SEPTIEME COLLOQUE SUR LE TRAITEMENT DU SIGNAL ET SES APPLICATIONS

NICE du 28 MAI au 2 JUIN 1979

COMPARAISON ENTRE DEUX METHODES RECENTES DE DETERMINATION DE L'ERREUR DE QUANTIFICATION

F.G. WITTLING

Université Claude Bernard - Ecole Centrale de Lyon - 501 Avenue du 8 Mai - 69300 CALUIRE

RESUME

A.B. SRIPAD et D.L. SNYDER dans leur article d'Octobre 1978 (1) analysent l'erreur de quantification (= E.Q.) de la "mantisse" d'un nombre X, exprimé à l'aide des puissances entières de la base 2. La densité de probabilité (= D.D.P.) de la "mantisse" est exprimée en faisant intervenir les fonctions caractéristiques du logarithme de X. Il est montré la D.D.P. de la 'mantisse" lui est inversement proportionnelle sous certaines conditions de nullité des fonctions caractéristiques. L'entrée aléatoire serait alors voisine de la fonction gaussienne.

F.G. WITTLING (2), Octobre 1978, a conduit un calcul basé sur le modèle doublement non linéaire d'une caractéristique de transfert en escalier avec application à la transformation logarithmique. La D.D.P. de l'E.Q. est calculée dans l'hypothèse d'une entrée gaussienne, en tenant compte de la faible valeur relative du pas de quantification dans les cas pratiques. La D.D.P. de l'erreur de quantification est alors uniforme, conclusion qui peut aussi être déduite des travaux de Sripad et Snyder.

SUMMARY

A.B. SRIPAD and D.L. SNYDER, in their paper of October 1978 (1), investigate the quantization error (= Q.E.) of the "mantissa" of the number X, written with the help of entire powers of the basis 2. The probability density (= P.D.) of the mantissa is expressed by the aid of the characteristic function of log X. It is shown that the P.D. of the "mantissa" is the reciprocal value of this last one, under certain conditions of zeroing characteristic functions. The random input is then approximately the gaussian function.

F.G. WITTLING (2), Octobre 1978, has performed a calculus based on the so-called double nonlinear staircase model. This model was applied to the logarithmical transformation. The P.D. of the Q.E. is computed when assuming a gaussian input and taking into account the small value of the reduced quantization step. Such is the cas in practical implementation. The P.D. of the quantization error is then uniform. Thie conclusion can also be deduced from the present work of Sripad and Snyder.

COMPARAISON ENTRE DEUX METHODES RECENTES DE DETERMINATION DE L'ERREUR DE QUANTIFICATION

1. INTRODUCTION

Nous exposerons dans une première partie les conditions adoptées pour l'expression de l'erreur de quantification d'un convertisseur analogique-digital non linéaire. Dans le cas particulier d'une conversion logarithmique il est possible, dans des conditions d'approximation conformes à l'expérience, d'obtenir une expression analytique de la densité de probabilité de l'erreur de quantification. Cette densité de probabilité est alors pratiquement constante. La deuxième partie sera consacrée à un compte-rendu succinct de la démarche suivie par SRIPAD et SNYDER. Ces auteurs ont publié un article (1) paru en synchronisme avec un compte-rendu de nos propres travaux (2). Leur but, très différent du nôtre, est d'analyser l'erreur de quantification existant sur la "mantisse" d'un nombre, la dite mantisse résultant de l'écriture d'un nombre sous forme d'un produit de deux facteurs, l'autre facteur étant une puissance entière, positive ou négative de la base choisie.

Nous montrerons ensuite que la densité de probabilité de la mantisse, telle qu'elle est établie par Sripad et Snyder, permet d'aboutir à notre propre conclusion concernant l'uniformité de la densité de probabilité de l'erreur de quantification dans le cas d'une conversion analogique digitale à sortie logarithmique.

2. RECHERCHE D'UNE EXPRESSION ANALYTIQUE DE LA D.D.P. DE L'ERREUR DE QUANTIFICATION

2.1. Position du problème

L'erreur de quantification, inhérente à la conversion analogique-digitale peut être définie de plusieurs manières équivalentes. La différence la plus marquante provient du principe de la conversion. On distingue les erreurs par arrondi d'une part, et les erreurs par troncature, d'autre part. Nous envisagerons le cas de l'erreur par troncature, puisque c'est ce type d'erreur de quantification qui existe sur nos propres montages (4, 5) et plus particulièrement dans le cas que nous avons étudié, de la conversion logarithmique (6).

L'analyse de l'erreur de quantification est rendue possible par l'utilisation de traitements numériques. De telles méthodes, actuellement très largement utilisées, exigent des moyens informatiques importants. L'investissement en calcul, toujours onéreux, peut être évité, au moins partiellement si l'on dispose d'une solution analytique. Il en découle un intérêt toujours renouvelé pour ce type de solution. Mais dans l'état actuel des choses, elles ne peuvent être obtenues que dans un nombre restreint de cas.

2.2. Existence et définition de l'erreur de quantification

Les définitions de l'erreur de quantification sont fondées sur l'existence de la transformation d'une grandeur continuement variable en une grandeur de variation discontinue. Les différentes définitions (3) sont très semblables et un choix convenable ne fait qu'introduire de légers allègements dans les calculs.

La conversion analogique-digitale, faisant intervenir une quantification est, dans son essence même, une transformation non linéaire et peut être représentée par une caractéristique de transfert en escalier. Dans ce cas le pas de quantification porté sur la grandeur analogique d'entrée est constant. Mais on peut envisager un cas plus complexe tel qu'il est représenté par la figure 2.1. La quantification, de pas constant Δ en sortie, fait intervenir des variations de la tension d'entrée U, de la forme $U_N - U_{N-1}$, différentes pour chaque valeur de l'indice entier N.

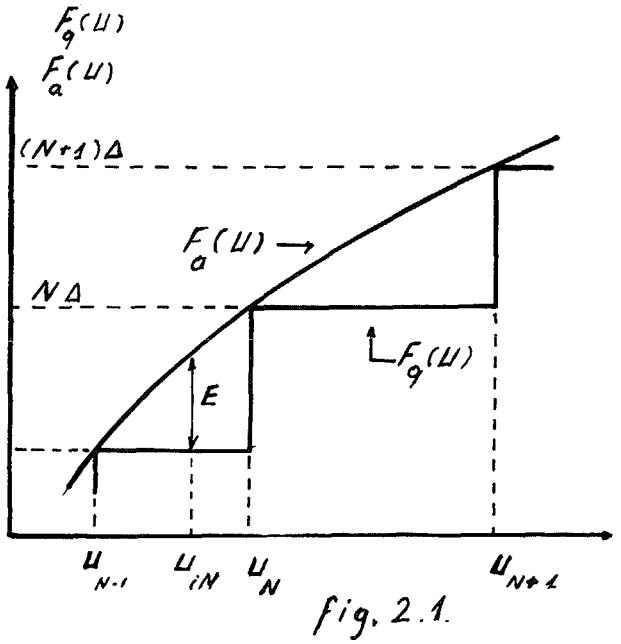

fig. 2.1.

Nous qualifierons une telle transformation de "doublement non linéaire". Tel sera, par exemple le cas de la conversion analogique digitale à sortie

COMPARAISON ENTRE DEUX METHODES RECENTES DE DETERMINATION DE L'ERREUR DE QUANTIFICATION

logarithmique. La caractéristique de transfert représentée par le diagramme de la figure 2.1. par la fonction $F_q(U)$ peut être obtenue à l'aide de deux opérations successives :

1 - Une transformation analogique-analogique $F_a(U)$ où $F_a(U)$ est une fonction non linéaire.

2 - Une quantification à pas constant Δ, opérée sur la fonction $F_a(U)$ prise comme variable d'entrée permettant l'obtention de $F_q(U)$.

Suivant les technologies utilisées, les deux opérations sont distinctes (6, 7) ou confondues (4).
La transformation doublement non linéaire est décrite par les relations :

$$N\Delta = F_a(U_N) \qquad (2.1)$$

$$F_q(U) = F_a(U_{N-1}) \qquad (2.2)$$

$$U_{N-1} \leqslant U < U_N \qquad (2.3)$$

Nous pouvons maintenant définir l'erreur de quantification en la définissant comme erreur absolue de troncature, toujours positive ou nulle. Par conformité aux relations 2.1 à 2.3 nous définirons l'erreur de quantification E par :

$$E = F_a(U) - F_q(U) \qquad (2.4)$$

$$0 \leqslant E < \Delta \qquad (2.5)$$

La figure 2.1. donne un exemple d'application des relations (2.1) à (2.5). Elle illustre le cas pratique de fonctionnement des circuits électroniques que nous avons eu l'occasion de mettre au point. Nous noterons cependant que pour ne pas allonger inutilement notre exposé nous ne traiterons pas le cas où la transformation présente un hystérésis, auquel cas l'erreur de quantification E varie entre $-\Delta$ et $+\Delta$. Le diagramme de la figure 2.2. rend compte de l'influence de l'hystérésis.

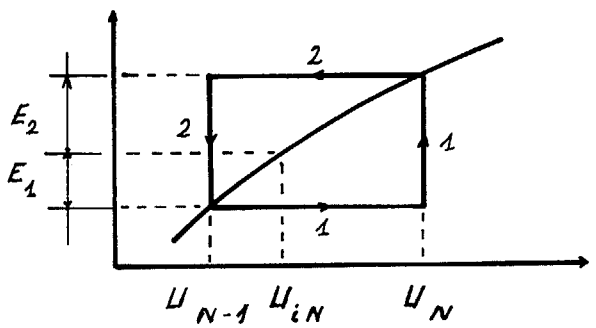

fig. 2.2.

2.3. Présentation générale du calcul

La densité de probabilité de l'erreur de quantification (D.D.P. de l'E.Q.) sera désignée par $f_E(E)$. Elle est calculable par extension du théorème de probabilité. Soit $f_{E,N}(E)$ la D.D.P. partielle de E dans l'intervalle (U_{N-1}, U_N). Nous pouvons écrire :

$$f_E(E) = \sum_{N=-\infty}^{\infty} f_{E,N}(E) \qquad (2.6)$$

A chaque valeur de E correspond une valeur U_{iN} de l'entrée U, prise dans chaque intervalle (U_{N-1}, U_N), et satisfaisant, d'après la relation (2.4), à :

$$E = F_a(U_{iN}) - (N-1)\Delta \qquad (2.7)$$

La valeur de U_{iN} s'obtient en utilisant la fonction inverse de $F_a(.)$, écrite $F_a^{-1}(.)$:

$$U_{iN} = F_a^{-1}(E + (N-1)\Delta) \qquad (2.8)$$

La D.D.P. partielle $f_{E,N}(E)$, relative à l'intervalle (U_{N-1}, U_N) se déduit de la probabilité partielle $f_U(U_{iN})$ en tenant compte du changement de variable de U en E.

$$f_{E,N}(E) = f_U(U_{iN})(dU/dE)_{U=U_{iN}} \qquad (2.9)$$

Dans cette dernière relation, l'élément différentiel dE est exprimable à l'aide de la relation (2.7)

$$dE = d\,F_a(U) \qquad (2.10)$$

A l'aide des relations (2.8) à (2.10), nous pouvons maintenant expliciter la D.D.P. de l'erreur de quantification, telle qu'elle figure en (2.6). On obtient $f_E(E)$ sous forme de sommation.

$$f_E(E) = \sum_{N=-\infty}^{\infty} f_U(F_a^{-1}(E+(N-1)\Delta).\ .(1/dF_a(U)/dU))_{U=U_{iN}} \qquad (2.11)$$

2.4. Cas de la conversion logarithmique avec entrée gaussienne centrée

La fonction aléatoire choisie a pour D.D.P.

$$f_U(U) = (1/\sigma\sqrt{2\Pi})\exp((-1/2\sigma^2)U^2) \qquad (2.12)$$

La fonction de conversion est définie à l'aide des relations (2.1) et (2.2) qui deviennent :

COMPARAISON ENTRE DEUX METHODES RECENTES DE DETERMINATION DE L'ERREUR DE QUANTIFICATION

$$N\Delta = F_a(U_N) = \log_\alpha U_N \qquad (2.13)$$

$$F_q(U) = \log_\alpha U_{N-1} \qquad (2.14)$$

L'erreur de quantification, qui rappelons-le est une erreur par troncature, devient, d'après (2.7) :

$$E = \log_\alpha (U_{iN}) - (N-1)\Delta \qquad (2.15)$$

La valeur de U_{iN}, déduite de la relation (2.15) est :

$$U_{iN} = \exp (\ell n\ \alpha((n-1)\Delta+E)) \qquad (2.16)$$

La D.D.P. de U_{iN} peut être exprimée en fonction de E en portant la valeur de U_{iN} dans la relation (2.13) :

$$f_U(U_{iN}) = (1/\sigma\sqrt{2\Pi}) \exp \{(-1/2\ \sigma^2) \exp 2(\ell n\ \alpha((N-1)\Delta + E\ \} \qquad (2.17)$$

La D.D.P. partielle $f_{E,N}(E)$, nécessite la connaissance de la dérivée $dF_a(U)/dU$.
D'après (2.13) cette dérivée est :

$$dF_a(U)/dU = 1/U\ \ell n\ \alpha \qquad (2.18)$$

Les relations (2.17) et (2.18) permettent d'exprimer $f_{E,N}(E)$:

$$f_{E,N}(E) = 1/\sigma\ \sqrt{2\Pi} \exp \{-1/2\ \sigma^2 \exp 2\{\ell n\ \alpha ((N-1)\Delta+E)\}\}\ell n\ \alpha \exp(\ell n\ \alpha((N-1)\Delta+E)) \qquad (2.19)$$

La densité de probabilité totale $f_E(E)$ est obtenue par sommation d'indice N selon la relation générale (2.11). Dans le but d'alléger l'écriture, la représentation suivante est adoptée pour $f_E(E)$:

$$K = \ell n\ \alpha/\sigma\ \sqrt{2\Pi} \qquad (2.20a)$$

$$Y(N) = ((N-1)\Delta+E)\ell n\ \alpha \qquad (2.20b)$$

Nous obtenons l'expression condensée suivante :

$$f_{E,N}(E) = K \exp(-1/2\ \sigma^2) \exp 2\ Y(N) \exp Y(N) \qquad (2.21)$$

Une nouvelle simplification d'écriture est introduite par le changement de variable

$$\exp Y(N) = z(N) \qquad (2.22)$$

L'expression générale (2.11) de la D.D.P. de l'E.Q. est alors écrite :

$$f_E(E) = K \sum_{N=-\infty}^{\infty} (\exp(-1/2\ \sigma^2)z^2(N))z(N) \qquad (2.23)$$

Dans l'expression ainsi obtenue, la sommation est remplacée par une intégration, prise par rapport à la variable N. La justification rigoureuse est longue, mais apporte des éléments nouveaux (9). Pour nous résumer, nous indiquerons simplement que l'on a intérêt à remplacer l'expression (2.23) par une expression en valeurs réduites, sur laquelle l'importance de la valeur de Δ est mise en évidence. Donnons simplement un exemple numérique. Pour une conversion de base e et 250 pas, l'erreur commise en remplaçant la sommation par une intégration, est inférieure en valeur relative à 0,07 %. En effectuant la transformation nous obtenons :

$$f_E(E) = K \int_{N=-\infty}^{+\infty} (\exp(-1/2\ \sigma^2)z^2(N))z(N)dN \qquad (2.24)$$

En tenant compte de la forme exponentielle de z(N), l'intégrale précédente est écrite :

$$(K/\Delta\ell n\ \alpha) \int_{z=0}^{+\infty} \exp((-1/2\ \sigma^2)z^2(N))dz \qquad (2.25)$$

L'intégrale représente alors, à un coefficient multiplicatif près, la fonction de répartition correspondant à la loi de D.D.P. gaussienne. Cette fonction de répartition est souvent désignée par le symbole Erf, c'est-à-dire fonction d'erreur. Après un changement de variable

$$t = z/\sigma \qquad (2.26)$$

On obtient, tous calculs effectués,

$$f_E(E) = (1/2\ \Delta)\ \mathrm{Erf}(x/\sqrt{2}) \qquad (2.27)$$

En imposant une conversion portant sur $|U|$ et en tenant compte du fait que x croit indéfiniment avec N nous aboutissons à la relation

$$f_E(E) = \frac{1}{\Delta} = \text{constante} \qquad (2.28)$$

L'uniformité de la densité de probabilité de l'erreur de quantification est ainsi mise en évidence pour un grand nombre de cas correspondant à la pratique courante.

COMPARAISON ENTRE DEUX METHODES RECENTES DE DETERMINATION DE L'ERREUR DE QUANTIFICATION

3. ERREUR DE QUANTIFICATION EN ARITHMETIQUE A VIRGULE FLOTTANTE

3.1. Objectifs de A.B. SRIPAD et D.L. SNYDER

Le titre du présent chapitre est la traduction du titre de l'article de Sripad et Snyder (1) dont nous proposons une analyse succincte.

Le principe du calcul, que ces auteurs avaient déjà présenté dans une communication antérieure (8), est le suivant : la densité de probabilité totale étant la somme des D.D.P. partielles, somme comprenant un nombre infini de termes, se présente comme une fonction périodique. La période considérée est la variation Δ dans le cas de la troncature ($0 \leqslant E < \Delta$). La périodicité de la fonction $f_E(E)$ permet d'en donner un développement en série de FOURIER, de la forme :

$$1/\Delta + \sum_{N\neq 0} \theta(2\Pi N)\exp(-2\Pi jNE) \qquad (3.1)$$

où $\theta(u)$ est la fonction caractéristique de l'entrée U. Si la condition (3.2) est satisfaite

$$\left.\begin{array}{l}\theta(2\Pi N) = 0 \\ N\neq 0\end{array}\right\} \qquad (3.2)$$

seul subsiste le terme constant. La D.D.P. de l'E.Q. est alors uniforme.

Ce calcul, publié en Octobre 1977, a été appliqué à une conversion analogique digitale (8).

3.2. Cas de l'arithmétique en virgule flottante

C'est le thème retenu par Sripad et Snyder pour leur deuxième communication, parue en Octobre 1978 (1). Elle concerne également une transformation que nous pourrions qualifier de proportionnel, à l'erreur de quantification près, par opposition à la conversion analogique digitale non proportionnelle, comme c'est le cas pour la conversion A-D à sortie logarithmique.
Le problème traité est le suivant : la valeur absolue Z d'une entrée U est mise sous la forme :

$$Z = |U| = M\,\alpha^N \qquad (3.3)$$

M est la "mantisse". Nous appellerons M la mantisse primitive par opposition à la mantisse du logarithme (log M), pris par rapport à la base α. On se propose de calculer la D.D.P. de l'erreur de quantification de la mantisse primitive M. En préambule nous pouvons faire les remarques suivantes :

1 - Les auteurs ont particularisé leurs calculs en choisissant la base $\alpha = 2$, tout en précisant que la base est quelconque mais toujours supérieure ou égale à 2. Nous nous inscrivons en faux contre cette affirmation : il suffit que la base α soit supérieure à l'unité. Cf. (4) par exemple.

2 - La mantisse M est soumise à la condition :

$$1/2 \leqslant M < 1 \qquad (3.4)$$

Nous pensons pour notre part, qu'il eut été plus judicieux de prendre

$$1 \leqslant M < 2$$

3 - Une erreur de quantification, donnée en valeur réduite, est définie par :

$$Y = W - W_q \qquad (3.5)$$

C'est une erreur par troncature, comprise soit entre 0 et +1, soit entre -1 et 0. Les auteurs optent pour :

$$-1 < Y < 0 \qquad (3.6)$$

4 - La suite de l'article de Sripad et Snyder montre cette erreur est prise sur l'exposant de l'expression :

$$Z = \alpha^W \qquad (3.7)$$

En utilisant la définition (3.5) de Y dans Z nous obtenons :

$$Z = \alpha^{Y+W_q} \qquad (3.8)$$

dont nous déduisons, par comparaison avec (3.3)

$$\left.\begin{array}{l}M = \alpha^Y \\ \alpha^{W_q} = \alpha^N\end{array}\right\} \qquad (3.9)$$

W_q est pris dans l'ensemble des entiers relatifs, et peut être remplacé par N. Nous pouvons maintenant exprimer la D.D.P. de l'E.Q. sur Y en suivant le même raisonnement qu'au chapitre 2.

$$f_Y(Y) = \sum_{N=-\infty}^{\infty} f_W(Y + N) \qquad (3.10)$$

Dans le cas présent le changement de variable n'entraine pas de modification d'écriture puisque dY = dW.
La périodicité de la fonction $f_Y(Y)$, la période étant ici l'unité, permet le développement en série de FOURIER, donné par :

$$f_Y(Y) = 1 + \sum_{N\neq 0} \Theta_W (2\Pi N) \exp(-2\Pi j\, NY)$$

Cette fonction se réduit à l'unité pour une condition identique à la condition (3.2), $\Theta_W(u)$ étant maintenant la fonction caractéristique de la variable aléatoire W, donc en définitif de $\log_\alpha$ Z. Nous retrouvons à nouveau une conditions d'uniformité de la D.D.P. de l'E.Q.

Pour procéder à un rapprochement avec nos calculs nous écrirons :

$$Z = \alpha^{Y+N} \qquad (3.12)$$

L'erreur de quantification Y se présente alors comme la valeur réduite E/Δ.
Sripad et Snyder déterminent une condition d'uniformité en particularisant la nature aléatoire du signal d'entrée par une condition du type (3.2) et indique dans la suite de leur communication que la grandeur aléatoire ainsi définie est sensiblement gaussienne. Rappelons que nous avions pris une entrée gaussienne et adopter une approximation résultant de la faible valeur relative du pas Δ. Les deux types d'approximation conduisent donc au même résultat. Nous ferons remarquer que la propriété d'uniformité déduite de la relation (3.11) ne figure pas chez Sripad et Snyder, puisque le but de ces auteurs est l'étude de l'erreur de quantification de la mantisse primitive M. Nous n'exposerons pas ce point de leur travail puisqu'il ne nous concerne pas.

REFERENCES

(1) SRIPAD A.B., SNYDER D.L. - Quantizing errors in floating-point arithmetic ; IEEE Trans. Acoust. Speech, signal processing ; vol. ASSP-26 ; n° 5, pp. 456-463, Oct. 1978.

(2) WITTLING F.G. - Estimation of quantizing noise in nonlinear following conversion ; Symp. Informatica 78 ; Bled ; Ch. 3 201, pp. 1-2, 1978.

(3) STEELE R. - Delta modulation systems; Pentech Press; London 1975.

(4) WITTLING F.G. - Conception d'un multiplicateur hybride ; Congrès AICA-IFIP; Munich ; pp. 811-816; 1970.

(5) WITTLING F.G. - Convertisseur analogique-digital à fonction de sortie préselectionnée ; Congrès AICA-IFIP ; Prague ; pp. 222-225 ; 1973.

(6) WITTLING F.G., BASSINSANA A, TAWIL E. - A new type of nonlinear signal processing ; 8ème Congrès IMEKO, Moscou ; Mai 1979 (à paraître).

(7) WITTLING F.G., HARATI W., ROTTHUT W. et MESNARD D. Convertisseur analogique numérique non linéaire fonctionnant en "suiveur" sans opération d'échantillonnage ; C.R. Acad. Sc. Paris ; t. 281 ; pp. 421-423, 1975.

(8) SRIPAD A.B., SNYDER D.L. - A necessary and sufficient condition for quantization errors to be uniform and white ; IEEE Trans. Acoust., Speech, signal processing ; vol. ASSP-25; pp. 442-448; Oct. 1977.

(9) WITTLING F.G. - Thèse de doctorat d'Etat - (soutenance 3ème trimestre 1979).