(1j)z^{-1}(i)+a(2j)z^{-2}(i)}\right] \qquad (C.10)$$

$$j = 1, 2, \ldots K, \quad i = 1, 2 \ldots M.$$

$$(3)\quad \frac{\partial|H(i)|}{\partial b(1j)} = \frac{1}{|H(i)|} Re\left[\overline{H(i)} \frac{\partial}{\partial b(1j)} \prod_{K=1}^{K} \frac{1+a(1K)z^{-1}(i)+a(2K)z^{-2}(i)}{1+b(1K)z^{-1}(i)+b(2K)z^{-2}(i)}\right]$$

$$= \frac{1}{|H(i)|} Re\left[\overline{H(i)} \left[\prod_{K=1, K\neq j}^{K} \frac{1+a(1K)z^{-1}(i)+a(2K)z^{-2}(i)}{1+b(1K)z^{-1}(i)+b(2K)z^{-2}(i)}\right] \frac{\partial}{\partial b(1j)} \frac{1+a(1j)z^{-1}(i)+a(2j)z^{-2}(i)}{1+b(1j)z^{-1}(i)+b(2j)z^{-2}(i)}\right]$$

$$= \frac{1}{|H(i)|} Re\left[\overline{H(i)} \left[\prod_{K=1, K\neq j}^{K} \frac{1+a(1K)z^{-1}(i)+a(2K)z^{-2}(i)}{1+b(1K)z^{-1}(i)+b(2K)z^{-2}(i)}\right] \frac{(-1)z^{-1}(i)\,1+a(1j)z^{-1}(i)+a(2j)z^{-2}(i)}{1+b(1j)z^{-1}(i)+b(2j)z^{-2}(i)}\right]$$

$$= \frac{1}{|H(i)|} Re\left[\overline{H(i)}\, H(i) \quad \frac{(-1)z^{-1}(i)}{1+b(1j)z^{-1}(i)+b(2j)z^{-2}(i)}\right]$$

et finalement :

$$\frac{\partial |H(i)|}{\partial b(1j)} = |H(i)| \, Re \left[\frac{-z^{-1}(i)}{1 + b(1j) z^{-1}(i) + b(2j) z^{-2}(i)} \right] \quad (C.11)$$

$$j = 1, 2, \ldots K \, , \; i = 1, 2, \ldots M .$$

(4) D'une manière identique, nous avons :

$$\frac{\partial |H(i)|}{\partial b(2j)} = |H(i)| \, Re \left[\frac{-z^{-2}(i)}{1 + b(1j) z^{-1}(i) + b(2j) z^{-2}(i)} \right] \quad (C.12)$$

$$j = 1, 2, \ldots K \, , \; i = 1, 2, \ldots M .$$

Résumons donc la méthode de calcul de $\hat{C}(\hat{\theta})$ et de grad $(\hat{C}(\hat{\theta}))$ pour $\hat{\theta}$ donné :

(1) Evaluer $H(i) = \prod_{k=1}^{K} \frac{1 + a(1k) z^{-1}(i) + a(2k) z^{-2}(i)}{1 + b(1k) z^{-1}(i) + b(2k) z^{-2}(i)}$, $i = 1, 2, \ldots M .$

(2) Evaluer $A^* = \dfrac{\sum_{i=1}^{M} |H(i)| \, R^d(i)}{\sum_{i=1}^{M} |H(i)|}$

(3) Evaluer $E(i) = A^* |H(i)| \, R^d(i)$, $i = 1, 2, \ldots M$

(4) Evaluer $\hat{C}(\hat{\theta}) = \sum_{i=1}^{M} E^2(i)$

(5) Evaluer $\frac{\partial |H(i)|}{\partial a(1j)}$, $\frac{\partial |H(i)|}{\partial a(2j)}$, $\frac{\partial |H(i)|}{\partial b(1j)}$, $\frac{\partial |H(i)|}{\partial b(2j)}$ pour

$j = 1, 2, \ldots K$ et $i = 1, 2, \ldots M$, à partir des relations (C,9), (C.10), (C.11), et (C.12).

(6) Evaluer $$\frac{\partial \hat{C}}{\partial \hat{\theta}(n)} = 2A^* \sum_{i=1}^{M} E(i) \frac{\partial H(i)}{\partial \hat{\theta}(n)}, \quad n = 1, 2, \ldots 4K$$

Notons que nous pouvons économiser du temps de calcul en évaluant les $z(i)^{-1}$ et $z(i)^{-2}$ pour i = 1, 2, ... M au début du programme et en les mémorisant.

A N N E X E D

Représentation des nombres réels dans une machine numérique

Dans une machine numérique, les nombres réels peuvent être représentés de plusieurs façons.
Par exemple :

- représentation en valeurs absolues plus le signe,
- représentation en complément à un
- représentation en complément à deux,

Ces représentations sont toutes équivalentes pour les nombres positifs. Par conséquent, les erreurs d'arrondi de ces nombres sont identiques pour ces représentations. La diférence essentielle relève de la représentation des nombres négatifs.

En général, dans les mini-ordinateurs travaillant en temps réel, on emploie l'arithmétique en virgule fixe, en complément à deux, et, par conséquent, les nombres réels sont représentés en complément à deux.

Ici, nous allons examiner cette représentation en détail.

Soit un nombre réel à représenter en complément à deux ; celui-ci peut être exprimé en fonction des coefficients binaires comme suit :

$$(A)_N = -g(0) + \sum_{j=1}^{N-1} g(j)2^{-j} \qquad (D.1)$$

où l'on a normalisé A tel que $|A| \leq 1$, et où l'on représente ce nombre en N bits binaires, signe compris. Le symbole $(X)_K$ est utilisé pour désigner un nombre réel X représenté sur K bits (K peut être infini).
Les g(j) sont soit 0, soit 1.

Si les mots de la machine avaient une infinité de bits, la représentation exacte de A serait :

$$(A)_\infty = -g(0) + \sum_{j=1}^{\infty} g(j)2^{-j} \qquad (D.2)$$

Puisque les nombres de bits de registres d'une machine numérique sont limités à N bits, on est amené soit à tronquer, soit à arrondir $(A)_\infty$ de façon à pouvoir représenter A en N bits. De ce fait, il y a une erreur introduite dans la représentation du nombre A dans la machine.
Calculons cette erreur e_T quand on tronque A aux N premiers bits :

$$e_T = (A)_\infty - (A)_N \qquad (D.3)$$

$$= (-g(0) + \sum_{j=1}^{\infty} g(j)2^{-j}) - (-g(0) + \sum_{j=1}^{N-1} g(j)2^{-j}) \qquad (D.4)$$

Si l'on a $g(j) = 0$ pour tout j tel que $N \leq j \leq \infty$ l'erreur est minimum et égale à zéro.

Par contre, si l'on a $g(j) = 1$ pour tout j tel que $N \leq j \leq \infty$ l'erreur est maximum. Elle est égale à :

$$e_{T\ max} = \sum_{j=N}^{\infty} g(j)\,2^{-j} \qquad (D.5)$$

$$= 2^{-N} + 2^{-(N-1)} + \ldots + 0$$

$$= \frac{2^{-N}}{(1-2^{-1})}$$

$$= 2^{-(N-1)}$$

et finalement $e_{T\ max} = 2^{-(N-1)}$ (D.6)

Evidemment, la valeur exacte de l'erreur dans cet intervalle $\left[0,\ 2^{-(N-1)}\right]$ dépend de la valeur de A. Mais, après la troncature, puisque l'on a perdu la valeur exacte de A, on a introduit une incertitude. C'est ce raisonnement intuitif qui nous conduit à faire l'hypothèse que l'erreur se comporte comme une variable aléatoire qui peut prendre toutes les valeurs dans cet intervalle continu $\left[0,\ 2^{-(N-1)}\right]$ avec une probabilité uniforme, égale à $2^{(N-1)}$. Avec cette hypothèse, on peut facilement évaluer les deux premiers moments de cette variable.

Posons $2^{-(N-1)} = q$ le pas de quantification

m_1, le moment d'ordre un de cette variable est donc :

$$m_1 = \frac{1}{q}\int_0^q x\ d\ x$$

$$= \frac{q}{2} \qquad (D.7)$$

et le moment d'ordre deux est :

$$m_2 = \frac{1}{q}\int_0^q x^2\ d\ x = \frac{q^2}{3} \qquad (D.8)$$

De la même façon, nous pouvons considérer la troncature du produit de deux nombres A et B, chacun étant représenté par N bits (pour simplifier l'écriture des formules, nous excluons le bit de signe dans ce qui suit). Le produit résultant de deux tels nombres A et B a, en général, 2N bits. Si l'on suppose que l'erreur dans la représentation de A et de B sur N bits est négligeable, on peut considérer l'erreur dûe à la troncature au premier n bits du produit résultant comme une variable aléatoire discrète, uniformément répartie sur un intervalle discret et fini. Evaluons les bornes inférieures et supérieures de l'erreur ainsi que ses deux moments m_1 et m_2 dans ce cas là.
Avec la définition de l'erreur e_T comme dans le cas précédent, nous avons :

$$e_T = (AB)_{2N} - (AB)_n \tag{D.9}$$

où le symbole $(x)_k$ a la même définition que dans (D.1)

et

$$(AB)_{2N} = -g(0) + \sum_{i=1}^{2N} g(i)2^{-i} \tag{D.10}$$

$$(AB)_n = -g(0) + \sum_{i=1}^{n} g(i)2^{-i} \tag{D.11}$$

d'où

$$e_T = \sum_{i=n+1}^{2N} g(i)2^{-i} \tag{D.12}$$

Posons : $2N = K$

En suivant les mêmes raisonnements que dans le cas précédent, nous avons les valeurs de l'erreur minimum et maximum, soit :

$$e_{T\ min} = 0 \tag{D.13}$$

$$e_{T\ max} = \sum_{i=n+1}^{K} 2^{-i} \tag{D.14}$$

$$= \frac{2^{-(n+1)} - 2^{(K+1)}}{(1-2^{-1})}$$

et finalement :

$$e_{T\ max} = 2^{-n} - 2^{-K} \tag{D.15}$$

Cette fois, l'erreur est définie sur l'intervalle discret $\left[0,\ 2^{-n} - 2^{-K}\right]$ et ne peut prendre qu'un nombre fini des valeurs discrètes. Soit $2^{(K-n)}$ valeurs, suivant la configuration des bits du produit $(AB)_{2N}$. La figure (D.1) montre la densité de probabilité de l'erreur.

Calculons les deux premiers moments m_1 et m_2 de cette variable aléatoire

Soit $\Phi_{e_j}(t)$ la fonction caractéristique de e_j et $p(e_j)$ la densité de probabilité de e_j. Nous avons :

$$p(e_j) = \frac{1}{2^{(k-n)}} \text{ pour } 0 \leq e_j \leq (2^{-n} - 2^{-K})$$

$$= 0 \qquad \text{pour } e_j \notin \left[0,\ (2^{-n} - 2^{-K})\right]$$

$$= \sum_{j=0}^{2^{(K-n)}-1} p(e_j)\ \exp(it e_j) \quad \text{avec } i = \sqrt{-1} \tag{D.16}$$

Posons $\alpha = 2^{-K}$

$$\beta = 2^{(K-n)}$$

donc $e_j = j\alpha$ pour $0 \leq j \leq \beta - 1$

ce qui donne : $$\Phi_{e_j}(t) = \frac{1}{\beta} \sum_{j=0}^{\beta-1} \exp(itj\alpha) \qquad (D.17)$$

$$= \frac{1}{\beta} \sum_{j=0}^{\beta-1} (1 + itj\alpha + (itj\alpha)^2 + \dots)$$

En effectuant la sommation indiquée, nous avons :

$$\Phi_{e_j}(t) = \frac{1}{\beta}\left[(1 + it\alpha + \frac{(it\alpha)^2}{2!} + \dots) + (1 + it2\alpha + \frac{(it2\alpha)^2}{2!} + \dots) + \dots + (1 + it(\beta-1)\alpha + \frac{(it(\beta-1)\alpha)^2}{2!} + \dots)\right]$$

$$= \frac{1}{\beta}\left[\beta + it\alpha + it2\alpha + \dots + it(\beta-1)\alpha + \left\{\frac{(it\alpha)^2}{2!} + \frac{(it2\alpha)^2}{2!} + \dots + \frac{(it(\beta-1)\alpha)^2}{2!}\right\} + \dots\right]$$

$$= \frac{1}{\beta}\left[\beta + it\alpha\frac{\beta(\beta-1)}{2} + \frac{(it\alpha)^2}{2!}\left\{1 + 2^2 + 3^2 + \dots + (\beta-1)^2\right\} + \dots\right]$$

et finalement :

$$\Phi_{e_j}(t) = \frac{1}{\beta}\left[\frac{it\alpha\beta(\beta-1)}{2} + \frac{(it\alpha)^2}{2!}\frac{\beta(\beta-1)(2\beta-1)}{6} + \dots\right]$$

Le moment d'ordre un m_1 est égal à :

$$im_1 = \left.\frac{d\Phi_{e_j}(t)}{dt}\right|_{t=0} = \frac{i\alpha\beta(\beta-1)}{2\beta}$$

d'où

$$m_1 = \frac{\alpha}{2}(\beta-1)$$

$$= \frac{2^{-K}}{2}(2^{(K-n)} - 1)$$

$$= \frac{2^{-n} - 2^{-K}}{2} \qquad (D.18)$$

Nous avons aussi, pour le moment d'ordre deux, m_2 :

$$(i)^2 m_2 = \left.\frac{d^2\Phi_{e_j}(t)}{dt^2}\right|_{t=0} = (i)^2 \frac{\alpha^2\beta(\beta-1)(2\beta-1)}{6\beta}$$

et $$m_2 = \frac{1}{6}\frac{(2^{K-n} - 1)(2^{K-n+1} - 1)}{2^{2K}} \qquad (D.19)$$

quand K tend vers l'infini, c'est à dire que les deux nombres A et B représentés sur une infinité de bits, nous avons :

$$m_{2\infty} = \frac{1}{6} 2^{-2n+1}$$

$$= \frac{1}{3} 2^{-2n} \qquad (D.20)$$

L'équation (D.20) est identique à l'équation (D.8)

Considérons le cas arrondi :

Quand on veut arrondir le produit représenté sur 2N bits au n bits poids fort, on teste le (n+1) ième bit. Si il vaut 1, on ajoute 1 au n ième bit. Si le (n+1) ième bit vaut 0, alors on élimine tous les bits à partir du (n+1) ième.
Nous avons donc pour l'erreur e_A :

$$e_A = (AB)_{2N} - (AB)_n = \left(-g(0) + \sum_{j=1}^{2N} g(j) 2^{-j}\right) - \left(-g(0) + \sum_{j=1}^{n} g(j) 2^{-j}\right) + g(n+1) 2^{-(n+1)} \qquad (D.21)$$

Ici, il existe quatre bornes de l'erreur. Soit

(i) tous les bits compris entre le (n+1) ième et 2Nième sont zéro. Donc, l'erreur est zéro.

(ii) Quand tous les bits, de (n+1) ième au 2N ième sont en 1, en posant 2N = K, nous avons d'après (D.21) :

$$(AB)_K - (AB)_n = e_{A\,min} = \sum_{j=n+1}^{K} 2^{-j} - 2^{-n} = 2^{-K} \qquad \text{(D.22)}$$

(iii) Quand, seul le (n+1) ième bit est un 1 et les autres sont zéro, nous avons :

$$(AB)_K - (AB)_n = e_{A\,max} = 2^{-(n+1)} - 2^{-n} = -2^{-(n+1)} = -\frac{q}{2} \qquad \text{(D.23)}$$

(iv) Quand le (n+1) ième bit est égal à zéro et tous les autres sont en 1, nous avons :

$$e_{A\,max} = \sum_{j=n+2}^{K} 2^{-j} = 2^{-(n+1)} - 2^{-K} = \frac{q}{2} - 2^{-K} \qquad \text{(D.24)}$$

En conclusion, nous pouvons considérer l'erreur dûe à l'arrondi à n bits du produit défini sur 2N bits comme une variable aléatoire discrète, uniformément répartie sur l'intervalle $\left[0, \frac{q}{2} - 2^{-K}\right]$ et l'intervalle $\left[-2^{-K}, -\frac{q}{2}\right]$

(où $q = 2^{-n}$) et $K = 2N$

et qui peut prendre une des $2^{(K-n)} - 1$ valeurs dans cet intervalle avec une probabilité de $\frac{1}{2^{(K-n)} - 1}$ (voir fig. D.2)

En faisant le calcul semblable au cas de troncature, on a pour les deux moments m_1 et m_2 :

$$m_1 = 2^{-(K+1)} \simeq 0 \qquad \text{(D.25)}$$

$$m_2 = \frac{1}{6} 2^{-2K} \left[2^{2(K-n)-1} - 2^{(K-n)} - 2^{K-n-1} + 6\,2^{K-n-2} + 1 \right] \qquad \text{(D.26)}$$

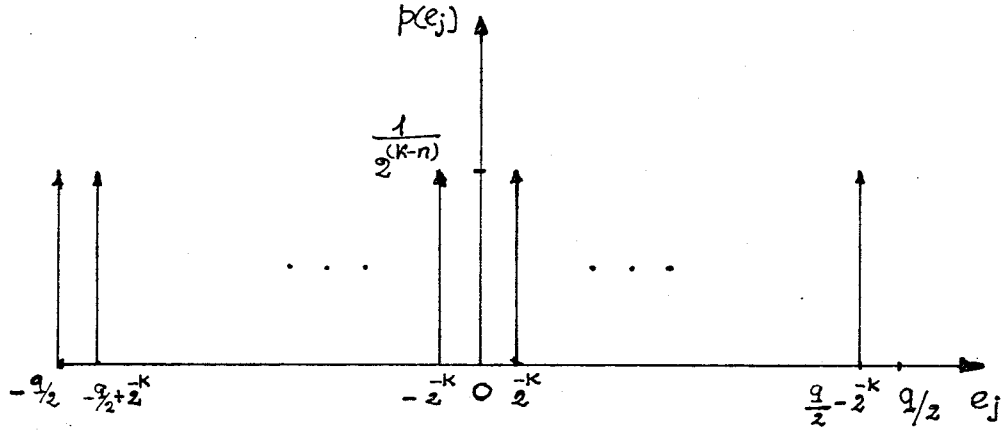


fig. D.2

Distribution discrête de l'erreur d'arrondi

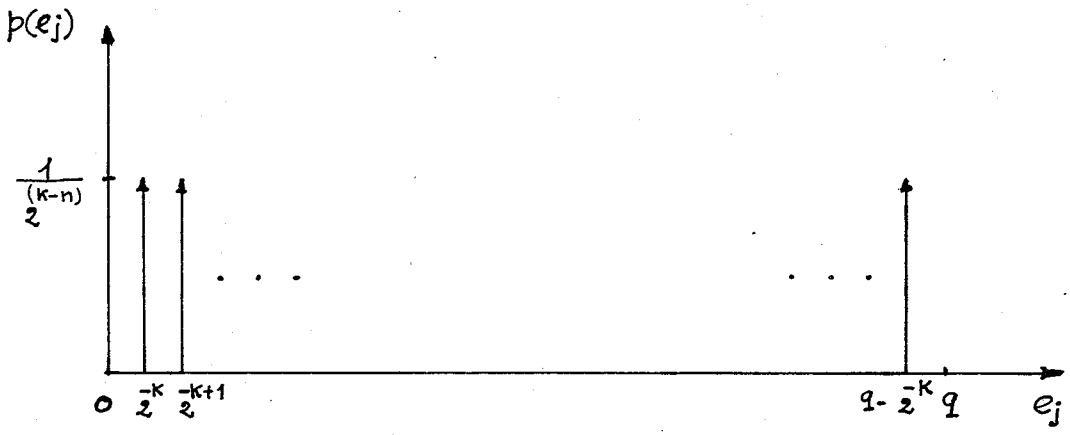


fig. D.1

Distribution discrête de l'erreur de troncature

Quand K tend vers l'infini, nous avons :

$$m_1 = 0 \tag{D.27}$$

et

$$m_2 = \frac{1}{6} 2^{-2n-1}$$

$$= \frac{1}{12} 2^{-2n} = \frac{q^2}{12} \tag{D.28}$$

Ce qui correspond au cas d'une variable aléatoire continue, uniformément répartie entre $\frac{-q}{2}$ et $\frac{q}{2}$

A N N E X E E

Contrainte de débordement

Nous nous proposons d'évaluer la sommation $\sum_{k=0}^{\infty} |h(k)|$ pour un filtre récursif de 2ème ordre.

Soit l'équation de récurrence d'un tel filtre :

$$y(n) = x(n) - b_1\, y(n-1) - b_2\, y(n-2) \qquad \text{(E.1)}$$

La solution générale de cette équation pour un système ayant un pôle complexe conjugué dans le plan z à $r \exp(\pm j\theta)$ est §5§ :

$$y(k) = A_1\, G_1^{\,k} + A_2\, G_2^{\,k} \qquad \text{(E.2)}$$

avec G_1, G_2, : pôles du système

A_1, A_2, : constants réels

La R.I./h(k)/est définie comme étant la sortie pour la séquence d'entrée /x(n)/ = (1, 0, 0, 0, ...)
Donc nous avons :

$$A_1 + A_2 = 1 \qquad \text{(E.3)}$$

$$A_1\, G_1 + A_2\, G_2 = -b_1 \qquad \text{(E.4)}$$

c'est à dire que :

$$A_1 r \exp(j\theta) + A_2 \exp(-j\theta) = -b_1 \qquad (E.5)$$

ou

$$(1-A_2) r \exp(j\theta) + A_2 \exp(-j\theta) = -b_1 \qquad (E.6)$$

ce qui donne :

$$-b_1 = r \exp(j\theta) + A_2 r (\exp(-j\theta) - \exp(j\theta)) \qquad (E.7)$$

qui se simplifie comme :

$$-b_1 = r \exp(j\theta) + A_2 \, r2j \sin\theta \qquad (E.8)$$

d'où

$$A_2 = \frac{b_1 + r\exp(j\theta)}{2j\, r \sin\theta}, \qquad j = \sqrt{-1} \qquad (E.9)$$

et $$A_1 = 1 - A_2 = 1 - \frac{b_1 + r\exp(j\theta)}{2j\, r \sin\theta}$$

Donc, finalement :

$$y(n) = \left[1 - \frac{b_1 + r\exp(j\theta)}{2j \sin\theta}\right] r^n \exp(jn\theta) + \frac{b_1 + r\exp(j\theta)}{2j\, r \sin\theta} \exp(-jn\theta) \qquad (E.10)$$

La relation (E.10) se simplifie comme :

$$y(n) = \frac{1}{2j\, r \sin\theta}\left[(2j\, r \sin\theta)\, r^n \exp(jn\theta) - 2j \sin(n\theta)\, r^n (b_1 + r\exp(j\theta))\right] \qquad (E.11)$$

D'autre part, $b_1 = -2r\cos\theta$

ce qui donne :

$$y(n) = \frac{1}{2jr\sin\theta}\left[(2jr\sin\theta)r^n exp(jn\theta) - r^{n+1}2j\sin n\theta(-2\cos\theta + exp(j\theta))\right] \quad (E.12)$$

$$= \frac{r^n}{\sin\theta}\left[\sin\theta\, exp(jn\theta) - \sin n\theta(-2\cos\theta + exp(j\theta))\right] \quad (E.13)$$

$$= \frac{r^n}{\sin\theta}\left[2\sin n\theta\cos\theta - exp(j\theta)\sin n\theta + \sin\theta\, exp(jn\theta)\right] \quad (E.14)$$

$$= \frac{r^n}{\sin\theta}\left[2\cos\theta\sin n\theta - \cos\theta\sin n\theta - j\sin\theta\sin n\theta + \cos n\theta\sin\theta + j\sin\theta\sin n\theta\right] \quad (E.15)$$

$$= \frac{r^n}{\sin\theta}\left[2\cos\theta\sin n\theta + \cos n\theta\sin\theta - \sin n\theta\cos\theta\right] \quad (E.16)$$

qui simplifie comme :

$$y(n) = \frac{r^n}{\sin\theta}\left[\sin(n+1)\theta\right] \quad (E.17)$$

La relation(E.17) donne, compte tenu du fait que $h(n) = y(n)$ dans notre cas :

$$\sum_{n=0}^{\infty}\left|h(n)\right| = \frac{1}{\sin\theta}\sum_{n=0}^{\infty} r^n\left|\sin(n+1)\theta\right| \quad (E.18)$$

D'où nous pouvons établir la borne supérieure suivante :

$$\sum_{n=o}^{\infty} \left| h(n) \right| < \frac{1}{\sin\theta} \sum_{n=o}^{\infty} r^n \qquad \text{(E.19)}$$

L'équation (E.19) nous donne une borne supérieure facilement évaluable. Mais cette borne ne sera jamais atteinte dans la pratique. La relation (E.19) montre également que pour petit (voisin de zéro) et r élevé (voisin de 1),la contrainte de débordement est très sévère.