Précision finie et non alignement en codage/décodage MICDA

Mériem JAIDANE-SAIDANE[1], Fathi CHERIF[1,2], Madeleine BONNET[3]

(1)*Laboratoire des Systèmes de Télécommunications, ENIT, Campus Universitaire, Tunis, Tunisie*
(2)*Ecole Normale Supérieure de l'Enseignement Technique, rue Tahar Hussein, 58961 Tunis, Tunisie*
(3)*Univ. René Descartes, UFR Mathématiques et Informatique, 45 rue des Saints Pères 75270 Paris Cedex 06*

RESUME : L'implantation d'une chaîne de codage/décodage MICDA est présentée en précision finie en utilisant une arithmétique en virgule flottante. Il est montré que l'alignement du décodeur sur le codeur est réalisé sous des conditions plus restrictives que celles trouvées en précision infinie. On introduit une mesure d'alignement montrant que celui-ci dépend de la longueur binaire des signaux et paramètres en jeu et aussi du type de signal en entrée : l'alignement est d'autant plus mauvais que les zéros du filtre prédicteur sont proches du cercle unité.

ABSTRACT : The behavior of an ADPCM coding/decoding system is investigated when implanted in finite precision arithmetic (floating point). It is shown that the adjustment of the decoder onto the encoder is reached within more restrictive conditions than that required when working in infinite precision. An adjustment measure is introduced. It comes that the misadjustment depends both upon the binary lenght and the kind of input signal. The adjustment becomes more bad as the zeroes of the predictor filter are close to the unit circle.

1 Introduction

Le codage MICDA (Modulation par Impulsions et Codage Différentiel Adaptatif), basé sur la réduction de redondance du signal par prédiction et quantification adaptatives couplées, est utilisé pour stocker ou transmettre à débit réduit des signaux tels que les signaux de parole ou d'images (voir par exemple [1]).

Le codeur MICDA réalise une prédiction de chaque échantillon de signal et, c'est l'erreur de prédiction (quantifiée) qui est est alors stockée ou transmise. Pour assurer l'*alignement* du décodeur sur le codeur, le décodeur doit réaliser l'*"inverse"* du codeur en reconstruisant, à partir de la seule erreur de prédiction quantifiée, un signal qui ne diffère du signal original, que par l'erreur de quantification.

L'analyse de la chaîne de codage/décodage a mis en évidence les situations dans lesquelles l'alignement n'est pas assuré [1][2]. La présente étude analyse un cas particulier de non alignement du décodeur sur le codeur, rencontré alors que les conditions sont celles qui théoriquement permettent l'alignement. Cette situation apparaît lorsqu'on implante la chaîne en précision finie. En effet, les études précédentes ne considéraient que le cas de la précision infinie.

Le codeur peut être considéré comme un filtre adaptatif destiné à minimiser une erreur quadratique moyenne de prédiction alors que le décodeur est un filtre évolutif destiné à *calquer* le comportement du codeur. L'analyse en précision finie de la chaîne codeur/décodeur est de ce fait complexe. Des résultats concernant la précision finie existent pour un filtre fixe et un schéma d'addition particulier [3]. Il est montré dans [4][5] comment la précision finie modifie les performances d'un filtre adaptatif, dans le cas d'une adaptation de type LMS ou RLS. Les résultats présentés ici concernent un filtre transverse adaptatif d'ordre N suivi de son inverse ; ils généralisent ceux obtenus pour un ordre 1 [6]

2 Alignement en précision infinie

Afin de mettre en évidence les problèmes introduits par l'implantation en précision finie, l'étude est faite pour une chaîne codage/décodage se trouvant dans les conditions les plus favorables à l'alignement : codeur et décodeur sont égaux à l'initialisation et il n'y a pas d'erreur de transmission. De plus, on omet le quantificateur, ce qui n'affecte pas les résultats qualitatifs.

2.1 Equations codeur/décodeur

En notant x_n le signal d'entrée, l'erreur de prédiction calculée au codeur et transmise au décodeur est alors, dans le cas d'un codeur transverse d'ordre N :

$$e_n = x_n - A_n^T X_n \qquad (1)$$

où $A_n = (a_1(n), ..., a_N(n))^T$ est le vecteur paramètre adaptatif du prédicteur et $X_n = (x_{n-1}, ..., x_{n-N})^T$ est le vecteur des échantillons passés.

Le signal $\tilde{x}_n$ reconstitué au décodeur est alors

$$\tilde{x}_n = e_n + A_n^T \tilde{X}_n \qquad (2)$$

où $\tilde{X}_n = (\tilde{x}_{n-1}, ..., \tilde{x}_{n-N})^T$; le vecteur paramètre du décodeur étant le même que celui du codeur.

Généralement l'adaptation du codeur est une version modifiée du LMS afin d'assurer l'alignement du décodeur sur le codeur [1]. Nous supposerons que l'on est dans un cas, que nous qualifierons d'évolutif, où le vecteur paramètre oscille autour d'une valeur optimale A_{opt} selon

$$A_n = A_{opt} + \Delta_n \tag{3}$$

avec $A_{opt} = (a_1, ..., a_N)^T$ et $\Delta_n = (\Delta_1(n), ..., \Delta_N(n))^T$.

La quantité Δ_n, supposée indépendante de X_n, est telle que $\mathrm{E}\{\Delta_n\} = 0$ et $\mathrm{E}\{\Delta_n \Delta_n^T\} = \sigma_\Delta^2 \mathbf{I}$. Par exemple pour l'adaptation LMS, la puissance des oscillations est donnée par $\sigma_\Delta^2 = \mu \mathrm{E}\{(x_n - A_{opt} X_n)^2\}$, où μ est le pas d'adaptation [7].

2.2 Conditions d'alignement

On note

$$dx_n = x_n - \tilde{x}_n \tag{4}$$

l'écart caractéristique du non alignement. Cet écart vérifie l'équation récurrente

$$dX_{n+1} = (M + \underline{\Delta}_n) dX_n \tag{5}$$

où $M = \begin{pmatrix} a_1 & a_2 & \cdots & a_{N-1} & a_N \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix}$

et

$$\underline{\Delta}_n = \begin{pmatrix} \Delta_1(n) & \Delta_2(n) & \cdots & \Delta_N(n) \\ 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 \end{pmatrix}$$

M est la matrice compagnon associée au filtre prédicteur de paramètre A_{opt} ; les valeurs propres λ_i de cette matrice sont les zéros du filtre codeur.

L'écart dx_n tend en valeur moyenne vers 0 lorsque, d'après l'équation (5), les zéros du codeur, donc les pôles du décodeur, sont à l'intérieur du cercle unité, puisqu'alors $\mathrm{E}\{dX_{n+1}\} = M\mathrm{E}\{dX_n\}$.

Nous montrons que l'écart quadratique $\mathrm{E}\{dx_n^2\}$ tend également vers 0 sous des conditions plus restrictives. Ainsi d'après (5) on a,

$$\mathrm{E}\{dX_{n+1}^T dX_{n+1}\} = \mathrm{E}\{dX_n^T M^T M dX_n\} + \sigma_\Delta^2 \mathrm{E}\{dX_n^T dX_n\} \tag{6}$$

En supposant que $M = Q\Lambda Q^{-1}$, où Λ est une matrice diagonale ($diag(\lambda_i)$), on peut écrire en utilisant les normes matricielles

$$\mathrm{E}\{dX_{n+1}^T dX_{n+1}\} \leq (\|Q\|^2 \|Q^{-1}\|^2 \|\Lambda\|^2 + \sigma_\Delta^2) \mathrm{E}\{dX_n^T dX_n\}. \tag{7}$$

Une condition suffisante de convergence est donc que

$$|\lambda_i|_{max} < \sqrt{\frac{1}{\|Q\|^2 \|Q^{-1}\|^2} - \sigma_\Delta^2} \tag{8}$$

Cette équation signifie, qu'au décodeur, le pôle de plus grand module doit être à une *distance finie* du cercle unité. Ce résultat caractérisant l'alignement à l'ordre 2, en précision infinie, est nouveau ; une généralisation de la notion d'alignement aux moments d'ordre supérieur est possible.

3 Cas de la précision finie

L'implantation en précision finie du système constitué par la chaîne codeur/décodeur modifie les valeurs théoriques trouvées en précision infinie et, en particulier les conditions d'alignement (8).

On considère maintenant une arithmétique en virgule flottante. Rappelons que la représentation d'un nombre x en virgule flottante s'exprime selon $(signe) 2^a b$, où a est l'exposant entier et b est la mantisse normalisée, codée sur B bits. La précision finie introduit des erreurs relatives lors de la représentation d'un nombre et lors des différentes opérations arithmétiques.

Si l'on note $(.)'$ la représentation du nombre $(.)$ en précision finie, on a en particulier $(x+y)' = (x+y)(1+\alpha)$ et $(x * y)' = (x * y)(1+\mu)$ où α et μ sont des variables aléatoires centrées, indépendantes de x et y. En général, $\mathrm{E}\{\alpha^2\} < \mathrm{E}\{\mu^2\}$ mais l'on suppose ici, sans perte de généralité, que ces écart-type sont égaux à $\sigma^2 = 0.18\ 2^{-2B}$ [4]. On supposera de plus, que dans l'ensemble des calculs il n'y a pas dépassement de capacité.

3.1 Equations codeur/décodeur

On note $\Delta e(n) = e_n - e'_n$ l'écart entre la sortie théorique du filtre codeur et sa valeur en précision finie, de même $\Delta\tilde{x}(n) = \tilde{x}_n - \tilde{x}'_n$ est l'écart sur la sortie du décodeur. En considérant un schéma d'additions successives : $x+y+z = ((x+y)+z)$, l'erreur de prédiction s'écrit, en précision finie

$$e'_n = x'_n \prod_{i=1}^{N} (1 + \alpha_i(n)) - A_n^T B_n X'_n \tag{9}$$

où $B_n = diag((1+\mu_i(n)) \prod_{j=i}^{N} (1+\alpha_j(n)))$. De même au décodeur, avec des notations similaires on trouve

$$\tilde{x}'_n = e'_n \prod_{i=1}^{N} (1 + \tilde{\alpha}_i(n)) + A_n^T \tilde{B}_n \tilde{X}'_n \tag{10}$$

où $\tilde{B}_n = diag((1+\tilde{\mu}_i(n)) \prod_{j=i}^{N} (1+\tilde{\alpha}_j(n))$.

3.2 Conditions d'alignement

L'écart caractérisant le non alignement est maintenant

$$dx'_n = x'_n - \tilde{x}'_n \tag{11}$$

En supposant que l'influence de la précision finie ne concerne que des variations du premier ordre, on a d'après (9) et (10)

$$x'_n = e'_n (1 - \sum_{i=1}^{N} \alpha_i(n)) + A_n^T (\mathbf{I} + \Delta C_n) X'_n \tag{12}$$

$$\tilde{x}'_n = e'_n (1 + \sum_{i=1}^{N} \tilde{\alpha}_i(n)) + A_n^T (\mathbf{I} + \Delta \tilde{C}_n) \tilde{X}'_n \tag{13}$$

avec

$$\Delta C_n = diag(\mu_i(n) - \sum_{j=1}^{i-1} \alpha_j(n)) \quad (14)$$

$$\Delta \tilde{C}_n = diag(\tilde{\mu}_i(n) + \sum_{j=i}^{N} \tilde{\alpha}_j(n)) \quad (15)$$

D'après les équations (12) et (13) l'écart dx' s'écrit alors

$$dx'_n = \epsilon_n + A_n^T(\mathbf{I} + \Delta\tilde{C}_n)dX'_n \quad (16)$$

où

$$\epsilon_n = -x'_n\Delta\alpha_n + A_n^T(\Delta C_n - \Delta\tilde{C}_n - \Delta\alpha_n\mathbf{I})X'_n \quad (17)$$

et

$$\Delta\alpha_n = \sum_{i=1}^{N}(\alpha_i(n) + \tilde{\alpha}_i(n)). \quad (18)$$

L'écart dx' satisfait l'équation récurrente

$$dX'_{n+1} = (M + \underline{\Delta}_n + \underline{\Delta}'_n)dX'_n + (\epsilon_n, 0, 0)^T \quad (19)$$

où

$$\underline{\Delta}'_n = \begin{pmatrix} \Delta'_1(n) & \Delta'_2(n) & \cdots & \Delta'_N(n) \\ 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 \end{pmatrix}$$

avec

$\Delta'_i(n) = a_i(\tilde{\mu}_i(n) + \sum_1^N \tilde{\alpha}_j(n))$.

L'équation (19) est à comparer à l'équation (5), trouvée pour la précision infinie $\underline{\Delta}'_n$). La valeur moyenne de ϵ_n étant nulle, l'écart dx'_n tend en moyenne vers zéro sous les mêmes conditions que celles trouvées pour la précision infinie, c'est à dire lorsque les pôles du décodeur sont à l'intérieur du cercle unité.

On montre par ailleurs, d'après (19), que

$$\mathrm{E}\{dX'^T_{n+1}dX'_{n+1}\} = \mathrm{E}\{\epsilon_n^2\} + \mathrm{E}\{dX'^T_n M^T M dX'_n\} +$$

$$\sigma_\Delta^2\mathrm{E}\{dX'^T_n dX'_n\} + \mathrm{E}\{dX'^T_n\mathrm{E}\{\underline{\Delta}'^T_n\underline{\Delta}'_n\}dX'_n\} \quad (20)$$

L'étude de la convergence de $\mathrm{E}\{dX'^T_{n+1}dX'_{n+1}\}$, donc du moment du deuxième ordre $\mathrm{E}\{dx'^2_n\}$, dépend de l'équation récurrente (20) sans le terme constant $\mathrm{E}\{\epsilon_n^2\}$. Selon la même démarche que celle adoptée en précision infinie, nous montrons que
$\mathrm{E}\{dX'^T_{n+1}dX'_{n+1}\} \leq$

$$(\|Q\|^2\|Q^{-1}\|^2\|\Lambda\|^2 + \sigma_\Delta^2 + \mathrm{E}\{\|\underline{\Delta}'_n\|^2\})\mathrm{E}\{dX'^T_n dX'_n\}. \quad (21)$$

Pour qu'il y ait convergence, il faut donc que

$$|\lambda_i|_{max} < \sqrt{\frac{1}{\|Q\|^2\|Q^{-1}\|^2} - \sigma_\Delta^2 - \sigma^2\sum_{i=1}^{N}(N-i+2)a_i^2} \quad (22)$$

Cette dernière équation, qui est à comparer à celle de la précision infinie (8), montre que la condition de convergence en précision finie est plus restrictive : le pôle de plus grand module doit être plus encore à l'intérieur du cercle unité et ceci d'autant plus que le nombre de bits, sur lequel se fait l'implantation, est faible. Par l'intermédiaire de $\mathrm{E}\{\underline{\Delta}'^T_n\underline{\Delta}'_n\}$ on montre que ce résultat dépend aussi du schéma d'addition de l'implantation.

3.3 Mesure d'alignement

Contrairement au cas de la précision infinie, le moment d'ordre 2, $\mathrm{E}\{dx'^2_n\}$ ne tend pas vers 0 mais vers une valeur constante. Ceci permet de définir une *mesure d'alignement* du décodeur sur le codeur, notée $\mathbf{m}_{aligne}$, définie de la façon suivante

$$\mathbf{m}_{aligne} = \frac{\mathrm{E}\{x'^2_n\}}{\mathrm{E}\{dx'^2_n\}} \quad (23)$$

L'équation (16) permet de comprendre comment évolue $\mathbf{m}_{aligne}$. En effet, la quantité dx'_n peut être considérée comme la sortie d'un filtre récursif de vecteur A_n perturbé par une quantité aléatoire dépendant uniquement de l'implantation en précision finie du décodeur. L'entrée de ce filtre récursif est ϵ_n qui peut elle même être considérée comme la filtrée transverse de vecteur paramètre A_n (lui aussi perturbé), d'un signal de faible amplitude $x'_n\Delta\alpha_n$

La puissance de ϵ_n est d'autant plus faible que le nombre de bits sur lequel s'effectue l'implantation de la chaîne de codage/décodage est faible. Cette puissance est proportionnelle à l'ordre du filtre, à la puissance d'entrée du signal et aussi à la puissance du bruit de calcul σ^2. Elle satisfait

$$\mathrm{E}\{\epsilon_n^2\} = N\gamma\sigma^2\mathrm{E}\{x'^2_n\} \quad (24)$$

avec $\gamma = [2 + 5\sum_{i=1}^N + \frac{1}{N}\sum\sum_{i\neq j} a_i a_j \rho_{i-j}(10i + 4j + 5N)]$ et où ρ_i est le coefficient de corrélation de x'_n. Le facteur γ dépend du schéma d'addition de l'implantation.

La puissance de dx'_n est d'autant plus élevée, donc l'alignement d'autant plus mauvais que les pôles du décodeur sont proches du cercle unité.

Un tel résultat est particulièrement important, il montre comment, alors que les conditions de transmission sont celles qui théoriquement devrait permettre l'alignement, l'écart entre le signal reconstitué au décodeur et le signal émis au codeur peut être très élevé du fait de l'implantation en précision finie du codeur et du décodeur.

3.4 Simulations

Afin d'illustrer les résultats précédents, nous considérons des filtres d'ordre 2. Les calculs sont effectués sur 32 bits. le codeur et le décodeur sont supposés fixes dans un premier temps.

La figure 1 montre l'évolution de $\mathbf{m}_{aligne}$ en fonction du module du zéro du codeur (on se place dans le cas complexe et l'on fait également varier Arg, l'argument de ce zéro). Plus ce module est proche du cercle unité plus l'alignement se dégrade. De faibles valeurs, de l'ordre de 20dB peuvent même être atteintes alors que l'on se trouve dans les meilleures conditions possibles d'alignement.

Dans le cas adaptatif, de tels résultats qualitatifs se conservent. On considère une adaptation LMS et l'on initialise

de la même façon codeur et décodeur. Le signal à transmettre est généré selon une filtrée récursive d'ordre 2 d'un bruit blanc et l'on fait varier le module du pôle complexe. La figure 2 illustre à la fois le cas évolutif et le cas adaptatif. Elle montre que l'hypothèse simplificatrice utilisée (cf éq (3)) est justifiée. L'effet de l'adaptation provoque toutefois une dégradation de l'alignement au voisinage du cercle unité, plus grande que dans le cas évolutif.

4 Conclusion

L'implantation d'une chaîne de codage/décodage MICDA montre que de nouveaux problèmes d'alignement apparaissent avec la précision finie. Il est montré qu'il n'y a plus d'alignement au sens classique où l'écart entre le signal reconstitué et le signal en entrée tend vers zéro. Cet alignement se mesure par un rapport signal sur bruit qui dépend à la fois de la représentation binaire des signaux et paramètres en jeu et du type de signal transmis. Tous les signaux tels que les zéros du prédicteur optimal sont proches du cercle unité conduisent à une forte dégradation de l'alignement.

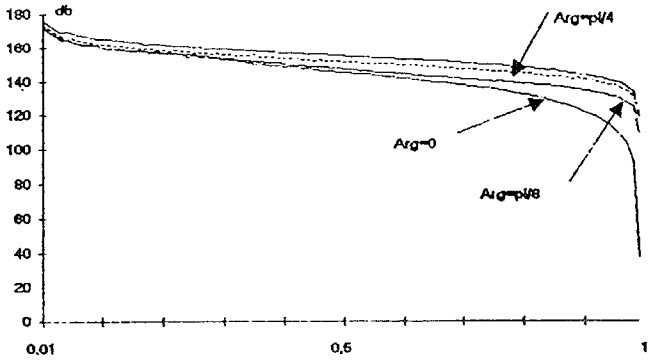

Figure 1 : Evolution de $\mathbf{m}_{aligne}$ et position du module des pôles du décodeur : cas fixe

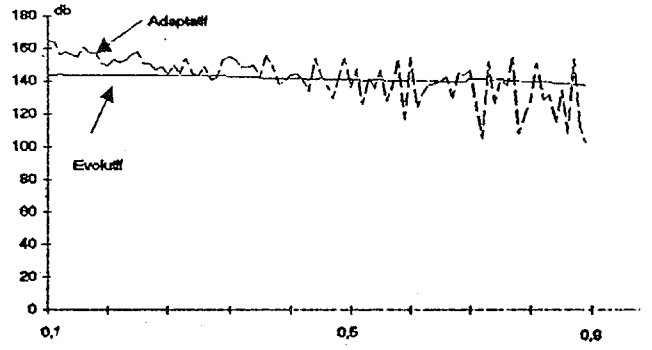

Figure 2 : Evolution de $\mathbf{m}_{aligne}$ et position du module des pôles du décodeur : cas adaptatif

References

[1] M. Bonnet, O. Macchi, M. Jaïdane, "*Theoretical Analysis of the ADPCM CCITT Algorithm*," *IEEE Trans. on Com.*, vol. 38, n°6, pp.847–858, June1990.

[2] M. Bonnet, O. Macchi, M. Jaïdane, "*Mistracking in successive PCM/ADPCM transcoders*," *IEEE Trans. on Com.*, vol. 37, pp.843–850, Aug 1989.

[3] B.D. Rao, "*Floating Point Arithmetic and Digital Filters*," *IEEE Trans. Signal Processing*, vol. 40 ,n°1, pp. 85–95, January 1992.

[4] C. Caraiscos, B. Liu, "*A Roundoff Error Analysis of the LMS Adaptive Algorithm*," *IEEE Trans. ASSP*, vol. 32, n°1, pp 34–41, Feb.1984.

[5] S. H. Ardalan, "*Floating point error analysis of recursive least-squares and least-mean-squares adaptive filters*" *IEEE Trans. on CAS*, vol. 33, n°1, pp 1192–1208, Dec.1986.

[6] F. Cherif, M. Jaïdane, "*Précision finie et Non Alignement en Codage MICDA pour le Stockage ou la Transmission des Signaux 1D et 2D*", *Numéro hors série des Annales Maghrébines de l'Ingénieur, Actes du Troisième Colloque Maghrébin sur les modèles numériques de l'ingénieur*, pp 241–248, Tunis, Nov. 1991.

[7] B. Widrow, S. D. Stearns, "*Adaptive Signal Processing*," Prentice Hall, Englewood Cliffs, New Jersey, 1985.