

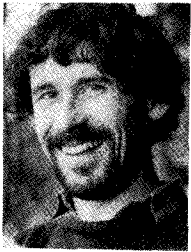
## Incertitudes de calcul

dans un processeur de Fourier rapide :

modélisation et expérience

Error analysis of a floating block FFT processor:

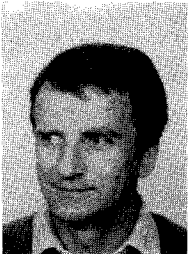
model and experiment



**Patrick FURON**

GERSIC, ISMRa, Esplanade de la Paix, 14032 CAEN CEDEX.

Ingénieur de l'École Nationale Supérieure d'Électronique et d'Électromécanique de Caen (ENSEEC 1977). Docteur-ingénieur en électronique de l'Université Paris-XI. Assistant d'automatique à l'ISMRa de Caen.



**Daniel BLOYET**

GERSIC, ISMRa, Esplanade de la Paix, 14032 CAEN CEDEX.

Docteur en Électronique (1970) et ès Sciences physiques de l'Université Paris Sud (1976). Actuellement professeur d'électronique à l'ISMRa de Caen. Dirige au sein du GERSIC l'activité « Instrumentation scientifique et médicale ».

### RÉSUMÉ

Cet article concerne l'analyse de la précision d'un processeur de Transformée de Fourier Rapide (TFR) réalisé autour de deux multiplieurs-accumulateurs et disposant d'une arithmétique flottante par blocs. Le modèle statistique de propagation de l'erreur d'étape y est donné. Par itération sur l'ensemble des étapes, cette analyse mène à l'estimation de l'énergie des trois contributions d'erreur — erreurs d'entrée, arithmétique et de coefficients — qui sont comparées à l'expérience.

#### MOTS CLÉS

TFR, analyse d'erreurs, erreur arithmétique, arithmétique flottante par blocs.

### SUMMARY

*This paper is dedicated to a floating block FFT error analysis. A statistical model of the error propagation in a pass is developed which considers separately the three following parts: the input, arithmetic roundoff and coefficient quantization errors. Each of them is discussed and experimentally verified.*

#### KEY WORDS

*FFT, error analysis, roundoff error, step by step scaling, block floating point arithmetic.*

## 1. Introduction

Lorsqu'une Transformée de Fourier Rapide (TFR) est effectuée, des erreurs de quantification interviennent, liées à la longueur nécessairement finie des valeurs numériques codées à l'intérieur de la machine sous la forme de mots binaires. Elles impliquent une indétermination sur le résultat qui est ainsi affecté d'un bruit.

Trois types de quantification ont effet sur la précision des résultats :

- une quantification d'entrée, liée à la résolution du convertisseur analogique-numérique, qui produit une « erreur d'entrée »;
- une quantification dans l'unité de calcul (multiplications, additions, décalages) qui produit une « erreur arithmétique »;
- une quantification des coefficients qui introduit une « erreur de coefficients ».

Ces trois types d'erreur se propagent de l'entrée vers la sortie lors du calcul d'une TFR. Dans de nombreux cas, il est raisonnable de traiter les effets de troncature et d'arrondi par un modèle statistique remplaçant chaque source d'erreur par un bruit blanc uniformément distribué en amplitude sur un domaine qui dépend du type de quantification. On admet de plus que les sources sont non corrélées et indépendantes du signal d'entrée. Ces hypothèses sont en général vérifiées pour une grande majorité de signaux comme les signaux aléatoires, la voix... [2].

Si l'erreur due aux coefficients est négligeable par rapport aux deux autres, ainsi que le rappellent Trân-Thông et Liu [6], elle a néanmoins été calculée [1-3, 7, 12-14]. Tufts et Hersey [16] ont montré qu'elle induit des lobes secondaires parasites dans le spectre. Les erreurs de coefficient sont corrélées entre elles, et une étude statistique ne peut suffire à les décrire totalement [2, 12-14].

Beaucoup plus importante est l'erreur arithmétique qui est généralement donnée [1-6, 8-11, 14] en valeur quadratique moyenne. Knight et Kaiser [7] en expriment un majorant, tandis que Moharir [15] montre que ce majorant dépend de la position de l'échantillon de sortie. Welch [1] puis Heute et Schuessler [14] donnent l'expression du terme d'incertitude dû à la quantification des données d'entrée.

En général, ces quantités sont exprimées en fonction de la dimension  $N$  de la TFR et de la quantification définie par le format des données à l'intérieur de la machine. Pour la structure flottante par blocs, les calculs de Welch [1] par exemple ont mené à la connaissance des limites supérieure et inférieure de l'erreur quadratique moyenne dans les deux seuls cas suivants : décalage à droite à chaque étape et aucun décalage. Seuls Knight et Kaiser donnent, pour une structure flottante par blocs, une borne supérieure de l'erreur quadratique moyenne connue *a posteriori*, c'est-à-dire calculée en même temps que la TFR. L'expression de l'erreur de calcul est ainsi fonction de l'agencement des opérations de décalage, mais elle est limitée aux seuls décalages à droite.

Inspirée de l'article de Welch [1] quant à la recherche d'une expression incluant les trois types d'erreur, la

méthode que nous proposons ici tient compte de manière très détaillée de l'architecture et des manipulations de données effectuées par un processeur TFR « flottant par blocs » dont la structure et les principales caractéristiques sont détaillées en [17-18]. En calculant l'erreur liée à une étape puis en étudiant la progression du maximum statistique de l'amplitude des données au cours du calcul du spectre, nous obtenons une équation récurrente qui permet de délimiter l'erreur finale : *a priori* en fonction de l'énergie du signal d'entrée, ou *a posteriori*, en fonction des décalages internes effectivement utilisés ou en fonction de l'exposant du spectre résultat.

Cette méthode peut être facilement généralisée au cas d'une structure TFR quelconque. Il s'avère cependant que le modèle d'erreur doit être précis et prendre en compte tous les détails de la réalisation matérielle. Ainsi, pour notre processeur, la prise en compte des particularités des différentes étapes (qui sont étudiées aux paragraphes 3.3.2 et 3.3.3) rapproche de 20 dB les évaluations théoriques et expérimentales de l'erreur.

C'est la raison pour laquelle l'étude présentée est volontairement limitée au cas de la structure réalisée ce qui nous permet de confronter de façon précise théorie et expérience.

## 2. Modélisation des incertitudes de calcul

### 2.1. L'ALGORITHME TFR ET SA RÉALISATION

Considérons un tableau de  $N$  données d'entrée représenté par un vecteur équivalent  $X_0$  de dimension  $N=2^n$ ; la transformation de Fourier rapide procède en  $n$  « étapes », qui le remplacent successivement par  $X_1, X_2, \dots, X_n$ , ce dernier vecteur étant le spectre résultat dont les composantes sont données par l'équation (1) de la transformée de Fourier discrète :

$$(1) \quad x_n(k) = \sum_{i=0}^{N-1} W_N^{ik} x_0(i)$$

avec

$$(2) \quad W_N = e^{-j2\pi/N}$$

Chaque étape se décompose en  $N/2$  opérations élémentaires nommées « papillons »; un papillon de la  $p$ -ième étape,  $p=0, 1, \dots, n-1$ , calcule deux composantes notées  $a_{p+1}$  et  $b_{p+1}$  du vecteur  $X_{p+1}$  à partir des deux anciennes  $a_p$  et  $b_p$  et du coefficient  $W_p$  dépendant des indices de passe et de papillon (l'indice de papillon est ici omis pour ne pas alourdir l'écriture). Pour l'algorithme de dédoublement temporel que nous avons retenu, les équations du papillon s'écrivent (fig. 1) :

$$(3) \quad a_{p+1} = a_p + W_p b_p$$

$$(4) \quad b_{p+1} = a_p - W_p b_p$$

Le processeur réalisé, qui sert de support à cet article, utilise une arithmétique flottante par blocs et exécute une TFR de  $N=1024$  points en 3 ms selon l'algo-

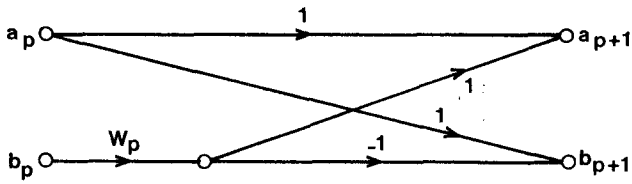


Fig. 1. - Graphe du papillon de l'étape d'indice  $p$  de l'algorithme de dédoublement temporel.

Fig. 1. - Graph of the  $p$  pass butterfly using the decimation in time algorithm.

rithme « de dédoublement temporel à géométrie constante ». Rappelons-en les points essentiels. Les 1024 données réelles d'entrée sont transmises par l'ordinateur hôte au processeur TFR selon un format en 16 bits signés représentés en notation « complètement à deux ». A la suite d'une opération de fenêtrage temporel exécuté par un processeur d'entrée-sortie, elles sont écrites (vecteur  $X_0$ ) dans la zone mémoire du processeur TFR selon le format 16 bits signés où cependant le bit de poids faible est nul, conséquence d'un mode de présentation particulier des résultats du multiplieur-accumulateur utilisé pour ce pré-traitement. Relues, elles sont ensuite traitées en dix étapes successives ( $p=0, 1, \dots, 9$ ).

Les multiplieurs-accumulateurs (MAC) opèrent en entrée avec des mots de 16 bits signés mais exécutent le calcul des papillons en pleine précision. Par contre, les résultats intermédiaires  $a_{p+1}$  et  $b_{p+1}$  sont stockés en mémoire suivant des mots de 17 bits signés puis renormalisés en cas de débordement par excès ou par défaut au format d'entrée des MAC. Cette renormalisation, qui consiste en une lecture des résultats de l'étape précédente avec un décalage approprié de gain  $G(p)$ , affecte la valeur de l'exposant que l'on notera  $e(p+1)$  du bloc résultat. Si l'opération de décalage précède en pratique la phase de calcul, nous considérerons, pour notre analyse, qu'elle la suit. La figure 2.1 représente le graphe d'un papillon qui inclut les opérations de décalage et de gestion de l'exposant où l'on a adopté les notations suivantes :

$$(5) \quad \begin{cases} a_p = x_p 2^{e(p)}, & a_{p+1} = x_{p+1} 2^{e(p+1)} \\ b_p = y_p 2^{e(p)}, & b_{p+1} = y_{p+1} 2^{e(p+1)} \end{cases}$$

et où l'opération de décalage se traduit par :

$$(6) \quad 2^{e(p+1)} = \frac{2^{e(p)}}{G(p)} \quad \text{avec } e(0) = 0$$

et  $G(p) = \{1/4, 1/2, 1, 2\}$  selon le décalage réalisé

### 2.2. LES ERREURS DE CALCUL DANS UN PAPILLON

Le traitement du papillon s'effectue en pratique sur des nombres complexes déjà quantifiés, le résultat l'étant également. Il se produit ainsi, par rapport au calcul théorique, des différences qui sont la cause d'erreurs aux origines localisées. Ainsi, trois types d'erreur interviennent dans le calcul d'un papillon de la  $p$ -ième étape (fig. 2.2) :

- les erreurs  $\varepsilon(x_p) 2^{e(p)}$  et  $\varepsilon(y_p) 2^{e(p)}$  sur les données présentées à l'entrée du papillon; elles sont définies comme la différence entre données réelles et données exactes;

- les erreurs  $\varepsilon(W_p)$  dues à la quantification des coefficients  $W_p$  de la table trigonométrique;

- les erreurs liées à la structure arithmétique du processeur (multiplieurs, additionneurs, décaleurs) qui font intervenir les liaisons entre ces éléments. Pour notre processeur qui utilise deux multiplieurs-accumulateurs travaillant en pleine précision, l'erreur de liaison entre multiplieurs et additionneurs est éliminée. Seules subsistent les erreurs  $\varepsilon_G(x_{p+1}) 2^{e(p)}$  et  $\varepsilon_G(y_{p+1}) 2^{e(p)}$  liées à la quantification des résultats.

Toutes ces erreurs sont, dans le cas général, des quantités complexes.

Par différence entre les équations associées aux graphes des figures 2.2 et 2.1, les mantisses des erreurs sur les données complexes de sortie d'un papillon s'écrivent :

$$(7) \quad \varepsilon(x_{p+1}) = G(p) \cdot [\varepsilon(x_p) + W_p \varepsilon(y_p) + y_p \varepsilon(W_p) + \varepsilon(W_p) \varepsilon(y_p)] + \varepsilon_G(x_{p+1})$$

$$(8) \quad \varepsilon(y_{p+1}) = G(p) \cdot [\varepsilon(x_p) - W_p \varepsilon(y_p) - y_p \varepsilon(W_p) - \varepsilon(W_p) \varepsilon(y_p)] + \varepsilon_G(y_{p+1})$$

### 2.3. MOMENT D'ORDRE 2 MOYEN DE L'ERREUR D'UNE ÉTAPE DE LA TFR

Pour bâtir un modèle d'analyse statistique de la propagation des erreurs dans le papillon, nous adoptons les hypothèses suivantes :

Le signal d'entrée de la TFR a les caractéristiques d'un bruit blanc centré. Il est uniformément distribué en amplitude sur l'intervalle  $[-\hat{X}(0), \hat{X}(0)]$  et son énergie est, en valeur moyenne d'ensemble, équirépar-

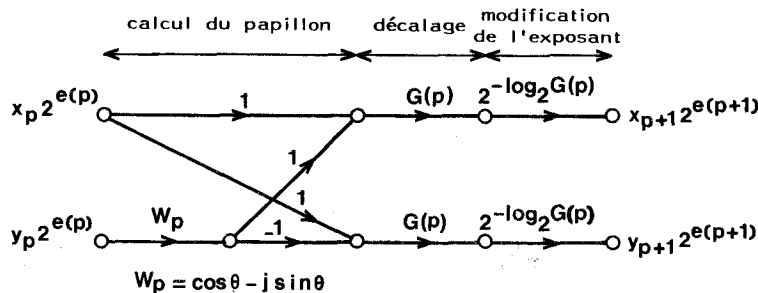


Fig. 2.1. - Graphe d'un papillon sans source d'erreur, associé à une arithmétique flottante par blocs où l'on a adopté les notations  $a_p = x_p 2^{e(p)}$ ,  $b_p = y_p 2^{e(p)}$ , ...

Fig. 2.1. - Simplified (error free) graph of a butterfly using a block floating point arithmetic ( $a_p = x_p 2^{e(p)}$ ,  $b_p = y_p 2^{e(p)}$ ).

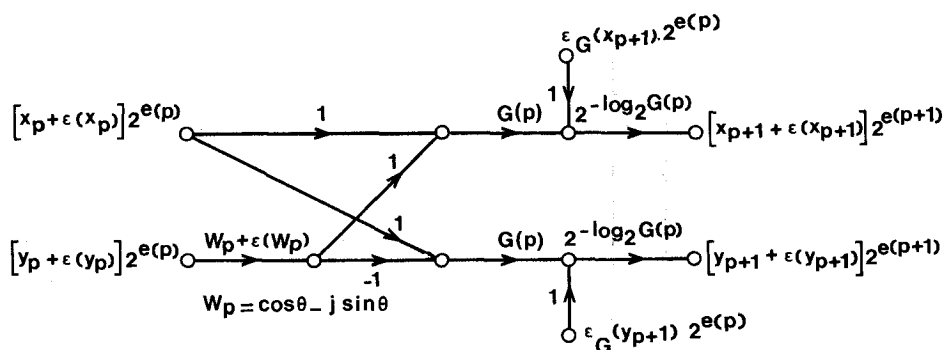


Fig. 2.2. — Graphe d'un papillon avec sources d'erreur, associé à une arithmétique flottante par blocs.

Fig. 2.2. — Complete (including sources) graph of a butterfly using a block floating point arithmetic.

tie sur le vecteur d'entrée. La structure de calcul de la TFR fait que cette propriété se conserve à chaque étape de calcul de la TFR pour cette classe de signaux d'entrée. Ainsi, les composantes du vecteur d'entrée d'une étape conservent les caractéristiques d'un bruit blanc, ce qui ne serait pas le cas dans l'hypothèse d'utilisation de signaux à spectre étroit. En outre, comme nous l'avons signalé en introduction, il est raisonnable d'assimiler ces erreurs à un bruit blanc et de considérer comme indépendants entre eux et avec  $W_p$  les échantillons du tableau d'entrée de chacune des étapes.

Munis de ces hypothèses, nous pouvons bâtir un modèle d'analyse statistique de la propagation des erreurs dans le papillon que nous compléterons par un modèle de propagation statistique du signal — énergie et valeur de crête — au cours de chacune des étapes.

Les moments d'ordre deux  $E[|\varepsilon(x_{p+1})|^2]$  et  $E[|\varepsilon(y_{p+1})|^2]$  des erreurs (7) et (8) s'expriment aisément. Dans leurs expressions les termes croisés disparaissent ainsi que les termes d'ordre 3 et 4 d'importance négligeable; de la relation  $|W_p|^2 = 1$ , il vient :

$$(9) \quad E[|\varepsilon(x_{p+1})|^2] \simeq G_p^2 \{ E[|\varepsilon(x_p)|^2] + E[|\varepsilon(y_p)|^2] + E[|y_p|^2] |\varepsilon(W_p)|^2 \} + E[|\varepsilon_G(x_{p+1})|^2]$$

$$(10) \quad E[|\varepsilon(y_{p+1})|^2] \simeq G_p^2 \{ E[|\varepsilon(x_p)|^2] + E[|\varepsilon(y_p)|^2] + E[|x_p|^2] |\varepsilon(W_p)|^2 \} + E[|\varepsilon_G(y_{p+1})|^2]$$

Les résultats (9) et (10) montrent que les moments d'ordre deux des erreurs aux deux points de sortie d'un papillon sont identiques. La grandeur intéressante pour le concepteur et l'utilisateur est le moment d'ordre deux moyen de l'erreur globale d'étape, noté  $E(p+1)$ , que l'on obtient par prise de moyenne des expressions (9) et (10) sur l'ensemble des  $N$  résultats constituant chaque étape :

$$(11) \quad E(p+1) \simeq G^2(p) [2 \cdot E(p) + E[|y_p|^2] |\varepsilon(W_p)|^2] + E_G(p+1)$$

où l'on a adopté les notations suivantes :

$$E(p+1) = E[|\varepsilon(x_{p+1})|^2] = E[|\varepsilon(y_{p+1})|^2]$$

$$E(p) = E[|\varepsilon(x_p)|^2] = E[|\varepsilon(y_p)|^2]$$

$$E_G(p+1) = E[|\varepsilon_G(x_{p+1})|^2] = E[|\varepsilon_G(y_{p+1})|^2]$$

$$|\varepsilon(W_p)|^2 = \frac{1}{N} \sum_{i=0}^{(N/2)-1} |\varepsilon(W_p)|^2$$

Ainsi l'énergie totale de l'erreur en sortie de la  $p$ -ième étape est la somme quadratique de trois contributions : la première résulte de la propagation et de l'amplification de l'erreur à l'entrée de l'étape, la seconde provient de la quantification des coefficients trigonométriques intervenant dans cette étape, la dernière est créée par la quantification du résultat par l'association MAC-décaleurs;

Avant de procéder à la confrontation expérimentale du résultat (11), nous allons analyser plus en détail les contributions individuelles des trois types d'erreurs.

### 3. Analyse théorique des trois types d'erreur

Afin d'exprimer l'effet des erreurs en termes de rapport signal sur bruit, nous allons d'abord exprimer l'énergie du tableau résultat en fonction des gains  $G(p)$  intervenant dans les différentes étapes du calcul. Les équations (3) et (4) du papillon, illustrées en figure 2.1 permettent de déterminer la progression de la mantisse de l'énergie des données à chaque étape de la TFR.

$$(12) \quad \frac{|x_{p+1}|^2 + |y_{p+1}|^2}{2} = 2 G(p)^2 \frac{|x_p|^2 + |y_p|^2}{2}$$

soit :

$$(13) \quad \frac{|x_p|^2 + |y_p|^2}{2} = 2^p \left\{ \prod_{i=0}^{p-1} G(i)^2 \right\} \frac{|x_0|^2 + |y_0|^2}{2}$$

La valeur efficace d'un tableau s'accroît donc de  $\sqrt{2}$  à chaque étape. L'énergie étant supposée statistiquement équirépartie sur les termes  $x_p$  et  $y_p$  soit  $|x_p|^2 = |y_p|^2$ , le terme  $E[|y_p|^2]$  intervenant dans l'expression (11) s'écrit finalement :

$$(14) \quad E[|y_p|^2] = 2^p \left\{ \prod_{i=0}^{p-1} G(i)^2 \right\} E[|y(0)|^2]$$

où  $E[|y(0)|^2]$  est l'énergie moyenne par point du tableau d'entrée.

L'énergie par point complexe du tableau de sortie s'écrit donc :

$$(15) \quad S = \sqrt{E[|y_{10}|^2] 2^{2e(10)}} \\ = \sqrt{NE[|y_0|^2]} = 32 \sqrt{E[|y_0|^2]}$$

### 3. 1. L'ERREUR D'ENTRÉE

En annulant dans la relation (11) l'erreur de coefficients  $|\varepsilon(W_p)|^2$  et le terme  $E_G(p+1)$  lié à la structure arithmétique, nous obtenons, par récurrence, la valeur efficace de l'erreur issue de la quantification des données d'entrée. Cette erreur peut être calculée directement et devient avec nos hypothèses :

$$(16) \quad E(p+1) = 2G^2(p)E(p)$$

En notant  $e$  la valeur de l'exposant final affecté à l'ensemble du résultat, la valeur efficace de l'erreur liée à l'entrée que nous notons  $B_E$  s'écrit :

$$(17) \quad B_E = \sqrt{E(10) 2^{2e}} = \sqrt{2^{10} \prod_{p=0}^9 G^2(p) E(0) 2^{2e}}$$

qui devient en tenant compte de la relation (6) étendue à l'ensemble de la TFR :

$$(18) \quad 2^e \cdot \prod_{p=0}^9 G(p) = 1$$

$$(19) \quad B_E = 32 \sqrt{E(0)}$$

La comparaison des équations (15) et (19) montre finalement que l'erreur relative sur la détermination du spectre est identique à l'erreur relative due à la quantification d'entrée.

### 3. 2. L'ERREUR DE COEFFICIENTS

En annulant dans la relation (11) l'erreur initiale  $E(0)$  ainsi que le terme  $E_G(p+1)$  qui provient de la structure arithmétique, nous obtenons, par récurrence, la valeur efficace de l'erreur sur le résultat

$$(20) \quad B_C = \sqrt{E(10) 2^{2e}}$$

introduite par la quantification des coefficients de la table trigonométrique.

Cependant, en faisant intervenir les relations (15) et (18),  $B_C$  s'écrit par récurrence :

$$(21) \quad B_C = \sqrt{E(10) 2^{2e}} = 16 \sqrt{E[|y_0|^2] \sum_{p=0}^9 |\varepsilon(W_p)|^2}$$

soit finalement :

$$(22) \quad \frac{B_C}{S} = \left[ \frac{1}{2} \sum_{p=0}^9 |\varepsilon(W_p)|^2 \right]$$

Le rapport signal sur bruit lié à l'erreur de coefficient ne dépend que de la qualité de la table trigonométrique mise en œuvre.

### 3. 3. L'ERREUR ARITHMÉTIQUE

L'erreur arithmétique  $B_A = E_G(10) 2^{2e}$  est nettement plus délicate à exprimer car elle fait intervenir, par le biais de l'amplitude maximale des données, la distribution des gains  $G(p)$  ainsi que les particularités de chacune des étapes. L'erreur arithmétique est donc liée de manière étroite à la réalisation matérielle des opérations d'addition, de multiplication et d'amplification des données (association MAC-décaleurs dans ce système). La comparaison à l'expérience montrera qu'une telle étude permet de cerner de manière nettement plus précise l'évolution de l'erreur arithmétique en fonction de l'énergie du tableau de données d'entrée.

#### 3. 3. 1. Évolution de l'amplitude maximale des données dans une étape

Le gain  $G(p)$  intervient explicitement dans l'équation générale de propagation des erreurs (11), mais également par l'intermédiaire de  $E_G(p+1)$ , via le processus de décalage qu'il déclenche.  $G(p)$  ne dépend que de la valeur maximale des éléments de sortie d'une étape, aussi avons-nous simulé, pour des données d'entrée uniformément distribuées en amplitude, l'évolution de l'amplitude maximale des données (parties réelle ou imaginaire) en fonction de l'indice de l'étape. Le résultat est illustré aux figures 3 et 4 où

$$(23) \quad \hat{R}(p) = \frac{\hat{X}(p+1)}{\hat{X}(p)}$$

avec

$$\hat{X}(p) \triangleq \max_{i=0 \rightarrow N-1} \{ |\operatorname{Re}(x_p(i))|, |\operatorname{Im}(x_p(i))|, \\ |\operatorname{Re}(y_p(i))|, |\operatorname{Im}(y_p(i))| \}$$

« gain en maximum » de la passe  $p$ , est le rapport des composantes maximales des vecteurs  $X_{p+1}$  et  $X_p$ , obtenues en imposant  $G(p) = 1$ .

Les tracés de la figure 4 montrent que la valeur moyenne de  $\hat{R}(p)$  vaut  $\sqrt{2}$ . De plus, une forte valeur

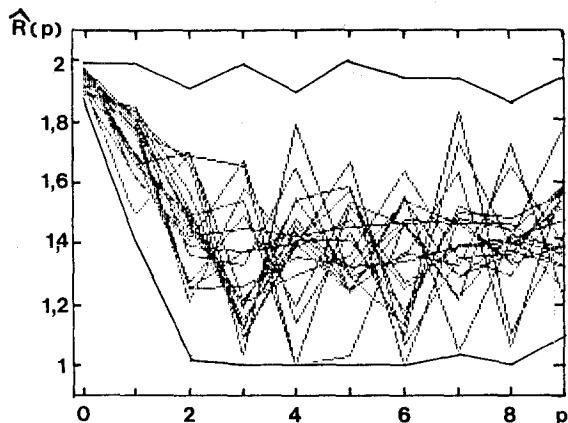


Fig. 3. — Représentation de la variation des « gains en maximum »  $\hat{R}(p)$  au cours de vingt tests (.....), tracé des valeurs maximales et minimales enregistrées sur un grand nombre de fichiers de données (—).

Fig. 3. — Maximum data pass amplification  $\hat{R}(p)$  for twenty different data files (.....). The continuous lines show the maximum and minimum  $\hat{R}(p)$  values.

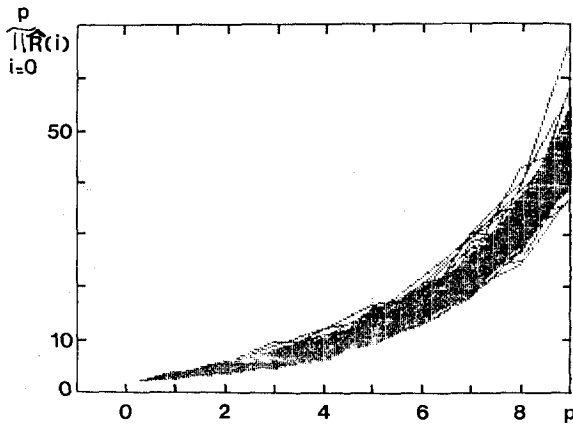


Fig. 4. — Tracé des produits successifs des « gains en maximums » et de leur enveloppe en fonction du numéro de l'étape.

Fig. 4. —  $\hat{R}(p)$  product evolution as a function of the pass index  $p$  for a large number of data files.

de  $\hat{R}(p)$  est généralement suivie d'une faible valeur de  $\hat{R}(p+1)$ . Les tracés reportés en figure 4 témoignent de ce fait en indiquant l'allure régulièrement croissante du produit des gains  $\hat{R}(p)$ . Ces résultats seront exploités lors de la confrontation à l'expérience de l'équation de propagation des différents types d'erreur (11) où l'approximation

$$(24) \quad \hat{R}(p) \approx \sqrt{2} \quad \text{pour } p=1, 2, \dots, 9 \text{ et } \hat{R}(0)=2$$

sera également utilisée.

Le rôle des décaleurs étant de maintenir  $0,5 \leq \hat{X}(p+1) < 1$ , l'amplitude maximale en sortie d'une étape sera donc liée à celle des entrées par la relation (25) qui permet de connaître l'évolution des gains  $G(p)$  :

$$(25) \quad \hat{X}(p+1) = G(p) \hat{R}(p) \hat{X}(p)$$

soit :

$$(26) \quad \hat{X}(p+1) = \hat{X}(0) \prod_{i=0}^p G(i) \hat{R}(i)$$

3.3.2. Les particularités des différentes étapes

Les dix étapes successives  $p=0, \dots, 9$  qui mènent à la connaissance du spectre d'un tableau d'entrée réel se distinguent par la valeur des coefficients  $W_p = \cos \theta - j \sin \theta$  utilisés. En particulier, pour  $W_p = 1$  et  $-j$ , les opérations effectuées par le multiplieur-accumulateur se résument à une addition ou à une soustraction. En outre, les données d'entrée étant réelles, la partie imaginaire du tableau complexe d'entrée est nulle et tous les résultats qui sont issus de  $W_p = 1$  sont aussi à partie imaginaire nulle.

Ces considérations illustrées en figure 5 pour  $N=16$ , montrent qu'il existe une proportion  $K_1$  de résultats d'étapes nuls :

$$(27) \quad K_1 = 2^{-(p+1)}$$

pour lesquels l'erreur de troncature est évidemment nulle et une proportion  $K_2$  de résultats non nuls :

$$(28) \quad K_2 = \text{Inf}(2^{p-1}, 3 \cdot 2^{-(p+1)})$$

obtenus par simples sommes ou différences pour lesquels le moment d'ordre deux moyen de l'erreur arithmétique noté  $E_G(p+1, K_2, Q_p)$  dépend par contre du nombre de bits  $Q_p$  de la partie fractionnaire des données d'entrée de la passe  $p$ .

3.3.3. L'erreur de quantification de l'association MAC-décaleur

Les vecteurs  $X_p$  successifs sont soumis aux opérations (3) et (4) de calcul des papillons effectués par les MAC. Leur format d'entrée est de  $(15+1)$  bits et leur résultat interne est donné avec 30 bits à droite du point décimal dont nous en relisons 14 obtenus par arrondi. Par ailleurs, dans le but de garantir la dynamique nécessaire aux calculs du papillon, nous faisons aussi provision de 3 bits à gauche de ce point décimal. Les décaleurs de notre arithmétique flottante par blocs ( $G(p)=1/4, 1/2, 1, 2$ ) ramènent ensuite l'ensemble des résultats au format d'entrée de  $(15+1)$  bits des MAC.  $\hat{X}(p+1)$  étant l'amplitude maximale des résultats de l'étape  $p$  avant décalage, les cas suivants peuvent se présenter :

- $\hat{X}(p+1) \geq 2$  : le gain utilisé est  $G(p)=1/4$  et les résultats subissent une opération d'arrondi à  $(16+1)$  bits suivie d'une troncature à  $(15+1)$  bits;
- $1 \leq \hat{X}(p+1) < 2$  : le gain employé est  $G(p)=1/2$ ; les résultats sont obtenus par arrondi à  $(15+1)$  bits;
- $0,5 \leq \hat{X}(p+1) < 1$  : le gain est égal à l'unité et les résultats sont arrondis à  $(14+1)$  bits, le 15-ième étant nul;
- $\hat{X}(p+1) < 0,5$  : le gain  $G(p)=2$  est alors appliqué, améliorant efficacement la dynamique de calcul (§ 5). Les résultats sont obtenus dans ce cas par arrondi à  $(13+1)$  bits, les 14 et 15-ième bits étant nuls.

Pour calculer les moments d'ordre 2 des erreurs de quantification associées à ces décalages, il faut également considérer les particularités des différentes passes évoquées au paragraphe précédent.

Ainsi, tout résultat obtenu par addition ou soustraction l'est à partir de données elles-mêmes obtenues par addition ou soustraction. Si ces données sont quantifiées sur  $(Q_p+1)$  bits, le résultat est au plus sur  $(Q_p+2)$  bits (soit  $Q_p$  bits à droite et 2 bits à gauche de la virgule binaire). Aussi, si  $Q_p$  est inférieur ou égal à 14, le résultat en sortie du MAC ( $14+3$  bits) n'est pas arrondi, tandis que si  $Q_p=15$ , il est amputé d'un bit par arrondi. Dans l'un et l'autre cas, son passage dans les décaleurs lui fait éventuellement subir, selon le gain  $G(p)$  utilisé, une troncature supplémentaire. Le tableau I exprime les valeurs des moments d'ordre 2 des erreurs correspondant aux situations rencontrées (annexe 1). La connaissance de l'erreur qui est associée aux  $K_2$  résultats implique celle du nombre de bits  $(Q_p+1)$  des données présentées à la  $p$ -ième étape, lié aux gains  $G(p)$  par les relations suivantes :

$$(29)$$

$$Q_{p+1} = \text{inf}(15, Q_{p+1}) \quad \text{si } G(p) = \frac{1}{4} \text{ ou } \frac{1}{2}$$

$$(30) \quad Q_{p+1} = Q_p \quad \text{si } G(p) = 1$$

$$(31) \quad Q_{p+1} = Q_p - 1 \quad \text{si } G(p) = 2$$

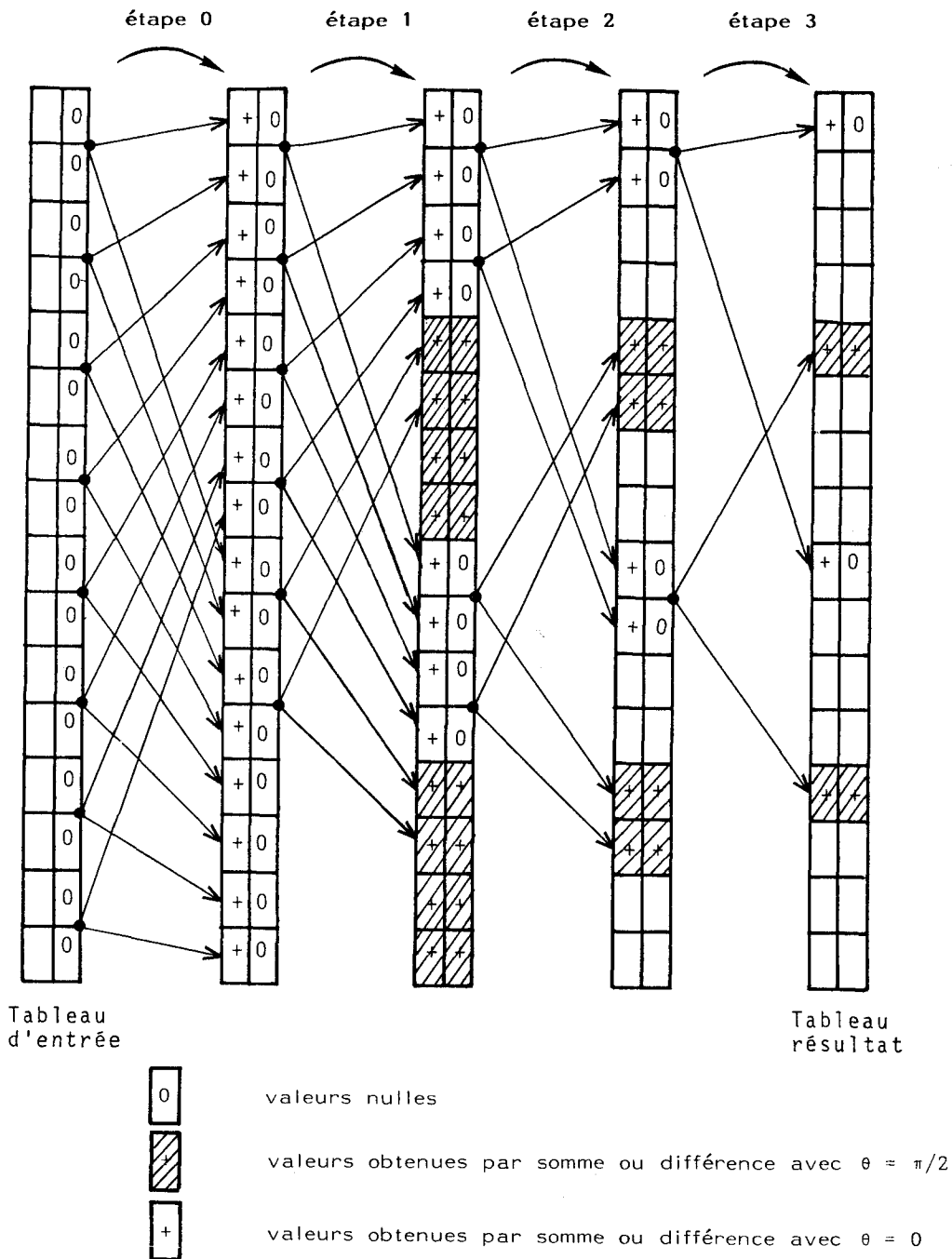


Fig. 5. — Évolution de la partie réelle (à gauche) et de la partie imaginaire (à droite) des données d'une TFR (N=16) à géométrie constante et dédoublement temporel pour des données d'entrée réelles.

Fig. 5. — Real (left) and imaginary (right) data passes using the DIT in place FFT algorithm. The considered input data block is real.

Munis de ces précisions, le terme d'erreur due aux décalages de l'équation (11) se réécrit :

$$(32) \quad E_G(p+1) = K_2 E_G(p+1, K_2, Q_p) + (1 - K_1 - K_2) E_G(p+1, 1 - K_1 - K_2)$$

où  $E_G(p+1, 1 - K_1 - K_2)$  représente le moment d'ordre deux moyen de l'erreur arithmétique pour les papillons non particularisés.

#### 4. Étude expérimentale des trois types d'erreur

La mesure des erreurs repose sur la méthode générale suivante. Un calculateur (HP85), opérant en virgule flottante, calcule « exactement » la transformée de Fourier rapide d'un tableau de données également présentées au processeur rapide. Les composantes de

TABLEAU I

Valeurs des moments d'ordre 2 des erreurs issues de l'association MAC-décaleurs.  $Q_p$  désigne le nombre de bits de la partie fractionnaire des données d'entrée de la passe  $p$  (§ 3.3.3) qui dépend de la valeur prise par  $G(p-1)$ .

Valeur crête	$G(p)$	$E_G(p+1, K_1)$	$E_G(p+1, K_2, Q_p)$			$E_G(p+1, 1-K_1-K_2)$
			$Q_p \leq 13$	$Q_p = 14$	$Q_p = 15$	
$\hat{X}(p+1) \geq 2$ . . . . .	$\frac{1}{4}$	0	0	$\frac{1}{8} 2^{-30}$	$\frac{3}{32} 2^{-30}$	$\frac{7}{48} 2^{-30}$
$1 \leq \hat{X}(p+1) < 2$ . . . . .	$\frac{1}{2}$	0	0	0	$\frac{1}{8} 2^{-30}$	$\frac{1}{12} 2^{-30}$
$0,5 \leq \hat{X}(p+1) < 1$ . . . . .	1	0	0	0	$\frac{1}{8} 2^{-28}$	$\frac{1}{12} 2^{-28}$
$\hat{X}(p+1) < 0,5$ . . . . .	2	0	0	0	$\frac{1}{8} 2^{-26}$	$\frac{1}{12} 2^{-26}$

bruit  $B_E$ ,  $B_C$  et  $B_A$  sont analysées en les rendant tour à tour dominantes par rapport aux deux autres par des interventions matérielles.

4.1. L'ERREUR D'ENTRÉE

L'erreur d'entrée a été évaluée en (19). Sa valeur dépend de la taille  $N$  du tableau d'entrée et de son moment d'ordre deux  $E(0)$ . Pour évaluer  $E(0)$ , rappelons que l'élément courant  $i$  du tableau d'entrée  $X_0$  est obtenu (§ 2.1) par multiplication d'une donnée  $d_i$  par le facteur de pondération correspondant  $f_i$ . Les quantités  $d_i$  et  $f_i$  sont quantifiées, de plus le produit  $d_i \cdot f_i$  est soumis à un arrondi lié au format de sortie des données du multiplieur accumulateur de prétraitement (§ 2.1) -  $Q_0=14$  -. Le moment d'ordre deux  $E(0)$  de l'erreur d'entrée du processeur de Fourier s'écrit donc :

$$(33) \quad E(0) = |F|^2 E(|\varepsilon(d_i)|^2) + |\varepsilon(f_i)|^2 E(|d_i|^2) + E(Q)$$

où  $E(|d_i|^2)$  est l'énergie moyenne par point du tableau présenté au préprocesseur et  $E(|\varepsilon(d_i)|^2)$  le moment d'ordre deux de l'erreur correspondant,

$|F|^2 = 1/N \sum_{i=0}^{N-1} |f_i|^2$  est la valeur quadratique moyenne de la fenêtre et  $|\varepsilon(f_i)|^2$  l'erreur quadratique moyenne correspondante,  $E(Q)$  est le moment d'ordre deux de l'erreur d'arrondi introduite par le processeur de prétraitement.

Pour vérifier les prévisions théoriques relatives à l'erreur d'entrée (19), nous avons comparé le spectre exact d'un tableau de données, obtenu à l'aide du micro-ordinateur HP85, avec les spectres du même tableau, obtenus par notre processeur TFR, sur des données quantifiées par arrondi à  $(n+1)$  bits, variant de 7 à 16. Les résultats, illustrés en figure 6, montrent un très bon accord pour  $(n+1) \leq 9$ . Par contre, pour  $n > 9$ , l'erreur arithmétique  $B_A$  vient à dominer.

4.2. L'ERREUR DE COEFFICIENTS

L'erreur de coefficient a été étudiée au paragraphe 3.2. Sa valeur dépend de l'énergie du tableau d'entrée et de la somme des moments d'ordre deux  $|\varepsilon_{W_p}|^2$  des erreurs de quantification affectant

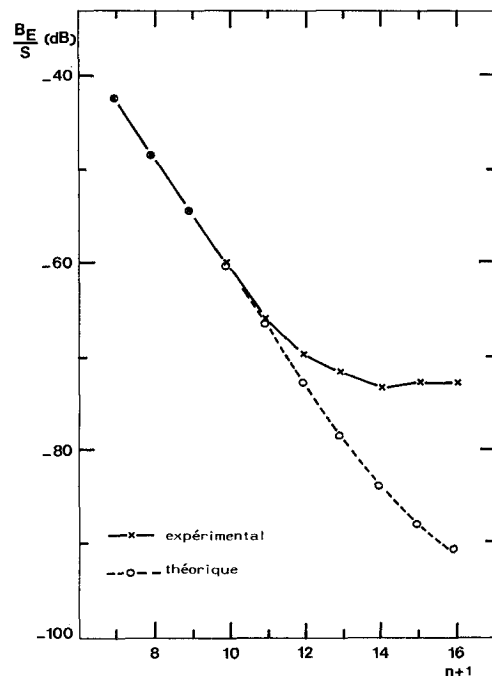


Fig. 6. - Valeurs théoriques et expérimentales de l'erreur d'entrée  $B_E$  en fonction du format des données. Le tableau D correspond à des données uniformément réparties sur la dynamique maximale du processeur de Fourier (la fenêtre  $F$  est rectangulaire de valeur unité).

Fig. 6. - Calculated and experimental input errors as a function of the data length  $n+1$  (rectangular window). The data set is distributed uniformly on the input range of the FFT processor.

les données trigonométriques intervenant dans chaque étape. Notre réalisation utilise une forme négative sur 15+1 bits de la table trigonométrique qui permet de réaliser exactement  $\cos 0$  et  $\sin(\pi/2)$  : les valeurs correspondantes de  $|\varepsilon(W_p)|^2$  sont reportées sur le tableau II.

TABLEAU II

Moment d'ordre deux  $|\varepsilon(W_p)|^2$  de l'erreur de quantification (à 15+1 bits) des coefficients trigonométriques intervenant en passe  $p$ .

$ \varepsilon(W_0) ^2 = 0$	$ \varepsilon(W_4) ^2 = 1,16 \cdot 10^{-10}$
$ \varepsilon(W_1) ^2 = 0$	$ \varepsilon(W_6) ^2 = 1,34 \cdot 10^{-10}$
$ \varepsilon(W_2) ^2 = 2,07 \cdot 10^{-10}$	$ \varepsilon(W_7) ^2 = 1,25 \cdot 10^{-10}$
$ \varepsilon(W_3) ^2 = 1,76 \cdot 10^{-10}$	$ \varepsilon(W_8) ^2 = 1,37 \cdot 10^{-10}$
$ \varepsilon(W_4) ^2 = 1,78 \cdot 10^{-10}$	$ \varepsilon(W_9) ^2 = 1,36 \cdot 10^{-10}$



Pour vérifier ce résultat théorique (22), nous avons successivement déconnecté les 8 bits de poids faible de la mémoire programmable contenant la table trigonométrique afin d'en augmenter l'erreur par troncature à  $m+1$  bits des coefficients initialement arrondis à 16 bits. Ayant la possibilité de relire les données d'entrée quantifiées  $X_0$  présentées au processeur TFR, nous avons pu comparer le spectre fourni par notre processeur avec le spectre « exact » des mêmes données  $X_0$ , obtenu par calcul avec des coefficients exacts. Ces résultats présentés en figure 7 montrent que dans le domaine où l'erreur arithmétique est négligeable, les courbes théoriques et expérimentales sont parallèles, avec cependant un décalage d'environ 3 dB. L'erreur de coefficient est mal définie par le modèle statistique utilisé, car des effets de corrélation amènent à l'apparition d'harmoniques (effets non linéaires) lors de l'analyse de formes sinusoïdales (16). Son influence étant négligeable devant les bruits d'entrée et arithmétique, nous n'avons pas poussé plus avant ces investigations.

### 4.3. L'ERREUR ARITHMÉTIQUE

Afin de vérifier la validité de l'erreur arithmétique exprimée en (11) et (32), nous avons procédé à des calculs de spectres sur des données quantifiées  $X_0$  d'énergie variable. Par comparaison entre le spectre calculé par notre processeur et le spectre « exact », obtenu comme pour l'erreur de coefficients par relecture des données présentées au processeur TFR, nous avons isolé l'erreur arithmétique expérimentale.

La figure 8, qui représente les évolutions des courbes théoriques et expérimentales, montre que le modèle fournit une excellente estimation de l'erreur quadratique moyenne. Elle montre également que l'erreur de coefficients est toujours négligeable tandis que l'erreur d'entrée contribue de manière dominante tant que

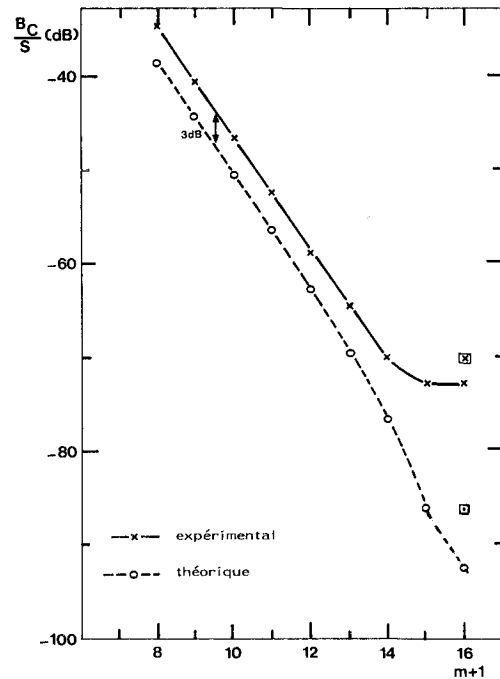


Fig. 7. — Valeurs théoriques et expérimentales de l'erreur de coefficients exprimée en fonction de la troncature à  $(m+1)$  bits d'une table «  $-\cos$ ,  $-\sin$  » déjà arrondie à 16 bits. L'inflexion de la courbe théorique pour les fortes valeurs de  $m+1$  correspond au passage d'une erreur de troncature à une erreur d'arrondi. Les signes □ et ⊠ représentent respectivement les erreurs théoriques et expérimentales liées à l'utilisation de la table «  $+\cos$ ,  $+\sin$  ».

Fig. 7. — Calculated and experimental coefficient errors as a function of the  $m+1$  bits truncation of the «  $-\cos$  » and «  $-\sin$  » trigonometric table at first rounded to 16 bits. The symbols □, ⊠ respectively refer to the calculated and experimental coefficient error making use of «  $+\cos$  » and «  $+\sin$  » trigonometric Table.

l'énergie du tableau à traiter est inférieure au centième de l'énergie maximale qu'il est susceptible de prendre. Il est ainsi possible, connaissant l'énergie du signal

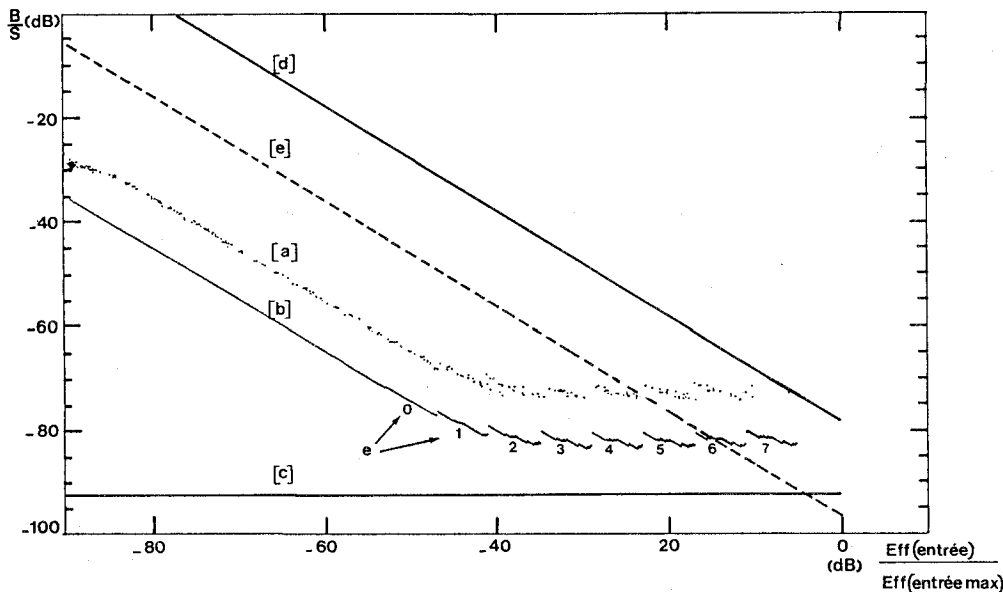


Fig. 8. — Évolution des composantes de bruit du système en fonction de la valeur efficace du signal d'entrée : bruit arithmétique ( $B_A/S$ ) : résultats expérimentaux (a) et théoriques (b) [pour plus de clarté, la courbe (b) est décalée de 10 dB vers le bas]; bruit d'arrondi de la table trigonométrique ( $B_C/S$ ) (c); bruit d'entrée ( $B_E/S$ ) pour des données quantifiées à 12 bits (d) et à 16 bits (e).

Fig. 8. — FFT noise components as a function of the input signal r. m. s. value. Curves (a) and (b) respectively refer to the arithmetic roundoff noise [for clarity, the (b) curve is 10 dB downward shifted]. Curve (c) shows the trigonometric coefficients noise. Curves (d) and (e) show the input noise effect for 12 bits and 16 bits data respectively.

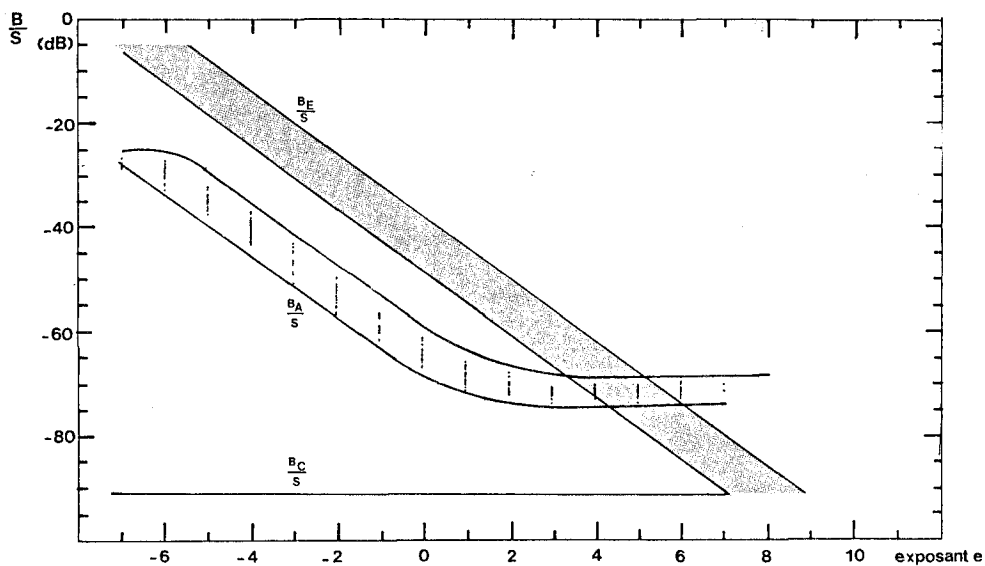


Fig. 9. — Évolution des rapports B/S associés aux erreurs de coefficients, arithmétique et d'entrée [fenêtre rectangulaire,  $(n+1)=16$ ] : représentation des enveloppes théoriques et de l'erreur arithmétique expérimentale.

Fig. 9. — Noise to signal ratio as a function of the output data block exponent ( $e$ ). The  $B_E/S$ ,  $B_A/S$ ,  $B_C/S$  curves refer respectively to the input, arithmetic roundoff and coefficient noise sources.

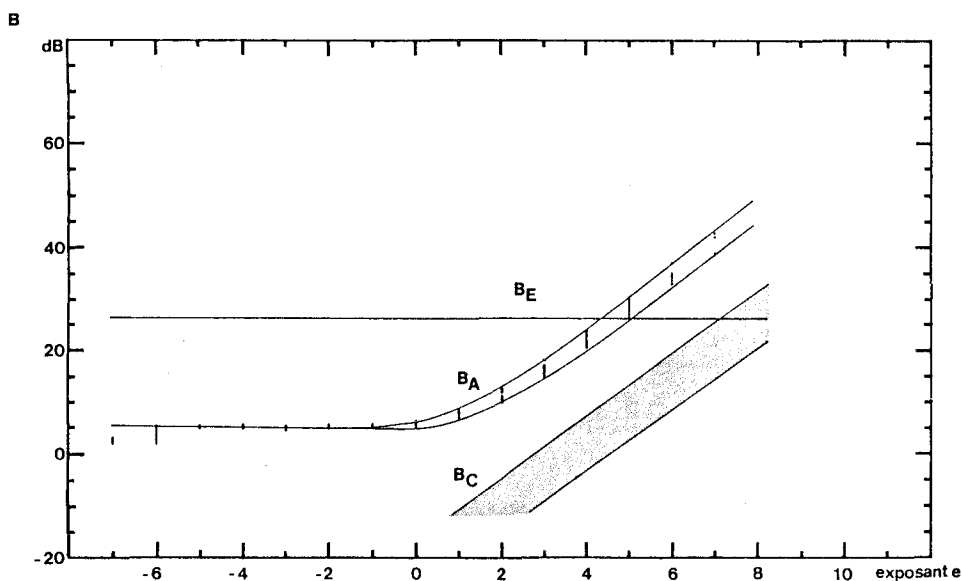


Fig. 10. — Évolution des erreurs de coefficients, arithmétique et d'entrée ( $n+1=16$ ); leur valeur efficace est normalisée par rapport au plus petit résultat fourni par le processeur.

Fig. 10. — Input, arithmetic and coefficient noise mean square values as a function of the output block exponent  $e$ . This noise is expressed in  $2^{-15}$  of the quantization level.

d'entrée d'estimer a priori le bruit de calcul qui en résultera. De même, il est intéressant de remarquer que l'exposant du résultat est un paramètre qui fournit, comme en témoignent les courbes présentées à la figure 9, une excellente estimation de la fourchette d'erreur. En effet, les valeurs expérimentales qui relient l'exposant de sortie  $e$  à la valeur efficace d'un signal d'entrée aléatoire centré forment une enveloppe large de 11 dB. Le tracé de l'erreur arithmétique absolue rapportée au pas de quantification du résultat (fig. 10) montre qu'elle est constante et faible pour  $e \leq -1$ , tandis qu'elle croît à raison de 6 dB pour une variation d'une unité de l'exposant pour  $e \geq 3$ . La largeur de l'enveloppe, qui est alors de 5 dB, corres-

pond à la dynamique du produit des gains en maximum utilisés (fig. 4),

$$\prod_{p=0}^8 \hat{R}(p)$$

la dernière passe de gain constant ne faisant pas intervenir  $\hat{R}(9)$ . Remarquons également que  $B_A$  est faible pour  $e < 1$  car l'exposant qui lui est associé est également. En réalité, c'est dans cette zone que la mantisse de l'erreur absolue est la plus grande; il en est de même de l'erreur relative (fig. 9).

L'allure de ces courbes s'explique aisément. Pour  $e$  faible, les opérations de décalage correspondent toutes

à des multiplications par 2 affectées d'un fort coefficient de propagation d'erreur et d'une erreur de quantification  $E_G(p)$  importante (tableau I). Lorsque la valeur efficace du signal d'entrée croît, le nombre de décalages de gain 2 décroît, accompagné d'erreurs moindres. Pour un signal d'entrée d'énergie suffisamment importante, les gains des différentes passes alternent entre les valeurs 1 et 1/2 et la mantisse de l'erreur absolue devient alors constante.

Afin de connaître l'importance de la dégradation de la précision du spectre d'un signal comportant une partie continue, nous avons étudié le comportement du processeur TFR envers un bruit blanc non centré et vérifié les points suivants :

- dans la mesure où la valeur moyenne du signal domine sa partie alternative, l'exposant est incrémenté d'une unité lorsque cette valeur moyenne double; il reflète l'évolution de la partie continue du signal;
- le bruit arithmétique  $B_A$  ne dépend que de l'exposant de sortie et donc de l'agencement des décalages. Sa variation en fonction de l'exposant est la même qu'en figure 10;
- il s'ensuit, ce que nous avons vérifié, qu'en dehors du domaine où la valeur moyenne  $\bar{x}$  du signal est petite comparée à sa partie alternative et pour lequel  $B_A$  et  $B_A/S$  sont constants, les bruits arithmétiques absolus et relatifs sur la partie alternative du signal croissent comme sa valeur moyenne à raison de 6 dB toutes les fois que cette dernière est doublée.

La structure de notre processeur TFR est celle d'un processeur à virgule flottante par blocs, amélioré par le fait qu'il amplifie un signal de faible énergie afin de perdre un minimum d'informations par effet de troncature. Sans prétendre atteindre les performances des systèmes opérant en virgule flottante, il est intéressant d'en comparer leurs erreurs arithmétiques respectives. Pour une TFR à virgule flottante disposant d'une mantisse de (15+1) bits, le rapport  $B_A/S$  est constant et égal à -84 dB [4]. Son intérêt est donc limité dans la mesure où l'erreur d'entrée vient la dominer.

## 5. L'effet de choix matériels sur le bruit de calcul

La représentation des nombres en complément à deux couvre la gamme  $(-1, 1-2^{-b})$  où  $2^{-b}$  est le pas de quantification. Pour simplifier la structure du processeur de Fourier, nous n'avons pas traité de façon particulière les papillons opérant sur les angles 0 et  $\pi/2$ . Aussi, les multiplications par  $\cos(0)$  et  $\sin(\pi/2)$  sont elles effectuées. L'adoption de tables « +cos, +sin » entache alors ces opérations d'erreurs; il n'en va pas de même pour la table négative (« -cos, -sin ») qui permet de réaliser exactement la multiplication par 1 via la multiplication par -1. Pour  $(m+1) = 16$ , l'erreur de coefficients issue de nos cal-

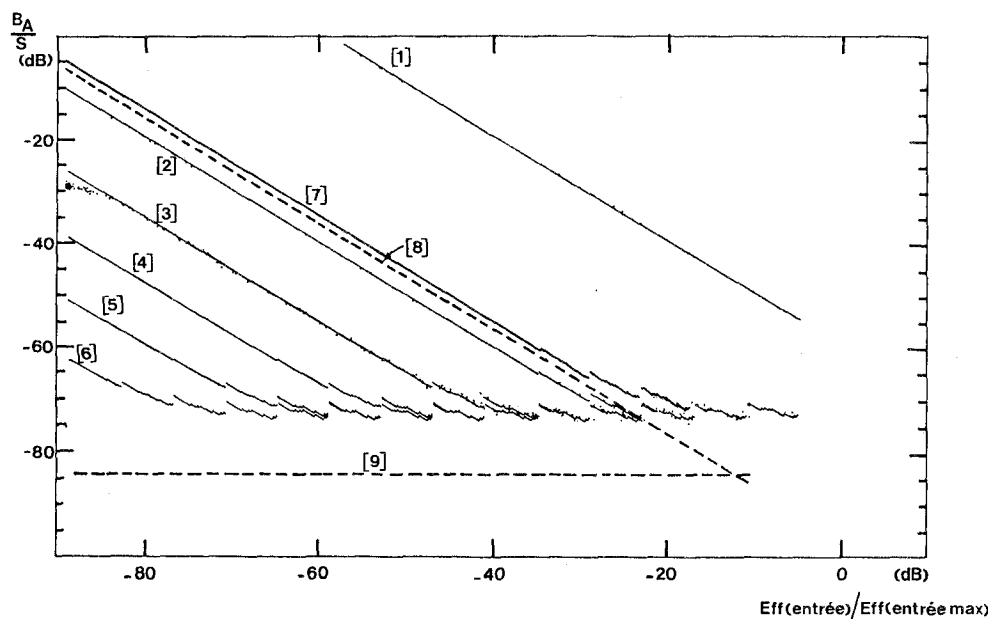


Fig. 11. — Superposition des erreurs arithmétiques théoriques et expérimentales, dans la gamme de variation accessible au gain d'étape  $G(p)$ . La discontinuité des courbes marque le changement d'exposant.

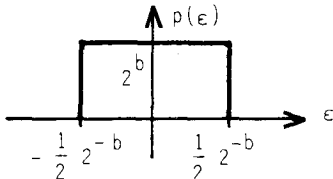
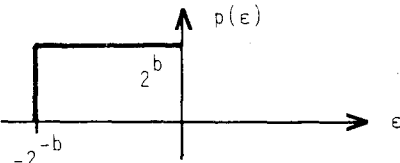
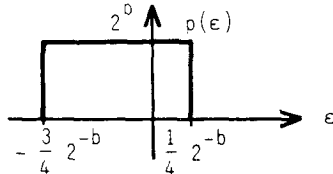
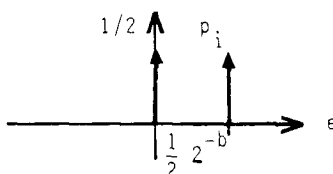
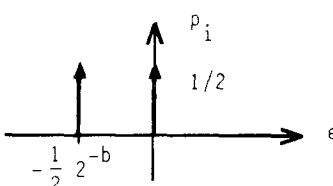
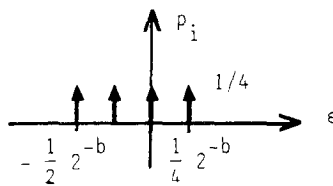
- [1]  $G(p) = \{1/4, 1/2\}$ ;
- [2]  $G(p) = \{1/4, 1/2, 1\}$ ;
- [3]  $G(p) = \{1/4, 1/2, 1, 2\}; \dots$ ;
- [6]  $G(p) = \{1/4, 1/2, 1, 2, 4, 8, 16\}$ ;
- [7]  $B_A/S$  avec table « +cos, +sin »;
- [8]  $B_E/S$  (16 bits);
- [9]  $B_C/S$  (16 bits).

Fig. 11. — Arithmetic roundoff noise  $B_A/S$  (calculated and experimental values) as a function of the input r. m. s. data value. The different curves show the influence of  $G(p)$ , the gain of a pass range :

- [1]  $G(p) = \{1/4, 1/2\}$ ;
- [2]  $G(p) = \{1/4, 1/2, 1\}$ ;
- [3]  $G(p) = \{1/4, 1/2, 1, 2\} \dots$

Curve [7] shows the contrary effect of a positive trigonometric table implementation. Curves [8] and [9] drawback the input ( $B_E/S$ ) and coefficient ( $B_C/S$ ) noise terms for 16 bits input data. The discontinuities of these curves mark a change of the output exponent value  $e$ .

TABLEAU III

Quantification initiale de M	Quantification effectuée sur M	Densités de probabilités $p(\epsilon)$ et $p_i$ associées	(Q)
Non quantifié .....	Arrondi à $(b+1)$ bits		$\frac{1}{12} 2^{-2b}$
Non quantifié .....	Troncature à $(b+1)$ bits		$4 \left( \frac{1}{12} 2^{-2b} \right)$
Non quantifié .....	Arrondi à $(b+2)$ bits, puis troncature à $(b+1)$ bits		$\frac{7}{4} \left( \frac{1}{12} 2^{-2b} \right)$
Quantifié sur $(b+2)$ bits .....	Arrondi à $(b+1)$ bits		$\frac{3}{2} \left( \frac{1}{12} 2^{-2b} \right)$
Quantifié sur $(b+2)$ bits .....	Troncature à $(b+1)$ bits		$\frac{3}{2} \left( \frac{1}{12} 2^{-2b} \right)$
Quantifié sur $(b+3)$ bits .....	Arrondi à $(b+2)$ bits, puis troncature à $(b+1)$ bits		$\frac{9}{8} \left( \frac{1}{12} 2^{-2b} \right)$

culs est ainsi 2 fois plus faible pour cette seconde solution (fig. 7), car pour une table trigonométrique positive, il n'existe plus aucune addition exacte, ce qui revient à annuler les coefficients  $K_2$  définis au paragraphe 4. Notre modèle d'erreur permet alors d'obtenir le résultat porté en figure 11 où nous constatons que l'erreur arithmétique devient aussi importante que l'erreur d'entrée.

La structure matérielle de notre processeur repose sur une dynamique de gain de passe de 8 :  $G(p) = \{1/4, 1/2, 1, 2\}$ . Pour justifier ce choix et vérifier la conformité de notre modèle d'erreur dans le cas d'autres structures, il nous était aisé, tant expérimentale-

ment qu'au niveau du modèle de modifier la gamme de variation de  $G(p)$ . La figure 11 qui présente les résultats obtenus nous montre que la dynamique  $\{1/4, 1/2\}$ , associée à l'absence de décalages (1), n'est pas acceptable car elle conduit à une erreur arithmétique beaucoup trop importante vis-à-vis de l'erreur d'entrée. La solution que nous avons retenue amène le bruit arithmétique à un niveau négligeable par rapport au bruit d'entrée minimal, elle est donc suffisante. Les configurations plus complexes (cinq décalages et plus) ne sont d'aucune utilité, la réduction de bruit qui leur est associée étant totalement masquée par l'erreur d'entrée.

## 7. Conclusion

Cette étude a montré l'importance de la connaissance exacte de la structure du processeur que l'on étudie et du format des données. Elle a permis de construire le modèle d'erreurs et d'obtenir des résultats théoriques et expérimentaux en accord presque parfait. Par modification du modèle, il a été possible de connaître les conséquences d'une évolution de la structure sur la précision du résultat. Pour la structure que nous avons réalisée, nous avons remarqué que dans le cas de signaux d'entrée aléatoires de valeur moyenne nulle, chacune des erreurs — erreur d'entrée, erreur arithmétique, erreur de coefficients — se situe à l'intérieur d'une enveloppe déterminée en fonction de l'exposant  $e$  affectant le résultat. L'analyse expérimentale de l'erreur arithmétique relative maximale (définie par le rapport erreur maximale/composante maximale du spectre) a permis de vérifier que cette erreur, exprimée en fonction de  $e$ , caractérise autant les signaux de valeur moyenne nulle que ceux de valeur moyenne non nulle. Enfin, nous avons noté que pour des signaux sinusoïdaux ou en bande étroite, les erreurs arithmétiques sont inférieures à celles que nous avons obtenues dans notre étude.

Ainsi, la valeur de l'exposant de sortie permet :

- de donner une estimation de l'énergie d'un signal d'entrée de valeur moyenne nulle;
- d'estimer individuellement les erreurs d'entrée, arithmétique et de coefficients et leur somme quadratique.

De ces trois erreurs, l'erreur de coefficient est la moindre. S'il est essentiel, sous peine d'augmenter considérablement l'erreur arithmétique, de choisir une table «  $-\cos$ ,  $-\sin$  », nous savons que l'erreur apportée par des coefficients arrondis à  $(15+1)$  bits n'a pas d'effet sur le résultat final. L'erreur arithmétique, plus importante, ajoute ses effets à l'erreur d'entrée qui est prépondérante pour des signaux fenêtrés de faible énergie, voire pour tous signaux quand la conversion analogique numérique se fait sur un nombre de bits « restreint » ( $\leq 12$ ).

Il apparaît que la cause fondamentale des erreurs d'une TFR est due à l'erreur d'entrée. Même dans le cas d'un calcul effectué avec une grande précision en virgule flottante, une erreur de quantification de moment d'ordre deux  $E(|\varepsilon(d_i)|)^2$ , donnée par la conversion analogique numérique, apportera toujours au résultat d'une TFR de dimension  $N$  une contribution d'erreur de valeur efficace au moins égale à  $NE(|\varepsilon(d_i)|)^2$ . Il est donc absolument nécessaire, lors de l'acquisition de données sous leur forme numérique, d'utiliser pleinement la dynamique du convertisseur, car le rapport bruit/signal augmente quand l'énergie du signal devient faible. Il y aurait avantage, dans le cas d'une TFR virgule flottante, à utiliser une conversion analogique numérique flottante, avec pleine résolution de conversion sur la mantisse [19]. Ainsi, seule une conversion analogique numérique de ce type peut réduire le bruit total de calcul du spectre d'un signal.

## Annexe 1

Cette annexe définit les différents types d'erreur de quantification rencontrés au cours du traitement de la TFR. Pour la représentation binaire en « virgule fixée » et en complément à 2, chaque nombre  $M$ ,  $-1 \leq M < 1$  est codé sur  $(b+1)$  bits et la virgule binaire est située immédiatement à droite du bit de poids le plus fort appelé « bit de signe ». Le poids  $2^{-b}$  du bit le plus faible représente le pas de quantification. Soit  $M_q$  la représentation numérique de  $M$ ; en considérant que l'erreur  $\varepsilon = M_q - M$  est de densité de probabilité  $p(\varepsilon)$  ou  $p_i$  constante, nous exprimons, au tableau III, son moment d'ordre 2 en fonction des types de quantification rencontrés.

*Manuscrit reçu le 28 mai 1986,  
version révisée le 23 juin 1987.*

## BIBLIOGRAPHIE

- [1] P. D. WELCH, A fixed-point fast Fourier transform error analysis, *IEEE Trans. Audio Electroacoust.* (Special Issue on Fast Fourier Transform), AU-17, 1969, p. 151-157.
- [2] A. V. OPPENHEIM et C. J. WEINSTEIN, Effects of finite register length in digital filtering and the fast Fourier transform, *Proc. IEEE (Invited Paper)*, 60, 1972, p. 957-976.
- [3] D. V. JAMES, Quantization errors in fast Fourier transform, *IEEE Trans. Acoustics, Speech Processing*, ASSP-23, 1975, p. 277-283.
- [4] C. J. WEINSTEIN, Roundoff noise in floating point fast Fourier transform computation, *IEEE Trans. Audio Electroacoust.*, AU-17, 1969, p. 209-215.
- [5] B. LIU et TRAN-THONG, Accumulation of roundoff errors in floating point FFT, *Transact. IEEE*, CAS-24, 1977, p. 132-143.
- [6] T. THONG et B. LIU, Fixed-points fast Fourier transform error analysis, *IEEE Trans. Acoustics, Speech, Signal Processing*, ASSP-24, 1976, p. 563-573.
- [7] W. R. KNIGHT et R. KAISER, A simple fixed-point error bound for the fast Fourier transformation, *IEEE-ASSP*, 27, 1979, p. 615-620.
- [8] D. R. REDDY et V. V. RAO, Error analysis of FFT of a Sparse Sequence, *Journal of Electrical and Electronics Engineering, Australia*, 2, 1982, p. 169-175.
- [9] S. PRAKASH et V. V. RAO, Fixed-point error analysis of Radix-4 FFT, *Signal Processing*, 3, 1981, p. 123-133.
- [10] K. WOLINSKI, Analysis of errors in mixed fast Fourier transform algorithms with decimation in time for fixed point arithmetic, *Signal Processing: Theories and application, EURASIP*, 1980, p. 589-594.
- [11] K. WOLINSKI, Analysis of errors in mixed fast Fourier transform algorithms with decimation in frequency for fixed point arithmetic, *ICASSP 82*, 3, 1982, p. 2089-2093.
- [12] U. HEUTE, A novel approach to DFT and FFT coefficient error analysis, *Arch. El. Uebertr. (AED)*, 33, 1979, p. 20-22.
- [13] U. HEUTE, Results of a deterministic analysis of FFT coefficient errors, *Signal Proces.*, 3, 1981, p. 321-331.

- [14] U. HEUTE et H. W. SCHUESSLER, FFT accuracy-new insights and a new point of view, *ICASSP 83*, 1983, p. 631-634.
- [15] P. S. MOHARIR, Bounds on error due to finite arithmetic in FFT computation, *Inst. Electronics and Telecom. Engrs.*, 6, 1983, p. 236-239.
- [16] D. W. TUFTS, H. S. HERSEY et W. E. MOSIER, Effects of FFT coefficient quantization on bin frequency response, *Proc. IEEE (Lett.)*, 60, 1972, p. 146-147.
- [17] P. FURON, Système multiprocesseur d'analyse spectrale en temps réel d'un flot continu de données acoustiques couplé au bus parallèle IEEE 488, *Thèse de docteur-ingénieur*, Université de Paris-Sud, Centre d'Orsay, 1984.
- [18] P. FURON et D. BLOYET, Système d'analyse de Fourier par recouvrement de blocs compatible IEEE 488, *Traitement du signal*, 2, n° 6, 1985, p. 487-495.
- [19] Micro Networks, *MN 5420*, 20 bit dynamic range floating-point A/D converter.