

Simulation comportementale pour la synthèse de convertisseurs analogique-numérique CMOS rapides

Hervé Petit

▶ **To cite this version:**

Hervé Petit. Simulation comportementale pour la synthèse de convertisseurs analogique-numérique CMOS rapides. domain_other. Télécom ParisTech, 2004. English. NNT : . pastel-00000868

HAL Id: pastel-00000868
https://pastel.hal.science/pastel-00000868

Submitted on 25 Nov 2004

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.





École Doctorale
d'Informatique,
Télécommunications
et Électronique de Paris

Thèse

présentée pour obtenir le grade de docteur
de l'École Nationale Supérieure des Télécommunications

Spécialité : **Électronique et Communications**

Hervé PETIT

Simulation comportementale pour la synthèse de convertisseurs analogique-numérique CMOS rapides

Soutenue le 11 octobre 2004 devant le jury composé de :

Georges Alquie	Président
Andreas Kaiser	Rapporteurs
Marie-Minerve Louerat	
Patrick Loumeau	Examinateurs
Dominique Morche	
Jean-Francois Naviner	Directeur de thèse

À la mémoire de mon Père

Remerciements

Je tiens à exprimer ici toute ma reconnaissance aux responsables du département COMELEC de l'ENST et à Monsieur Bernard Robinet, directeur de l'EDITE, qui ont rendu possible l'accomplissement de cette thèse.

Elle n'aurait peut-être pas eu lieu si Rachid Bouchakour, aujourd'hui professeur à l'Ecole Polytechnique Universitaire de Marseille, ne m'avait pas incité dans cette démarche et associé à sa recherche. Qu'il trouve ici toute ma gratitude.

J'adresse un vif remerciement à mes collègues de travail pour leur soutien permanent et spécifiquement à Jean-François Naviner, qui a accepté de diriger cette thèse.

Je remercie le professeur Georges Alquié pour m'avoir fait l'honneur de présider la commission d'examen ainsi que tous les membres du jury pour avoir accepté de juger ce travail.

Cette thèse est le résultat d'un parcours inhabituel qui a mêlé ma vie professionnelle et ma formation pendant de nombreuses années. Aussi, je tiens à remercier tout particulièrement mon épouse pour la patience exceptionnelle dont elle a dû faire preuve. je dois certainement à mon père la persévérance nécessaire à l'accomplissement de ce travail. Il n'a malheureusement pas pu voir son aboutissement. Je suis sûr qu'il en aurait été très fier et je lui dédie ce document.

Notations

Sigles

CAN	Convertisseur Analogique-Numérique
CNA	Convertisseur Numérique-Analogique
CMOS	*Complementary Metal Oxyde Semiconductor*
	Technologie basée sur l'utilisation de transistors MOS complémentaires
DEM	*Dynamic Element Matching*
	Technique de linéarisation par sélection d'éléments unitaires dans un CNA
DWA	*Data Weighted Averaging*
	Méthode de sélection cyclique des éléments d'un CNA
ENBW	*Equivalent Noise BandWidth*
	Bande équivalente de bruit associée à une fenêtre d'analyse spectrale
FFT	*Fast Fourier Transform*
	Algorithme de transformée de Fourier rapide
IP	*Intellectual Property*
	Bloc ou cellule préconçue d'un circuit intégré
SoC	*System on chip*
	Système intégré sur une puce
TFD	Transformée de Fourier discrète
VLSI	*Very Large Scale Integration*
	Très haute densité d'intégration

Symboles physiques

k_B	Constante de Boltzmann	$k_B = 1,38\ 10^{-23}\ J/K$
T_a	Température absolue en Kelvin	

Echantillonnage

F_s	Fréquence d'échantillonnage
B	Bande de fréquences positives support du signal
	$f \in [-B, B]$ ou $f \in [-f_u, -f_l] \cup [f_l, f_u]$ avec $B = f_u - f_l$
R	Rapport de suréchantillonnage (ou OSR :*Oversampling Ratio*)
	$R = \frac{F_s}{2\,B}$

Note : On peut exprimer la bande des fréquences positives en fonction de la fréquence d'échantillonnage sous la forme :

$$f \in [(k-1)\,\frac{F_s}{2}, k\,\frac{F_s}{2}]$$

On dit alors que le signal est dans la $k^{ième}$ zone de Nyquist.

Quantification

FSR	*Full Scale Range* Domaine non saturé de l'entrée d'un quantificateur linéaire
n	Nombre de bit d'un quantificateur ($n = \log_2 N_Q$ pour un quantificateur ayant N_Q niveaux)
Δ	Pas de quantification idéal d'un quantificateur linéaire ($\Delta = \frac{FSR}{2^n}$)
Q	Pas de quantification moyen d'un quantificateur linéaire
DNL	*Differential Non Linearity* Non-linéarité différentielle
INL	*Integral Non Linearity* Non-linéarité intégrale
LSB	*Least-significant bit* bit de plus faible poids [1]

Caractéristiques dynamiques des convertisseurs

DR	*Dynamic Range* Rapport entre le signal d'entrée maximum exploitable et le signal minimum (généralement pris égal au bruit)
SFDR	*Spurious Free Dynamic Range*) Dynamique libre de raies parasites
SINAD	*Signal to Noise and Distortion Ratio* Rapport signal sur bruit plus distorsion
SNR	*Signal to Noise Ratio* Rapport signal sur bruit
SNHR	*Signal to Non-Harmonic Ratio* Rapport signal sur bruit sans prise en compte des harmoniques
THD	*Total Harmonic Distortion* Distorsion harmonique
IMD	*Intermodulation Distortion*) Distorsion d'intermodulation
NPR	*Noise Power Ratio* Rapport de puissance de bruit

[1] Losque ce terme se réfère à l'entrée d'un quantificateur, il est équivalent au pas de quantification Δ. Lorsqu'il se réfère au code de sortie, il désigne le bit de plus faible poids.

Technologie CMOS

L_{min}	Longueur minimum de canal
L	Longueur de canal
W	Largeur de canal
V_{dd}	Tension d'alimentation
V_{ij}	Différence de potentiel entre les électrodes i et j $(i,j) \in \{d,g,s,b\}$)
	d : drain, g : grille, s : source, b : substrat (*bulk*)
μ	Mobilité des porteurs : électrons (μ_n) ou trous (μ_p)
t_{ox}	Epaisseur de l'oxyde de silicium sous la grille
C_{ox}	Capacité surfacique de l'oxyde de silicium
V_T	Tension de seuil
A_{VT}	Paramètre de dispersion de la tension de seuil
A_β	Paramètre de dispersion du gain
I_{ds} Courant drain-source g_m	Transconductance
g_{ds}	Conductance drain-source
C_{gs}	Capcité grille-source
C_{jd}	Capacité de jonction drain

Table des matières

Introduction 1

I Caractérisation et test des CAN 5

1 Echantillonnage et quantification 7
- 1.1 Introduction . . . 7
- 1.2 Echantillonnage . . . 8
  - 1.2.1 Suréchantillonnage . . . 9
  - 1.2.2 Echantillonnage des signaux à bande étroite . . . 9
- 1.3 Quantification . . . 10
  - 1.3.1 Erreur quadratique moyenne et condition d'optimalité . . . 12
  - 1.3.2 Quantification uniforme . . . 13

2 Caractérisation et limites des CAN 21
- 2.1 Introduction . . . 21
- 2.2 Caractérisation des CAN . . . 21
  - 2.2.1 Caractéristiques statiques . . . 22
  - 2.2.2 Caractéristiques dynamiques . . . 24
- 2.3 Limites fondamentales . . . 25
  - 2.3.1 Dynamique libre de raies parasites . . . 25
  - 2.3.2 Incertitude d'échantillonnage . . . 26
  - 2.3.3 Métastabilité . . . 27
  - 2.3.4 Bruit thermique . . . 29

3 Test dynamique des CAN 33
- 3.1 Introduction . . . 33
- 3.2 Le choix du signal de test . . . 34
  - 3.2.1 Echantillonnage synchrone . . . 34
  - 3.2.2 Echantillonnage aléatoire . . . 35
- 3.3 Analyse spectrale . . . 35
- 3.4 Test statistique . . . 37
  - 3.4.1 Calcul des seuils de transitions . . . 38
  - 3.4.2 Limitations . . . 39
- 3.5 Conclusions . . . 39

II CAN CMOS rapides **41**

4 Convertisseur flash **43**
4.1 Introduction . . . 43
4.2 Limitations . . . 44
4.3 Limitations . . . 45
4.4 Codage et métastabilité . . . 47
4.5 Filtrage spatial . . . 47
4.5.1 Facteur de mérite . . . 48
4.5.2 Conditions aux limites . . . 49
4.6 Techniques de comparaison et résolution . . . 50
4.7 Interpolation . . . 51
4.8 Pré-traitement analogique . . . 51
4.8.1 Interpolation entre les blocs de repliement . . . 53
4.9 Convertisseurs flash 6 bit . . . 54

5 Convertisseur pipeline **57**
5.1 Introduction . . . 57
5.2 Codage redondant . . . 61
5.3 Sources d'erreurs dans le pipeline . . . 64
5.3.1 Dispersion sur les gains . . . 64
5.3.2 Erreurs liées aux CNA . . . 65
5.3.3 Calibrage . . . 65
5.4 Dimensionnement des étages . . . 66
5.5 Résolution par étage . . . 67
5.6 Utilisation du parallélisme . . . 68
5.6.1 Erreurs liées au traitement parallèle . . . 69
5.7 Réalisations récentes . . . 73

6 Convertisseurs ΣΔ **77**
6.1 Introduction . . . 77
6.2 Codeur prédictif et modulation Δ . . . 78
6.3 Modulateur Σ Δ du premier ordre . . . 79
6.3.1 Entrée constante et décodeur de type moyenne . . . 80
6.3.2 Réalisation à base d'un intégrateur . . . 81
6.4 Architecture générale d'un modulateur Σ Δ . . . 82
6.4.1 Erreurs liées au CNA . . . 86
6.4.2 Critères de performance des modulateurs . . . 87
6.5 Architecture à un seul quantificateur . . . 88
6.5.1 Modèle linéaire et stabilité . . . 89
6.5.2 Modulateur du premier ordre . . . 90
6.5.3 Modulateur d'ordre supérieur ou égal à deux . . . 90
6.6 Architecture cascade . . . 94
6.6.1 Principe . . . 94
6.6.2 Relations de couplage dans la cascade . . . 94
6.6.3 Structures cascades du troisième et quatrième ordre . . . 98
6.6.4 Utilisation de plusieurs niveaux de quantification . . . 99
6.7 Choix des paramètres d'un modulateur . . . 100
6.8 Convertisseurs ΣΔ CMOS récents . . . 100

7 Contraintes technologiques **103**
7.1 Introduction . . . 103
7.2 Réduction de la tension d'alimentation . . . 104
7.3 Circuits à capacités commutées . . . 105
7.3.1 Commutateur CMOS . . . 105
7.4 Conclusions . . . 110

III Simulation comportementale des CAN **113**

8 Simulation comportementale **115**
8.1 Introduction . . . 115
8.2 Généralités sur la synthèse . . . 116
8.3 Sélection d'une architecture de CAN . . . 118
8.4 Classes C++ pour la simulation de CAN . . . 119

9 Modèle linéaire et ΣΔ **125**
9.1 Introduction . . . 125
9.2 Estimation des paramètres du quantificateur . . . 126
9.3 bruit de quantification . . . 127
9.4 Optimisation des coefficients d'un modulateur . . . 129
9.5 Prise en compte des effets linéaires . . . 132
9.5.1 Effet du gain fini dans un intégrateur . . . 132
9.5.2 Analyse statistique . . . 133
9.6 Classes C++ associées . . . 133
9.7 Conclusions . . . 134

10 Application aux capacités commutées **137**
10.1 Introduction . . . 137
10.2 Blocs de base des CAN à capacités commutées . . . 137
10.2.1 Sources d'erreur . . . 138
10.3 Charge non linéaire des capacités . . . 139
10.4 Exploration des performances dynamiques . . . 141
10.4.1 Modèle dynamique pour le transfert de charge . . . 141
10.4.2 Principe de la méthode . . . 142
10.4.3 Application au CAN ΣΔ . . . 145
10.4.4 Application au CAN pipeline . . . 147

Conclusions et perspectives **151**

Annexes **153**

A Couplage dans le convertisseur flash **155**

B Codage redondant des nombres **159**
B.1 Ecriture des nombres avec chiffres signés . . . 159
B.1.1 Numération simple de position . . . 159
B.1.2 Extension à des chiffres signés . . . 159
B.2 Algorithme de division et chiffres signés . . . 161
B.2.1 Sélection d'un élément du quotient . . . 161

C Fonctions de transfert des modulateurs ΣΔ **163**
C.1 Fonctions de transfert des modulateurs à un seul quantificateur 163
C.1.1 Fonctions de transfert complètes 163
C.1.2 Fonctions de transfert simplifiées 164

D Optimisation convexe **167**

E Modèle dynamique de l'intégrateur à capacités commutées **169**
E.1 Intégrateur en phase de transfert 169
E.1.1 Equations de conservations des charges 170
E.1.2 Influence des paramètres R_{sw} et I_{sat} 171
E.1.3 Analyse simplifiée 171
E.2 Intégrateur dans la phase de maintien 174
E.3 Réponse de l'intégrateur sur une période complète 175

F Caractéristiques d'un étage amplificateur pipeline **177**
F.1 Calcul de la capacité d'entrée du flash 178
F.2 Bande passante 178
F.3 Bruit thermique échantillonné 179

Bibliographie **181**

Introduction

Depuis le début des années 80, la combinaison du haut niveau d'intégration de la technologie VLSI CMOS et de méthodes structurées de conception a permis de concevoir des systèmes numériques de plus en plus complexes avec des capacités de traitement toujours plus élevées conduisant à la notion de système sur une puce (*SoC : System On Chip*). Cette potentialité à fait naître des applications nouvelles où le besoin de communiquer avec l'extérieur est vite apparu. Paradoxalement, cette nécessité d'échanger l'information a conduit à l'intégration simultanée de circuits de traitement mixtes (analogiques et numériques) dans une technologie essentiellement orientée vers la conception de circuits purement numériques. Il est ainsi possible de faire cohabiter sur un même substrat de silicium des parties radio traitant des signaux très faibles et sensibles aux interférences avec des parties de traitement numérique complexes. Le circuit Bluetooth constitue un exemple d'un tel système radio complet sur une puce [126]. L'introduction de tels systèmes sur le marché est de plus en plus rapide et nécessite une forte automatisation des tâches de conception. Si celle-ci est déjà très efficace pour les circuits VLSI numériques, il n'en est pas de même pour les circuits mixtes où de nombreuses tâches restent manuelles et effectuées par des concepteurs expérimentés. Chaque nouvelle conception est basée sur l'expérience acquise dans les réalisations précédentes et introduit des modifications réduites afin de minimiser le risque. Les circuits mixtes sont en effet beaucoup plus sensibles aux variations du procédé de fabrication que les circuits numériques. Dans ce contexte, la réutilisation de blocs préconçus s'avère importante pour réduire le temps de conception. On utilise alors le terme d'IP (Intellectual Property) pour désigner ces blocs. Il est souhaitable que ceux-ci soient paramétrables pour pouvoir migrer d'une technologie à une autre et s'adapter aux spécifications d'une application donnée [31]. Cette tâche, quasi automatique pour les circuits numériques, est très délicate à réaliser pour les circuits mixtes dont les caractéristiques sont très dépendantes de la technologie. D'autre part, les contraintes de fiabilité liées à la miniaturisation des dispositifs ont conduit à une réduction continue de la tension d'alimentation. Les concepteurs analogiciens doivent trouver de nouvelles techniques de traitement pour faire face à cette réduction de la tension d'alimentation. Ils sont d'autre part confrontés à des effets du second ordre importants du fait de l'inadéquation de la technologie avec les contraintes fortes de précision imposées aux fonctions analogiques.

Elément frontière entre deux modes de représentation du signal, les convertisseurs analogique-numérique (CAN) et numérique-analogique (CNA) conditionnent les performances globales du système, le partitionnement entre ces deux types de traitement étant très dépendant de leurs caractéristiques de résolution, de vitesse et de consommation. La conversion analogique-numérique est la plus critique de ces interfaces. La figure 1 donne les performances que doivent satisfaire les CAN pour quelques applications récentes [59]. Les lignes obliques marquent les frontières de la consommation actuellement réalisables en technologie CMOS. Cette contrainte est particulièrement

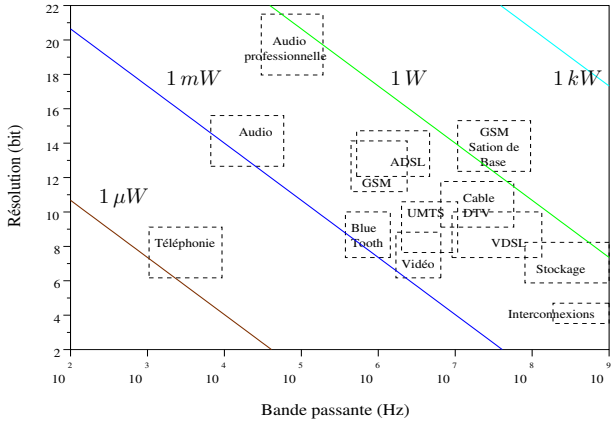


FIG. 1 – Performances pour quelques applications récentes.

forte pour le secteur des systèmes portables, miniaturisés et alimentés par batterie qui connaît une forte croissance. Ces systèmes nécessitent une gestion efficace de l'énergie disponible. Le coût, le poids et la fiabilité sont en effet fortement influencés par cette gestion (poids, coût et durée de vie des batteries, échauffement,...). A elle seule, cette limite rend par exemple irréaliste une conversion directe du signal radio dans un récepteur pour les systèmes les plus performants représentés sur la figure 1. Une fréquence d'échantillonnage supérieure au GHz et une dynamique de l'ordre de 100 dB, compatible avec ces systèmes, conduirait alors à une consommation de plusieurs centaines de watts. Ces performances sont de toute façon inaccessibles dans la technologie CMOS actuelle, seule la technologie des supraconducteurs permet de s'en approcher [55].

Un grand nombre d'architectures de CAN et CNA utilise des composants qui doivent avoir des caractéristiques aussi voisines que possible et ceci d'autant plus que la résolution du convertisseur est importante. En pratique, l'écart entre la valeur réelle et la valeur nominale du paramètre d'intérêt est fonction de différents facteurs. Cette erreur peut être occasionnée par des variations du processus de fabrication (telles que le désalignement des masques, un gradient sur l'épaisseur d'oxyde ou une non uniformité du dopage), par des variations de l'environnement (par exemple un gradient de température), ou par le stress de certains composants lié à leurs conditions de fonctionnement. Une grande précision est donc difficile à garantir du fait des variations dans le procédé de fabrication. La solution consistant à corriger ces erreurs après fabrication est très coûteuse. De plus, les caractéristiques des composants doivent être garanties dans le temps et sous certaines conditions de température. Il est préférable de choisir des techniques de conversion robustes et utilisant le fort potentiel du traitement numérique des technologies actuelles pour améliorer les performances. Les architectures de CAN de type flash, pipeline et $\Sigma\Delta$ ont montré une grande robustesse face à l'évolution de la technologie CMOS. Elles couvrent relativement bien l'espace (résolution-fréquence d'échantillonnage) de la figure 1. Etant donné un point de cet espace, le choix particulier d'une architecture plutôt qu'une autre est souvent basé sur l'expérience et prend peu en compte l'évolution technologique [129]. Faute de méthodes et d'outils, la validité de ce choix nécessite quelquefois une synthèse complète jusqu'au niveau des masques du circuit. Les efforts en matière de synthèse des circuits mixtes se sont concentrés essentiellement au niveau des cellules élémentaires. La croissance de ce type de circuit nécessite des méthodes compatibles avec celles utilisées dans le domaine

numérique. En particulier, l'association de synthèse et de la vérification aux étapes de plus haut niveau architectural peut permettre de limiter les vérifications post-synthèse très coûteuses. Ceci peut également permettre d'explorer plus efficacement l'espace des configurations possibles.

Cette thèse est une contribution à la synthèse architecturale des CAN. Nous proposons d'utiliser la simulation comportementale rapide avec différents niveaux d'abstraction pour l'évaluation des performances. Celle-ci ouvre différentes voies telles que l'optimisation des paramètres ou l'analyse statistique, qui nécessitent généralement des temps de calcul prohibitifs pour des circuits complexes. Un outil prototype basé sur des classes C++ a été réalisé pour la description et la simulation rapide des CAN à partir de modélisations spécifiques. Une méthode originale a ainsi été développée pour le modulateur $\Sigma\Delta$ où l'analyse de performances est étroitement couplée à la simulation pour réduire le coût de calcul. Une modélisation propre aux circuits à capacités commutées à également été établie. Elle permet l'exploration des performances dynamiques pour une technologie donnée et fournit une estimation de la consommation.

Ce manuscrit est divisé en trois parties :

La première partie est consacrée à la caractérisation et au test des convertisseurs. Une attention particulière est portée à la modélisation linéaire des quantificateurs de faible résolution qui est exploitée au chapitre 9. On y rappelle également les limites fondamentales de performances des CAN.

La seconde partie est dédiée à l'étude des architectures de convertisseurs CMOS rapides de type flash, pipeline et $\Sigma\Delta$ avec, pour chacun d'eux, un état de l'art des réalisations actuelles. Les techniques de redondance qui ont permis de rendre ces architectures robustes vis à vis de la technologie seront mises en évidence (filtrage spatial pour le flash, redondance dans le codage pour le pipeline et modulation du bruit pour le $\Sigma\Delta$). Cette partie se termine par un examen des contraintes liées à l'évolution technologique sur les circuits échantillonnés.

La simulation comportementale fait l'objet de la troisième partie. La description structurelle des CAN à partir de classes C++ est présentée au chapitre 8. Une application à la simulation rapide du modulateur $\Sigma\Delta$ est donnée au chapitre 9. Elle permet l'optimisation des coefficients sous des contraintes de faible tension d'alimentation. Le chapitre 10 est consacré à la modélisation spécifique des circuits à capacités commutées. Un modèle simple pour le transfert de charge est utilisé pour l'exploration des performances dynamiques des convertisseurs pipeline et $\Sigma\Delta$.

Nous terminons ce manuscrit par les conclusions et les perspectives de recherche suscitées par cette approche.

Première partie

Caractérisation et test des CAN

Chapitre 1

Echantillonnage et quantification

1.1 Introduction

L'opération de numérisation du signal est en général effectuée en trois étapes qui sont :

- Le *filtrage* du signal pour remplir la condition de Nyquist sur l'opération suivante qu'est l'échantillonnage. Cette limitation du spectre vise également à réduire la dynamique du signal en particulier lorsque des signaux voisins indésirables sont de forte amplitude (cas fréquent en réception radio-fréquences).
- L'*échantillonnage* du signal pour faciliter l'opération suivante de conversion. Cette échantillonnage est souvent intimement lié à l'opération de conversion. C'est par exemple le cas des convertisseurs pipeline et $\Sigma\Delta$ qui utilisent principalement la technique des capacités commutées et qui sont par nature des circuits échantillonnés.
- La *conversion* analogique numérique qui attribue un code spécifique à chaque échantillon du signal de manière à pouvoir le traiter efficacement par le système numérique. L'efficacité est liée aux types d'opérations désirées. Dans le domaine des communications par exemple, le but ultime est en général d'extraire du signal un flux de symboles avec un taux d'erreur minimal. C'est alors principalement la forme des signaux qui importe plus que la connaissance précise de leur valeur à un instant donné (comme c'est par exemple le cas en métrologie).

Si toutes ces opérations sont nécessaires pour obtenir des échantillons discrets et quantifiés, l'ordre dans lequel elles sont effectuées est théoriquement indifférent. En pratique, l'échantillonnage procure l'avantage de s'affranchir du temps de propagation du signal dans certains blocs élémentaires comme les amplificateurs ou les comparateurs. Il permet également le traitement pipeline des opérations itératives afin d'accroître la vitesse de traitement. Si la dynamique des signaux le permet l'opération de filtrage peut être effectuée après la conversion. Cette technique est largement utilisée dans les convertisseurs $\Sigma\Delta$ qui utilisent une fréquence d'échantillonnage nettement supérieure à deux fois la bande du signal. Le filtrage des signaux parasites peut alors être efficacement réalisé dans le domaine numérique en même temps que l'opération de décimation qui vise à réduire la cadence d'échantillonnage et le bruit de quantification.

La réalisation d'un convertisseur analogique-numérique de résolution élevée nécessite en général des étapes de *quantifications intermédiaires* qui sont de résolution bien plus réduite. Le cas extrême est celui du convertisseur $\Sigma\Delta$ où la résolution des quantificateurs peut être de seulement un bit alors que la résolution globale peut être supérieure à 16 bit. Par ailleurs ces niveaux de quantification intermédiaires utilisent en général des convertisseurs rapides dont la complexité et la consommation impose une faible résolution. Nous attacherons donc beaucoup d'importance à la description précise de la quantification avec une faible résolution.

Nous allons voir plus en détails les opérations d'échantillonnage et de quantification sous leurs aspects théoriques. Les limitations fondamentales liées à leur implantation matérielle seront étudiées dans le chapitre 2.

1.2 Echantillonnage

L'opération d'échantillonnage consiste à transformer un signal $x(t)$ continu dans le temps en une suite d'échantillons pris à des instants discrets. Lorsque ces échantillons sont régulièrement espacés et que la transformée de Fourier $X(f) = F[x(t)]$ est nulle en dehors de la bande [-B,+B] le signal peut être parfaitement reconstitué à partir de la décomposition de Shannon :

$$x(t) = \sum_{n \in Z} x(n\,T) W(t - n\,T) \tag{1.1}$$

$$\text{avec} \qquad W(t) = \frac{sin(\pi t/T)}{\pi t/T} \qquad T = \frac{1}{2B}$$

Le spectre périodique $X_d(f)$ du signal échantillonné est obtenu par translation du spectre $X(f)$ avec une période 2B. Si le spectre du signal n'est pas nul en dehors de la bande [-B,+B] le signal échantillonné a toujours pour transformée de Fourier :

$$X_d(f) = \sum_{k \in Z} X(f - k\,F_s) \quad \text{avec} \quad F_s = 2\,B \tag{1.2}$$

mais dans ce cas on a un phénomène de *recouvrement de spectre* (ou aliasing) et le support [-B,+B] contient une version distordue du spectre du signal (figure 1.1).

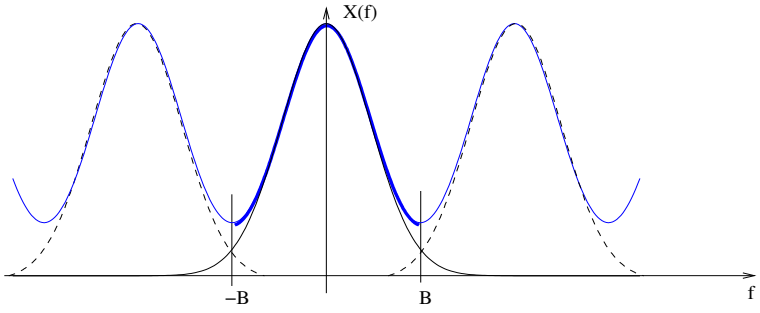


FIG. 1.1 – Recouvrement de spectre

Dans le cas d'un signal aléatoire, la densité spectrale de puissance du signal échantillonné est obtenue en remplaçant dans la formule 1.2, $X(f)$ par $S(f)$, densité spectrale du signal d'entrée. C'est par exemple le cas du bruit thermique large bande présent

dans le système d'échantillonnage. La puissance totale de l'erreur qui est accumulée par le phénomène de recouvrement spectral est alors :

$$P_{rec} = \int_{-\infty}^{-\frac{F_s}{2}} S(f)df + \int_{\frac{F_s}{2}}^{\infty} S(f)df \tag{1.3}$$

Le filtrage préalable du signal et le choix d'une fréquence d'échantillonnage F_s suffisante peuvent permettre de rendre cette erreur négligeable.

1.2.1 Suréchantillonnage

Le filtrage analogique très sélectif étant une opération coûteuse, on utilise très souvent une fréquence d'échantillonnage nettement supérieure à deux fois la bande du signal. La limitation de la bande du signal se fait alors dans le domaine numérique. La figure 1.2 montre la place respective du filtrage et de la conversion analogique-numérique (CAN).

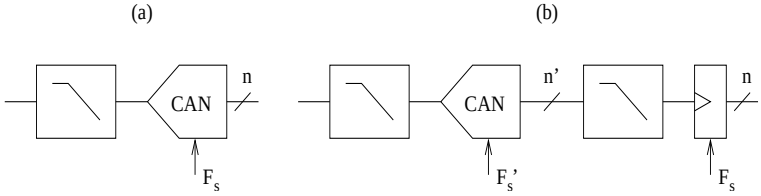


FIG. 1.2 – Choix de la fréquence d'échantillonnage : (a) $F_s = 2\,B$ (b) $F_s \gg 2\,B$

Dans le cas (b) on définit le rapport de suréchantillonnage :

$$R = \frac{F_s^{'}}{F_s}$$

La bande de transition du filtre analogique étant élargie, cette configuration permet d'en réduire l'ordre et de simplifier ainsi sa réalisation. Le filtrage numérique après conversion peut être réalisé efficacement dans les technologies actuelles et autorise par ailleurs une plus grande souplesse de traitement. Le système (b) est également intéressant pour la réduction du bruit de quantification qui sera étudié dans la seconde partie. En effet, celui-ci peut généralement être considéré comme un bruit blanc qui occupe la bande $[-\frac{F_s^{'}}{2}, \frac{F_s^{'}}{2}]$. La réduction de la bande passante dans le rapport R va réduire la puissance du bruit de quantification dans les mêmes proportions. Ainsi le rapport signal sur bruit de quantification est doublé (augmente de 3 dB) lorsque l'on double la fréquence d'échantillonnage. On peut obtenir un gain bien plus conséquent en effectuant une mise en forme spectrale du bruit de quantification avant filtrage. Ceci est possible en associant plus étroitement le filtrage analogique et la quantification. C'est le principe de la conversion $\Sigma\Delta$ qui sera étudiée au chapitre6.

1.2.2 Echantillonnage des signaux à bande étroite

Pour les signaux à bande étroite, la condition d'échantillonnage précédente est trop restrictive. Pour ces signaux le spectre est limité à une bande $B = [f_l, f_u]$ centrée en

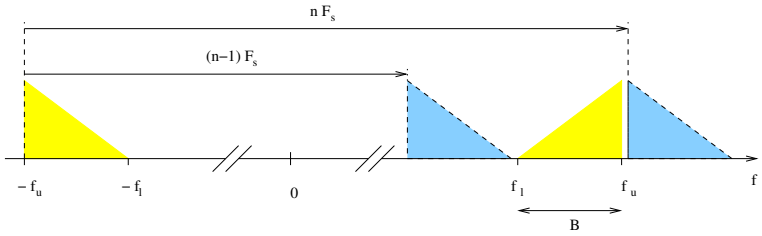


FIG. 1.3 – Signal passe-bande échantillonné.

F_o et telle que $\frac{B}{F_o} \ll 1$. La figure 1.3 montre un exemple d'un tel signal ainsi que les spectres translatés voisins dus à l'échantillonnage.

Les conditions de non-recouvrement des spectres translatés peuvent être facilement obtenues à partir de la figure 1.3 :

$$\begin{aligned} -f_l + (n-1)\,F_s &\leq f_l \\ -f_u + n\,F_s &\geq f_u \end{aligned} \tag{1.4}$$

Ces relations conduisent à la condition suivante sur la fréquence d'échantillonnage [127] :

$$\begin{aligned} \frac{2\,f_u}{n} \;\leq\; F_s \;\leq\; & \frac{2\,f_l}{n-1} \\ 1 \;\leq\; n \;\leq\; & \left\lfloor \frac{f_u}{B} \right\rfloor \end{aligned} \tag{1.5}$$

où $\lfloor x \rfloor$ désigne la partie entière de x.

Lorsque $\frac{f_u}{B}$ est un entier il existe un minimum de la fréquence d'échantillonnage :

$$F_s \geq 2\,B \tag{1.6}$$

Ce minimum est cependant purement théorique, car il est nécessaire de prévoir une certaine variation de la fréquence d'échantillonnage. Les contraintes sur la précision de cette fréquence sont d'ailleurs d'autant plus fortes que l'indice n de la formule 1.5 est élevé. Ceci est clairement visible sur la figure 1.4 qui montre les régions permises pour les premiers indices [127].

Une limitation importante de l'échantillonnage passe-bande est liée à l'accroissement du bruit. Le bruit du système d'échantillonnage est en effet présent sur les $(n-1)$ bandes intermédiaires comme on peut le voir sur la figure 1.3. Après échantillonnage, la puissance totale de ce bruit est affectée à la bande $[-\frac{F_s}{2}, \frac{F_s}{2}]$. La dégradation en rapport signal sur bruit sera donc d'autant plus grande que l'indice n est élevé. Ceci conduit, en général, à une fréquence d'échantillonnage bien supérieure à la limite imposée par la formule 1.6.

1.3 Quantification

La quantification est l'opération de discrétisation de l'amplitude du signal qui consiste à transformer un ensemble continu de valeurs en une suite finie de valeurs discrètes. C'est une opération non linéaire qui s'accompagne nécessairement d'une perte d'information. La définition d'un quantificateur scalaire peut se décomposer en trois étapes [85] :

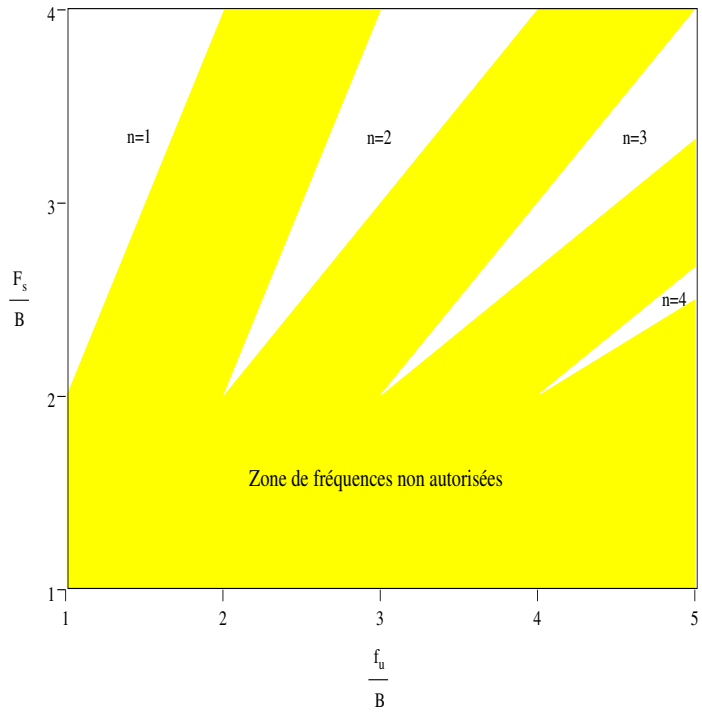


FIG. 1.4 – Fréquences permises pour l'échantillonnage passe-bande.

1. Une partition du domaine X de l'entrée en intervalles X_i disjoints et tels que $\bigcup_i X_i = X$.
2. L'attribution d'un code spécifique C_i à chaque intervalle.
3. La définition d'un représentant y_i pour chaque intervalle X_i avec $y_i \in X_i$.

La donnée de l'ensemble des partitions est équivalente à celle des points de transitions T_i entre deux intervalles X_i et X_{i+1}. Ainsi un quantificateur scalaire est complètement définit par le vecteur des transitions et celui des représentants. On notera le caractère statique de cette définition qui ne prend pas en compte l'effet des variations du signal sur la caractéristique du quantificateur. Dans la pratique, les seuils de transitions peuvent évoluer avec le signal (hysteresis des comparateurs, bruit thermique...). Ces effets sont en général considérés comme un bruit supplémentaire qui va se superposer à l'erreur de quantification déduite du modèle statique.

Les opérations élémentaires décrites précédemment sont, en pratique, le résultat de l'association de deux composants : Le convertisseur analogique-numérique (CAN) et le convertisseur numérique-analogique (CNA) (figure 1.5).

L'association d'un code C à une partition de l'entrée E est réalisée par un CAN. C'est en général un convertisseur rapide de type flash que l'on étudiera en détail dans le chapitre 4. Le CNA associe une valeur réelle au code numérique C. Dans les circuits à capacités commutées qui seront envisagés dans la suite cette conversion est efficacement réalisée par une pondération de charges. Nous en verrons des exemples dans le chapitre 5.

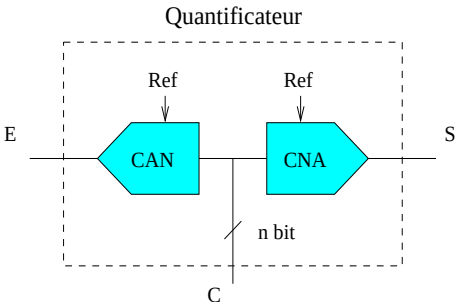


FIG. 1.5 – Réalisation pratique du quantificateur.

1.3.1 Erreur quadratique moyenne et condition d'optimalité

La mesure de l'erreur quadratique moyenne introduite par le quantificateur peut être calculée à partir de la densité de probabilité $p(x)$ du signal d'entrée :

$$D = \sigma_q^2 = \sum_{i=1}^{N_Q} \int_{X_i} (x - y_i)^2 p(x) dx \tag{1.7}$$

où N_Q est le nombre de niveaux de quantification.

La recherche du quantificateur optimum consiste à minimiser D sur l'ensemble des partitions et des représentants associés.

Lorsque les représentants y_i sont donnés, les transitions optimales sont obtenues pour :

$$T_i = \frac{y_i + y_{i+1}}{2} \tag{1.8}$$

De même, lorsque l'ensemble des partitions est donné, les meilleurs représentants sont tels que :

$$y_i = \frac{\int_{X_i} x\ p(x) dx}{\int_{X_i} p(x) dx} \tag{1.9}$$

C'est seulement dans le cas d'une distribution gaussienne pour $p(x)$ que les conditions 1.8 et 1.9 garantissent l'optimalité du quantificateur.

Dans l'hypothèse d'une grande résolution du quantificateur où p(x) peut être supposée constante sur un intervalle X_i la formule 1.7 devient :

$$D = \sigma_q^2 = \sum_{i=1}^{N_Q} p_i \frac{\|X_i\|^2}{12} \tag{1.10}$$

où p_i représente la probabilité que le signal appartienne à l'intervalle X_i et $\|X_i\|$ est la mesure de l'intervalle X_i.

On peut également chercher à maximiser la quantité moyenne d'information générée par la source discrète analogique à la sortie du quantificateur. Celle-ci est indépendante de l'échantillon si la source est stationnaire. C'est l'entropie H de cette source :

$$H = -\sum_{i=1}^{N_Q} p_i log_2(p_i) \tag{1.11}$$

Elle est maximum lorsque toutes les probabilités p_i sont égales à $\frac{1}{N_Q}$.

Ces deux approches (minimiser D ou maximiser H) peuvent conduire à des résultats similaires pour une classe particulière de densité de probabilité [81].

1.3.2 Quantification uniforme

En l'absence d'une connaissance à priori de la loi de probabilité du signal ou lorsque la quantification doit être appliquée à différents types de signaux le choix le plus simple consiste à effectuer une partition du domaine de l'entrée en intervalles de même largeur Δ. Ce choix conduit par ailleurs a une réalisation pratique plus simple et sera le seul cas étudié dans la suite. La sortie s en fonction de l'entrée e d'un tel quantificateur est représentée à la figure 1.6 ainsi que l'erreur de quantification $e_q = e - s$. Pour cet

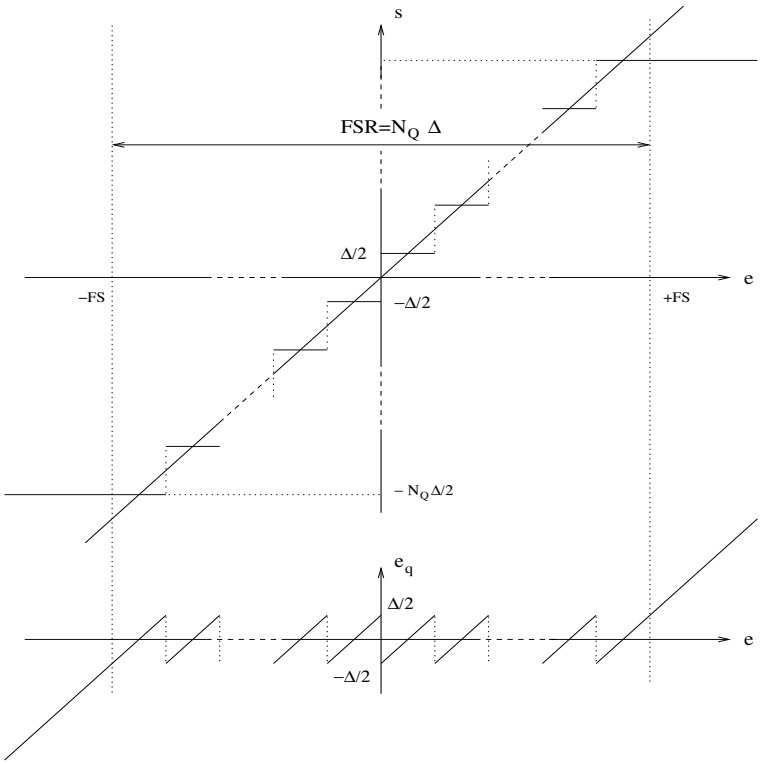


FIG. 1.6 – Quantificateur uniforme

exemple la caractéristique est symétrique autour de l'origine et le nombre de niveaux de quantification N_Q est pair. Il existe un domaine de l'entrée du quantificateur où l'erreur de quantification e reste inférieure à $\frac{\Delta}{2}$. La largeur FSR (*full-scale range*) de cette plage est reliée au pas de quantification Δ par la relation : $FSR = N_Q\,\Delta$. Dans cette région chaque niveau y_k est associé au centre de l'intervalle (semi-ouvert) :

$$X_k = [T_0 + k\,\Delta,\, T_0 + (k+1)\,\Delta)$$

où T_0 est un décalage qui vaut $-N_Q\frac{\Delta}{2}$ dans le cas symétrique de la figure 1.6. En dehors de ce domaine l'erreur croit de manière monotone et on dit que l'on a atteint la saturation du quantificateur. Un quantificateur uniforme est ainsi complètement caractérisé par le nombre N_Q de pas de quantification, la pleine échelle FSR et le

décalage T_0 de sa caractéristique. Lorsque $T_0 = 0$ on dit que le quantificateur est unipolaire. Il est dit bipolaire lorsque la caractéristique est symétrique par rapport à l'origine. Les signaux de communications étant en général signés, seul ce dernier cas sera envisagé. On notera que la valeur $y_k = 0$ est présente uniquement lorsque le nombre de niveaux de quantification est impair. On ne considérera que le cas N_Q pair, celui-ci étant en général une puissance de deux pour un codage efficace des niveaux de quantification. Un quantificateur de n bit aura ainsi $N_Q = 2^n$ niveaux.

densité de probabilité et fonction caractéristique

La densité de probabilité et la fonction caractéristique du signal quantifié jouent un rôle analogue à celui du signal et de sa transformée de Fourier dans l'échantillonnage temporel. En notant $f_q(x)$ la fonction indicatrice de l'ensemble $[0, \Delta)$, la probabilité d'avoir le niveau y_k est donnée par :

$$P(y_k) = \int_{X_k} p_x dx = \int f_q(x - k\,\Delta) p_x dx = f_q \star p_x(k\,\Delta) \tag{1.12}$$

La densité de probabilité du signal quantifié p_q est donc un échantillonnage de période Δ de la fonction $f_q \star p_x$ qui résulte de la convolution de la densité de probabilité du signal par la fonction f_q soit en notant Π_Δ le peigne de Dirac de période Δ :

$$p_q = [p_x \star f_q] \cdot \Pi_\Delta \tag{1.13}$$

En prenant la transformée de Fourier de l'équation (1.13) on obtient la fonction caractéristique ($\Phi_\nu(u) = E\{e^{-2j\pi\,\nu\,u}\}$) du signal quantifié Φ_q en fonction de celle de l'entrée Φ_x :

$$\Phi_q(u) = [\Phi_x(u) \cdot e^{j\pi u\Delta} \cdot \frac{sin(\pi u\Delta)}{\pi u\Delta}] \star \Pi_{\frac{1}{\Delta}} \tag{1.14}$$

On note une grande analogie avec le spectre du signal échantillonné. La fonction caractéristique du signal quantifié est également périodique de période $\frac{1}{\Delta}$. La reconstitution du signal à partir de sa version quantifiée nécessite également l'absence de recouvrement des translatées de la fonction caractéristique :

$$\Phi_x(u) = 0 \quad \text{pour} \quad |u| > \frac{1}{2\Delta} \tag{1.15}$$

Ceci constitue le premier théorème de quantification dû à Widrow [132].

On peut tirer profit de la grande analogie qui existe entre le recouvrement spectral dans le processus d'échantillonnage et celui des fonctions caractéristiques dans le processus de quantification pour déterminer un moyen de satisfaire plus efficacement la contrainte précédente. En effet, dans le processus d'échantillonnage un moyen efficace de limiter le recouvrement spectral est d'appliquer un pré-filtrage, c'est à dire de pondérer le spectre du signal. Dans le domaine des fonctions caractéristiques, ceci revient à multiplier la fonction caractéristique du signal par une autre moins étalée. Le produit au niveau des fonctions caractéristiques correspond à une convolution au niveau des densités de probabilité. Par ailleurs, la densité de probabilité de la somme de deux variables aléatoires indépendantes est obtenue par la convolution de leurs densités respectives. La solution est donc d'ajouter au signal d'entrée un bruit blanc qui va limiter l'étendue de sa fonction caractéristique. Cette source de bruit (*dither*) est très souvent associée au quantificateur. Pour l'étude de l'effet de différentes sources de bruit additif sur la résolution du quantificateur on pourra consulter la référence [16].

Dans le cas d'un convertisseur analogique-numérique cette source de bruit existe naturellement, c'est le bruit thermique ou toute autre source de bruit naturelle qui n'est pas liée au signal.

Modèle linéaire du quantificateur

Le modèle linéaire de quantificateur de la figure 1.7 consiste à approximer la caractéristique de la figure 1.6 par un gain K et un bruit additif uniforme, blanc et indépendant du signal. L'avantage de cette modélisation est de permettre l'utilisa-

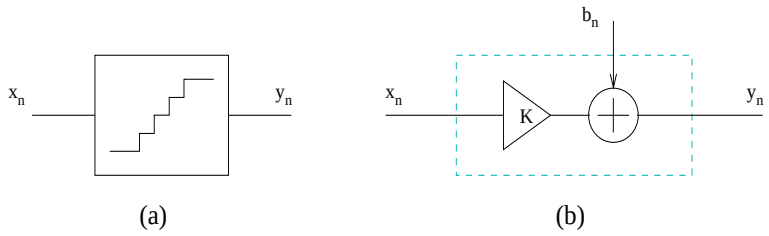


FIG. 1.7 – Modèle linéaire de quantificateur.

tion des méthodes de traitement linéaire pour estimer le rapport signal sur bruit. Elle justifie l'utilisation fréquente du terme *bruit de quantification* alors qu'il serait plus judicieux de parler *d'erreur de quantification*. Bennett [9] a montré la validité de ce modèle dans les conditions limites suivantes :

- grande résolution (Δ petit devant l'amplitude du signal).
- absence de saturation du quantificateur.
- la densité de probabilité conjointe de deux échantillons d'entrée est suffisamment régulière.

Sous cette hypothèse de grande résolution, la puissance du bruit de quantification peut être déduite de la formule (1.10) avec $\|X_i\| = \Delta$, $\forall i$ soit $P_e = \frac{\Delta^2}{12}$. Malheureusement ces conditions sont rarement satisfaites pour un quantificateur de faible résolution. Widrow a cependant montré que ce modèle est valide si des conditions moins restrictives (formule 1.15) sont vérifiées [132].

Afin de montrer la prudence qui doit accompagner l'utilisation de cette modélisation nous considérons le cas d'un signal sinusoïdal de phase aléatoire uniforme qui est très utilisé pour le test des convertisseurs.

Puissance de bruit Le bruit de quantification (normalisé par rapport à Δ^2) dans le cas d'une entrée sinusoïdal d'amplitude A avec un pas de quantification Δ est donné par [47] :

$$E\{e_n^2\} = \frac{1}{12} + \frac{1}{\pi^2}\sum_{k=1}^{\infty}\frac{1}{k^2}J_o(2\pi k\frac{A}{\Delta}) \tag{1.16}$$

où J_o est la fonction de Bessel de première espèce et d'ordre 0. Le second terme de cette expression est représenté à la figure 1.8. Lorsque l'amplitude augmente, ce second terme tend vers zéro et l'erreur quadratique tend vers $\frac{1}{12}$ qui correspond bien au cas d'une densité de probabilité uniforme du bruit de quantification.

Densité spectrale de puissance Dans la zone non-saturée de la figure 1.6 l'erreur de quantification est périodique et admet un développement en séries de Fourier. Le spectre du bruit de quantification pour une séquence sinusoïdale d'amplitude A et

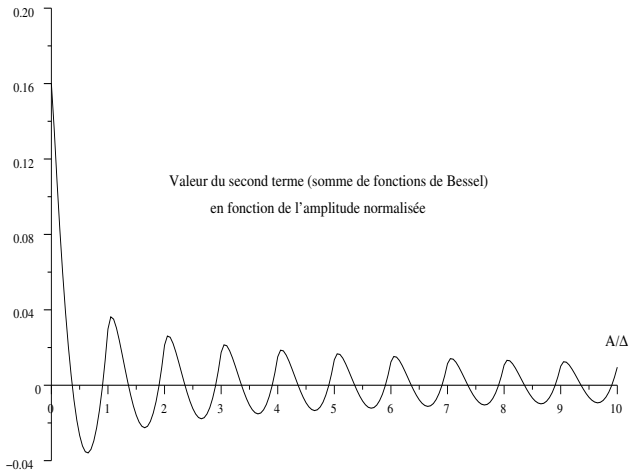


FIG. 1.8 – Ecart par rapport au modèle théorique du bruit de quantification.

de fréquence f_o (normalisée par rapport à la fréquence d'échantillonnage) peut être obtenu à partir de ce développement [47] :

$$S_e(f) = \sum_{m=-\infty}^{\infty} S_m \delta(f - f_m)$$
$$\text{avec} \quad S_m = (\frac{1}{\pi} \sum_{k=1}^{\infty} \frac{1}{k} J_{2m-1}(2\pi k \frac{A}{\Delta}))^2 \tag{1.17}$$

C'est un spectre de raies de fréquences $f_m =< (2m-1)\, f_o >$. Seules les harmoniques impaires du signal sont présentes et la partie fractionnaire $<>$ est liée à la nature échantillonnée du signal où toutes les composantes sont dans la bande de Nyquist. La figure 1.9 représente le rapport signal sur distorsion en ne considérant que la première raie $f_1 = 3f_o$. Celui-ci présente de grandes variations en fonction de l'amplitude du signal. On peut borner inférieurement sa valeur à 9n (dB) où n est le nombre de bit du quantificateur [94]. Cette borne est également représentée sur la figure 1.9.

On remarque que le spectre est bien différent du bruit blanc prévu par le modèle linéaire. La présence de ces raies fixe d'ailleurs une limite à la dynamique libre de toute raie parasite (SFDR). Ce paramètre est important lorsque l'on veut numériser des signaux de faible amplitude en présence de signaux dont l'amplitude est proche de la pleine échelle du quantificateur et qui sont susceptibles de provoquer des raies parasites dans la bande du signal désiré. La limite établie pour $f_1 = 3f_o$ reste pratiquement valable pour l'ensemble des composantes spectrales du bruit de quantification (bien que théoriquement certaines raies d'ordre élevé puissent avoir une amplitude supérieure à cette limite, le seul bruit thermique provoque un étalement du spectre qui les rend généralement négligeables [94]).

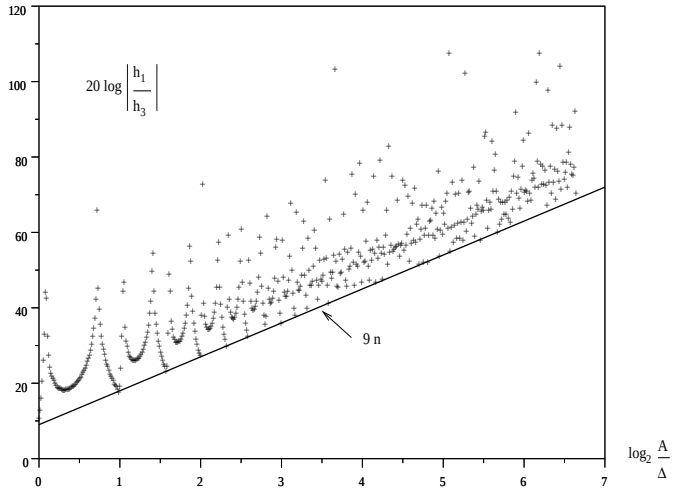


FIG. 1.9 – Rapport signal sur distorsion liée à la première harmonique $f = 3\,f_o$ en fonction de l'amplitude normalisée.

Rapport signal sur bruit et gain du quantificateur

Les performances dynamiques du quantificateur linéaire sont déterminées par le maximum du rapport signal sur bruit qui peut être obtenu pour une résolution donnée du quantificateur et pour un type particulier de signal. Ainsi, dans le cas d'une grande résolution avec $N_Q = 2^n$ où n est le nombre de bit du quantificateur et pour un signal de densité de probabilité uniforme sur $[-FS, FS]$ le rapport signal sur bruit est :

$$SNR = 10\,Log_{10}\frac{P_s}{P_q} = 20\,Log_{10}\frac{FSR}{\Delta} \simeq 6\,n\ dB \tag{1.18}$$

Ce rapport signal sur bruit a bien peu de chance d'être atteint en pratique où la nature des signaux est plus souvent de type gaussien. Pour un tel signal, avec un écart-type $\sigma = \frac{FS}{4}$ pour s'affranchir de toute saturation, on a seulement :

$$SNR = 10\,Log_{10}\frac{\frac{FSR^2}{64}}{\frac{\Delta^2}{12}} \simeq 6\,n - 7,3\ dB \tag{1.19}$$

Enfin pour un signal sinusoïdal $A\,sin(\omega t + \phi)$ avec $A = FS$ on a :

$$SNR = 10\,Log_{10}\frac{\frac{FSR^2}{8}}{\frac{\Delta^2}{12}} \simeq 6\,n + 1,76\ dB \tag{1.20}$$

Le fait de considérer particulièrement ce signal est lié a son utilisation privilégiée dans le test des convertisseurs qui sera étudié au chapitre 3.

Nous avons jusqu'ici considéré que le gain du quantificateur était unitaire dans le modèle de bruit additif. Cette hypothèse est très bien vérifiée lorsque la résolution est élevée et l'amplitude du signal suffisante. Elle n'est plus valide dès que l'amplitude du

signal est supérieure à $\frac{FS}{2}$ ou inférieure au pas de quantification Δ [24]. Ce dernier cas est particulièrement important pour les convertisseurs $\Sigma\Delta$ qui seront analysés dans le chapitre 6. En effet, ceux-ci peuvent être construits à partir de quantificateurs n'ayant qu'un seul bit, c'est à dire deux niveaux $\pm\frac{\Delta}{2}$. Le gain du quantificateur doit alors être choisi pour minimiser l'erreur de quantification. Ce gain est alors dépendant de l'amplitude du signal, ce qui est la caractéristique d'un composant fortement non linéaire. Pour prendre en compte ce phénomène nous reprenons la formule (1.7) en y introduisant le gain linéaire g correspondant à une densité de probabilité $p(x)$ donnée pour le signal :

$$\sigma_q^2 = \sum_{i=1}^{N_Q} \int_{X_i} (g \cdot x - y_i)^2 p(x) dx \tag{1.21}$$

Le développement du carré nous donne :

$$\sigma_q^2 = g^2\sigma_x^2 - 2 \cdot g \cdot v + p \tag{1.22}$$

avec
$$v = \sum_{i=1}^{N_Q} y_i \int_{X_i} x \cdot p(x) dx = E\{x \cdot y\} \tag{1.23}$$

$$p = \sum_{i=1}^{N_Q} {y_i}^2 \int_{X_i} p(x) dx = \sum_{i=1}^{N_Q} {y_i}^2 \cdot P(y_i) = E\{y^2\} \tag{1.24}$$

On remarque que l'expression p représente la variance totale en sortie. D'après le théorème de projection, l'erreur quadratique minimale est orthogonale au signal, d'où :

$$p = g^2 \cdot \sigma_x^2 + \sigma_q^2 \tag{1.25}$$

La combinaison de (1.22) et (1.25) conduit au gain :

$$g = \frac{v}{\sigma_x^2} \tag{1.26}$$

On peut également exprimer le rapport signal sur bruit à partir des expressions (1.23) et (1.24) :

$$SNR = \frac{v^2}{p \cdot \sigma_x^2 - v^2} \tag{1.27}$$

Les figures 1.10 et 1.11 montre respectivement l'évolution du rapport signal sur bruit et du gain pour un signal gaussien centré d'écart type σ et un nombre de bit variant de 1 à 8. Le maximum du rapport signal sur bruit en fonction de la résolution est également reporté dans le tableau 1.1 avec l'évaluation obtenue par la formule 1.19. On remarque un écart d'autant plus important que la résolution diminue. Ceci est dû au fait que pour ces faibles résolutions le gain effectif du quantificateur est très dépendant de l'amplitude d'entrée et s'écarte de l'unité. Le cas extrême est celui du quantificateur 1 bit dont le gain subit les plus fortes variations. Pour être valide, le modèle linéaire doit donc inclure ce paramètre, en particulier lorsque la résolution du quantificateur est très faible.

n	1	2	3	4	5	6	7	8
$SNR_{max}(dB)$	2,435	8,7	14,1	19,32	24,54	29,8	35,16	40,56
Formule (1.19)	-1,3	4,7	10,7	16,7	22,7	28,7	34,7	40,7

TAB. 1.1 – Maximum du SNR en fonction de la résolution

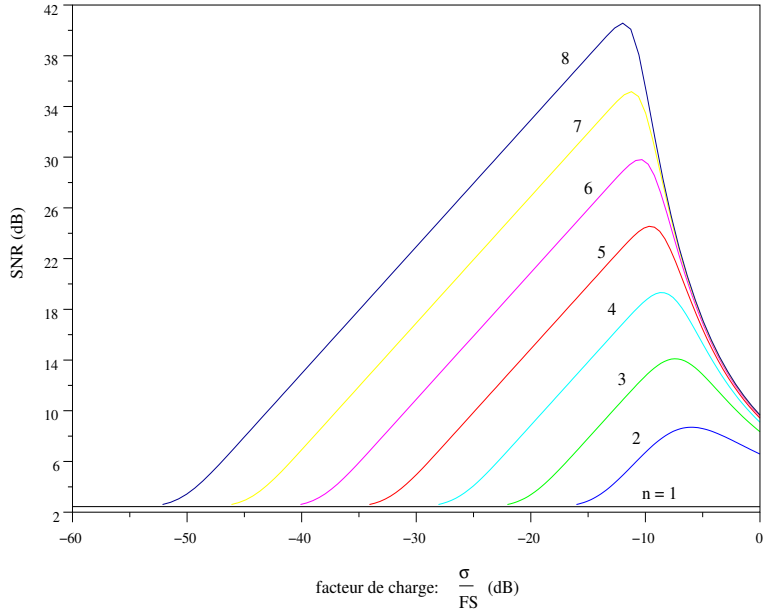


FIG. 1.10 – Rapport signal sur bruit en fonction du facteur de charge $\frac{\sigma}{FS}$ et du nombre de bit n du quantificateur.

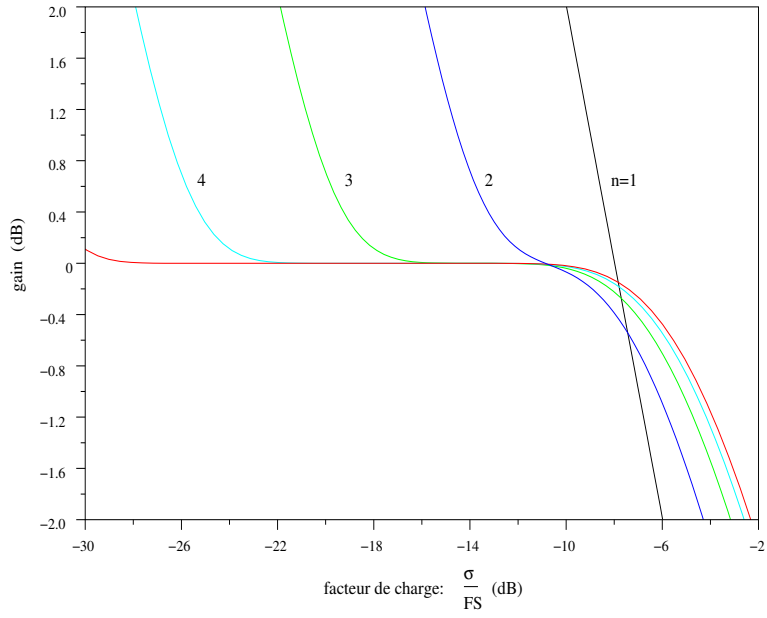


FIG. 1.11 – Gain du quantificateur en fonction du facteur de charge $\frac{\sigma}{FS}$.

Chapitre 2

Caractérisation et limites fondamentales des CAN

2.1 Introduction

Les paramètres qui décrivent un CAN sont liés à un environnement (température, fréquence d'échantillonnage, tension d'alimentation, tension de référence, source d'excitation...) et au résultat d'un ensemble de mesures qui sont effectuées sur le composant (puissance dissipée, nombre de bit effectifs...). Dans ce chapitre nous donnons les paramètres essentiels qui peuvent être obtenus à partir de ces mesures. Pour une description complète des différentes spécifications on pourra consulter le standard IEEE-1241 [56] consacré à la terminologie et au test de ces composants.

Nous étudions ensuite les erreurs fondamentales qui fixent les limites théoriques sur la résolution et la fréquence d'échantillonnage du convertisseur.

2.2 Caractérisation des CAN

Le signal d'entrée (grandeur analogique continue) et de sortie (code discret) d'un convertisseur analogique-numérique étant de nature différente, la caractérisation d'un tel composant nécessite une nouvelle conversion pour comparer des grandeurs de même nature. Cette conversion du code vers une grandeur analogique est supposée idéale. L'ensemble de ces deux éléments définit un quantificateur dont les seuls défauts sont associés à la conversion analogique-numérique.

La définition d'un quantificateur idéal qui a été faite au chapitre 1 est une description purement statique qui considère une caractéristique parfaitement définie et indépendante du signal. Dans un quantificateur réel, deux types de défauts vont apparaître. D'une part, la caractéristique statique va s'écarter de sa valeur nominale (par exemple du fait des décalages dans les comparateurs). Ceci définit les paramètres statiques du convertisseur. D'autre part, les variations rapides du signal peuvent introduire des erreurs supplémentaires liées, par exemple, à la bande passante finie des amplificateurs, à l'introduction de temps de propagation ou à l'hystérésis des comparateurs. Les paramètres dynamiques prennent en compte cette deuxième source d'erreurs.

2.2.1 Caractéristiques statiques

Comme on l'a vu au chapitre 1 un quantificateur uniforme parfait est complètement défini par le nombre de pas de quantification, la pleine échelle et la position de sa caractéristique par rapport à l'origine. En pratique, il existe une dispersion des seuils de transitions par rapport à leurs positions idéales. Les paramètres statiques ont pour objet de définir cette caractéristique statique en prenant en compte cette dispersion.

- Pleine échelle (FSR)
  C'est la différence entre la valeur maximum et la valeur minimum du signal d'entrée. Pour un convertisseur de n bit on a $FSR = 2^n\Delta$ où Δ est le pas de quantification nominal du convertisseur.
- Seuils de transitions ($T(k)$) et pas de quantification ($W(k)$)
  On note$T(k)$ le niveau de transition entre le code $k-1$ et le code k. C'est la valeur du signal d'entrée telle que la probabilité d'avoir le code k soit égale à la probabilité d'avoir le code $k-1$.
  On note $W(k) = T(k+1) - T(k)$ le pas de quantification lié au code k. Idéalement celui-ci est égal à Δ. Ces deux définitions sont illustrées à la figure 2.1 pour un convertisseur de n bit.

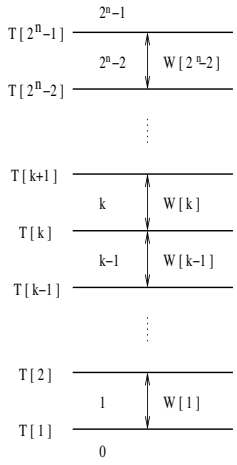


FIG. 2.1 – Niveaux de transitions et pas de quantification

- Gain (G) et décalage (D)
  Ces valeurs sont telles qu'elles minimisent au sens des moindres carrés l'erreur $\epsilon[k]$ pour l'ensemble des transitions dans l'équation suivante :

$$G \cdot T[k] + D + \epsilon[k] = (k-1)\,Q + T_1 \tag{2.1}$$

  où T_1 est la valeur nominale de la première transition et Q est le pas de quantification moyen :

$$Q = \frac{1}{2^n - 2} \sum_{k=1}^{2^n-2} W(k) = \frac{T[2^n - 1] - T[1]}{2^n - 2} \tag{2.2}$$

  Idéalement on a $G = 1$ et $D = 0$. On notera qu'il existe d'autres définitions pour ces paramètres selon le critère retenu pour l'erreur $\epsilon[k]$. On peut, par exemple annuler $\epsilon[k]$ pour les points extrêmes ou rechercher le minimum de l'erreur absolue $|\epsilon[k]|$ sur tous les k.

- Non linéarité intégrale (INL) et différentielle (DNL)
  La non linéarité intégrale est directement obtenue à partir de la formule 2.1. C'est en fait l'erreur $\epsilon[k]$ qui est en général normalisée par rapport au pas de quantification :

$$INL(k) = \frac{\epsilon[k]}{Q} \qquad k \in [1, .., 2^N - 1] \tag{2.3}$$

  La non-linéarité différentielle mesure l'écart du pas de quantification réel par rapport à sa valeur moyenne. Elle est définie par :

$$DNL(k) = \frac{W(k) - Q}{Q} \qquad k \in [1, .., 2^N - 2] \tag{2.4}$$

Erreur de quantification et non-linéarité

La non linéarité intégrale se traduit par un déplacement des seuils de transitions qui va provoquer un accroissement de l'erreur de quantification. L'erreur quadratique diffère également de la valeur théorique :

$$\sigma_q^2 = E\{e_q^2\} = \frac{\Delta^2}{12}$$

obtenue dans l'hypothèse d'une grande résolution.

Pour déterminer cet accroissement, considérons une portion de la caractéristique où seule l'erreur de non-linéarité intervient (figure 2.2). L'erreur ϵ_k représente la non-

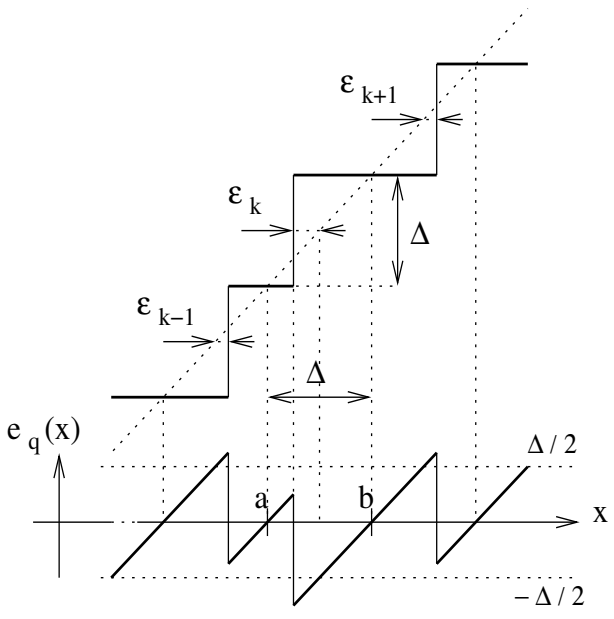


FIG. 2.2 – Erreur de quantification et non-linéarité

linéarité intégrale (formule 2.1 avec $G = 1$ et $V_{os} = 0$). On peut calculer l'erreur quadratique sur un pas de quantification Δ en supposant une distribution de probabilité uniforme du signal sur cet intervalle. Considérons par exemple l'intervalle X_k

comprise entre les points a et b de la figure 2.2. Ces deux points sont séparés par un quantum et l'erreur quadratique est donnée par :

$$\sigma_k^2 = \int_{-\frac{\Delta}{2}}^{-\epsilon_k} \left(x + \frac{\Delta}{2}\right)^2 \frac{dx}{\Delta} + \int_{-\epsilon_k}^{\frac{\Delta}{2}} \left(x - \frac{\Delta}{2}\right)^2 \frac{dx}{\Delta} = \frac{\Delta^2}{12} + \epsilon_k^2 \tag{2.5}$$

Si p_k est la probabilité que le signal soit dans l'intervalle X_k l'erreur totale devient :

$$\sigma_q^2 = \sum_k \sigma_k^2 \, p_k \tag{2.6}$$

où la somme s'étend à tous les intervalles couverts par le signal et qui ne sont pas dans la zone de saturation du quantificateur. Dans le cas d'une distribution uniforme du signal sur la pleine échelle, l'équation précédente conduit à :

$$\sigma_q^2 = \frac{\Delta^2}{12} + \frac{1}{2^n - 1} \sum_{k=1}^{2^n - 1} \epsilon_k^2 \tag{2.7}$$

Le second terme représente la norme quadratique de la non-linéarité intégrale. C'est, pour cette distribution particulière, l'erreur additionnelle due à la répartition non-uniforme des seuils de transitions.

2.2.2 Caractéristiques dynamiques

Une grande partie des caractéristiques des CAN est établie pour une source d'excitation sinusoïdale. Ceci est dû à la possibilité de générer pratiquement ce type de signal avec une grande pureté et à sa faculté à mettre en évidence les phénomènes de distorsion. Sauf mention contraire, on supposera que le signal d'excitation est une sinusoïde pure dont l'amplitude et la fréquence sont spécifiées, les différentes caractéristiques étant en général dépendantes de ces deux paramètres. Les signaux présents à l'entrée d'un CAN sont cependant généralement de natures plus complexes, en particulier dans les applications de communication. La mise en évidence des caractéristiques non linéaires nécessitent alors des excitations plus proches des signaux naturellement présent à l'entrée de ces systèmes tels que la combinaison de signaux sinusoïdaux, de signaux aléatoires large bande ou de signaux impulsifs.

- Nombre de bit effectifs (ENOB)
  C'est le nombre de bit de résolution effectifs en considérant le bruit total en sortie (quantification, distorsion, gigue d'horloge,...). Si σ_{tot} et σ_q représentent respectivement l'écart type du bruit total et celui de quantification (théorique) et si n est la résolution idéale (absence d'imperfection et de bruit) le nombre de bit effectif est donné par :

$$ENOB = n - \log_2 \frac{\sigma_{tot}}{\sigma_q} \tag{2.8}$$

- Rapport signal sur bruit plus distorsion (SINAD)
  Rapport entre puissance du signal et celle du bruit total en sortie. On emploie souvent le terme de SNR (*Signal to Noise Ratio*) pour la même mesure. Ce terme est cependant ambigu car également utilisé lorsque le bruit ne contient pas la distorsion (le standard 1241 [56] à introduit le terme SNHR pour ce cas particulier). On a la relation suivante entre le SINAD et le nombre de bit effectifs (cf formule 1.20) :

$$SINAD(dB) = 6\,ENOB + 1,76 \tag{2.9}$$

- Dynamique libre de raie parasite (SFDR)
  Rapport entre l'amplitude du fondamental et celle de la raie (pas nécessairement harmonique) parasite de plus forte amplitude dans la bande de Nyquist.
- Distorsion harmonique (THD)
  Rapport entre la puissance des M premières raies harmoniques et la puissance du fondamental (f_o). On devra prendre en compte un éventuel repliement spectral pour déterminer la fréquence de ces raies. Du fait de l'échantillonnage celles-ci sont en effet telles que :

$$f_h = (k \cdot f_o) \ \ modulo \ \ f_s \qquad k = \pm 2, \pm 3, \cdots, \pm M \tag{2.10}$$

- Intermodulation (IMD)
  L'intermodulation caractérise la non linéarité du convertisseur en présence de plusieurs signaux sinusoïdaux. Le cas le plus simple est celui de deux signaux aux fréquences f_1 et f_2. Ceux-ci donnent naissance aux produits d'intermodulations :

$$f_{im} = (i \cdot f_1 + j \cdot f_2) \ \ modulo \ \ f_s \tag{2.11}$$

  où i et j sont des entiers relatifs. Un produit d'intermodulation d'ordre k est tel que $|i| + |j| = k > 1$. Le paramètre d'intermodulation est alors défini comme le rapport entre le puissance relative à ces différents produits jusqu'à un ordre donné et celle du signal d'entrée.
- Rapport de puissance de bruit (NPR)
  Ce paramètre permet de caractériser le convertisseur en présence d'un signal large bande. L'entrée est un bruit blanc filtré par un réjecteur de bande étroite $B = f_2 - f_1$. Ce paramètre est donné par le rapport (en sortie du convertisseur) entre la puissance de bruit dans une bande de même largeur mais en dehors de celle du filtre réjecteur sur la puissance de bruit dans la bande $[f_1, f_2]$. Ce paramètre est particulièrement important pour caractériser la dégradation liée à l'intermodulation pour un système multi-canaux. La bande $[f_1, f_2]$ correspond dans ce cas à un canal du système et l'ensemble des autres canaux est assimilé au bruit blanc filtré.

2.3 Limites fondamentales

Les paramètres tels que la fréquence d'échantillonnage et la résolution sont très dépendants de la technologie et de l'architecture du convertisseur. On peut cependant dégager certaines limites fondamentales qui affecterons pratiquement tout type de convertisseur et qui font apparaître les barrières à lever pour améliorer les performances de ces composants.

2.3.1 Dynamique libre de raies parasites

Ce paramètre est très important dans les systèmes ou le signal utile, de faible amplitude, peut être accompagné d'un signal perturbateur, de forte amplitude, sensé être éliminé par filtrage numérique après conversion. Du fait des non linéarités introduites par le perturbateur, des raies parasites peuvent apparaître dans la bande utile et dégrader fortement le rapport signal sur bruit. On a vu au chapitre 1 que pour un quantificateur idéal de n bit, il existe une limite théorique de $SFDR \simeq 9\,n\ dB$ à ce paramètre. Cette grandeur est illustrée sur la figure 2.3 qui représente le spectre d'un

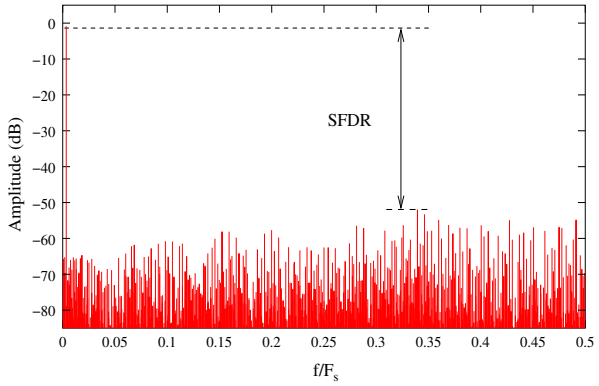


FIG. 2.3 – Dynamique libre de raie parasite.

convertisseur idéal de 6 bit. Les non linéarités additionnelles conduisent cependant à une valeur plus faible du $SFDR$. Ce paramètre, comme le SINAD, évolue avec la fréquence du signal du fait des erreurs dynamiques et ces deux paramètres sont généralement tracés en fonction de cette fréquence. Une étude due à Walden [130] faite sur un grand nombre de convertisseurs montre que l'accroissement de performances en terme de SFDR est très lent, correspondant à une amélioration de l'ordre de 6 dB en huit ans pour une fréquence d'échantillonnage donnée. Les besoins spécifiques aux applications de communications ont cependant suscités des travaux récents [94] pour améliorer ce paramètre.

2.3.2 Incertitude d'échantillonnage

L'opération d'échantillonnage qui a été décrite au chapitre 1 suppose une horloge parfaite, dont la période est bien définie et stable dans le temps. Le bruit intrinsèque aux dispositifs qui vont produire cette horloge (oscillateur fixe ou variable, boucle à verrouillage de phase,...) introduit cependant une certaine incertitude t_j (*jitter*) sur l'instant exact d'échantillonnage :

$$t = kT + t_j(kT) \tag{2.12}$$

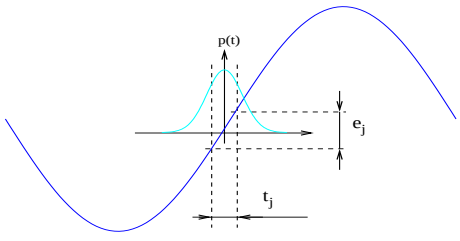


FIG. 2.4 – Incertitude d'échantillonnage.

Lors de l'échantillonnage d'un signal $x(t)$ avec la période T et en supposant que l'incertitude temporelle est suffisamment petite, on a au premier ordre une erreur

correspondante sur l'amplitude (figure 2.4) :

$$e_j(kT) = t_j(kT)\frac{\partial x}{\partial t}|_{kT} = t_j(kT) \cdot x^{'}(kT) \tag{2.13}$$

Pour un signal aléatoire, avec une incertitude d'échantillonnage indépendante et de moyenne nulle, on peut déterminer la fonction d'autocorrélation de cette erreur :

$$R_{ej}(kT) = R_{x'}(kT) \cdot R_{tj}(kT) = -R_x^{''}(kT) \cdot R_{tj}(kT) \tag{2.14}$$

où on a utilisé le fait que l'autocorrélation de la dérivée d'un processus est la dérivée au second ordre de l'autocorrélation du processus [26]. La densité spectrale correspondante s'en déduit par transformation de Fourier :

$$S_{ej}(f) = [(2\pi f)^2 S_x(f)] \star S_{tj}(f) \tag{2.15}$$

où $(\star)$ désigne l'opération de convolution. La formule (2.14) permet d'obtenir une expression générale pour le rapport signal sur bruit dû à l'incertitude d'échantillonnage :

$$SNR_j(dB) = 10\ log_{10}\frac{R_x(0)}{R_{ej}(0)} = 10\ log_{10}\frac{R_x(0)}{-R_x^{''}(0) \cdot R_{tj}(0)} \tag{2.16}$$

La connaissance de l'autocorrélation R_x du signal et de la variance $\sigma_j^2 = R_{tj}(0)$ de l'incertitude temporelle sont ainsi suffisantes pour évaluer les performances de l'échantillonnage en terme de rapport signal sur bruit [26].

Pour un signal sinusoïdal $x(t) = Acos(\omega t + \phi)$ où ϕ est une phase aléatoire uniforme sur $[0, 2\pi]$ on a :

$$R_x(\tau) = \frac{A^2}{2}cos(\omega\tau) \qquad R_x^{''}(\tau) = -\frac{A^2\omega^2}{2}cos(\omega\tau) \tag{2.17}$$

Le rapport signal sur bruit est alors obtenu à partir de la formule (2.16) :

$$SNR_j(dB) = -20\ log_{10}(\omega\sigma_j) \tag{2.18}$$

D'après la formule 2.18 et pour une incertitude σ_j donnée, le rapport signal sur bruit chute d'environ 6 dB pour une octave supplémentaire de la fréquence d'échantillonnage, soit une perte de résolution de 1 bit. La figure 2.5 montre l'évolution de ce rapport en fonction de la fréquence du signal pour deux valeurs de l'incertitude d'échantillonnage : $\sigma = 1\ ps$ et $\sigma = 10\ ps$. Le SINAD de quelques convertisseurs CMOS récents est également reporté sur le même graphe. L'incertitude d'échantillonnage constitue une limite importante pour ces convertisseurs rapides [130]. L'amélioration des performances étant largement déterminée par la possibilité de fournir un signal d'horloge avec une grande stabilité [27].

2.3.3 Métastabilité

Une autre limite importante à l'utilisation d'une fréquence d'échantillonnage élevée est liée au dispositif de comparaison qui doit effectué une décision logique en un temps limité. Pour des valeurs du signal d'entrée proche de son seuil de décision, un comparateur peut ne pas atteindre un niveau de sortie suffisant pour être décodé comme un état logique valide. L'état du comparateur est dit métastable. On peut modéliser ce fait par une plage d'incertitude δ_{me} de l'entrée pour laquelle l'état de sortie est une variable aléatoire prenant l'état logique "0" ou "1" avec une égale probabilité. La

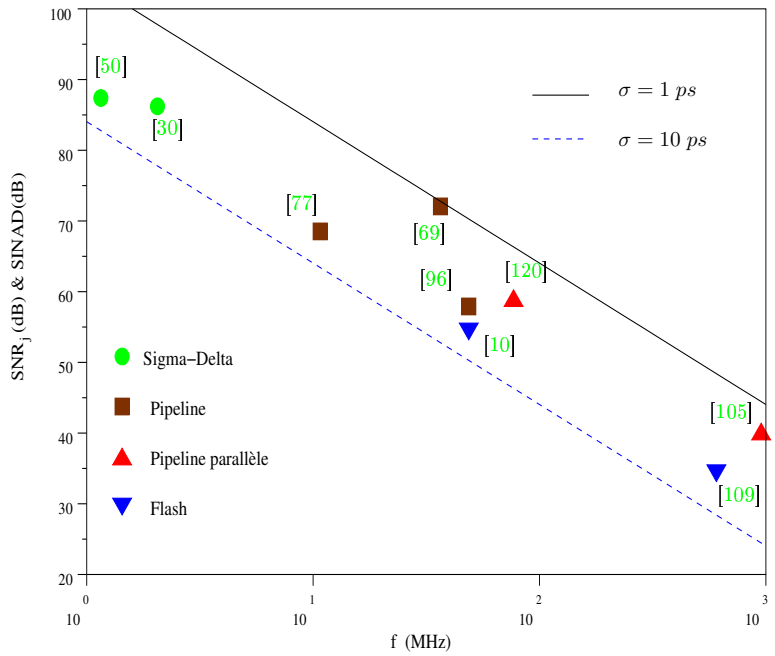


FIG. 2.5 – Rapport signal sur bruit dû à l'incertitude d'échantillonnage.

métastabilité peut être considérée comme une source de bruit supplémentaire qui va dégrader le rapport signal sur bruit global. La valeur de l'erreur de sortie résultant de la prise de décision des comparateurs est fortement dépendante de la logique de codage du convertisseur. Ce phénomène affectant essentiellement les convertisseurs très rapides, nous considérons uniquement le cas des convertisseurs de type Flash qui seront étudiés en détail au chapitre 4. Pour ceux-ci, et afin de limiter l'effet de la métastabilité, on utilise généralement un code intermédiaire de Gray. Dans ce code, deux mots consécutifs ne diffèrent que par un seul bit. L'erreur résultante de la métastabilité d'un comparateur est alors d'un quantum et la probabilité associée à cette erreur est de $\frac{1}{2}$ (Le bit obtenu a une chance sur deux d'être faux). La variance de l'erreur liée à la métastabilité d'un comparateur est donc :

$$\sigma_{me}^2 = P_{me} \cdot \frac{\Delta^2}{2} \tag{2.19}$$

La probabilité d'avoir un état métastable P_{me} dépend de la statistique du signal. Pour une distribution uniforme de celui-ci, elle est égale à $P_{me} = \frac{\delta_{me}}{\Delta}$. On peut expliciter cette probabilité en fonction des paramètres du comparateur [103] :

$$P_{me} = \frac{2\ V_L}{G_c\ \Delta} e^{\frac{-t_r}{\tau}} = \frac{2^{(n+1)}\ V_L}{G_c \cdot FSR} e^{-\frac{t_r}{\tau}} \tag{2.20}$$

où V_L est l'excursion nécessaire en sortie pour définir un état logique valide, G_c est le gain du comparateur dans l'état transparent, t_r est le temps de comparaison et τ est la constante de temps de régénération du comparateur.

Pour une entrée sinusoïdale pleine échelle ($A = \frac{FSR}{2}$) le rapport signal sur bruit lié à la métastabilité est obtenu à l'aide des formules 2.19 et 2.20 :

$$SNR_{me} = \frac{\frac{A^2}{2}}{\sigma_{me}^2} = \frac{2^{2n}}{4\ P_{me}} = 2^n\ \frac{G_c \cdot FSR}{8\ V_L}\ e^{\frac{t_r}{\tau}} \tag{2.21}$$

Soit en dB :

$$\begin{aligned} SNR_{me}[dB] &\approx 6\ n + 10\ log_{10}\ \tfrac{G_c \cdot FSR}{8\ V_L} + 4,3\ \tfrac{t_r}{\tau} \\ &= SNR_q - 1,76 + 10\ log_{10}\ \tfrac{G_c \cdot FSR}{8\ V_L} + 4,34\ \tfrac{t_r}{\tau} \end{aligned} \tag{2.22}$$

où SNR_q est le bruit de quantification correspondant à n bit de résolution. Le troisième terme est difficile à évaluer avec une bonne précision en raison du paramètre V_L qui est très dépendant des dispersions technologiques [103]. En prenant pour simplifier $V_L = FSR$ et en supposant un gain $G_c = 2$ ce terme vaut environ -6dB. Le dernier terme que nous appellerons SNR_{reg} est particulièrement important pour l'impact de la métastabilité sur le rapport signal sur bruit. Il est directement lié à la fréquence d'échantillonnage par l'intermédiaire du temps t_r de régénération. Si celui-ci est fixé à la demi période de l'horloge (l'autre demi période étant réservée à l'initialisation du comparateur) on a :

$$SNR_{reg}[dB] = 4,34\ \frac{1}{2\ F_s\ \tau} = \frac{2,17}{F_s \cdot \tau} \tag{2.23}$$

La figure 2.6 donne l'évolution de ce terme en fonction de la fréquence d'échantillonnage pour différentes constantes de temps de régénération. Une constante de temps de 100

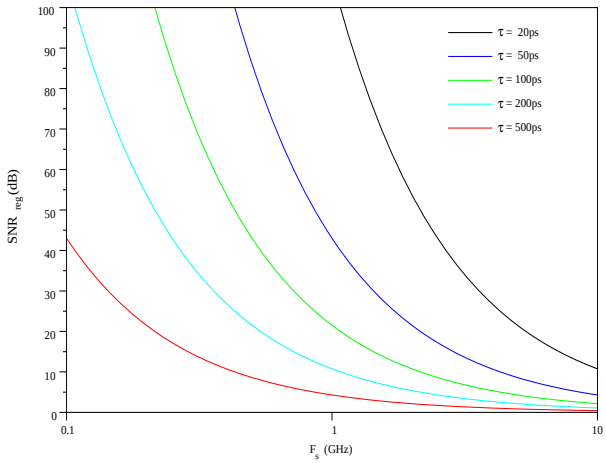


FIG. 2.6 – Rapport SNR_{reg} en fonction de la fréquence d'échantillonnage.

ps conduit à un gain de 21dB pour une fréquence d'échantillonnage de 1GHz. D'après la formule 2.22, cette marge est suffisante pour que le rapport signal sur bruit total soit dominé par le bruit de quantification. On voit qu'avec un codage efficace du signal après la comparaison la métastabilité ne constitue une limite que pour des valeurs très élevée ($>$ 1GHz) de la fréquence d'échantillonnage.

2.3.4 Bruit thermique

Pour déterminer l'influence du bruit thermique nous considérons le circuit très simple de la figure 2.7 qui est à la base de nombreux blocs fonctionnels échantillonnés.

Il existe bien d'autres sources de bruit dans ces blocs (bruit en 1/f des transistors MOS, bruit de grenaille dans les jonctions, bruit lié aux commutations logiques,...) mais celui de l'échantillonnage d'entrée fixe une borne inférieure au bruit total. Le calcul de la

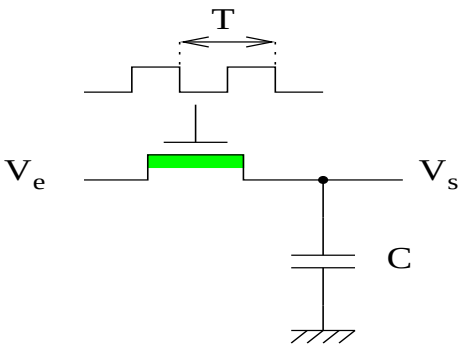


FIG. 2.7 – Echantillonneur MOS.

densité spectrale de bruit échantillonné peut être fait simplement en considérant que le transistor MOS est équivalent à une résistance de valeur R_{on}. Cette approximation est justifiée juste avant la coupure du canal car la tension aux bornes de l'interrupteur est nécessairement faible à cet instant pour un échantillonnage correct du signal. La résistance du canal est le siège d'une tension de bruit thermique dont la densité spectrale de puissance et la fonction d'auto-corrélation sont donnés par :

$$S_v(f) = N_o = 2k_B \cdot T_a \cdot R_{on} \qquad R_v(u) = N_o\delta(u) \tag{2.24}$$

où k_B $(1,38\ 10^{-23}\ J/K)$ est la constante de Boltzmann et T_a est la température absolue. Ce bruit subit alors un filtrage linéaire du premier ordre de constante de temps $\tau = R_{on}C$ avant l'échantillonnage idéal du signal avec la période T_s comme indiqué à la figure 2.8.

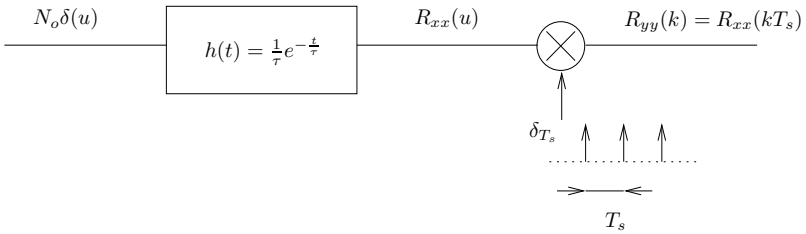


FIG. 2.8 – Bruit thermique échantillonné.

On déduit de ce modèle la fonction d'auto-corrélation en sortie de l'échantillonneur :

$$\begin{aligned} R_{yy}(k) = R_{xx}(kT_s) &= N_o\ \delta(u) \star h(u) \star h(-u)|_{u=kT_s} \\ R_{yy}(k) = \tfrac{N_o}{2\tau}e^{-|k|\frac{T_s}{\tau}} &= \tfrac{k_B \cdot T_a}{C}e^{-|k|\frac{T_s}{\tau}} \end{aligned} \tag{2.25}$$

La densité spectrale de puissance s'obtient par transformée de Fourier de $R_{yy}(k)$:

$$S_y(f) = \frac{1}{F_s} \sum_{k=-\infty}^{\infty} R_{yy}(k) e^{-j\,2\pi\,k\,f\,T_s}$$

$$S_y(f) = \frac{k_B \cdot T_a}{C \cdot F_s} \sum_{k=-\infty}^{\infty} e^{-|k|\frac{T_s}{\tau}} e^{-j\,2\pi\,k\,f\,T_s} = \frac{k_B \cdot T_a}{C \cdot F_s} \frac{sinh(\frac{T_s}{\tau})}{cosh(\frac{T_s}{\tau}) - cos(2\pi\,f\,T_s)} \tag{2.26}$$

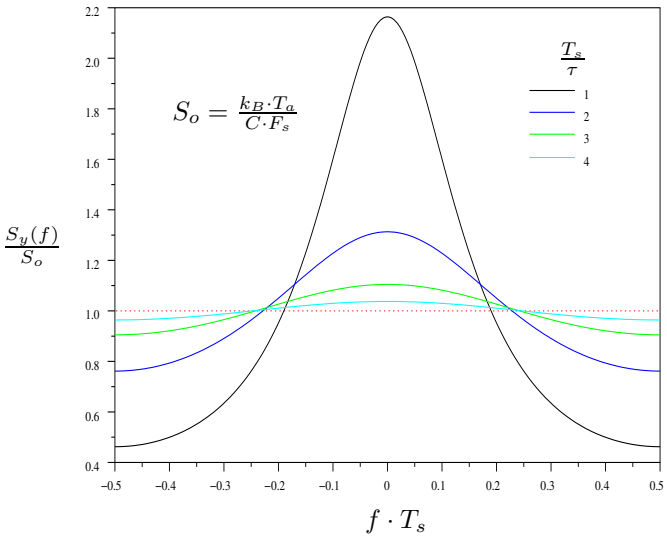


FIG. 2.9 – Densité spectrale du bruit thermique.

Cette densité est représentée sur la figure 2.9 en fonction du paramètre $\frac{T_s}{\tau}$. Celle-ci est sensiblement constante dès que ce paramètre est supérieur à 4, ce qui correspond à une précision de l'ordre de 2% sur l'établissement en fin d'échantillonnage. On peut donc considérer que le bruit échantillonné a une puissance totale :

$$P_{th} = \frac{k_B \cdot T_a}{C} \tag{2.27}$$

uniformément répartie sur la bande $[-\frac{F_s}{2}, \frac{F_s}{2}]$.

Dans le cas d'un suréchantillonnage du signal avec un facteur R, la puissance résultante après filtrage numérique est, comme pour le bruit de quantification, réduite de ce même facteur.

Si on limite la puissance du bruit thermique à un quart de celle du bruit de quantification (ce qui correspond à une perte d'approximativement 1 dB en rapport signal sur bruit) on en déduit une valeur minimale de la capacité d'échantillonnage C :

$$P_{th} = \frac{k_B \cdot T_a}{C} \leq P_q = \frac{1}{4}\frac{\Delta^2}{12} = \frac{FSR^2}{2^{2n} \cdot 48}$$

$$C \geq \frac{48 \cdot k_B \cdot T_a \cdot 2^{2n}}{FSR^2} \tag{2.28}$$

On obtient par exemple, pour n=10 bit et FSR=1 V, une capacité minimum de 0,2 pF. On peut donc toujours choisir une capacité d'échantillonnage qui rende le bruit thermique suffisamment faible vis à vis du bruit de quantification. Une valeur importante de la capacité se traduira cependant par un accroissement de la consommation pour une fréquence donnée du signal. Pour un signal sinusoïdal, le courant maximum à fournir par la source de signal est en effet directement proportionnel à la capacité d'échantillonnage. Il en va de même du courant d'alimentation via le rendement de cette source.

Chapitre 3

Test dynamique des CAN

3.1 Introduction

Nous avons examiné dans le chapitre précédent les paramètres essentiels pour la description du comportement d'un CAN. Ce chapitre présente quelques méthodes utilisées pour obtenir ces paramètres. Une description complète de ces méthodes est fournie par le standard IEEE-1241 [56] dédié au test de ces composants. Notre propos n'est pas de faire une étude exhaustive de ces méthodes mais uniquement de présenter les tests dynamiques les plus usuels et en particulier ceux qui seront utilisés pour la simulation comportementale des CAN. Ces tests utilisent une source variable dans le temps à l'entrée du convertisseur. Un certain nombre de codes sont ensuite stockés en mémoire puis analysés.

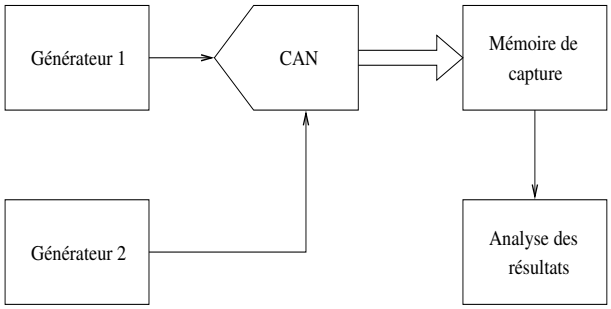


FIG. 3.1 – Environnement simplifié de test.

Un environnement de test simplifié d'un CAN est donné à la figure 3.1. Il fait apparaître le fait que les mesures sont conditionnées par le signal d'entrée (générateur 1) et le signal d'horloge (générateur 2). On exige de ce dernier une grande stabilité et une gigue de phase compatible avec la précision du CAN (équation 2.18).

Le test étant grandement dépendant du choix du signal d'entrée nous examinons dans un premier temps les contraintes liées à ce signal.

3.2 Le choix du signal de test

Le signal le plus simple à générer avec une grande pureté spectrale est le signal sinusoïdal. Il présente également l'avantage de pouvoir facilement être extrait du bruit et des non linéarités introduites par le composant. Un tel signal est caractérisé par son amplitude A, sa fréquence F et sa phase à l'origine ϕ :

$$V = A \sin(2\,\pi\,F\,t + \phi) \tag{3.1}$$

La fréquence de ce signal peut être synchrone avec celle de l'horloge ou indépendante. Nous considérons séparément ces deux cas qui donne lieu à une modélisation différente des propriétés du signal.

3.2.1 Echantillonnage synchrone

Dans le cas de l'échantillonnage synchrone, la suite des échantillons s'écrit :

$$V(n) = A \sin(2\,\pi\,n\,F\,T_s + \phi) = A \sin(2\,\pi\,n\frac{L}{N} + \phi) \tag{3.2}$$

où $T_s = \frac{1}{F_s}$ est la période d'échantillonnage et L est le nombre de périodes du signal d'entrée dans l'enregistrement. Ces différents paramètres sont liés par la relation :

$$N \cdot F = L \cdot F_s \tag{3.3}$$

Il est important que les nombre L et N soit premiers entre eux pour garantir des échantillons différents sur toute la longueur de l'enregistrement. Ceci est illustré sur la figure 3.2 qui représente un cas particulier avec 512 échantillons. Dans le cas (a),

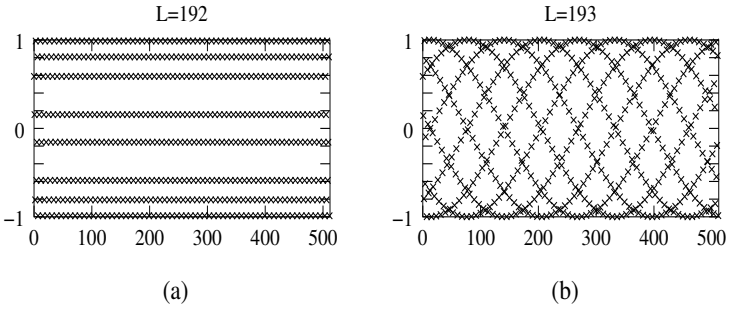


FIG. 3.2 – Cas particulier d'échantillonnage synchrone ($\phi = \frac{\pi}{5}$).

seulement 8 valeurs différentes ($\frac{192}{512} = \frac{3}{8}$) de l'amplitude d'entrée sont utilisées alors que dans le cas (b) on exploite les 512 valeurs permises par la longueur de l'enregistrement. La longueur de l'enregistrement étant généralement une puissance de deux pour faciliter l'analyse, il suffit de choisir L impaire pour garantir des échantillons différents. Cette condition fixe en même temps le nombre minimal d'échantillons pour le test. En effet, si l'on veut couvrir tous les codes d'un convertisseur de n bit, cela impose $N > 2^n$. Pour certains types de test, ce nombre peut s'avérer très insuffisant. C'est, par exemple, le cas du test statistique qui sera étudié au 3.4.

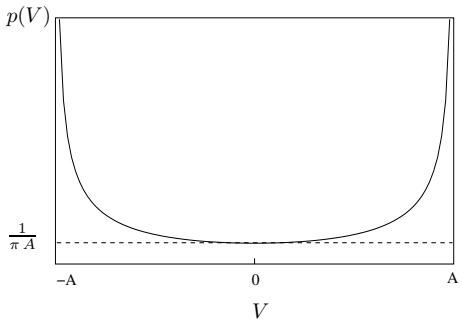


FIG. 3.3 – Densité de probabilité du signal d'entrée.

3.2.2 Echantillonnage aléatoire

Lorsque le signal source et le signal d'horloge sont indépendants, la phase ϕ est considérée comme une variable aléatoire uniforme sur $[0, 2\,\pi]$. La densité de probabilité associée au signal 3.1 est dans ce cas (figure 3.3) :

$$p(V) = \frac{1}{\pi\,\sqrt{A^2 - V^2}} \tag{3.4}$$

et la probabilité que celui-ci soit compris dans l'intervalle $[V_a, V_b]$ est obtenue par intégration :

$$P(V_a, V_b) = \int_{V_a}^{V_b} p(V)dV = \frac{1}{\pi}\left(\arcsin\frac{V_b}{A} - \arcsin\frac{V_a}{A}\right) \tag{3.5}$$

3.3 Analyse spectrale

Les paramètres dynamiques tels que le SINAD ou le THD, peuvent être déterminés efficacement par une analyse spectrale des codes de sorties du convertisseur. Cette méthode consiste à effectuer une transformée de Fourier discrète (TFD) sur une séquence y de N codes pondérée par un fenêtre w :

$$Y(k) = \sum_{n=0}^{N-1} y(n) \cdot w(n)\; e^{-j\,2\pi\,k\frac{n}{N}}$$

Théoriquement, le fenêtrage n'est pas nécessaire si l'échantillonnage est synchrone. En pratique, cette condition est difficile à vérifier parfaitement et il est préférable d'appliquer une fenêtre qui évite les problèmes de discontinuités liés à la limitation de la durée d'observation. L'effet de la troncature de la séquence est illustré sur la figure 3.4 qui correspond à l'analyse d'un signal sinusoïdal avec A=1 et N=32. Dans ce cas, la fenêtre est rectangulaire avec $w(n) = 1$ pour tous les échantillons. Le spectre du signal est convolué avec la TFD de cette fenêtre (le résultat de cette convolution est représenté en pointillé sur la figure 3.4). Dans le cas (a), la fréquence du signal coïncide avec un point de calcul de la TFD. Seule la troisième raie est présente avec la bonne amplitude (au facteur 1/N près lié à la définition de la TFD). Dans le cas (b), la fréquence du signal est entre deux points de calcul de la TFD et l'opération de convolution précédente se traduit par un étalement de l'amplitude sur tous les points de la TFD. En dehors de la résolution fréquentielle limitée ($\frac{F_e}{N}$), on introduit

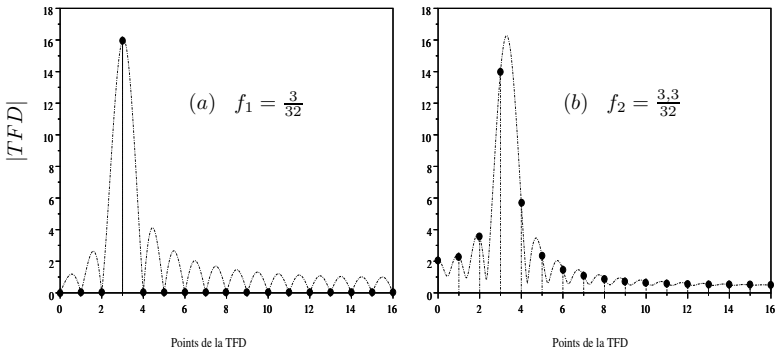


FIG. 3.4 – Effet de la troncature de la séquence d'observation (N=32).

une incertitude importante sur la mesure de l'amplitude du signal. Pour remédier à ce problème on doit utiliser une autre fenêtre temporelle. Il en existe de nombreuses qui sont dédiées à l'analyse spectrale. Parmi les propriétés importantes d'une fenêtre on peut citer l'atténuation des lobes secondaires de sa TFD par rapport au lobe central et la bande équivalente de bruit. L'atténuation des lobes secondaires doit être suffisante pour distinguer les composantes spectrales liées à la distorsion introduite par la non linéarité du CAN, des raies qui sont le produit de l'étalement associé à la fenêtre d'observation. Afin d'examiner l'effet du fenêtrage sur le bruit, considérons le cas où y est un bruit blanc de variance σ^2. La puissance de bruit associée à un point de la TFD est alors :

$$P_n = \sigma^2 \sum_{n=0}^{N-1} w(n)^2$$

En présence d'une séquence $y(n) = A \exp(j2\pi\, n\frac{k}{N})$, la puissance associée au point k de la TFD est :

$$P_s = A^2 \left[\sum_{n=0}^{N-1} w(n) \right]^2$$

Le rapport signal sur bruit correspondant à ces deux signaux est donné par :

$$SNR = \frac{A^2}{\sigma^2} \frac{N}{ENBW} \tag{3.6}$$

où ENBW est la bande équivalente de bruit (*Equivalent Noise BandWidth*) définie par :

$$ENBW = \frac{N \sum_{n=0}^{N-1} w(n)^2}{\left[\sum_{n=0}^{N-1} w(n) \right]^2} \tag{3.7}$$

Celle-ci mesure la dégradation du rapport signal sur bruit lié à la fenêtre d'analyse. Le tableau 3.1 donne quelques exemples de fenêtre d'analyse [115]. Une atténuation importante des lobes secondaires s'accompagne d'un accroissement de la bande de bruit. Cette atténuation est cependant indispensable pour mesurer de faibles taux de distorsion. La perte en rapport signal sur bruit peut par ailleurs être compensée par un accroissement du nombre de points de la TFD (formule 3.6).

Fenêtre	Atténuation des lobes secondaires	ENBW
Rectangulaire	13,28 dB	1
Hanning	31,47 dB	1,5
Blackman-Harris(4 termes)	92,01 dB	2,004
Blackman-Harris(7 termes)	191,45 dB	2,632

TAB. 3.1 – Exemples de fenêtres d'analyse

Lorsque la fréquence du signal sinusoïdal ne coïncide pas avec un point de la TFD, son amplitude doit être estimée à partir des points voisins centrés sur le lobe principal de la fenêtre. Lorsque celle-ci est décrite par une expression du type :

$$w(n) = \sum_{k=0}^{K} (-1)^k \, a_k \, \cos(2\pi \cdot k \cdot \frac{n}{N}) \tag{3.8}$$

on peut estimer l'amplitude A de la sinusoïde par [115] :

$$A = \sqrt{\frac{2}{N \sum_{n=0}^{N-1} w(n)^2} \sum_{k \in B, B^*} |TFD(k)|^2} \tag{3.9}$$

où B et B^* sont des ensembles de points centrés sur la fréquence de la sinusoïde et sur son image. Ceux-ci sont constitués de $2\,K + 1$ points pour une fenêtre du type 3.8 comportant K coefficients.

A partir de cette expression, il est aisé de déterminer les composantes du signal et des harmoniques sur les ensembles B_i et B_i^* qui leurs sont associés. Le bruit est évalué de la même manière sur tous les points de la TFD complémentaires à ces ensembles. Les calculs du SINAD ou du THD se déduisent alors directement de ces mesures. On notera que le bruit est sous évalué par cette technique puisqu'une partie de celui-ci est également présente sur les ensembles B_i, B_i^* et n'est pas comptabilisée. Cette erreur peut cependant être rendue négligeable en choisissant un nombre de points de la TFD suffisamment grand.

3.4 Test statistique

Dans ce test, nous considérons les N codes issus du convertisseur comme une suite de variables aléatoires discrètes et indépendantes. La probabilité d'occurrence de ces codes est ensuite utilisée pour estimer les seuils de transition du convertisseur. A partir de ces seuils, il est alors possible de calculer des caractéristiques telles que l'INL et la DNL à l'aide des formules 2.3 et 2.4. Le signal de test est a priori quelconque dès l'instant que sa densité de probabilité est parfaitement connue. Pour les raisons évoquées au 3.2, le choix d'un signal sinusoïdal est le plus fréquent. La densité de probabilité de ce signal est donnée par 3.4 lorsque sa phase est une variable aléatoire uniforme. Cette situation est en pratique approchée par un signal synchrone de l'horloge et respectant la condition 3.3 avec L et N premiers entre eux. L'amplitude crête-crête du signal est égale à la pleine échelle (FSR) du convertisseur. Les caractéristiques d'INL et de DNL étant définies de manières statiques, la source est a priori lentement variable. Lorsque cette condition est réalisée, ces mesures coïncident avec les valeurs statiques mais il est cependant possible d'utiliser une fréquence élevée pour mettre en évidence des effets dynamiques. En présence de ces effets, le seuil de transition entre deux codes n'est pas

parfaitement défini du fait, par exemple, de la métastabilité des comparateurs. Dans le calcul qui suit nous supposerons que ces phénomènes sont absents et que la définition d'un seuil de transition est conforme à celle du chapitre précédent.

3.4.1 Calcul des seuils de transitions

Nous notons $CH(k)$ l'ensemble de tous les codes inférieurs ou égaux au code k pour les N échantillons. Si l'on considère un code c quelconque, la probabilité que celui-ci soit inférieur ou supérieur au code k est telle que :

$$P(c \leq k) = P(V < V_k) = p \quad \text{et} \quad P(c > k) = P(V \geq V_k) = 1 - p$$

où V_k est la transition supérieure du code k. La probabilité de réalisation de l'événement $CH(k) = M$ suit une loi de Bernouilli :

$$P(CH(k) = M) = C_N^M p^M (1-p)^{N-M}$$

dont la moyenne et la variance sont donnés par :

$$E\{CH(k)\} = N\,p \qquad \sigma^2_{CH(k)} = N\,p\,(1-p)$$

L'histogramme cumulé $ch(k) = \frac{CH(k)}{N}$ est caractérisé par :

$$E\{ch(k)\} = p = P(V < V_k) \qquad \sigma^2_{ch(k)} = \frac{p\,(1-p)}{N} \tag{3.10}$$

A partir de 3.10 et de 3.5 on peut estimer les seuils de transitions :

$$V_k = -A\,cos(\pi\,ch(k))$$

et la variance sur ces transitions :

$$\sigma^2_{V_k} = \sigma^2_{ch(k)}\,A^2\,\pi^2\,[sin(\pi\,ch(k))]^2$$

Le maximum de l'erreur est obtenu pour $ch(k) = 0,5$ soit :

$$\max\{\sigma_{V_k}\} = \frac{A\,\pi}{2\,\sqrt{N}} \tag{3.11}$$

Prenons par exemple un convertisseur de 6 bit. Pour que l'écart-type sur les seuils soit inférieur à $\frac{1}{10}$ du quantum, on doit avoir :

$$\frac{A\,\pi}{2\,\sqrt{N}} < \frac{\Delta}{10} = \frac{2\,A}{10\,(2^6-1)}$$

c'est à dire un nombre N d'échantillons de l'ordre de 250000. Ce nombre est très supérieur au nombre de codes possibles générés par le CAN et croît avec le carré de celui-ci. Le coût de ce test augmente ainsi très rapidement avec la résolution. La figure 3.5 montre un exemple de test d'histogramme d'un convertisseur 6 bit et les caractéristiques d'INL et DNL associées.

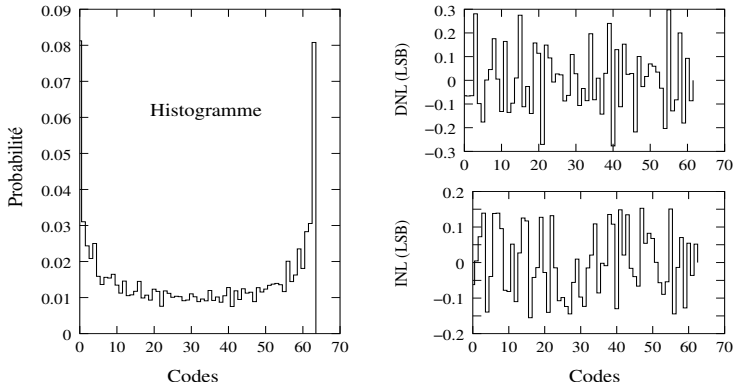


FIG. 3.5 – Test d'histogramme d'un CAN 6 bit.

3.4.2 Limitations

Le test d'histogramme suppose la monotonicité de la caractéristique de transfert. Certaines erreurs qui ne respectent pas cette condition peuvent ne pas être détectées. En présence du bruit interne au composant la probabilité d'occurrence des codes est modifiée. Il est possible de limiter cet effet en choisissant l'amplitude du signal légèrement supérieure à la pleine échelle du convertisseur [56]. Cependant, en présence d'un bruit important, le résultat du test peut être très optimiste [43] par rapport aux performances réelles du composant.

3.5 Conclusions

Nous avons examiné deux types de test dynamiques des convertisseurs. Le premier, l'analyse spectrale, est un test très efficace pour fournir des informations globales telles que le SINAD, le THD ou le SFDR. Le temps de calcul est en effet réduit lorsque la TFD est réalisée par l'algorithme de transformée de Fourier rapide (FFT). L'utilisation de plusieurs sources sinusoïdales peut également permettre la mesure des produits d'intermodulation (IMD). Le second test, l'analyse statistique, permet d'obtenir des informations complémentaires sur la caractéristique de transfert. Celui-ci doit cependant être appliqué avec prudence en raison des limites évoquées en 3.4.2. Pour une résolution élevée du convertisseur, le nombre d'échantillons devient vite très important (équation 3.11) et pose un problème de mémorisation des données et de durée du test.

L'analyse spectrale s'avère très intéressante pour les besoins d'évaluation rapide des CAN. Lorsque la bande utile du signal est réduite par rapport à la fréquence d'échantillonnage, le coût de calcul peut cependant être important. En effet, l'analyse (FFT) doit alors être faite sur les N points alors qu'une fraction seulement est utilisée pour mesurer les caractéristiques du CAN. C'est le cas du convertisseur $\Sigma\Delta$ étudié au chapitre 6. Nous verrons au chapitre 9 une méthode d'analyse spécifique pour ce convertisseur qui permet d'évaluer rapidement certaines performances à partir d'une simulation d'un modèle comportemental.

Deuxième partie

CAN CMOS rapides

Chapitre 4

Convertisseur flash

4.1 Introduction

Le convertisseur flash exploite la relation directe de définition d'un quantificateur sans mémoire :

$$\hat{x} = \sum_i y_i 1_{S_i} \qquad i \in C \tag{4.1}$$

où 1_A est la fonction indicatrice de l'ensemble A. S_i représente une partition de l'espace des valeurs de x en intervalles disjoints dont l'union représente le domaine de x. Dans le cas d'un quantificateur linéaire, ces intervalles sont tous de longueur égale et cette longueur est le quantum Δ du convertisseur. L'indice i représente le code associé à l'intervalle et y_i est un représentant de cet intervalle. Le choix de ce représentant à partir du code constitue le cadre de la conversion numérique-analogique alors que la conversion analogique-numérique consiste à déterminer l'intervalle de valeurs S_i auquel x appartient et à restituer le code i. Ce problème est résolu lorsque l'on sait générer la fonction indicatrice associée à un intervalle donné. Si $A = [a, b)$ est cet intervalle et si $Q(x) = 1_{x \geq 0}$ alors la fonction indicatrice de A est donnée par :

$$1_A = Q(x - a)\overline{Q(x - b)} \tag{4.2}$$

où $\overline{1_B}$ est la fonction indicatrice complémentaire de B. Cette relation se traduit par le schéma électronique de la figure 4.1. Le convertisseur flash exploite directement

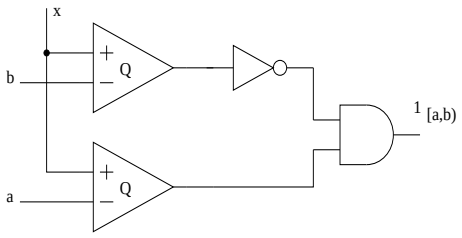


FIG. 4.1 – Réalisation électronique de la fonction indicatrice $1_{[a,b)}$.

cette correspondance avec la mise en commun d'un comparateur Q pour deux intervalles contigus. Les tensions de références peuvent être générées à partir d'un diviseur

potentiométrique. On arrive ainsi au schéma de la figure 4.2 pour un convertisseur CMOS de 3 bits de résolution (2^3 intervalles). Le code fourni par les sorties de comparateurs est idéalement constitué d'une suite de "1" pour tous les intervalles inférieurs ou appartenant à la tension d'entrée puis d'une suite de "0" pour les intervalles supérieurs. Du fait de cette configuration ce code est généralement dénommé code thermométrique par analogie entre la tension d'entrée et la température d'un thermomètre à réservoir. Le décodeur fournit idéalement une seule sortie à "1" à partir du code thermométrique. Cette sortie définit le code associé à l'intervalle d'appartenance de l'entrée. Dans le schéma de la figure 4.2 chaque sortie adresse une mémoire qui donne une représentation binaire équivalente du code. Ce codage est en général utilisé pour sa représentation économique qui est bien adaptée au traitement numérique réalisé sur les échantillons du signal. La résolution de ce type de convertisseur est limitée par la précision avec

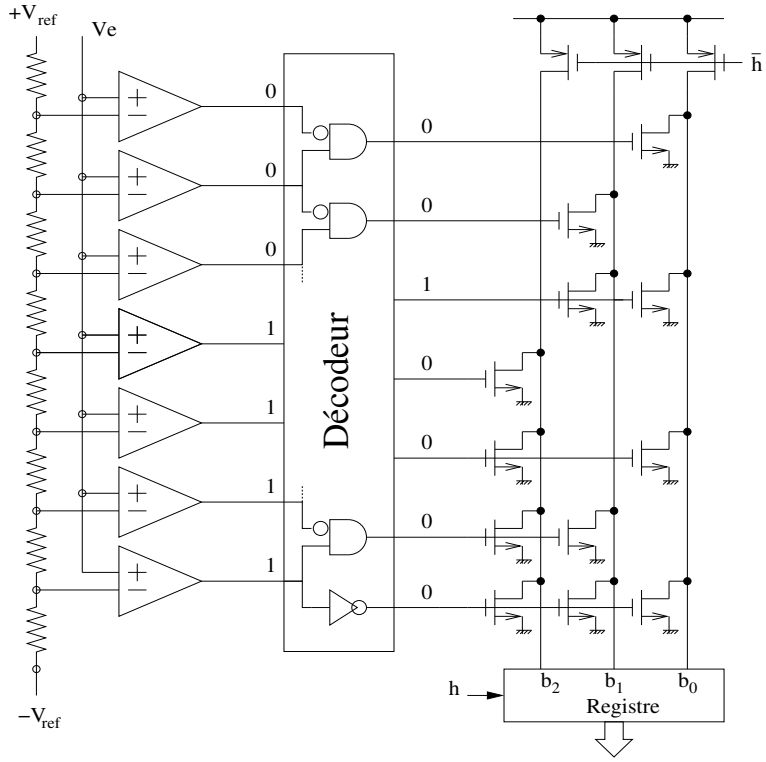


FIG. 4.2 – Flash.

laquelle sont définies les tensions de référence et par les erreurs statiques introduites dans le circuit de comparaison.

4.2 Limitations dues à la précision sur les résistances

Dans l'exemple de la figure 4.2 la précision avec laquelle sont obtenues les tensions de référence dépend de l'erreur sur les valeurs des résistances qui constituent le diviseur potentiométrique. Cette erreur peut être de nature déterministe ou de nature aléatoire. Dans le premier cas, des dispositions particulières dans le dessin des masques peuvent être efficaces pour en réduire l'effet [113]. En ce qui concerne les sources aléatoires

d'erreur, la dispersion est en général réduite par un accroissement de la surface des composants. Dans le cas du diviseur potentiométrique, si chaque résistance est décrite par une loi normale de moyenne R et d'écart type σ_R les potentiels de références suivent une loi qui est proche d'une loi normale dont la moyenne et l'écart type sont donnés par [71] :

$$V_j = \frac{j}{N}V_R \qquad \sigma_{V_j} = \sqrt{\frac{(j/N)(1-j/N)}{N}}(\frac{\sigma_R}{R})V_R \tag{4.3}$$

où V_R est la différence de potentiel aux bornes de l'échelle et $j = 1, 2, ..., N-1$ correspond aux différents points intermédiaires entre les N résistances. Cette erreur est maximale pour le potentiel central ($j = \frac{N}{2} - 1$) avec un écart type :

$$\sigma^V_{max} = \frac{1}{2\sqrt{N}}(\frac{\sigma_R}{R})V_R \tag{4.4}$$

Pour un convertisseur de n bit et une non-linéarité intégrale inférieure à $\frac{\Delta}{2}$, l'erreur sur la résistance doit être telle que $\frac{\sigma_R}{R} < \frac{1}{\sqrt{2^n}}$.

4.3 Limitations dues aux tensions de décalages

Les tensions de décalage dans le circuit de comparaison sont une limitation importante dans le convertisseur flash. Celles-ci affectent directement les seuils de transition et peuvent à la limite entraîner la non-monotonicité de sa caractéristique statique. Cette limite est atteinte lorsque la non-linéarité différentielle définie à partir des seuils de transition V_j par $DNL = V_{j+1} - V_j - \Delta$ est telle que $DNL < -\Delta$. Si la dispersion sur les seuils de transition est caractérisée par une loi normale d'écart type σ, la probabilité p de non-monotonicité locale est donnée par :

$$p = \frac{1}{\sqrt{2\pi}}\int_z^\infty e^{\frac{-u^2}{2}}du \quad \text{avec} \quad z = \frac{\Delta}{\sigma\sqrt{2}} \tag{4.5}$$

Le rendement Y [98], c'est à dire la probabilité de monotonicité globale du convertisseur est tel que le défaut de monotonicité soit absent sur les $2^n - 2$ couples de seuils adjacents soit $Y = (1-p)^{2^n-2}$. La figure 4.3 représente ce rendement pour différentes résolutions du convertisseur. On a, par exemple, pour une résolution de 6 bit et $\sigma = \frac{\Delta}{4}$, un rendement de 86,5%. Avec le même écart type, le rendement n'est plus que de 55,1% pour une résolution de 8 bit et 9,1% pour 10 bit.

Dans le schéma de la figure 4.2, chaque comparateur peut être réalisé de manière plus ou moins complexe selon la résolution désirée. Pour limiter l'effet de sa tension de décalage, celui-ci est en général précédé d'un pré-amplificateur. Un exemple de configuration est donné à la figure 4.4. La tension de décalage effective à l'entrée de l'amplificateur est dans ce cas égale à $e_{da} + \frac{e_{dc}}{G_a}$ où G_a est le gain de l'amplificateur, e_{da} est la tension de décalage de l'amplificateur et e_{dc} celle du comparateur. Supposons un écart type des tensions de décalage de $\sigma(e_{dc}) = 50\ mV$ pour le comparateur et de $\sigma(e_{da}) = 10\ mV$ pour l'amplificateur. Un gain $G_a = 5$ conduit alors à un écart type global de la tension de décalage de $\sigma(ed) \simeq 14\ mV$. Cette amélioration serait bien entendu obtenue au prix d'un accroissement important de la surface du convertisseur sans l'utilisation des techniques de repliement et d'interpolation qui seront étudiées plus loin et qui sont rendues possibles par cette pré-amplification du signal. En considérant une excursion de signal de 2 V ($-1\ V < V_e < 1\ V$) on voit d'après la figure 4.3

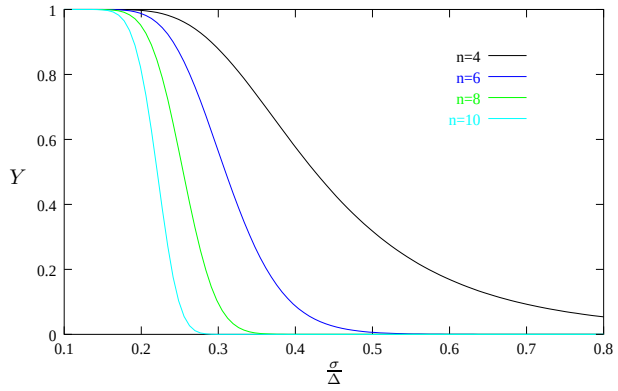


FIG. 4.3 – Rendement en fonction de la dispersion sur les seuils.

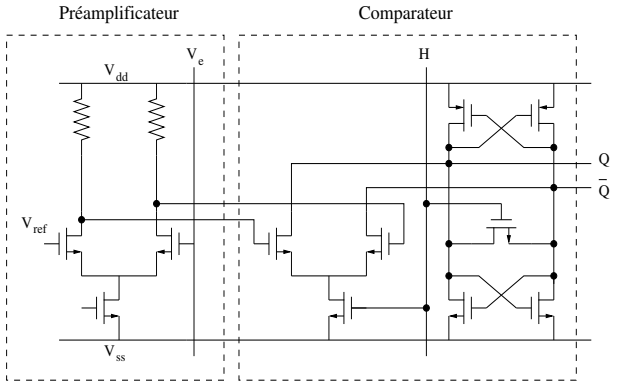


FIG. 4.4 – Exemple de circuit de comparaison.

qu'un convertisseur de 10 bit nécessiterait un écart type sur la tension de décalage inférieur à 0,5 mV. Afin de pouvoir atteindre cette résolution, des techniques particulières doivent être utilisées pour limiter l'effet de cette tension de décalage. Nous examinerons dans la suite le filtrage spatial qui consiste à pondérer les sorties d'amplificateurs qui est aujourd'hui très répandu [23] [38] [109].

Nous avons précédemment uniquement considéré les limitations statiques du convertisseur flash. Celui-ci est cependant dédié à des vitesses de conversion élevée et les limitations dynamiques sont tout aussi importantes. Elles sont essentiellement dues à la bande passante finie des amplificateurs et à une caractéristique propre du comparateur : la métastabilité. Le temps nécessaire pour atteindre un niveau logique bien défini en sortie du comparateur est fortement dépendant de sa tension d'entrée et peut être supérieur à la période de l'horloge. Ceci affecte localement le code thermométrique et peut conduire à des erreurs grossières en sortie du convertisseur en fonction de la méthode de codage utilisée.

4.4 Codage et métastabilité

Le passage du code thermométrique au code binaire peut être effectué par l'intermédiaire d'une mémoire ROM (figure 4.2) ou par une utilisation de portes logiques (figure 4.5 (b)). La première méthode présente un avantage de régularité dans le dessin des masques mais nécessite un temps de précharge qui peut limiter la vitesse de fonctionnement du convertisseur. Quelque soit la méthode utilisée, le codage doit prendre en compte le phénomène de métastabilité des comparateurs. La figure 4.5 (a) montre un exemple de configuration d'erreur. Un état indéfini (X) à la sortie du comparateur peut se propager sur deux lignes de la mémoire occasionnant une erreur qui dans ce cas est égale à la moitié de la pleine échelle du convertisseur. Comme on l'a vu au

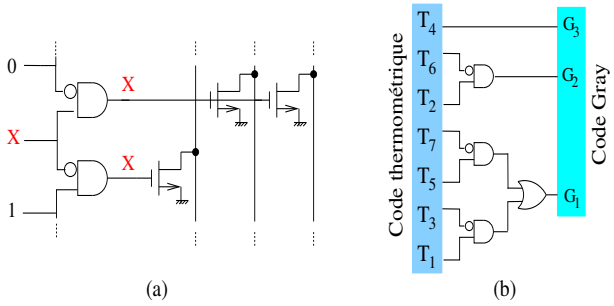


FIG. 4.5 – Codage basé sur une ROM (a) et codage par portes logiques (b).

chapitre 2, la réduction du phénomène de métastabilité du comparateur est contradictoire avec une vitesse de fonctionnement élevée. En effet celui-ci dépend du temps τ de régénération du comparateur qui est limité par la technologie. Les conséquences d'une erreur de métastabilité sur le code de sortie peuvent cependant être fortement réduites par un codage approprié. Ceci peut être réalisé par un code intermédiaire de Gray (figure 4.5 (b)), chaque sortie de comparateur n'affectant qu'un seul bit de ce code [109]. Dans ce dernier, deux mots consécutifs ne diffèrent que d'un seul bit et une erreur simple de métastabilité se traduit par une erreur d'un seul quantum du convertisseur.

4.5 Filtrage spatial

Le principe de ce filtrage est d'effectuer une somme pondérée des sorties d'amplificateurs voisins [68]. On peut faire une analogie complète entre la répartition spatiale des sources constituées par les amplificateurs et les échantillons temporels à l'entrée d'un filtre dont la réponse impulsionnelle est donnée par les coefficients de pondération associés [65] (à ceci près que le filtre n'est pas assujetti à la contrainte de causalité). Avec cette analogie, la dispersion des tensions de décalage peut être considérée comme un bruit blanc spatial sur lequel le filtrage va être appliqué. La moyenne effectuée sur la répartition spatiale du signal et du bruit peut alors être considérée comme un problème de filtrage optimal visant à améliorer le rapport signal sur bruit. La figure 4.6 donne une configuration du réseau de résistances qui peut être utilisé à cette fin. Les résistances R_0 sont les résistances de sortie des pré-amplificateurs. Chaque pré-amplificateur fournit un courant $I(n)$ qui est idéalement proportionnel à sa tension d'entrée. Du fait des tensions de décalage, ce courant est également le siège d'une composante aléatoire. Les

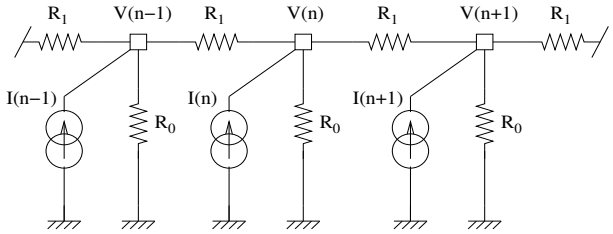


FIG. 4.6 – Circuit de pondération des sources.

résistances R_1 sont introduites entre chaque étage pour effectuer une moyenne sur les sorties de manière à atténuer l'effet de cette composante aléatoire. La tension en un noeud particulier peut s'exprimer en fonction de ces voisins :

$$V(n) = \alpha\left[V(n-1) + V(n+1)\right] + (1-2\alpha)R_0 I(n) \tag{4.6}$$

avec

$$\alpha = \frac{R_0}{R_1 + 2R_0} \qquad \alpha \in [0; 0,5)$$

Cette relation caractérise un filtrage spatial à une dimension de fonction de transfert :

$$T(f) = \frac{V_n}{R_0 I_n}(f) = \frac{1-2\alpha}{1-2\alpha cos(2\pi f)} \tag{4.7}$$

La réponse spatiale de ce filtre est la solution symétrique de l'équation 4.6 [65] :

$$h(n) = \frac{V(n)}{V(0)} = r^{|n|} \quad \text{avec} \quad r = \frac{2\alpha}{1+\sqrt{1-4\alpha^2}} \tag{4.8}$$

Elle est représentée à la figure 4.7 pour différentes valeurs du paramètre α.

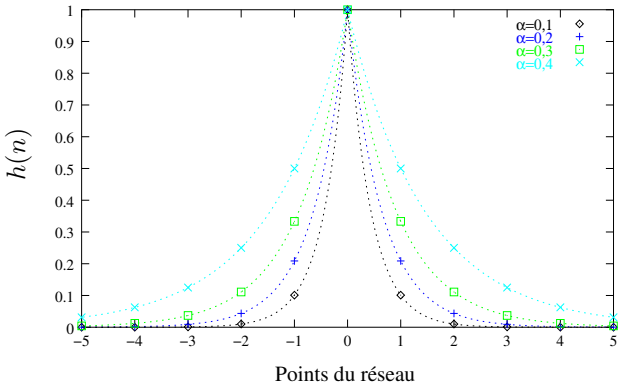


FIG. 4.7 – Filtre spatial : coefficients de pondération.

4.5.1 Facteur de mérite

On peut définir un facteur de mérite lié au circuit de pondération. Celui-ci est donné par l'amélioration du rapport signal sur bruit apporté par le filtrage spatial. L'impact

sur le signal étant peu important, ce facteur détermine pratiquement la réduction de l'offset :

$$ECF = \frac{\sum_{-W_s}^{W_s} h(k)}{\sqrt{\sum_{-W_n}^{W_n} |h|^2(k)}} \tag{4.9}$$

W_s représente la fenêtre spatiale associée au signal. C'est le nombre de points du réseau où le gain de l'amplificateur est supposé linéaire. Le gain est considéré comme nul en dehors de cette fenêtre (saturation de l'amplificateur). W_n représente la fenêtre spatiale associée au bruit constitué par la dispersion des tensions de décalage des amplificateurs. Dans le cas d'un réseau infini (ce qui revient à considérer $W_n = \infty$ dans la formule précédente) et avec la réponse 4.8, le facteur de mérite devient :

$$ECF = \frac{1 + r - 2\,r^{W_s+1}}{(1-r)\sqrt{\frac{1+r^2}{1-r^2}}} \tag{4.10}$$

Dans ce cas, un optimum existe sur le facteur de mérite en fonction du rapport des résistances R_0 et R_1. La figure 4.8 montre son évolution en fonction du rapport R_1/R_0. On remarque que l'efficacité du filtrage spatial est d'autant plus grande que le nombre

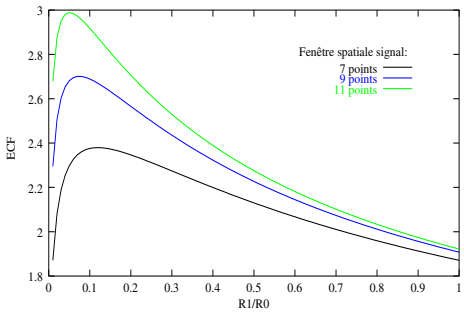


FIG. 4.8 – Filtre spatial : facteur de mérite.

de points de la fenêtre signal est important. L'analyse précédente suppose néanmoins que le réseau est infini et les conditions aux limites des tensions de références doivent être considérées.

4.5.2 Conditions aux limites

Afin de conserver un voisinage identique pour les éléments extrémités, il est nécessaire d'adjoindre un certain nombre d'étages neutres. D'autre part, la résistance de terminaison doit être déterminée pour "simuler" un réseau infini (figure 4.9). Sa valeur est donnée par la formule 4.11.

$$R_T = \frac{R_1 + \sqrt{R_1^2 + 4R_1R_0}}{2} \tag{4.11}$$

L'ajout d'étages neutres supplémentaires a pour conséquence une augmentation de la surface du circuit et une réduction de l'excursion maximale permise pour le signal. Un nombre optimum de ces étages peut être déterminé pour satisfaire une contrainte donnée de non-linéarité différentielle [94].

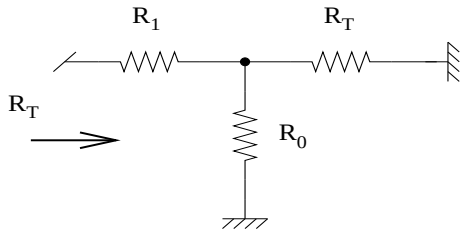


FIG. 4.9 – Résistance de terminaison.

4.6 Techniques de comparaison et résolution

Les techniques de comparaison utilisées dépendent des performances de la technologie et de la résolution recherchée. La réduction croissante des tensions d'alimentation entraîne par ailleurs une réduction de l'excursion du signal qui accroît l'importance de la dispersion des tensions de décalage. Pour de très faibles résolutions ($\sim 4bit$) l'utilisation d'un simple comparateur dynamique peut être suffisante. L'utilisation d'amplificateurs et la technique du filtrage spatial s'avèrent très rapidement nécessaires pour des résolutions supérieures. La référence [94] donne une borne maximale sur la résolution en technologie CMOS pour un écart type de l'offset de $\frac{\Delta}{4}$ et l'utilisation du filtrage spatial :

$$n_{max} = \log_2(\frac{0,385.FSR.\sqrt{A}}{4A_{VT}}) \tag{4.12}$$

Dans cette formule FSR représente l'excursion totale du signal, A est la surface des transistors réalisant l'amplificateur d'entrée et A_{VT} est un paramètre technologique caractérisant la déviation de la tension de seuil des transistors [97]. Le paramètre A_{VT} est pratiquement proportionnel à l'épaisseur d'oxyde t_{ox}. Pour une longueur de grille de 0,5 μm une valeur typique de ce paramètre est de 10 $mV.\mu m$ [14]. Avec cette valeur du paramètre A_{VT} et une excursion du signal $FSR = 2V$, la surface nécessaire pour les transistors de l'amplificateur est égale à 175 μm^2 pour une résolution de 8 bits. Celle-ci est 16 fois plus grande soit 2800 μm^2 pour un convertisseur de 10 bits d'après la formule 4.12. L'accroissement de la surface devient vite prohibitif pour des résolutions supérieures à 8 bit. Cet accroissement de surface est lié à l'effet combiné de la complexité exponentielle du convertisseur flash ($2^n - 1$ amplificateurs et comparateurs) et de la surface nécessaire à la réalisation de l'amplificateur qui dépend de la résolution. Cet accroissement de la surface de l'amplificateur se traduit également par une capacité d'entrée importante qui va directement influencer le comportement dynamique du convertisseur. L' impédance d'entrée du convertisseur est en effet constituée des $2^n - 1$ capacités d'entrée C_p des comparateurs reliées aux 2^n résistances R du diviseur potentiométrique. Ces capacités introduisent un couplage direct entre le signal et les $2^n - 1$ potentiels de référence. Ce couplage est étudié en annexe A. La contrainte sur le retour à l'équilibre des potentiels de référence fixe la valeur maximum de la constante de temps RC_p en fonction de la fréquence de conversion F_{clk} et de la résolution n du convertisseur :

$$RC_p < \frac{\pi^2}{n\,2^{2n}\,\ln(2)\,F_{clk}} \tag{4.13}$$

4.7 Interpolation

La pondération introduite par le filtrage spatial permet d'éliminer de manière régulière des amplificateurs ce qui présente l'avantage de réduire la complexité et la consommation du convertisseur. La figure 4.10 montre un exemple d'interpolation linéaire d'ordre 4 avec un diviseur résistif.

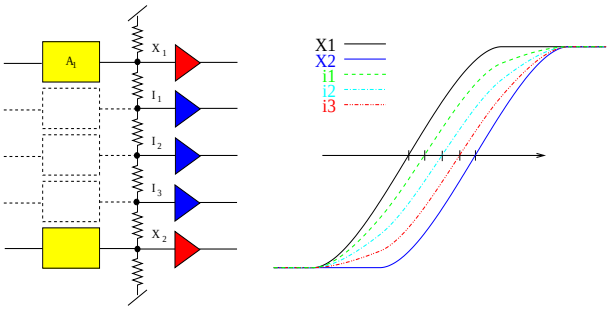


FIG. 4.10 – Interpolation d'ordre 4.

Cette réduction du nombre d' amplificateurs permet également de réduire la tension de décalage. En effet, soient n amplificateurs constitués de transistors de dimensions W et L. Si l'aire totale A_T est fixée, on a $A_T = n \cdot W \cdot L$. D'autre part, l'écart type sur la tension de décalage est tel que [97] :

$$\sigma_{ed} = \frac{A_{VT}}{\sqrt{W \cdot L}} = A_{VT}\sqrt{\frac{n}{A_T}} \tag{4.14}$$

A surface totale constante, la tension de décalage varie comme la racine carrée du nombre total d'amplificateurs.

Le facteur d'interpolation est cependant limité par la non-linéarité des amplificateurs, la plage de valeurs interpolées devant être exempte de toute saturation. Le choix du nombre optimal d'amplificateurs résulte ainsi d'un compromis entre la non linéarité introduite par la tension de décalage et celle qui résulte de la saturation des amplificateurs.

La figure 4.10 fait également apparaître que l'interpolation introduit une constante de temps supplémentaire entre les amplificateurs et les comparateurs. Ceci se traduit par une limitation de la bande passante du système d'autant plus grande que le facteur d'interpolation est important.

4.8 Pré-traitement analogique

La technique de repliement est une technique de pré-traitement du signal qui permet de réduire le nombre de comparateurs dans la structure flash. Elle consiste en un traitement modulaire qui divise le domaine de conversion en sous-domaines par une conversion grossière. La figure 4.11 donne le principe du traitement effectué dans le cas d'un convertisseur de 5 bit avec un facteur de repliement de 4. La sortie du bloc de repliement F est quantifiée sur 3 bit qui constituent les bit de poids faible du convertisseur. L'ambiguïté qui résulte du repliement est résolue par une quantification directe sur 2 bit du signal d'entrée. Ces deux bit constituent les poids forts du convertisseur. Le nombre total de comparateurs est de 7+3=10 alors qu'une structure flash directe en

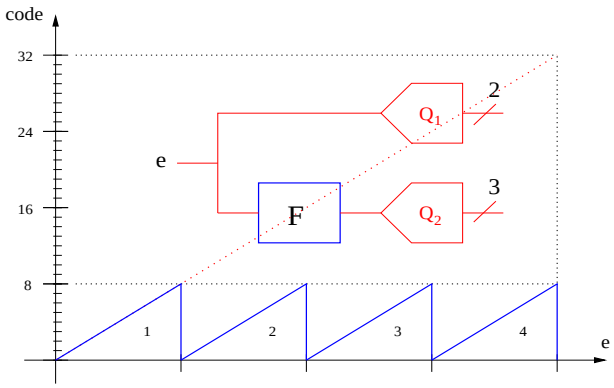


FIG. 4.11 – Technique de repliement.

aurait nécessité 31. La réalisation d'une telle forme de signal pour le pré-traitement est cependant très délicate. On préfère en pratique utiliser autant de circuits de repliement que de seuils nécessaires pour les bit de poids faible comme indiqué à la figure 4.12. Dans l'exemple précédent, 8 blocs de repliement seraient utilisés, chacun étant associé à un comparateur.

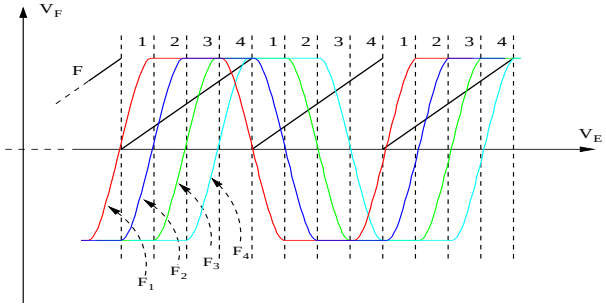


FIG. 4.12 – Signaux de repliement utilisés en pratique.

Dans le cas général, pour un convertisseur de $n = p + m$ bit, p bit sont attribués à la conversion grossière et le facteur de repliement est $F_r = 2^p$. Si chaque seuil de transition de la quantification fine est associé à un signal, ceci nécessite $N_F = 2^m$ blocs de génération pour ces signaux. La figure 4.13 montre un exemple de circuit de pré-traitement avec un facteur de repliement $F_r = 2$. On remarque que le nombre de paires différentielles est égale à $F_r + 1$. Le facteur de repliement est en pratique limité par des considérations statiques et dynamiques. Un facteur de repliement élevé, associé au faible gain des paires différentielles, entraîne en effet une forte sensibilité aux tensions de décalage. D'autre part, l'augmentation du nombre de paires différentielles se traduit par une capacité de sortie accrue du bloc de repliement qui entraîne une réduction de la bande passante du convertisseur.

La séparation en résolution grossière et fine n'est pas le seul moyen de décomposer le codage introduit par le pré-traitement analogique pour en réduire la complexité. On peut en effet, d'après le théorème du reste chinois, factoriser le code X en M résidus de manière unique $X \rightarrow \{x_1, x_2, ..., x_M\}$. Les résidus x_i sont congrus à X modulo m_i.

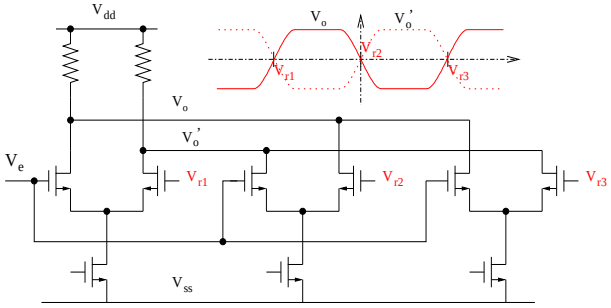


FIG. 4.13 – Bloc de repliement (F=2).

Les modules m_i sont M entiers premiers entre eux et la plage de codage est donnée par leur produit $\prod_{i=1}^{M} m_i$. La figure 4.14 donne un exemple d'un tel codage à l'aide de deux modules (m_1, m_2)=(4,3).

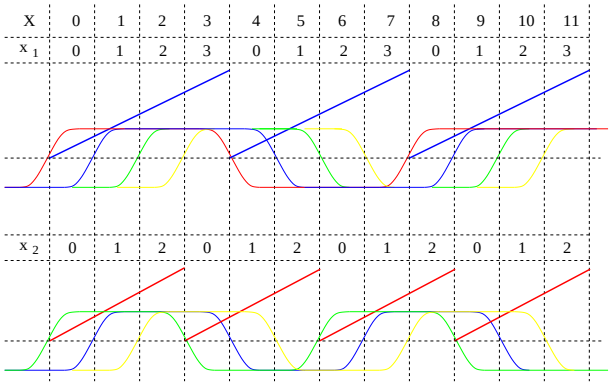


FIG. 4.14 – Exemple de traitement résiduel (m_1, m_2)=(4,3).

L'intérêt d'un tel codage est de permettre le traitement en parallèle sur les différents résidus sans retenues. Les opérations d'addition, de soustraction et de multiplication peuvent ainsi être réalisées de manière rapide et économique (les opérateurs sont de taille réduite) directement sur le code issu du convertisseur. Le problème de synchronisation des différents codes résiduels est cependant délicat. La référence [93] donne un exemple de code résiduel robuste vis à vis de ce problème. Celui-ci se traduit cependant par une réduction des codes effectivement exploitables et de l'excursion d'entrée du convertisseur par rapport à la tension d'alimentation.

4.8.1 Interpolation entre les blocs de repliement

Afin de réduire le nombre d'amplificateurs nécessaires au pré-traitement du signal, on peut effectuer une interpolation entre les sorties des blocs de repliement. Une interpolation d'ordre N_I permet alors de réduire le nombre de ces blocs à $N_F = \frac{2^m}{N_I}$. Le principe est identique à celui de la figure 4.10 où chaque amplificateur est ici remplacé par un bloc de repliement.

4.9 Convertisseurs flash 6 bit

Le tableau 4.1 donne quelques réalisations récentes de convertisseurs flash de 6 bit de résolution (années 1998-2002). Le choix de cette résolution est dicté par le nombre important de publications liées aux domaines d'application (circuit de lecture pour disque dur et DVD, équipement de test,...) qui nécessitent une résolution moyenne et une grande fréquence d'échantillonnage. La fréquence F_{sig} correspond à la mesure du nombre effectif de bit ($ENOB$). Pour certains convertisseurs, cette fréquence est bien inférieure à la demi-fréquence d'échantillonnage (F_{clk}) et les performances en terme de rapport signal sur bruit se dégradent rapidement avec la fréquence du signal.

Lmin (μm)	Vdd (V)	P (mW)	ENOB (bit)	F_{sig} (MHz)	F_{clk} (MHz)	A (mm^2)	Ec (pJ)	Ref
0,6	5,0	380	5,0	100	200	2,7	59,4	[28]
0,6	3,0	330	4,6	50	500	5,3	27,2	[133]
0,5	3,3	110	5,3	10	75	1,0	37,2	[87]
0,5	3,2	200	5.2	10	400	0,6	13,6	[35]
0,4	3,3	400	5,5	125	500	2,4	17,6	[121]
0,35	3,3	225	5,1	200	400	0,8	16,4	[79]
0,35	3,3	1155	4,4	141	1000	0,8	54,7	[124]
0,35	3,3	545	5,0	650	1300	0,8	13,1	[23]
0,35	3,3	300	5,4	450	900	0,3	7,9	[38]
0,35	3,0	190	5,0	100	400	1,2	14,8	[123]
0,25	3,3	187	5,5	136	700	0,5	5,9	[88]
0,25	3,3	400	5,0	200	800	1,7	15,6	[119]
0,25	2,5	193	5,1	100	600	0,2	9,4	[108]
0,25	2,2	150	4,7	200	400	1,2	14,4	[32]
0,18	1,95	328	5,0	660	1500	0,12	6,8	[109]
0,18	1,8	70	4,8	100	400	0,9	6,3	[92]
0,13	0,8	0,48	5,2	10	22	0,3	0,6	[75]

TAB. 4.1 – flash 6 bit

L'évolution de la technologie CMOS pour cette période est significative. On note en effet,que la longueur de canal et la tension d'alimentation sont divisées par un facteur de l'ordre de 5 dans cette période. Même si la dispersion reste très importante, l'impact de la diminution des dimensions sur la surface A du circuit est notable pour les réalisation les plus récentes [109]. Afin de comparer l'efficacité en puissance consommée P pour ces différentes réalisations, l'énergie par conversion et par quantum E_c [75] est utilisée :

$$E_c = \frac{P}{2^{ENOB}.F_{clk}} \tag{4.15}$$

La figure 4.15 représente l'évolution de cette énergie en fonction de la longueur de canal. Une régression linéaire sur les données du tableau 4.1 est également représentée. D'après cette approximation, l'énergie par conversion a également été réduite par un facteur de l'ordre de 5 pour la période considérée, sa valeur actuelle étant inférieur à 10 pJ.

Afin de mieux appréhender l'influence des différents paramètres sur la consommation, nous pouvons utiliser le modèle empirique suivant inspiré de la référence [129] :

$$P = F_{clk}^{a_1} \cdot {V_{dd}}^{a_2} \cdot {L_{min}}^{a_3} \cdot a_4 \tag{4.16}$$

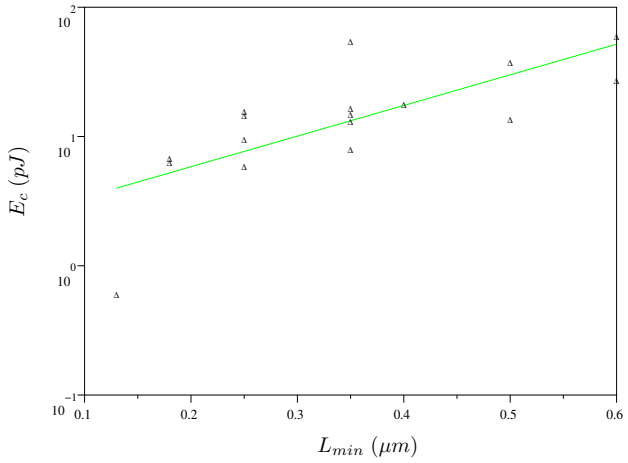


FIG. 4.15 – Evolution de l'énergie par conversion en fonction de la longueur du canal.

Le tableau 4.2 donne les paramètres de ce modèle, issus d'un ajustement quadratique sur les données précédentes (les unités étant celles du tableau 4.1).

a_1	a_2	a_3	a_4
0,93	1,7	0,76	0,23

TAB. 4.2 – Coefficients du modèle 4.16

On note l'importance de la tension d'alimentation V_{dd}. Une consommation purement dynamique liée aux commutations des circuits numériques conduirait à une dépendance en ${V_{dd}}^2$ (soit un paramètre $a_2 = 2$). La part de consommation des amplificateurs étant sensiblement proportionnelle à V_{dd} [129], ce facteur varie entre 1 et 2 en fonction de la consommation respective de ces deux types de circuits. Dans les deux cas, la consommation est sensiblement proportionnelle à la fréquence d'échantillonnage ou à celle du signal, ce qui se traduit par un coefficient a_1 voisin de 1. La valeur du coefficient a_3, ainsi que la figure 4.15, montrent que pour les longueurs de canal considérées la réduction des dimensions est favorable à la structure flash.

Chapitre 5

Convertisseur pipeline

5.1 Introduction

Le principe du convertisseur pipeline ou des convertisseurs dits algorithmiques (ou cycliques) repose sur l'algorithme de division récurrente. La figure 5.1 illustre une conversion de ce type sur n bit en base 2.

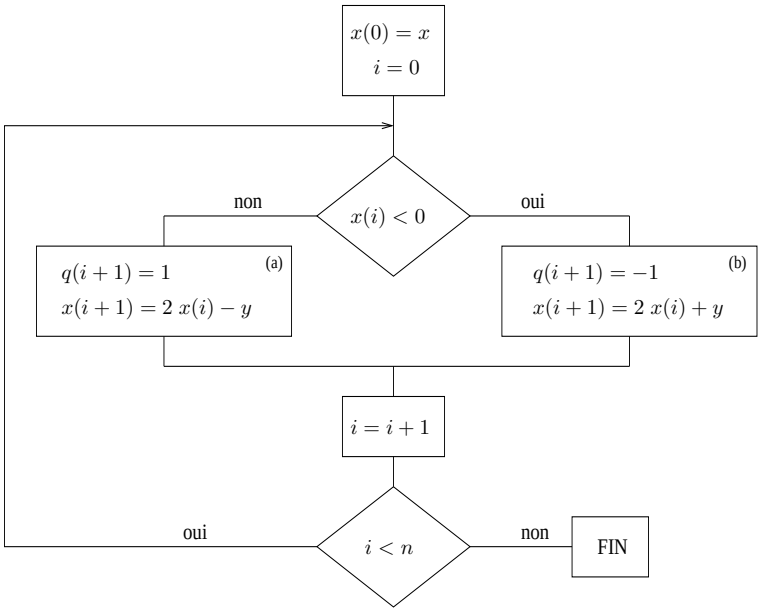


FIG. 5.1 – Conversion algorithmique.

Montrons qu'à l'issue des n itérations, les chiffres $q(i)$ contiennent une approximation numérique du quotient $\frac{x}{y}$.

Les étapes (a) et (b) assurent en effet que si $|x| < y$, tous les restes partiels $x(i)$ sont tels que $|x(i)| < y$. Par ailleurs, celles-ci peuvent s'écrire sous la forme condensée :

$$x(i+1) = 2\,x(i) - q(i+1)\,y \tag{5.1}$$

On a alors par récurrence :

$$\begin{aligned} x(0) &= x \\ x(1) &= 2\ x - q(1)\ y \\ x(2) &= 2^2\ x - (2\ q(1) + q(2))\ y \\ &\cdots \\ x(n) &= 2^n\ x - (\sum_{i=1}^{n} q(i)\ 2^{n-i})\ y \end{aligned} \tag{5.2}$$

Ce qui donne finalement :

$$\frac{x}{y} = \sum_{i=1}^{n} q(i)\ 2^{-i} + \frac{x(n)}{y}\ 2^{-n} \tag{5.3}$$

Le dernier terme étant strictement inférieur à 2^{-n}, on voit que les chiffres $q(i)$ constituent bien une représentation approchée du quotient $\frac{x}{y}$. Nous avons choisi de faire intervenir des "chiffres signés", les éléments du quotient pouvant être positifs ou négatifs. L'utilisation de chiffres signés est particulièrement intéressante car elle permet d'introduire une notation redondante des nombres. L'annexe B décrit un tel système de nombres et son utilisation dans l'algorithme de division. L'emploi de la redondance est pratiquement systématique dans tous les convertisseurs basés sur la division récurrente. Elle permet, comme nous allons le voir, de corriger certaines erreurs inhérentes aux imperfections des composants.

L'algorithme précédent fait intervenir une multiplication par deux à chaque étapes de la récursion. On peut supprimer cette multiplication en faisant intervenir le calcul d'un autre reste partiel $z(i) = 2^{-i}\ x(i)$. La relation 5.1 devient alors :

$$z(i+1) = z(i) - q(i+1)\ 2^{-(i+1)}\ y \tag{5.4}$$

La procédure de la figure 5.1 peut également être utilisée avec cette définition du reste partiel dans les étapes (a) et (b). La multiplication par deux n'est plus nécessaire mais on voit d'après la relation 5.4 que l'on doit disposer des multiples $w(i) = 2^{-(i+1)}\ y$ du diviseur. Cette procédure est utilisée dans la conversion dite à "approximations successives". Le diviseur est constitué de la tension de référence du convertisseur et les différents multiples $w(i)$ sont obtenus à partir d'un CNA. Les n poids $w(i)$ sont évalués en partant du poids fort par comparaison de la sortie du CNA et de l'entrée (figure 5.2). Ce type de convertisseur est donc de réalisation particulièrement simple et permet d'obtenir de très faibles consommations. La vitesse est cependant limitée par le temps d'établissement du CNA qui augmente avec la résolution du convertisseur. Pour un système du premier ordre de constante de temps τ l'établissement avec une précision relative de 2^{-n} (n bit) requiert un temps supérieur à $n \cdot \tau \cdot \log 2$. Comme il faut n itérations pour une conversion, on voit que le temps nécessaire varie en n^2.

La récursion décrite par l'équation 5.1 se prête également à une réalisation très simple en technique échantillonnée. La figure 5.3 en donne le schéma de principe basé sur deux phases ϕ_1 et ϕ_2. Durant ϕ_1, la tension d'entrée charge les capacités C_1 et C_2 et le signe de V_x est mémorisé dans le bit d. Durant ϕ_2, la charge est redistribuée en fonction de la tension de référence et du bit de signe :

$$V_r = (1 + \frac{C_2}{C_1}) \cdot V_x - q \cdot \frac{C_2}{C_1} \cdot V_{ref} \tag{5.5}$$

avec

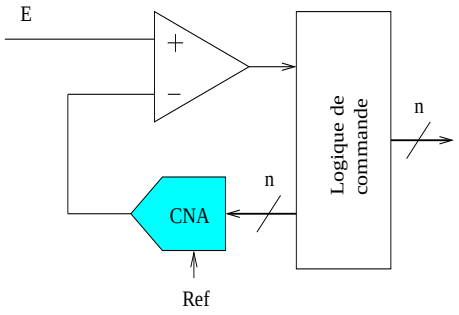


FIG. 5.2 – Convertisseur à approximation successives

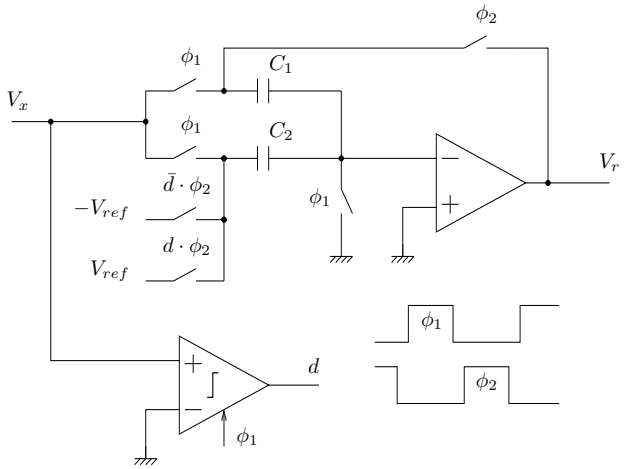


FIG. 5.3 – Schéma de principe réalisant la récursion décrite par l'équation 5.1.

$$q = \begin{cases} +1 & \text{si } d_i = \text{“1”} \\ -1 & \text{si } d_i = \text{“0”} \end{cases}$$

En choisissant $C_1 = C_2$, on obtient bien l'équation 5.1, le diviseur y étant remplacé par la tension de référence V_{ref} du convertisseur.

Pour effectuer une conversion sur n bit, nous pouvons utiliser le circuit précédent pendant n cycles. Dans le premier cycle, V_x représente la tension à convertir. Dans les cycles suivants V_x est égal au résidu V_r précédent. Autrement dit, on boucle la sortie du circuit précédent sur son entrée. La complexité du convertisseur ainsi obtenu (appelé convertisseur cyclique) est donc très réduite mais le temps de conversion est lié à la résolution désirée. Pour réduire le temps de conversion, le convertisseur pipeline utilise plusieurs blocs élémentaires de faible résolution pour construire le code de sortie. Dans notre exemple précédent, une résolution de n bit serait obtenue à partir de n blocs identiques à celui de la figure 5.3. Ces blocs forment une chaîne et sont activés à chaque période de l'horloge H (figure 5.4). Le résultat du traitement élémentaire est constitué d'une donnée analogique : le résidu r_i et d'un code intermédiaire d_i. C'est la combinaison de ces codes intermédiaires qui forme le mot de sortie du convertisseur. Chaque code étant associé à un poids et à un retard qui correspond à son rang dans la structure. Pour m blocs élémentaires, le traitement complet d'un échantillon du signal nécessite m périodes d'horloge. Ceci constitue le temps de latence du convertisseur. Les

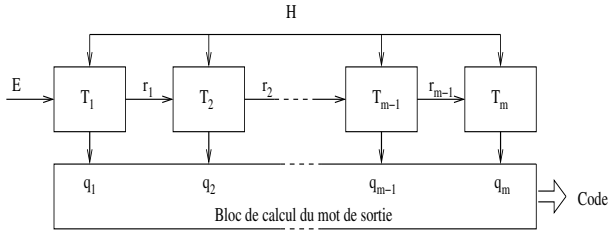


FIG. 5.4 – Schéma de principe du convertisseur pipeline

m blocs fonctionnant en parallèle, un échantillon du signal est cependant disponible à chaque période d'horloge. Dans le cas général, le traitement d'un bloc est constitué d'une quantification qui fournit le code élémentaire, d'une soustraction et d'un gain comme indiqué sur la figure 5.5. On notera que la génération du dernier résidu n'est pas nécessaire puisque celui-ci n'est pas exploité. Le dernier étage du pipeline est donc constitué d'un simple CAN. La résolution réduite et la recherche d'un temps de conversion minimum conduisent à utiliser une structure flash pour le CAN de tous les étages.

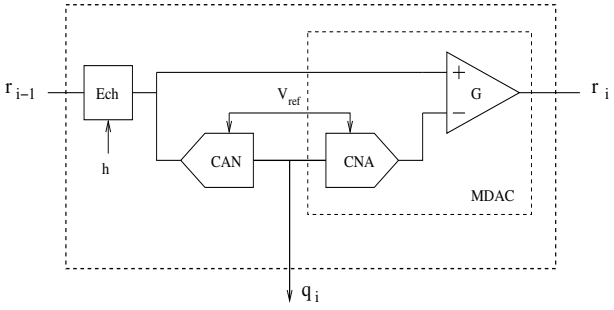


FIG. 5.5 – Etage d'un convertisseur pipeline

Le CNA et l'amplificateur sont généralement regroupés en un seul élément que l'on nomme MDAC. Ceci est dû à l'utilisation fréquente de la pondération de charges pour le CNA comme c'est le cas dans la figure 5.3 pour une résolution de 1 bit. La figure 5.6 donne un exemple de MDAC pour une résolution de n bit.

La conservation de la charge entre les deux phases ϕ_1 et ϕ_2 donne :

$$-\sum_{i=0}^{n} C_i\, V_e \quad = \quad -\sum_{i=1}^{n} q_i\, C_i\, V_{ref} - C_0\, V_s \tag{5.6}$$

Avec $C_i = 2^{i-1}C_0\,,\ i = 1,..,n$ on obtient pour la sortie V_s :

$$V_s = 2^n\,[V_e - \frac{V_{ref}}{2^n}\,\sum_{i=1}^{n} q_i\, 2^{i-1}] = G \cdot (V_e - V_q) \tag{5.7}$$

Ceci correspond bien au schéma de la figure 5.5 où le gain G et la sortie V_q du CNA sont donnés par le formule 5.7. Aux imperfections du CNA près, la sortie V_q est une approximation de l'entrée directement obtenue à partie du code q. En notant $G(i)$ et $V_q(i)$ les valeurs correspondantes pour le bloc i, la relation de récurrence générale pour les différents résidus est telle que :

$$V_r(i) = G(i) \cdot (V_r(i-1) - V_q(i)) \tag{5.8}$$

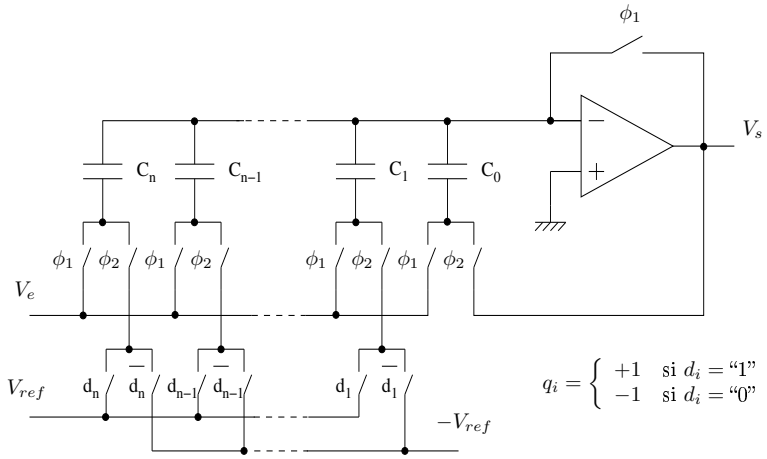


FIG. 5.6 – MDAC à pondération de charges

Pour m étages la tension d'entrée s'écrit :

$$V_e = \sum_{i=1}^{m} \frac{V_q(i)}{\prod_{j=1}^{i-1} G(j)} + \frac{V_r(n)}{\prod_{j=1}^{m} G(j)} \tag{5.9}$$

Dans le cas idéal, les différents gains $G(i)$ sont égaux à 2^{n_i} ou n_i est le nombre de bit par étage. La valeur du résidu étant bornée à $[-V_{ref}, V_{ref}]$ le dernier terme est, en valeur absolue, inférieur ou égal à $\frac{V_{ref}}{2^n}$ qui correspond au demi quantum du convertisseur.

5.2 Codage redondant

Une condition de convergence de l'équation de récursion dans l'algorithme de division est que le reste partiel reste borné par le diviseur. Dans la suite des étages du pipeline ceci est équivalent au fait que le résidu reste borné par la tension de référence. En pratique, les seuils de comparaison vont s'écarter de leurs positions idéales du fait de la tension de décalage des comparateurs. Il n'est alors plus possible de garantir que le résidu reste inférieur ou égal en valeur absolue à la tension de référence. La figure 5.7 illustre ce phénomène avec un convertisseur de 3 bit constitué de 3 étages de 1 bit similaires à celui de la figure 5.3. Les résidus r_1 et r_2 des deux premiers étage du pipeline sont représentés ainsi que la valeur associée au code de sortie par le quantum du convertisseur. Un décalage V_{os} sur le seuil de comparaison entraîne un débordement du résidu et affecte directement le code de sortie du convertisseur. On est alors confronté au même problème qu'avec le convertisseur flash ou la précision requise sur les seuils de comparaison est une fraction du quantum du convertisseur qui dépend des spécifications de non linéarité.

Pour relâcher cette contrainte sur la précision des seuils de comparaison, on peut introduire une certaine redondance dans la représentation des codes intermédiaires. En reprenant l'exemple d'un étage avec 1 bit de résolution et le code $q_i \in \{-1, 1\}$, nous pouvons par exemple introduire un code supplémentaire avec $q_i \in \{-1, 0, 1\}$. Conformément à l'annexe B, ceci conduit en base 2 à une redondance maximale dans la représentation du code intermédiaire (on a dans ce cas le facteur de redondance

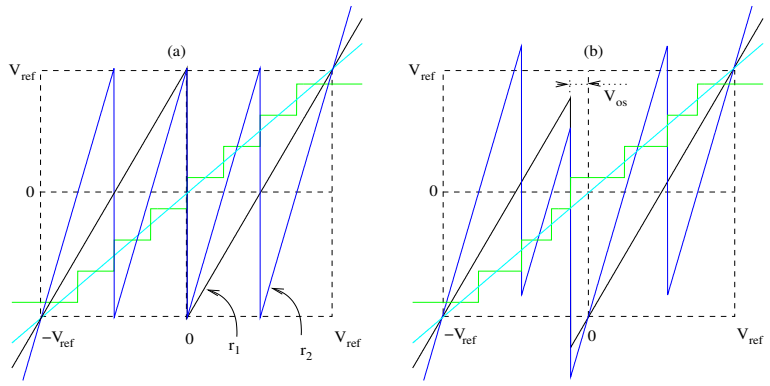


FIG. 5.7 – Caractéristique entrée-sortie et résidus d'un pipeline 3 bit à 1 bit par étage : (a) avec seuils idéaux (b) avec décalage des seuils de comparaison .

$\rho = \frac{m}{b-1}$ qui est égal à 1). Le diagramme de Robertson (cf annexe B) qui donne le résidu r_{j+1} en fonction du résidu courant r_j pour ce cas particulier est donné à la figure 5.8.

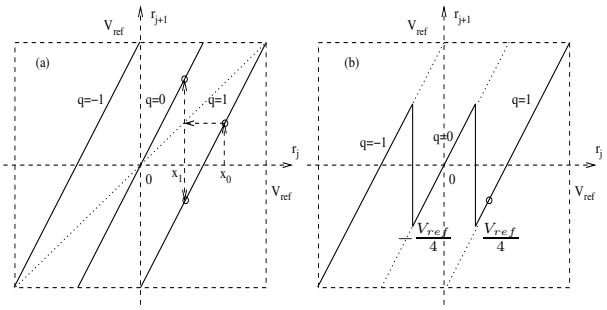


FIG. 5.8 – Diagramme de Robertson.

On voit sur la figure 5.8(a) que pour un résidu donné, il existe plusieurs choix possibles pour le code. Pour le résidu x_1 par exemple les choix $q = 0$ et $q = 1$ sont tous les deux valides. Prenons par exemple $x_0 = 0,7\,V_{ref}$, on a après 4 itérations les deux résultats suivants qui sont possibles [1] :

$$\hat{V}_e = 0,1011\,V_{ref} = 0,11\bar{1}1\,V_{ref} = \frac{11}{16}V_{ref}$$

La figure 5.8(b) montre comment nous pouvons exploiter cette redondance dans le code. Les seuils de comparaison sont placés au centre des zones de recouvrement des codes soit $\{-\frac{V_{ref}}{4}, \frac{V_{ref}}{4}\}$. D'après le diagramme 5.8(a), ces seuils peuvent varier de $\pm\frac{V_{ref}}{4}$ sans altérer le processus de division. Les deux seuils de comparaison nécessitent l'emploi de deux comparateurs et les 3 niveaux $\{-1, 0, 1\}$ sont représentés par deux bit du code intermédiaire. Chaque étage de ce type ne contribue cependant que pour un seul bit à la résolution totale du convertisseur. On rencontre souvent le terme d'étage à 1,5 bit dans la littérature pour désigner ce type de codage ($\log_2(3) \sim 1,585$).

[1] le symbole $\bar{1}$ représentele chiffre -1

Reprenons l'exemple du pipeline de trois étages de un bit (figure 5.7) en utilisant la structure redondante précédente dans les deux premiers étages. La figure 5.9 montre l'influence de l'introduction d'un décalage des seuils de comparateurs dans ces étages.

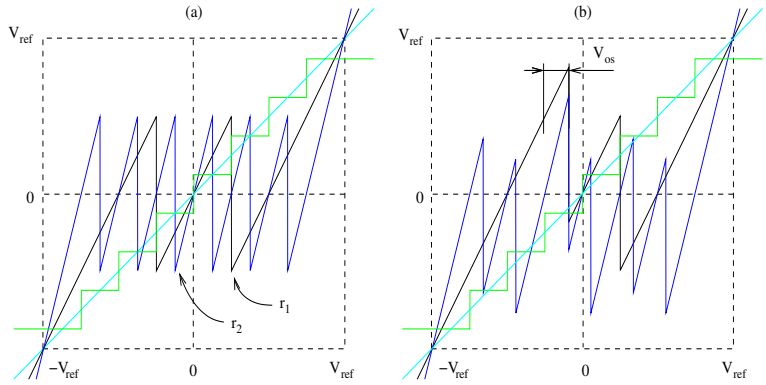


FIG. 5.9 – Caractéristique entrée-sortie et résidus avec redondance : (a) avec seuils idéaux (b) avec décalage des seuils de comparaison .

On remarque que ce décalage est sans effets sur la caractéristique du convertisseur tant que le résidu reste inférieur à la pleine échelle du convertisseur (décalage de $\pm\frac{V_{ref}}{4}$). On notera que l'introduction de redondance est inutile dans le dernier étage, puisque dans celui-ci le résidu n'est pas exploité. Le décalage sur les seuils des comparateurs de cet étage n'affecte cependant que les bits de poids faible du convertisseur. Grâce à l'utilisation de la redondance, la contrainte sur la précision des seuils est donc fortement relâchée tant que la résolution par étage est suffisamment faible. En effet, d'après l'annexe B, la zone de recouvrement maximum des codes est égale à $\frac{V_{ref}}{b}$. Doubler la base b, c'est à dire ajouter un bit de résolution par étage entraîne une tolérance moitié moindre sur les seuils de comparaison. Le nombre de comparateurs nécessaires est quant à lui égal à $2\,(b-1)$ soit deux fois plus qu'en absence de redondance. Un bit supplémentaire est également nécessaire pour la représentation des codes intermédiaires. Pour la réalisation du MDAC, une structure similaire à celle de la figure 5.6 peut être utilisée. L'introduction des niveaux supplémentaires est possible par l'utilisation d'une position supplémentaire pour chaque commutateur durant la phase ϕ_2. La base de chaque capacité est alors commutée sur V_{ref}, 0 ou $-V_{ref}$ en fonction de la valeur respective du chiffre $q_i = \{+1, 0, -1\}$ composant le code. En pratique, le réseau de capacités constituant le MDAC est souvent construit à partir d'une capacité élémentaire $C_{min} = C_0$. Les capacités C_1 à C_n de la figure 5.6 sont alors remplacées par $b-1$ capacités élémentaires qui sont contrôlée à partir du code thermométrique issu du convertisseur flash.

En résumé, on a les propriétés suivantes pour un convertisseur pipeline utilisant la redondance :

- Si n_i est le nombre de bit effectifs pour l'étage i ($b_i = 2^{n_i}$) et n la résolution totale du convertisseur, le nombre total de comparateurs n_c et le nombre m d'étages sont tels que :

$$n_c = \left(\sum_{i=1}^{m-1} 2\,(2^{n_i} - 1)\right) + (2^{n_m} - 1) \qquad n = \sum_{i=1}^{m} n_i \tag{5.10}$$

– Le décalage maximum δ sur chaque seuil de comparaison de l'étage i est égal à :

$$\delta = \pm \frac{V_{ref}}{2^{(n_i+1)}} \tag{5.11}$$

La résolution de 1 bit effectif par étage constitue un cas particulier très répandu en pratique pour sa simplicité. On a dans ce cas $n_c = 2\,n$: le nombre de comparateurs croît linéairement avec la résolution du convertisseur. En contre partie, la latence est alors maximum puisqu'il faut dans ce cas n périodes de l'horloge pour que le code d'un échantillon du signal d'entrée soit disponible à la sortie.

5.3 Sources d'erreurs dans le pipeline

Les sources d'erreurs dans le pipeline proviennent des imperfections liées aux CAN et aux MDAC. La section précédente a montré que, grâce au codage redondant, l'erreur sur les niveaux de décision du flash était sans conséquence sur la fonction du convertisseur tant que ceux-ci restent dans la plage de tolérance donnée par 5.11. Il n'en va pas de même pour les erreurs du MDAC qui affectent directement la caractéristique du pipeline. Celles-ci sont essentiellement liées à l'erreur de gain et aux erreurs propres à la conversion analogique-numérique. Dans un MDAC à transfert de charges tel que celui de la figure 5.6 ces grandeurs sont couplées car elles proviennent de la dispersion des valeurs de capacités qui interviennent dans le gain et dans le poids des coefficients du CNA. Le transfert des charges est également dépendant des caractéristiques de l'amplificateur tel que son gain statique et son temps d'établissement. Chaque étage du pipeline est également le siège d'un bruit qui provient du bruit thermique des commutateurs ou du bruit propre de l'amplificateur (bruit thermique et bruit en 1/f).

5.3.1 Dispersion sur les gains

D'après les relations 5.8 et 5.9 et en supposant le CNA idéal, on voit que si les gains G_i ne sont pas égaux aux bases respectives b_i des différents étages l'interprétation de la tension d'entrée à partir des codes intermédiaires sera fausse (chaque code intermédiaire étant défini par $\frac{V_q(i)}{\Delta_i}$ où Δ_i est le quantum local). Dans le cas ou ces gains sont connus et n'entraînent pas de saturation sur les résidus, il est possible d'interpréter directement V_e à partir des ces relations sans commettre d'erreur (autre que la quantification liée à cette représentation). Les bases de représentation du code sont alors non entières et égales aux gains réels des étages. Dans le cas contraire la caractéristique entrée-sortie peut présenter des discontinuités importantes. Ceci est illustré par la figure 5.10 où la caractéristique d'un pipeline redondant de trois étages de 2 bit est représentée. Le gain est idéalement de 4, valeur de la base. Le gain du premier étage est fixé à 5 pour mettre en évidence l'erreur introduite sur la caractéristique.

La figure 5.11 représente les courbes de non linéarité différentielle et intégrale du même convertisseur obtenues à partir d'un test d'histogramme. Celles-ci mettent en évidence les discontinuités de l'ordre d'un quantum dans ce cas particulier.

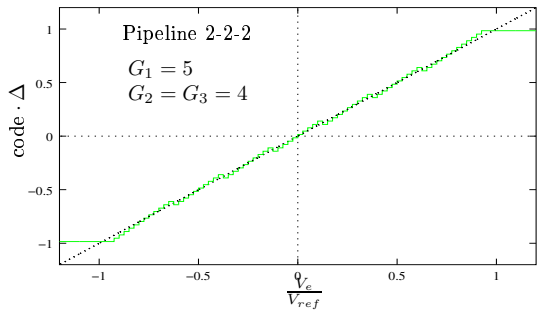


FIG. 5.10 – Caractéristique entrée-sortie avec défaut de gain.

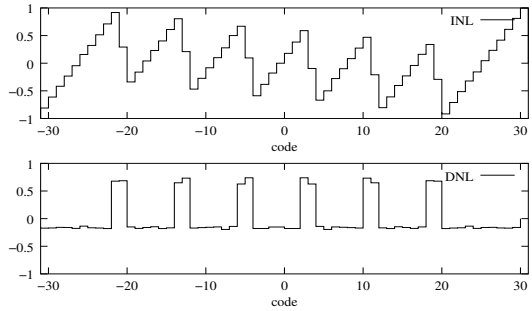


FIG. 5.11 – Non linéarité intégrale et différentielle avec défaut de gain.

5.3.2 Erreurs liées aux CNA

En notant ϵ_i l'erreur additive associée à la conversion numérique-analogique on a d'après 5.8 et 5.9 une erreur totale ϵ à l'entrée donnée par :

$$\epsilon = \epsilon_1 + \frac{\epsilon_2}{G_1} + \frac{\epsilon_3}{G_1 \cdot G_2} + \cdots = \sum_{i=1}^{m} \frac{\epsilon_i}{\prod_{j=1}^{i-1} G(j)} \tag{5.12}$$

Celle-ci est pondérée par les gains du pipeline et son influence est prédominante dans les premiers étages. Ainsi tout décalage sur les niveaux du premier CNA se traduit par un décalage identique sur la caractéristique entrée-sortie du convertisseur. Ceci est illustré sur la figure 5.12 qui reprend l'exemple du pipeline 2-2-2 où l'on a introduit une erreur égale au dixième de la tension maximum d'entrée sur certains niveaux du premier CNA.

5.3.3 Calibrage

Si l'utilisation d'un codage redondant relâche les contraintes sur le convertisseur flash, celle-ci ne modifie pas les exigences sur le MDAC. En particulier la linéarité du premier étage doit être compatible avec la résolution totale du convertisseur (cf formule 5.12). L'appariement des composants n'autorise généralement que des linéarités de 10 à 12 bit pour le MDAC, limitant ainsi la résolution du pipeline à cette valeur. Pour atteindre des résolutions plus élevées, un calibrage est nécessaire. Celui-ci peut être

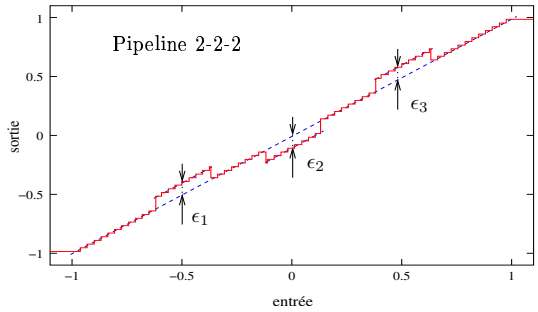


FIG. 5.12 – Caractéristique entrée-sortie avec une erreurs sur le premier CNA.

effectué de manière analogique en utilisant essentiellement des CNA supplémentaires de résolution supérieure à la résolution d'un étage [44]. Une phase de calibrage mesure les erreurs qui sont stockées dans une mémoire associée à chaque code. La lecture de cette mémoire est alors utilisée en fonctionnement normale pour commander les CNA qui vont corriger la sortie des MDAC. Afin de réduire le coût et le temps de conception liés à l'introduction de composants de précision supplémentaires, des techniques de correction numériques ont été proposées [67]. Dans leur principe, ces méthodes utilisent les derniers étages du pipeline pour mesurer les non linéarités des premiers étages. On peut par exemple dans le cas du pipeline 2-2-2 considéré précédemment utiliser les étages 2 et 3 pour mesurer les discontinuités de la caractéristique qui ont été mises en évidence en 5.3.1 et 5.3.2. Les coefficients de correction sont toujours stockés en mémoire mais sont appliqués directement sur les codes intermédiaires dans le domaine numérique. Qu'elles soient numériques ou analogiques, les méthodes précédentes nécessitent une phase de calibrage où le convertisseur est indisponible. Afin de pouvoir effectuer la correction en cours de conversion on peut ajouter au signal à convertir un signal de calibrage non corrélé à ce dernier. Des technique de corrélation permettent alors d'extraire l'information liée au signal de calibrage. La référence [83] est un exemple d'une telle réalisation qui utilise une séquence pseudo-aléatoire ± 1 comme signal de calibrage.

5.4 Dimensionnement des étages

A côté des erreurs systématiques, telle celle du gain fini de l'amplificateur, il existe des sources d'erreur aléatoires non corrélées telles que la dispersion des valeurs de capacités ou le bruit thermique. Pour un MDAC à pondération de charges, l'importance de ces sources dépend essentiellement de la capacité totale associée à l'étage du pipeline. Ainsi pour une capacité, la variance de l'erreur relative est généralement inversement proportionnelle à sa valeur :

$$\epsilon = \frac{\Delta C}{C} \qquad \sigma_\epsilon^2 \propto \frac{1}{C} \tag{5.13}$$

Pour un rapport fixé de deux capacités, la variance de l'erreur est donc inversement proportionnelle à la capacité minimum utilisée. Il en est du même du bruit thermique échantillonné dont la variance est égale à $\frac{k_B T_a}{C}$. Globalement, la variance des sources d'erreur associée à un étage du pipeline est donc inversement proportionnelle à sa capacité minimum (ou à sa capacité totale qui est déduite de la capacité minimum par

la valeur de la base). La puissance consommée étant liée à la valeur de la capacité totale, il peut être intéressant de dimensionner celle-ci de manière à réduire la consommation. Ceci ce fera au prix d'une légère augmentation du bruit. Pour le montrer, considérons le cas particulier ou tous les étages sont identiques avec un gain G et la puissance est proportionnelle à C^2. Cette dernière hypothèse résultant d'une bande passante constante $B \propto \frac{g_m}{C}$ et d'une transconductance $g_m \propto \sqrt{I}$ de l'amplificateur ($P \propto I \propto C^2$). On note P_0 la puissance consommée par le premier bloc et σ_0^2 la variance du bruit ramenée à l'entrée de ce bloc. De même, on note P la puissance totale consommée et σ^2 la variance totale du bruit à l'entrée. En appliquant un facteur d'échelle k sur la capacité entre deux étages :

$$C_{i+1} = \frac{C_i}{k} \quad , \quad i = 0, .., m-1 \tag{5.14}$$

La contribution de l'étage i au bruit et à la puissance totale sont respectivement :

$$\sigma_i^2 = \sigma_0^2 \left[\frac{k}{G^2}\right]^i \qquad P_i = P_0 \left[\frac{1}{k^2}\right]^i \tag{5.15}$$

D'où le bruit et la puissance normalisés par ceux du premier étage :

$$\frac{\sigma^2}{\sigma_0^2} = \sum_{i=0}^{m-1} \left[\frac{k}{G^2}\right]^i \qquad \frac{P}{P_0} = \sum_{i=0}^{m-1} \left[\frac{1}{k^2}\right]^i \tag{5.16}$$

En reprenant l'exemple du pipeline 2-2-2 le tableau 5.1 donne le bruit et la puissance consommée pour quelques facteurs d'échelle. On remarque sur cet exemple qu'il est

k	1	$\sqrt{2}$	2
σ/σ_0	1,03	1,05	1,07
P/P_0	3	1,75	1,31

TAB. 5.1 – bruit et puissance consommée en fonction du facteur d'échelle

possible d'obtenir une réduction importante de la consommation pour un très faible accroissement du bruit.

5.5 Résolution par étage

Il existe différents choix pour la répartition de la résolution sur les m étages d'un pipeline. La contrainte essentielle étant que la somme des résolutions effectives n_i (sans prise en compte des bit de redondance) soit égale à la résolution totale n du convertisseur (formule 5.10). La solution qui consiste à prendre n étages identiques avec un bit de résolution (et un pour la redondance) est très utilisée. Elle a l'avantage de conduire à des structures très simples et répétitives. D'autre part, ceci conduit à un gain minimum de deux par étage. Si le produit gain bande des amplificateurs est fixé, cette solution permet d'obtenir la plus grande bande passante. La contrepartie est que le nombre d'amplificateurs est alors maximum et que ce composant intervient pour une grande part de la consommation totale. Prenons par exemple un convertisseur non redondant de résolution n avec un nombre de bit x identique par étage. Si la puissance consommée par un amplificateur est a fois la puissance P_{comp} d'un comparateur, la puissance totale est alors donnée par :

$$P = \left(\frac{n}{x}[(2^x - 1) + a]\right) P_{comp} \tag{5.17}$$

Le tableau 5.2 donne la résolution optimum x^* par étage correspondant au minimum de consommation pour quelque valeur de a. Celle-ci augmente avec la consommation relative de l'amplificateur.

a	10	20	50	200
x^*	3	4	5	6

TAB. 5.2 – Résolution optimum

L'analyse précédente suppose que la consommation de l'amplificateur est indépendante de la résolution de l'étage. En réalité, celle-ci augmente avec la résolution car la précision requise pour un temps d'établissement donné augmente. Ceci est dû par exemple à l'utilisation de capacités de plus grandes valeurs pour augmenter la précision des rapports et à l'augmentation de la capacité d'entrée du CAN flash qui est également fonction de sa résolution. D'autre part la bande passante d'un étage est fortement dépendante de sa résolution. Ce point est examiné plus en détail à l'annexe F.

5.6 Utilisation du parallélisme

Afin d'augmenter la cadence d'échantillonnage F_s on peut utiliser un réseau de M convertisseurs pipelines en parallèle. Chaque convertisseur est échantillonné à une cadence $\frac{F_s}{M}$ selon le schéma de principe indiqué à la figure 5.13. Les séquences d'échantillonnage sont entrelacées et la recombinaison des différentes sortie est effectuée par un multiplexeur.

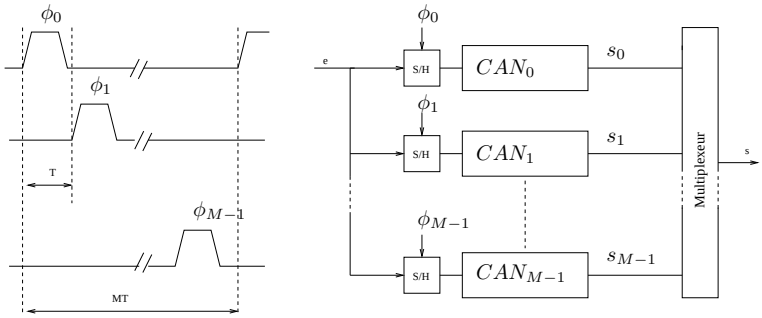


FIG. 5.13 – Utilisation du parallèlisme.

Ce schéma de principe est en fait applicable à différents types de convertisseurs mais la complexité réduite du convertisseur pipeline le rend particulièrement bien adapté à ce type de traitement. On peut ainsi atteindre des fréquences d'échantillonnage inaccessibles avec un seul convertisseur. Cette technique a ainsi récemment permit de réaliser un convertisseur 8 bit avec une fréquence d'échantillonnage de 20 GHz en technologie CMOS 0,18 μm utilisant 80 pipelines échantillonnés à 250 MHz [104].

La résolution est théoriquement identique à celle d'un seul des convertisseurs pipeline. En pratique, la dispersion sur le gain, le décalage ou le temps d'échantillonnage propre à chacune des voies va introduire une erreur supplémentaire dans le traitement parallèle qui se traduit par une perte de résolution.

5.6.1 Erreurs liées au traitement parallèle

Erreurs de gain et de décalage

Pour analyser l'effet des dispersions sur les différentes voies, on considère que chaque traitement est modélisé par un gain α_i et un décalage β_i, $i = 0, .., M-1$. Ceci traduit le comportement global de chaque convertisseur sans prendre en compte le bruit de quantification propre qui est supposé additif et indépendant. Les sorties s_0 à s_{M-1} qui proviennent de l'entrelacement temporel sont alors données par :

$$s_i(t) = u_i(t) \cdot \Pi_i(t) \tag{5.18}$$

avec

$$u_i(t) = \alpha_i\, e(t) + \beta_i \quad , \quad \Pi_i(t) = \sum_{n=-\infty}^{\infty} \delta\left[t - \left(n + \frac{i}{M}\right) MT\right]$$

La transformée de Fourier d'une séquence est obtenues par la convolution de :

$$U_i(f) = \alpha_i\, E(f) + \beta_i\, \delta(f)$$

transformée de Fourier de $u_i(t)$) et de celle du peigne $\Pi_i(t)$:

$$\mathcal{F}\{\Pi_i(t)\} = \frac{1}{MT} \sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{MT}\right) e^{-j2\pi \frac{i \cdot n}{M}} \tag{5.19}$$

soit

$$S_i(f) = \frac{1}{MT} \sum_{n=-\infty}^{\infty} U_i\left(f - \frac{n}{MT}\right) \cdot e^{-j2\pi \frac{i \cdot n}{M}} \tag{5.20}$$

La transformée de Fourier de la sortie s'obtient en sommant sur les M voies :

$$S(f) = \sum_{i=0}^{M-1} S_i(f) \tag{5.21}$$

$$S(f) = \tfrac{1}{T} \sum_{n=-\infty}^{\infty} \left[A_n\, E\left(f - \tfrac{n}{MT}\right) + B_n\, \delta\left(f - \tfrac{n}{MT}\right)\right]$$

avec

$$A_n = \frac{1}{M} \sum_{i=0}^{M-1} \alpha_i \cdot e^{-j2\pi \frac{i \cdot n}{M}} \quad , \quad B_n = \frac{1}{M} \sum_{i=0}^{M-1} \beta_i \cdot e^{-j2\pi \frac{i \cdot n}{M}}$$

Au facteur $\frac{1}{M}$ près, les coefficients A_n et B_n sont les transformées de Fourier discrètes respectives des coefficients α et β. Dans le cas idéal, ceux-ci sont respectivement égaux à l'unité et à zéro. A_n est nul sauf pour les multiple de M pour lesquels il est égal à l'unité et B_n est identiquement nul. Le spectre en sortie est alors donné par :

$$S(f) = \sum_{i=0}^{M-1} S_i(f) = \frac{1}{T} \sum_{n=-\infty}^{\infty} E\left(f - \frac{n}{T}\right) \tag{5.22}$$

qui correspond bien à un échantillonnage simple de période T du signal d'entrée.

D'après la formule 5.21, on remarque que l'erreur liée au décalage introduit sur chaque voie est indépendante du signal d'entrée. Dans la bande de Nyquist du convertisseur, le spectre associé à cette erreur est composé de M raies dont la puissance totale, d'après la relation de Parseval, est donnée par :

$$P_b = \sum_{n=0}^{M-1} |B_n|^2 = \frac{1}{M} \sum_{i=0}^{M-1} |\beta_i|^2 \tag{5.23}$$

Si les β_i sont décrits par des variables aléatoires indépendantes, centrées et d'écart type commun σ_b, la puissance moyenne de bruit liée au décalage est alors [2] :

$$\bar{P_b} = \sigma_b^2 \tag{5.24}$$

Le rapport signal sur bruit peut alors être facilement déterminé, dès lors que la puissance du signal est connue. On notera que celui-ci est indépendant du nombre de canaux utilisés.

Dans le cas d'une excitation complexe $E(f) = a \cdot \delta(f - f_o)$ où seule l'erreur de gain est présente, le spectre du signal de sortie devient :

$$S(f) = \frac{a}{T} \sum_{n=-\infty}^{\infty} A_n \, \delta\left(f - f_o - \frac{n}{MT}\right) \tag{5.25}$$

En utilisant une relation semblable à 5.23 sur les coefficients A_n :

$$G_a = \sum_{n=0}^{M-1} |A_n|^2 = \frac{1}{M} \sum_{i=0}^{M-1} |\alpha_i|^2 \tag{5.26}$$

la puissance totale du signal et du bruit s'exprime simplement par $P_a = a^2 \cdot G_a$. En notant $\alpha_i = 1 + \xi_i$ où ξ_i est une variable aléatoire centrée qui représente l'erreur sur le gain, la puissance totale moyenne $\bar{P_a}$ devient :

$$\bar{P_a} = a^2 \cdot \left[1 + \frac{1}{M} \sum_{i=0}^{M-1} \bar{\xi_i^2}\right] \tag{5.27}$$

Dans le cas où les erreurs de gain sont indépendantes avec un écart type commun σ_a, cette puissance devient $\bar{P_a} = a^2 \cdot \left(1 + \sigma_a^2\right)$. Dans la bande de Nyquist ($B = \frac{1}{T}$) du convertisseur, le spectre donné par 5.25 est composé du fondamental ($n = 0$) et de $M - 1$ raies parasites ($n = 1, .., M - 1$). On a, pour la puissance moyenne du fondamental :

$$\bar{P_s} = a^2 \cdot \bar{A_0^2} = \frac{a^2}{M^2} \sum_{i=0}^{M-1} \sum_{j=0}^{M-1} \overline{\alpha_i \, \alpha_j} = a^2 \cdot \left(1 + \frac{\sigma_a^2}{M}\right) \tag{5.28}$$

Ce qui conduit au rapport signal sur bruit :

$$SNR = 10 \cdot \log_{10}\left(\frac{P_s}{P_a - P_s}\right) = 10 \cdot \log_{10}\left(\frac{1 + \frac{\sigma_a^2}{M}}{\sigma_a^2 \left(1 - \frac{1}{M}\right)}\right) \tag{5.29}$$

Le tableau 5.3 donne la valeur de ce rapport en fonction de M pour $\sigma_a = 0,005$. On remarque une faible dépendance en fonction du nombre de canaux en particulier pour

M	2	4	8	16	32
SNR (dB)	49	47,3	46,6	46,3	46,16

TAB. 5.3 – Rapport signal sur bruit lié à la dispersion des gains

les valeurs importantes de M. Cette erreur sur les gains représente une limite forte sur la résolution globale du convertisseur et ceci indépendamment de la résolution propre à chaque canal. Cette dispersion des gains représente également une contrainte sur les performances en terme de SFDR. On peut montrer [102] que pour une excitation sinusoïdale le spectre est composé de $M-1$ raies parasites dont l'amplitude moyenne par rapport au fondamental est donnée par :

$$\mu = \frac{\sigma_a}{2}\sqrt{\frac{\pi}{M}} \tag{5.30}$$

Lorsque M augmente, le nombre de raies parasites augmente également alors que les amplitudes moyennes décroissent. Pour $\sigma_a = 0,005$ et $M = 4$, on a par exemple 3 raies parasites dont l'amplitude moyenne est $-20\,log_{10}(\mu) = 53,1\,dB$ en dessous du niveau du fondamental (figure 5.14).

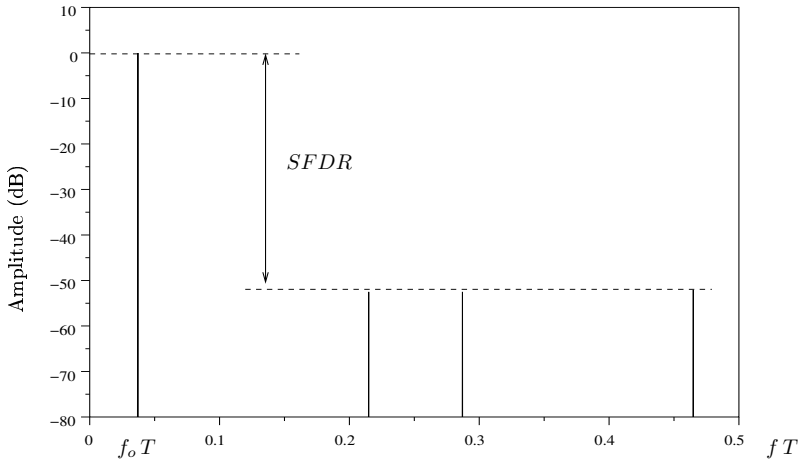


FIG. 5.14 – Spectre moyen (100 échantillons) avec $\sigma = 0,005$ et $M = 4$.

Erreur sur l'instant d'échantillonnage

On suppose dorénavant que le traitement sur chaque voie est identique mais que l'instant d'échantillonnage est légèrement décalé par rapport à sa position idéale :

$$t_i = n \cdot MT + i \cdot T + \delta_i \cdot T$$

La variable δ_i représentant l'erreur relative par rapport à la période T pour le canal i. La suite des échantillons en sortie du multiplexeur étant supposée régulière, la formule 5.18 devient :

$$s_i(t) = \Pi_i(t) \cdot e(t - \delta_i \cdot T) \tag{5.31}$$

[2]La barre supérieure représente ici la moyenne statistique

La transformée de Fourier est obtenue par le produit de convolution :

$$S_i(f) = \mathcal{F}\{\Pi_i(t)\} \star \left\{E(f) \cdot e^{-j2\pi f\, \delta_i \cdot T}\right\} \tag{5.32}$$

soit

$$S_i(f) = \frac{1}{T} \sum_{n=-\infty}^{\infty} \left[\frac{1}{M} e^{-j2\pi\, \delta_i\, T\left(f-\frac{n}{MT}\right)} \cdot e^{-j2\pi \frac{i \cdot n}{M}}\right] \cdot E\left(f - \frac{n}{MT}\right) \tag{5.33}$$

La transformée de Fourier de la sortie est alors donnée par :

$$S(f) = \sum_{i=0}^{M-1} S_i(f) = \frac{1}{T} \sum_{n=-\infty}^{\infty} C_n \cdot E\left(f - \frac{n}{MT}\right) \tag{5.34}$$

avec

$$C_n = \frac{1}{M} \sum_{i=0}^{M-1} e^{-j2\pi\, \delta_i\, T\left(f-\frac{n}{MT}\right)} \cdot e^{-j2\pi \frac{i \cdot n}{M}}$$

Pour une excitation complexe $E(f) = a \cdot \delta(f - f_o)$, on a en en sortie le spectre de raies :

$$S(f) = \frac{a}{T} \sum_{n=-\infty}^{\infty} D_n \cdot \delta\left(f - f_o - \frac{n}{MT}\right) \tag{5.35}$$

avec

$$D_n = \frac{1}{M} \sum_{i=0}^{M-1} e^{-j2\pi\, \delta_i\, f_o\, T} \cdot e^{-j2\pi \frac{i \cdot n}{M}}$$

Cette expression est identique à celle des A_n (5.21) si l'on pose $\alpha_i = e^{-j2\pi\, \delta_i\, f_o\, T}$. On a en particulier, d'après 5.26 :

$$\sum_{n=0}^{M-1} |D_n|^2 = \frac{1}{M} \sum_{i=0}^{M-1} |\alpha_i|^2 = 1 \tag{5.36}$$

La puissance totale signal plus bruit est alors égale à $P_t = a^2$ alors que la puissance du fondamental est $Ps = a^2 \cdot |D_0|^2$. Ceci permet d'exprimer le rapport signal sur bruit sous la forme :

$$SNR = 10 \cdot \log_{10}\left(\frac{P_s}{P_t - P_s}\right) = 10 \cdot \log_{10}\left(\frac{|D_0|^2}{1 - |D_0|^2}\right) \tag{5.37}$$

avec

$$D_0 = \frac{1}{M} \sum_{i=0}^{M-1} e^{-j2\pi\, \delta_i\, f_o\, T}$$

L'analyse précédente suppose que les décalages temporels déterministes δ_i sont quelconques. Si maintenant ceux-ci sont considérés comme des variables aléatoires centrées normalement distribuées et d'écart type σ_r, on peut montrer [62] que le rapport signal sur bruit peut être approximé par :

$$SNR = 20 \cdot \log_{10}\left(\frac{1}{\omega_o\, \sigma_r\, T}\right) - 10 \cdot \log_{10}\left(\frac{M-1}{M}\right) \tag{5.38}$$

où $\omega_o = 2\,\pi\,f_o$ est la pulsation du signal d'entrée. Le premier terme est similaire à l'expression 2.18 obtenue lors de l'étude de l'échantillonnage avec incertitude temporelle en posant $\sigma_j = \sigma_r\,T$. Le rapport signal sur bruit est par ailleurs pratiquement déterminé par celui-ci dès que M est suffisamment grand. La tolérance sur l'incertitude temporelle propre à chaque canal est donc pratiquement équivalente à celle que l'on aurait avec un seul échantillonnage à la cadence $F_s = \frac{1}{T}$. Pour réduire la contrainte sur l'horloge de chaque canal, on utilise en général un échantillonneur supplémentaire, en amont de tous les autres, qui élimine ainsi l'effet des dispersion sur les différentes voies [118].

Correction numérique

Les erreurs précédentes liées au traitement parallèle conduisent à une résolution plus faible que celle autorisée par un seul des convertisseurs. Ceci peut être toléré pour des applications utilisant des résolutions de l'ordre de 8 bit qui sont compatibles avec les tolérances des composants intégrés, mais cette limitation est trop contraignante pour un grand nombre d'applications de communications utilisant des résolutions supérieures ou égale à 10 bit avec une fréquence d'échantillonnage élevée. On peut profiter des potentialités du traitement numérique des technologies actuelles pour corriger les erreurs de décalage, de gain ou de dispersion sur l'instant d'échantillonnage. La référence [61] donne un exemple de réalisation d'un convertisseur pipeline parallèle de 10 bit de résolution qui fait appel au techniques de traitement numérique adaptatif du signal.

5.7 Réalisations récentes

Le tableau 5.4 donne quelques exemples de réalisations de convertisseurs pipeline CMOS. Les résolutions varient de 8 à 15 bit, la valeur de 10 bit correspondant au plus grand nombre de réalisations.

La figure 5.15 représente l'évolution de l'énergie par conversion E_c (cf 4.15) en fonction de la longueur de canal à partir des données du tableau 5.4. Afin d'évaluer la tendance avec L_{min}, une régression linéaire est également représentée. Comme pour les convertisseurs flash, l'énergie par conversion a été réduite par un facteur de l'ordre de 5 pour la période considérée, sa valeur actuelle étant proche de 1 pJ. Cette faible valeur justifie l'utilisation fréquente du convertisseur pipeline pour des résolutions moyennes.

Indépendamment de la technologie, la figure 5.16 représente la puissance normalisée $P_r = \frac{P}{2^{ENOB}}$ en fonction de la fréquence d'échantillonnage F_s pour les différentes résolutions du tableau 5.4. La figure fait également apparaître une droite de régression qui conduit à l'estimation suivante pour la puissance :

$$P(mW) \approx 5\,10^{-4} \cdot 2^{ENOB} \cdot Fs(MHz)^{1,6} \tag{5.39}$$

Cette estimation globale est très approximative comme on peut le voir sur la figure 5.17 qui représente la puissance estimée à partir de 5.39 et la puissance mesurée donnée par le tableau 5.4. Ceci réduit beaucoup l'intérêt d'un modèle de régression basé sur un ensemble de réalisations pour décider d'un choix particulier d'architecture. En effet, pour être précis un tel modèle nécessite un grand nombre de données homogènes. Ceci est contradictoire avec l'évolution très rapide de la technologie et des techniques mises en oeuvre pour s'adapter à cette environnement. C'est particulièrement vrai pour la conception des convertisseurs pipelines qui, pour palier aux imprécisions de la technologie VLSI utilise de plus en plus des techniques de correction numérique. Le coût

n (bit)	F_s (Ms/s)	P (mW)	L_{min} (μm)	V_{dd} (V)	ENOB (bit)	Ec (pJ)	ref
8	75	70	.5	3.3	6.9	7.8	[13]
8	40	61	.5	2.7	6.5	16.8	[117]
8	80	250	.5	3	7	24.4	[83]
9	14	8.2	.5	1.5	8	2.3	[131]
9	5	1.6	.5	1	8	1.3	[131]
10	14	36	.6	1.5	8.7	6.2	[1]
10	40	55	.35	3	9.6	1.8	[80]
10	20	43	.25	1.4	8	8.4	[22]
10	100	105	.35	3	9.3	1.7	[89]
10	50	65	.25	1.5	9.2	2.2	[52]
10	100	80	.18	1.8	9.2	1.4	[96]
10	30	16	.3	2	9.2	.9	[84]
10	150	1200	.6	3	9.2	13.6	[120]
10	120	234	.35	3.3	9	3.8	[61]
12	50	850	.6	3.3	10.3	13.5	[95]
12	65	480	.5	5	11.3	2.9	[114]
12	50	200	.35	3.3	9.8	4.5	[63]
12	10	338	.5	3.3	10.8	19	[57]
12	21	35	.6	3	10.8	.9	[72]
12	50	115	.18	1.8	11	1.1	[77]
13	40	800	.5	5	10.8	11.2	[21]
14	10	118	.35	3.3	11.5	4.1	[134]
14	20	720	.5	5	11.9	9.4	[19]
14	15	380	.5	5	12.2	5.4	[19]
14	10	200	.5	5	12.2	4.3	[19]
14	75	340	.35	3	11.8	1.3	[69]
15	10	320	.35	3	12.8	4.5	[48]

TAB. 5.4 – Exemples de réalisations CMOS (1998-2002)

supplémentaire en surface et en consommation qu'elles génèrent à un instant donné est très vite compensé par le relâchement des contraintes sur les parties analogiques.

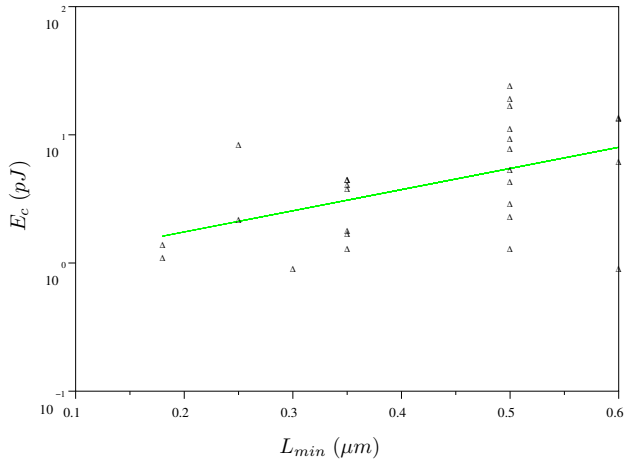


FIG. 5.15 – Evolution de l'énergie par conversion en fonction de la longueur du canal.

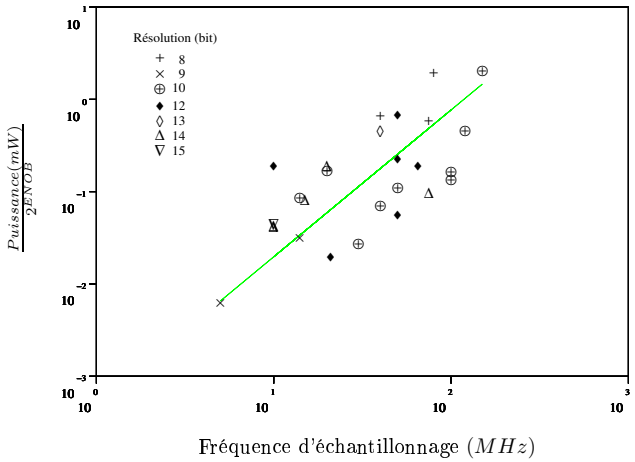


FIG. 5.16 – Puissance normalisée en fonction de la fréquence d'échantillonnage.

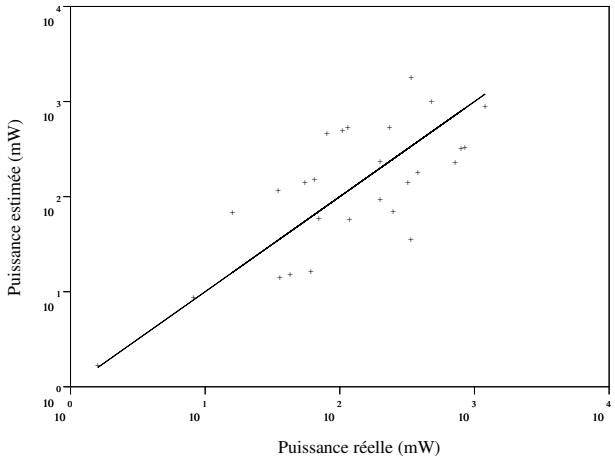


FIG. 5.17 – Puissance estimée en fonction de la Puissance mesurée.

Chapitre 6

Convertisseurs $\Sigma\Delta$

6.1 Introduction

La technique de conversion Σ Δ est une des principales méthodes envisagées pour l'intégration des systèmes mixtes. Ceci est principalement dû à sa grande robustesse et à son adaptabilité. Elle autorise en effet un échange entre résolution et rapidité de conversion et le changement de technologie est facilité par la réduction de la complexité des parties analogiques. Cette technique est basée sur le sur-échantillonnage du signal dont la bande d'intérêt B est bien inférieure à la fréquence d'échantillonnage F_s. Cette dernière étant limitée par la technologie, le taux de sur-échantillonnage $R = \frac{F_s}{2B}$ fixe la bande passante du convertisseur. Le signal sur-échantillonné en sortie du convertisseur nécessite un filtrage numérique et une réduction de la cadence d'échantillonnage (*décimation*) dans le rapport R. Bien que la technique soit connue dans les années 60 [116], il a fallu attendre le milieu des années 80 et la technologie VLSI pour que ce filtrage puisse être réalisé de manière économique. L'utilisation de ces convertisseurs dans le domaine des signaux audio est aujourd'hui très fréquente. On peut en effet, pour ces signaux, obtenir très facilement des taux de sur-échantillonnage très importants (>100) avec une résolution élevée (>16 bit) sans nécessiter de composants de grande précision. Un taux de sur-échantillonnage élevé présente également l'avantage de réduire les contraintes de filtrage avant conversion, la bande de transition de ce filtre étant directement liée à ce facteur. Ceci est particulièrement intéressant pour les applications de radiocommunications où un faible signal utile peut être accompagné de différents signaux indésirables et de forte amplitude (canaux radio adjacents, signaux de blocage, ...). Il est alors possible de combiner la sélection de canal et la décimation de manière souple et économique. L'extension de la technique $\Sigma\Delta$ aux applications large bande conduit à réduire le taux de sur-échantillonnage. Pour éviter que cette réduction s'accompagne d'une perte trop importante de résolution, on devra alors faire appel à des circuits plus complexes.

La conversion Σ Δ fait partie de la classe des quantificateurs à mémoire pour lesquels le codage d'un échantillon est fonction des échantillons précédents du signal. On distingue généralement deux types fondamentaux de quantificateur à mémoire : les codeurs prédictifs et les codeurs à minimisation de bruit. Le convertisseur Σ Δ fait partie de la seconde catégorie. Les codeurs prédictifs sont largement décrits dans la littérature en particulier pour leur intérêt en compression de l'information. Bien qu'ils aient été les premiers à être utilisés en conversion analogique-numérique, ceux-ci se révèlent cependant moins robustes que les convertisseurs Σ Δ pour cette application. Une version très simple de codeur prédictif, le modulateur Δ [60], va nous permettre

de mieux appréhender ce fait.

6.2 Codeur prédictif et modulation Δ

La modulation Δ est un cas particulier du codage prédictif dont le schéma de principe est donné à la figure 6.1(a). Le bloc P représente un filtre linéaire dont le

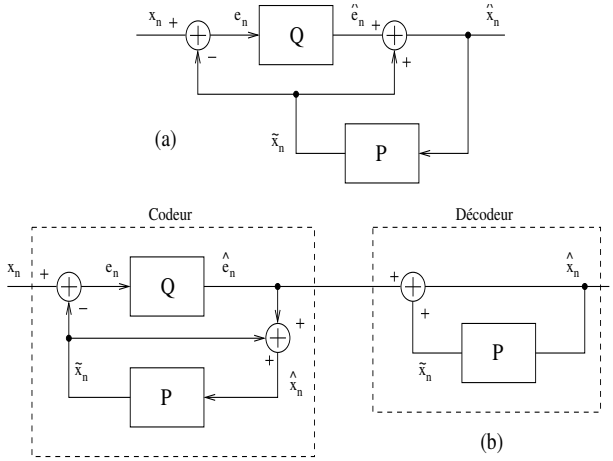


FIG. 6.1 – Schéma de principe d'un codeur prédictif.

but est de fournir une estimation du signal à convertir. Cette estimation est utilisée pour réduire la dynamique à l'entrée du quantificateur représenté par le bloc Q. Pour cela on effectue une soustraction du signal et de son estimé $\tilde{x}$ avant la quantification. L'estimation du signal est ensuite ajoutée au résultat de la quantification.

Le codeur et le décodeur obtenus à partir de ce schéma sont donnés à la figure 6.1(b). Le décodeur est constitué d'une réplique du bloc de prédiction P. Un cas particulier (*Differential Pulse Code Modulation*) est obtenu lorsque le filtre de prédiction est un simple retard : $P = \alpha z^{-1}$. L'erreur de prédiction est définie par $e_n = x_n - \tilde{x}_n$ et vaut dans ce cas $e_n = x_n - \alpha\tilde{x}_{n-1}$. L'erreur quadratique moyenne correspondante est :

$$E\{e_n^2\} = E\{(x_n - \alpha\tilde{x}_{n-1})^2\} = E\{x_n^2\} - 2\,\alpha E\{x_n\,\tilde{x}_{n-1}\} + \alpha^2\,E\{\tilde{x}_{n-1}^2\}$$

Si $\tilde{x}_n$ est une estimation non biaisée de x_n et en notant R_e et R_x les fonctions d'autocorrélation respectives de l'erreur de prédiction et du signal, la relation précédente devient :

$$R_e(0) = R_x(0) - 2\,\alpha R_x(1) + \alpha^2 R_x(0)$$

La minimisation de l'erreur de prédiction en fonction de α conduit à :

$$\frac{d}{d\alpha}R_e(0) = 0 \quad \Longrightarrow \quad \alpha = \frac{R_x(1)}{R_x(0)}, \quad R_e(0) = R_x(0)\,(1-\alpha^2)$$

Si la fréquence d'échantillonnage est suffisamment élevée par rapport à la bande du signal, les échantillons successifs sont fortement corrélés et α tend vers 1. Le gain de prédiction G_p défini comme $G_p = R_x(0)/R_e(0)$ est alors grand et la dynamique du signal à l'entrée du quantificateur est plus faible. Pour une même résolution globale, le

nombre de pas du quantificateur peut ainsi être réduit. En prenant $\alpha = 1$ et un quantificateur de 1 bit on obtient la modulation Δ qui conduit à une réalisation très simple. La partie linéaire (bloc P) correspond en effet dans ce cas à un simple intégrateur dans le circuit de retour sur le quantificateur. La figure 6.2 donne un exemple de réponse du modulateur Δ avec une entrée sinusoïdale. Il existe deux types d'erreurs dans le codeur

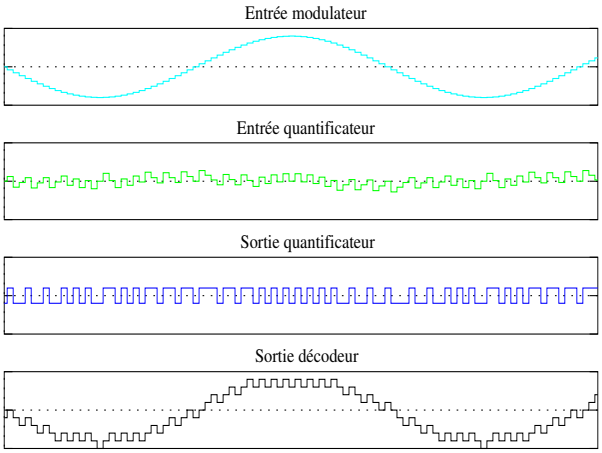


FIG. 6.2 – Modulateur Δ avec une entrée sinusoïdale.

Δ. Lorsque le signal varie lentement, la sortie du quantificateur passe alternativement de $+\Delta/2$ à $-\Delta/2$ pour un pas de quantification Δ. Ce bruit "granulaire" peut être réduit par une diminution de Δ. Cette réduction du pas de quantification est par contre limitée par les variations rapides du signal. La pente maximale du signal de sortie du codeur est en effet nécessairement inférieure à Δ/T_s où T_s est la période de l'horloge. Une solution à ce problème consiste à adapter le pas de quantification au signal mais ceci introduit une complexité supplémentaire. Un autre point important dans ce type de codeur est que le signal quantifié est également filtré et que les composantes basses fréquences atténuées par le système nécessitent un filtrage inverse dans le décodeur. Dans le cadre de la conversion analogique-numérique ce filtrage inverse, effectué dans le domaine numérique est nécessairement différent du filtrage subi par le signal dans le codeur qui sera quant à lui analogique. Pour palier à ce problème, Inose, Yasuda et Murakami [58] ont proposé de placer un intégrateur dans le chemin du signal avant le codeur Δ d'où le nom de codeur sigma-delta (Σ Δ) utilisé maintenant couramment pour décrire toute une classe de convertisseurs basés sur ce principe. Nous commençons par l'étude du plus simple d'entre eux : le modulateur Σ Δ du premier ordre.

6.3 Modulateur Σ Δ du premier ordre

Le modèle générique d'un quantificateur ΣΔ du premier ordre est donné à la figure 6.3. Par rapport au modulateur Δ, c'est une prédiction élémentaire (retard z^{-1}) de l'erreur de quantification qui est soustraite du signal d'entrée. On considère un quantificateur Q à deux niveaux :

$$q(u_n) = \begin{cases} +\Delta/2 & \text{si } u_n \geq 0 \\ -\Delta/2 & \text{si } u_n < 0 \end{cases}$$

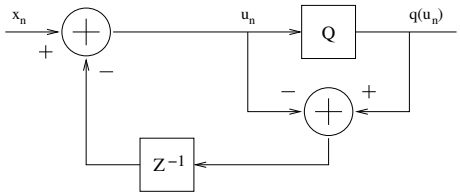


FIG. 6.3 – ΣΔ du premier ordre.

Le signal d'entrée est limité en amplitude à l'intérieur de l'unique pas de quantification : $|x_n| < \frac{\Delta}{2}$.

On a l'équation de récurrence :

$$u_n = x_n + u_{n-1} - q(u_{n-1}) \tag{6.1}$$

En notant $\epsilon_n = q(u_n) - u_n$ l'erreur de quantification locale on a :

$$q(u_n) = x_n + \epsilon_n - \epsilon_{n-1} \tag{6.2}$$

L'erreur de quantification locale subit un filtrage passe-haut du premier ordre à la sortie du modulateur. La relation (6.1) implique que l'état u_n du modulateur est tel que $|u_n| < \Delta$. Il en résulte que l'erreur de quantification ϵ_n est telle que $|\epsilon_n| < \frac{\Delta}{2}$.

Afin de comparer la résolution potentielle de cette structure élémentaire avec un quantificateur sans mémoire, nous envisageons dans un premier temps le cas d'une entrée constante et d'un décodeur qui effectue une simple moyenne de la sortie quantifiée.

6.3.1 Entrée constante et décodeur de type moyenne

Supposons que l'entrée x soit maintenue constante pendant N échantillons et que la sortie quantifiée soit constituée de la somme :

$$y_N = Q(x) = \frac{1}{N} \sum_{i=0}^{N-1} q(u_i) \tag{6.3}$$

Cette somme ne peut prendre que N+1 valeurs qui constituent l'alphabet de reproduction du quantificateur :

$$y_N = k\frac{\Delta}{N} - \frac{\Delta}{2} \qquad k = 0, 1, ..., N$$

En combinant les relations (6.2) et (6.3) on a :

$$Q(x) = x + \frac{\epsilon_N}{N} - \frac{\epsilon_0}{N}$$

L'erreur de quantification globale $e_q = Q(x) - x$ est telle que $|e_q| < \frac{\Delta}{N}$. Dans le cas où l'entrée est une variable aléatoire l'estimation de l'erreur quadratique moyenne $E\{e_q^2\}$ est beaucoup plus délicate. Du fait du nombre réduit de pas de quantification les hypothèses asymptotiques de bruit de quantification uniforme ne sont pas applicables. L'erreur locale de quantification ϵ_n est fortement corrélée au signal et l'alphabet de reproduction ne satisfait pas au critère d'optimalité d'un quantificateur linéaire (le

niveau de reproduction y_i est tel que $y_i = E\{X|X \in S_i\}$ où S_i est le segment associé au niveau y_i). En utilisant un modèle linéaire dans une arithmétique modulo Δ, Gray [46] a établit la borne supérieure suivante pour l'erreur quadratique moyenne :

$$E\{(Q(x) - x)^2\} \leq \frac{\Delta^2}{3N^2}$$

Cette borne peut être nettement réduite par l'utilisation d'un décodeur plus complexe qu'une simple moyenne sur N échantillons qui n'exploite pas tous les codes issus du modulateur. En particulier des décodeurs linéaires [51] où non-linéaire [53] pour le modulateur du premier ordre ont été développés et conduisent à une erreur en $O(N^{-3})$.

6.3.2 Réalisation à base d'un intégrateur

La figure 6.3 met en évidence le traitement effectué sur l'erreur de quantification $\epsilon = q(u) - u$. Pour la réalisation pratique du modulateur, on préfère cependant utiliser le schéma de la figure 6.4 qui fait appel à un intégrateur. A un retard près sur le signal, les équations de récurrences sont identiques et la fonction intégratrice admet une réalisation simple en traitement analogique. Sous cette forme, le convertisseur ΣΔ

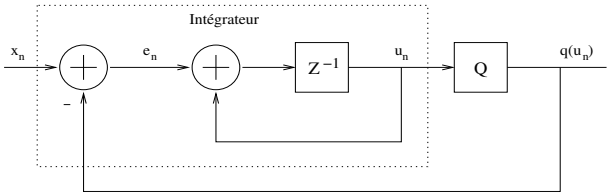


FIG. 6.4 – ΣΔ du premier ordre utilisant un intégrateur.

peut permettre d'atteindre de grandes résolutions sans le recours à des composants précis. Il s'avère également robuste quant aux imperfections du quantificateur. La résolution est cependant limitée par les pertes de l'intégrateur qui limitent son gain statique. La figure 6.5 montre la caractéristique entrée sortie du quantificateur pour un intégrateur imparfait dont la relation de récurrence est $u_{n+1} = \alpha\, u_n + e_n$. La sortie représente la moyenne sur N=1000 valeurs de sortie du quantificateur. On note une précision fortement dégradée par rapport au cas idéal avec en particulier une zone autour du zéro ou l'entrée est sans influence sur la sortie.

Le spectre de l'erreur de quantification locale ϵ_n pour une entrée constante est un spectre de raies dont l'amplitude a et la fréquence f sont donnés par [47] :

$$a = \frac{1}{(2\pi n)^2} \qquad f = < n(\frac{1}{2} + \frac{x}{\Delta}) > \qquad \text{n=1,2,...}$$

où <x> représente la partie fractionnaire de x.

L'architecture de la figure 6.3 est un cas particulier où l'opération linéaire effectuée sur le bruit de quantification est un simple retard et où la résolution du quantificateur est de 1 bit. L'utilisation d'un filtrage non récursif d'ordre M et d'un quantificateur linéaire de résolution N généralise l'exemple précédent au cas d'un modulateur Σ Δ d'ordre M [116].

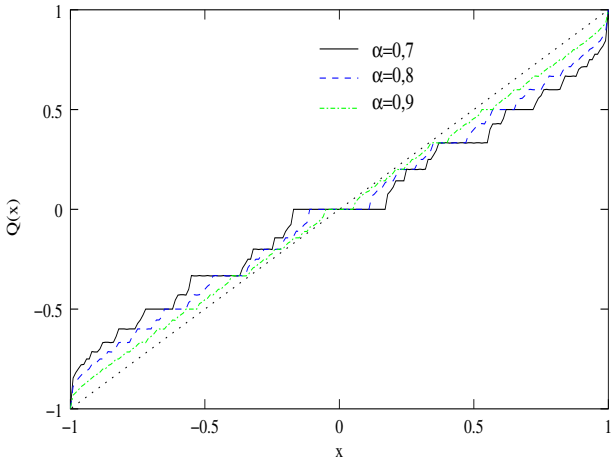


FIG. 6.5 – Effet des pertes dans l'intégrateur.

6.4 Architecture générale d'un modulateur Σ Δ

La forme générique d'un modulateur Σ Δ est donnée à la figure 6.6. Le traitement linéaire sur le bruit de quantification est effectué par le filtre $H(z)$.

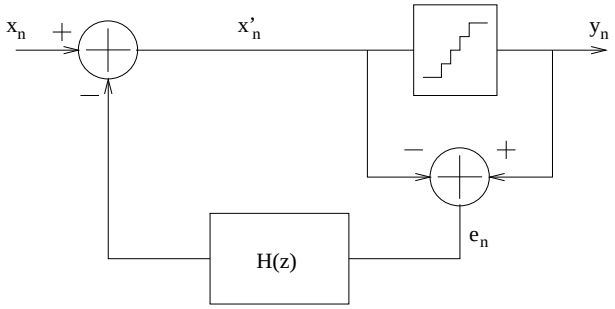


FIG. 6.6 – Forme générique d'un convertisseur ΣΔ.

On définit l'erreur $e_n^{'}$ sur le signal d'entrée et l'erreur de quantification e_n telles que (figure 6.6) :

$$y_n = x_n^{'} + e_n = x_n + e_n^{'}$$

en prenant la transformée en z, on a

$$X^{'}(z) = X(z) + E^{'}(z) - E(z) = X(z) - H(z)E(z)$$

d'où

$$E^{'}(z) = [1 - H(z)]E(z)$$

La densité spectrale de puissance de l'erreur de quantification $S_E(\omega)$ est mise en forme par le filtre :

$$S_{E'}(\omega) = \|1 - H(j\omega)\|^2 S_E(\omega)$$

On cherche à minimiser la puissance résultante avec une fenêtre de pondération $W(\omega)$:

$$P = \int_{-\pi}^{\pi} S_{E'}(\omega) W(\omega) d\omega$$

Si le filtre est non récursif et comporte au moins un retard, on peut écrire celui-ci sous la forme :

$$1 - H(z) = \sum_{n=0}^{M} a(n) z^{-n} \qquad avec \qquad a(0) = 1$$

La puissance de bruit en sortie devient :

$$P = \int_{-\pi}^{\pi} \| \sum_{n=0}^{M} a(n) e^{-j\omega n} \|^2 S_E(\omega) W(\omega) d\omega$$

Cette puissance est la même que celle qui résulterait du filtrage d'un signal $u(n)$ de densité spectrale de puissance $S_E(\omega)W(\omega)$ par un filtre de réponse impulsionnelle $a(k)$. D'après le théorème de Parseval on a :

$$P = \sum_{-\infty}^{\infty} (\sum_{k=0}^{M} a(k) u(n-k))^2$$

La minimisation de cette puissance entraîne :

$$\frac{\partial P}{\partial a(k)} = 0 \qquad k = 1 \dots M$$

Ceci conduit au système d'équations dites normales :

$$a = -R^{-1} v = (a(1), a(2), ..., a(M))^T$$

$$v = (r(1), r(2), ..., r(M))^T$$

$$R = \begin{bmatrix} r(0) & r(1) & \cdots & r(M-1) \\ r(1) & r(0) & & r(M-2) \\ \vdots & & \ddots & \vdots \\ r(M-1) & & \cdots & r(0) \end{bmatrix}$$

où $r(k)$ est la fonction d'autocorrélation associée au signal $u(n)$:

$$r(k) = \sum_{n=-\infty}^{\infty} u(n)\, u(n-k)$$

Si le bruit de quantification est supposé blanc, les coefficients du filtre sont entièrement contrôlés par la fenêtre de pondération spectrale $W(\omega)$. La fonction d'autocorrélation étant obtenue dans ce cas par transformée de Fourier inverse :

$$r(k) = \int_{-\frac{1}{2}}^{\frac{1}{2}} W(f) e^{j2\pi fk} df \tag{6.4}$$

Cette valeur optimale des coefficients nécessite une connaissance du spectre du signal pour distribuer le bruit de manière efficace. Dans le domaine audiofréquence par

exemple un modèle psychoacoustique peut permettre de rendre le bruit de quantification le moins audible possible. Les paramètres de ce modèle évoluant au cours du temps il est nécessaire de mettre à jour les coefficients du filtre en fonction de cette évolution.

Pour des applications de conversion analogique-numérique rapide la mise en oeuvre de cette technique est délicate étant donné le nombre important de calculs qu'elle nécessite et la complexité matérielle introduite au niveau du modulateur. Lorsque le sur-échantillonnage du signal est suffisant on peut choisir une fenêtre de pondération égale à 1 sur une bande B très inférieure à la demi-fréquence d'échantillonnage du signal. Dans ce cas, un développement limité de la fonction d'autocorrélation conduit au filtre optimal suivant [122] :

$$\sum_{n=0}^{M} a(n) z^{-n} = (1 - z^{-1})^M \tag{6.5}$$

En supposant un bruit de quantification blanc de puissance $P_E = S_E F_s$ on peut calculer la puissance de bruit en sortie du modulateur :

$$P = \frac{P_E}{F_s} \int_{-B}^{B} \| (1 - z^{-1} \|^{2\,M}_{z = e^{2j\pi \frac{f}{F_s}}} df \approx \frac{P_E}{\pi (2\,M + 1)} \left(\frac{\pi}{R} \right)^{2\,M+1} \tag{6.6}$$

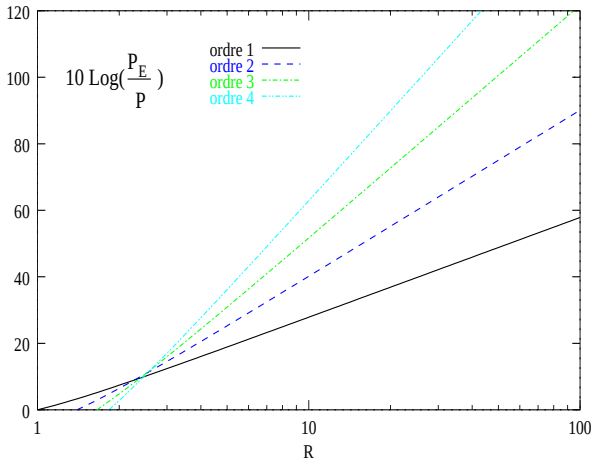


FIG. 6.7 – Accroissement du rapport signal sur bruit en fonction de R.

Le rapport entre la puissance de bruit du quantificateur et celle obtenue en sortie du modulateur dans une bande B est reporté à la figure 6.7 en fonction de $R = \frac{F_s}{2\,B}$ pour des ordres de modulateur allant de 1 à 4. Il correspond au gain en rapport signal sur bruit dans la bande utile du signal. D'après la formule 6.6 le gain en rapport signal sur bruit lorsque R est doublé est de $(6\,M + 3)\,dB$. On note l'intérêt de combiner l'accroissement de l'ordre du filtre et du facteur de suréchantillonnage.

Si le rapport R n'est pas très supérieur à l'unité, le fait de placer tous les zéros du filtre de mise en forme du bruit à l'origine (formule 6.5) ne constitue plus une solution optimale. On peut, dans ce cas distribuer les zéros à l'intérieur de la bande passante et obtenir des performances supérieures à celles indiquées à la figure 6.7 [110]. Cette solution a cependant l'inconvénient de rendre le filtre spécifique à une application

donnée et rend sa réalisation plus complexe. Nous nous limiterons dans la suite au cas simple du filtrage donné par la formule 6.5.

La forme générique précédente du modulateur ΣΔ utilise un filtrage non récursif et peut être utilisée directement pour une implantation numérique où en conversion numérique-analogique. Pour la conversion analogique-numérique, la réalisation récursive du filtre analogique conduit à la forme générale de la figure 6.8.

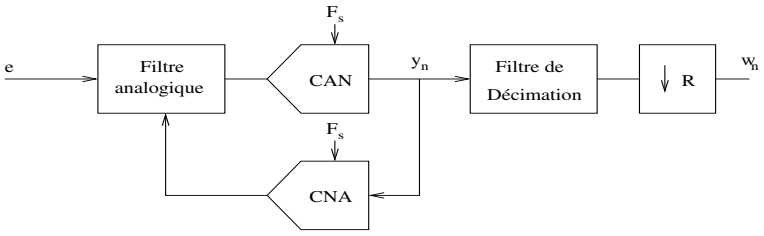


FIG. 6.8 – Modulateur ΣΔ.

Le filtre linéaire reçoit le signal d'entrée et le signal quantifié par l'intermédiaire du couple CAN-CNA. Le but de ce filtre est de rejeter le bruit de quantification hors de la bande du signal utile. Ce dernier doit par ailleurs subir le minimum de perturbations lors du traitement dans le modulateur. Le filtrage numérique effectué après le modulateur a pour but de limiter le spectre du bruit de quantification afin de réduire la cadence d'échantillonnage dans le rapport R. Cette opération est en général décomposée en plusieurs étapes pour en réduire la complexité.

La description mathématique précédente supposait implicitement une représentation discrète dans le temps des signaux qui est appropriée au traitement échantillonné ou numérique. Le principe de fonctionnement précédent reste cependant valable même pour une représentation continue du signal dans le filtre analogique de la figure 6.8. Etant donné les limitations pratiques sur la fréquence d'échantillonnage, l'utilisation d'un filtrage continu peut permettre de traiter des signaux de fréquences plus élevées [20]. Ainsi, si le filtre est constitué de résonateurs, il est possible d'appliquer le principe précédent à des signaux à bande étroite, une utilisation importante d'un tel modulateur passe-bande étant la numérisation en fréquence intermédiaire dans un récepteur radio. L'étude qui suit sera limitée au modulateur passe-bas échantillonné. L'extension des résultats, en terme d'architecture, au traitement continu peut être faite en utilisant des transformations appropriées [2]. les problèmes d'implantation tels que le retard et la forme de l'impulsion de retour du CNA sont toutefois spécifiques de ce traitement.

L'architecture de la figure 6.8 fait apparaître une boucle autour d'un élément non linéaire constitué par le quantificateur. Pour un ordre élevé du filtre, ce système peut être le siège d'instabilités qui vont réduire son efficacité voir le rendre inopérant. Ce problème est d'autant plus crucial que le nombre de pas de quantification est réduit. En particulier, pour un quantificateur à deux niveaux (largement utilisé dans ce type d'architecture pour sa linéarité parfaite) les contraintes de stabilité conduisent à une fonction de filtrage qui ne sera pas optimum du point de vue de la mise en forme du bruit. De ce fait, les performances en terme de rapport signal sur bruit seront en général moins bonnes que celles de la figure 6.7.

Le bruit de quantification que nous avons considéré supposait implicitement que le CNA était parfait. Cette hypothèse n'est plus justifiée pour un quantificateur à plus de deux niveaux qui va introduire des non-linéarités du fait des dispersions des

composants. Comme nous allons le voir, les erreurs liées au CNA constituent une limite importante en termes de linéarité du convertisseur ΣΔ.

6.4.1 Erreurs liées au CNA

L'utilisation d'un quantificateur linéaire à plus de deux niveaux permet d'augmenter la résolution du modulateur mais on doit dans ce cas prendre en compte les erreurs introduites par la dispersion du pas de quantification. La figure 6.9 fait apparaître les erreurs liées aux conversions A/N et N/A. Le filtrage linéaire est représenté par les fonctions $H_1(z)$ et $H_2(z)$. On peut exprimer la sortie Y du modulateur en fonction de

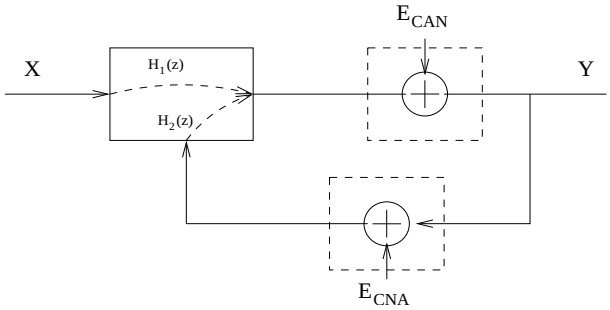


FIG. 6.9 – Erreur introduite par le CNA.

l'entrée X et de ces erreurs :

$$Y = \frac{E_{CAN}}{1 - H_2(z)} + \frac{H_2(z)}{1 - H_2(z)} E_{CNA} + \frac{H_1(z)}{1 - H_2(z)} X$$

Le filtrage passe-haut du bruit de quantification implique que H_2 soit très grand dans la bande passante du signal. Il en résulte que le deuxième terme de cette expression est pratiquement égal au signe près à l'erreur introduite par le CNA. Cette erreur affecte donc directement la sortie du modulateur contrairement à l'erreur introduite par le CAN qui subit le filtrage passe-haut. Les exigences de linéarité sur la conversion N/A sont donc identiques à celles demandées pour le convertisseur complet. Il existe différentes méthodes de correction pour réduire les effets liés aux non linéarités du CNA :

- Calibrage
- Quantification avec différentes résolutions : Leslie-Singh, cascade
- Linéarisation par sélection d'éléments unitaires (*Dynamic Element Matching*).

Dans le cas du calibrage, les erreurs sur les différents niveaux du CNA sont mesurées et stockées en mémoire. Elles sont ensuite compensées en fonctionnement normal du convertisseur dans le domaine analogique ou numérique. L'architecture de Leslie-Singh [74] utilise le bit de plus fort poids du CAN pour la quantification du signal de retour. Un filtrage numérique est ensuite réalisé pour éliminer l'effet de cette quantification sur un bit et ne conserver que l'erreur provenant du CAN, mise en forme par le modulateur. Dans l'architecture cascade qui sera étudiée à la section 6.6, on profite du fait que le bruit de quantification du dernier étage est mis en forme par les modulateurs qui le précèdent pour y effectuer une quantification sur plusieurs bit. Celle-ci est théoriquement la seul erreur visible après filtrage numérique des sorties

de la cascade. Enfin, la linéarisation par sélection d'éléments unitaires effectue un tirage pseudo-aléatoire des éléments d'un CNA. Celui-ci est réalisé avec N éléments de poids identiques ($N = 2^n$ pour un CNA de n bit). La sortie correspondant à un code K consiste à choisir K éléments parmi N et il y a C_K^N manières de faire ce choix. Si l'erreur sur le poids de chaque élément peut être considérée comme une variable aléatoire indépendante et centrée, on peut exploiter ce degré de liberté dans le choix pour sélectionner les éléments de manière telle que, en moyenne, l'erreur introduite soit proche de zéro. Cette erreur est ainsi déplacée vers les fréquences élevées, en dehors de la bande utile du signal. Un exemple simple de sélection correspond à prendre les éléments de manière cyclique. On utilise un pointeur qui est incrémenté par le code. Avec cette méthode (Data Weighted Averaging [7]), les éléments sont sélectionnés à partir de la position du pointeur modulo le nombre d'éléments. Ceci est illustré sur

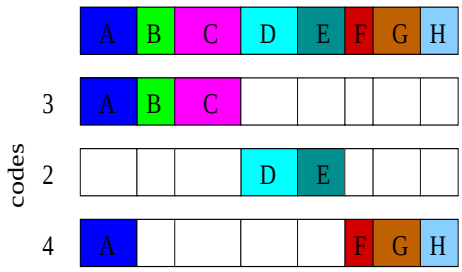


FIG. 6.10 – Sélection cyclique des éléments d'un CNA.

la figure 6.10 avec 8 éléments (ceux-ci sont représentés avec une taille inégale pour matérialiser le fait que les poids sont différents). La figure 6.11 donne le résultat de l'application de ce traitement sur un quantificateur 6 bit dont les 64 éléments du CNA sont affectés d'une erreur suivant une loi normale avec un écart type $\sigma = \frac{\Delta}{4}$. Le signal d'entrée est sinusoïdale avec une amplitude proche de la pleine échelle. La courbe (a) correspond à une sélection systématique des k premiers éléments pour un code k en entrée du CNA. Le spectre de raies lié à la distorsion introduite est dans ce cas, concentré vers les basses fréquences. Pour un rapport de suréchantillonnage R égal à 8 un grand nombre de ces raies sont dans la bande B du signal et le rapport signal sur bruit est égal à 36,6dB. La courbe (b) correspond à une sélection cyclique (DWA) des éléments du CNA. Les harmoniques sont déplacées vers les fréquences élevées et le bruit dans la bande B est nettement réduit conduisant à un rapport signal sur bruit de 45,4 dB. On peut montrer [90] que cette méthode de sélection procure une mise en forme au premier ordre en $(1 - z^{-1})$ du bruit.

D'autres techniques que la sélection cyclique précédente ont été développées pour limiter l'effet des non linéarités dans le CNA. Elles visent en général à augmenter l'ordre de la mise en forme du bruit au prix d'une complexité matérielle accrue. On trouvera une comparaison de ces différentes méthodes de correction dans la référence [42].

6.4.2 Critères de performance des modulateurs

Les spécifications importantes sont liées à la dynamique des signaux qui peuvent être traités par le modulateur. Le maximum est déterminé par le niveau du signal d'entrée qui préserve le fonctionnement correct de celui-ci (c'est à dire préserve sa stabilité) et qui garantit un taux de distorsion donné. La caractéristique la plus simple à utiliser est le rapport signal sur bruit plus distorsion ($SINAD$). En particulier l'évolution de

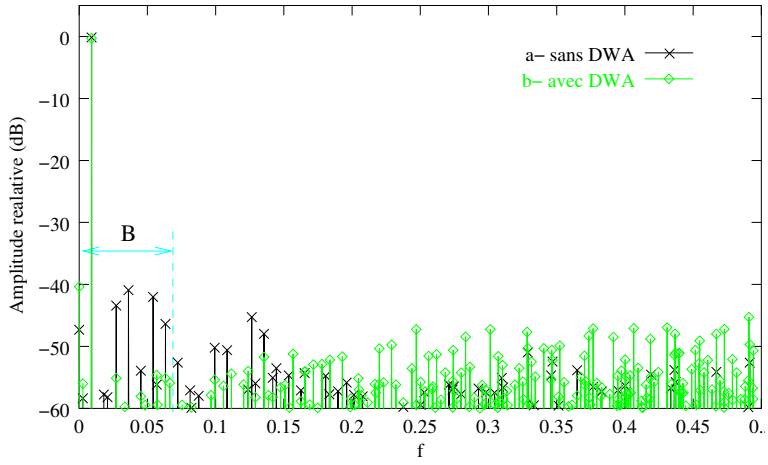


FIG. 6.11 – Spectre en sortie du CNA avec (a) et sans (b) DWA.

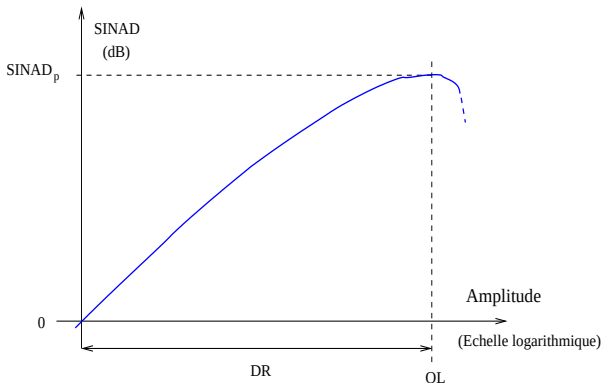


FIG. 6.12 – Paramètres dynamiques d'un modulateur.

ce dernier en fonction de l'amplitude du signal d'entrée permet de déterminer la valeur maximum de ce rapport que l'on notera $SINAD_p$ et l'amplitude OL du signal d'entrée pour laquelle ce maximum est atteint. Le minimum du signal est généralement pris égal au bruit total dans le modulateur. La plage d'entrée comprise entre ce minimum et l'amplitude OL correspondant au maximum du $SINAD$ est appelée dynamique du modulateur (DR : *Dynamic Range*). Ces caractéristiques sont résumées sur la figure 6.12. Elles sont en fait générales et peuvent être utilisés pour d'autres convertisseurs. Une particularité du modulateur ΣΔ est cependant liée à l'amplitude d'entrée maximale OL. Celle-ci peut en effet être nettement inférieure à la tension de référence du quantificateur pour garantir la stabilité.

6.5 Architecture à un seul quantificateur

L'avantage de la mise en forme du bruit donnée par la formule 6.5 est de conduire à une réalisation simple du filtre à partir d'intégrateurs.

La figure 6.13 montre une réalisation du filtre avec M intégrateurs disposés en

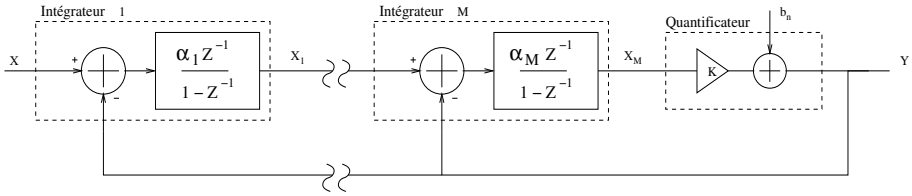


FIG. 6.13 – Modulateur d'ordre M à base d'intégrateur.

cascade. Le quantificateur est représenté par son modèle linéaire où b_n représente le bruit de quantification. En notant $I_i = \alpha_i\, z^{-1}/(1-z^{-1})$ la fonction de transfert de l'intégrateur idéal et $STF(z)$ et $NTF(z)$ les fonctions de transfert respectives du signal et du bruit de quantification :

$$Y(z) = STF(z)\, X(z) + NTF(z)\, B(z) \tag{6.7}$$

on a, à partir de la figure 6.13 :

$$STF(z) = \frac{K \prod_{i=1}^{M} I_i(z)}{1 + K \sum_{i=1}^{M} \prod_{j=i}^{M} I_j(z)} \qquad NTF(z) = \frac{1}{1 + K \sum_{i=1}^{M} \prod_{j=i}^{M} I_j(z)} \tag{6.8}$$

Si le rapport de sur-échantillonnage est suffisamment élevé la bande du signal est concentrée à l'origine et on a approximativement pour les expressions précédentes :

$$STF \approx 1 \qquad NTF \approx \frac{(1-z^{-1})^M}{K \prod_{i=1}^{M} \alpha_i} \tag{6.9}$$

On obtient bien la mise en forme souhaitée à un coefficient près qui fait intervenir le produit $K \prod_{i=1}^{M} \alpha_i$ des gains des intégrateurs et du gain équivalent du quantificateur. Dans le cas d'un quantificateur à deux niveaux (comparateur) le gain α_M du dernier intégrateur est indifférent car seul le signe de x_M intervient dans le fonctionnement du modulateur. Pour minimiser le bruit de quantification, les gains précédents le dernier intégrateur doivent donc être maximisés sous la contrainte de préserver la stabilité du modulateur. Pour un ordre élevé et une faible résolution du quantificateur cette contrainte limite fortement la valeur du maximum de ces gains et les performances de la figure 6.7 ne pourrons pas être atteintes.

6.5.1 Modèle linéaire et stabilité

On considère que le modulateur est stable quand l'évolution des états $x_1, .., x_M$ reste bornée pour une classe particulière des signaux d'entrée. Du fait de la non linéarité introduite par le quantificateur, l'étude dynamique directe du modulateur $\Sigma\Delta$ est complexe et on a souvent recours à des modèles analytiques simplifiés [3, 6] ou numériques [111].

En particulier, l'approximation qui consiste à remplacer le quantificateur par un gain linéaire K et un bruit blanc additif et indépendant du signal, est très utilisée car elle permet d'évaluer avec une bonne précision le bruit de quantification. Il a récemment été montré [137] que ce modèle linéaire pouvait également être utilisé pour

étudier la stabilité d'un modulateur dans le cas d'un quantificateur à deux états ($\pm\frac{\Delta}{2}$). On a en effet dans ce cas une relation très simple en exprimant de deux manières différentes la puissance du signal en sortie du modulateur avec une excitation sinusoïdale d'amplitude A :

$$P_{tot} = \frac{\Delta^2}{4} = \frac{A^2}{2} + \frac{\Delta^2}{12} \int_{-\frac{1}{2}}^{\frac{1}{2}} \|NTF(K,z)\|^2_{z=e^{2j\pi f}} df$$

où on a rappelé que la fonction de transfert de bruit était dépendante du gain K du quantificateur. Celle-ci doit satisfaire la relation :

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} \|NTF(K,z)\|^2_{z=e^{2j\pi f}} df = P_o = 3\,(1-2\,a^2) \tag{6.10}$$

où $a = \frac{A}{\Delta}$ est l'amplitude réduite que l'on suppose restreinte à l'intervalle $[0, \frac{1}{2}]$. La valeur P_o est donc comprise entre 1,5 et 3. Pour une valeur fixée de P_o dans cet intervalle, il doit exister un gain K du quantificateur et une fonction de transfert de bruit stable qui satisfait (6.10). Sur la base d'un grand nombre de simulations, les auteurs [137] stipulent que dans ce cas, le modulateur est stable.

6.5.2 Modulateur du premier ordre

La configuration du modulateur du premier ordre est celle de la figure 6.4 étudiée en 6.3.2. Nous avons déjà établi en 6.3, que le seul état x_1 du modulateur avec un quantificateur à deux niveaux $\pm\frac{\Delta}{2}$ était inférieur à Δ si l'entrée est inférieure à $\frac{\Delta}{2}$. Le gain α_1 est arbitraire puisque dans ce cas $Q(\alpha_1\, x_1) = Q(x_1)$. En choisissant $\alpha_1 = \frac{1}{2}$ on peut donc contraindre l'état x_1 à la même dynamique que le signal d'entrée.

La stabilité, comme définie précédemment, est donc assurée pour le modulateur du premier ordre. Utilisée seule, cette structure s'avère cependant peu efficace pour le codage de signaux dynamiques avec un quantificateur de faible résolution. Ceci est dû au fait que le bruit de quantification est dans ce cas fortement corrélé au signal et essentiellement composé de raies. L'étude pour une excitation sinusoïdale et un quantificateur à deux niveaux, montre que l'amplitude et la fréquence de ces raies dépend de celle du signal d'entrée d'une manière complexe[47]. La figure 6.14 donne un exemple de spectre obtenu pour une entrée sinusoïdale de $0,1\,\Delta$ ainsi que la densité spectrale qui résulterait d'un bruit de quantification blanc. Dans la bande passante du modulateur, la présence de ces raies dégrade fortement le rapport signal sur bruit et la dynamique des signaux qui peuvent être traités. Le modulateur du premier ordre est donc peu approprié pour une réalisation qui n'utilise qu'un seul quantificateur de faible résolution. Celui-ci pourra cependant être utilisé dans une chaîne de modulateurs qui sera étudiée plus loin. L'entrée d'un étage est alors constituée du bruit de quantification de l'étage précédent qui est suffisamment décorrélé du signal à convertir.

6.5.3 Modulateur d'ordre supérieur ou égal à deux

Au début de la section 6.5, nous avons considéré un gain identique pour les deux entrées d'un bloc intégrateur. Cette hypothèse minimise le nombre de coefficients pour le modulateur, ce qui en simplifie la réalisation. La valeur optimum de ces coefficients peut être obtenue à partir de simulations comportementales de ces architectures. La référence [42] fournit les tableaux de coefficients ainsi obtenus pour différentes résolutions des quantificateurs et des modulateurs d'ordre inférieur à 4. Dans le cas d'une

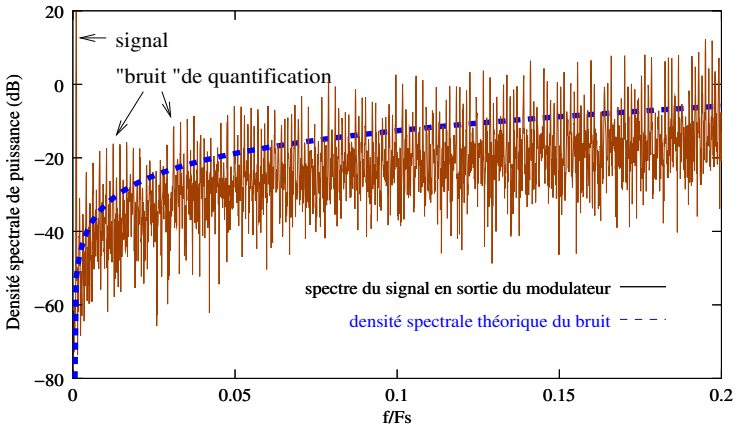


FIG. 6.14 – Spectre de sortie du modulateur du premier ordre.

faible résolution, et en particulier lorsque le quantificateur a seulement deux niveaux, il peut être difficile de contrôler l'amplitude des sorties d'intégrateurs tout en recherchant un optimum du rapport signal sur bruit. Prenons comme exemple le modulateur du second ordre avec un bit de résolution et un rapport de suréchantillonnage R de 16. Le choix classique consiste à prendre des coefficients α_1 et α_2 identiques égaux à $\frac{1}{2}$ pour les deux intégrateurs. On a alors un optimum du rapport signal sur bruit $SINAD_p = 41dB$ pour une amplitude $OL = 0,7\,V_{ref}$. A cette amplitude le SINAD théorique ($NTF = (1 - z^{-1})^2$) est de 46 dB et le gain moyen du quantificateur (voir chapitre 1, équation 1.26) est de 2,1. L'écart par rapport à la valeur théorique est complètement justifié par la formule 6.9. L'histogramme des sorties d'intégrateurs à cet optimum est donné à la figure 6.15. L'amplitude de l'état du premier intégrateur

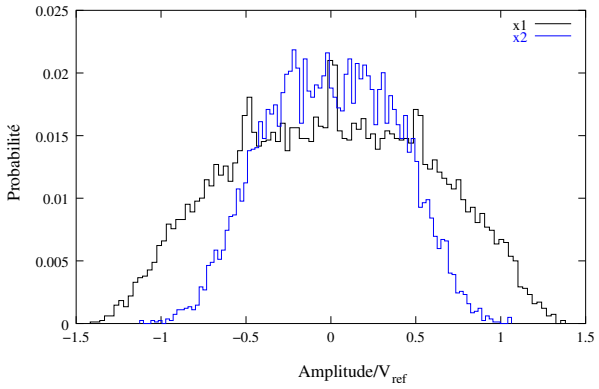


FIG. 6.15 – Histogramme des états du modulateur d'ordre 2 avec $\alpha_1 = \alpha_2 = 0,5$.

est environ 1,5 fois plus grande que la tension de référence. Ceci représente une limite importante pour une réalisation dans les technologies actuelles. Comme on le verra dans le chapitre 7, la réduction des tensions d'alimentation impose des contraintes fortes sur l'amplitude des sorties d'amplificateurs. Nous décrirons dans le chapitre 9 une méthode de synthèse des coefficients qui permet de limiter l'amplitude des états tout en conservant un SINAD proche de l'optimum. Ceci est obtenu au prix d'un ac-

croissement du nombre de coefficients du modulateur qui sont alors différents pour les deux entrées d'un intégrateurs. Le tableau 6.1 donne les coefficients obtenus pour le modulateur d'ordre 2 et 3 avec $R = 16$ [1].

ordre	α_1	β_1	α_2	β_2	α_3	β_3
2	2/7	2/7	3/4	1/2		
3	1/5	1/5	1/3	1/3	2/3	4/9

TAB. 6.1 – Coefficients des intégrateurs avec quantificateur 1 bit

Pour l'ordre 2 et avec cette valeur des coefficients, l'optimum du rapport signal sur bruit est de 40,6 dB, toujours à une amplitude $OL = 0,7\,V_{ref}$. L'avantage de ce choix réside dans l'excursion limitée des états des intégrateurs dont l'histogramme est donné à la figure 6.16.

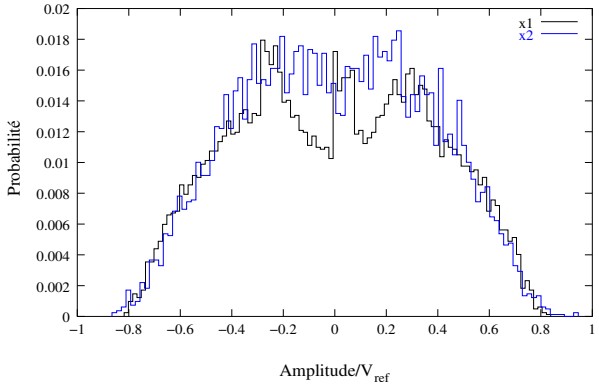


FIG. 6.16 – Histogramme des états du modulateur d'ordre 2 avec les coefficients du tableau 6.1.

Pour le modulateur du troisième ordre, le rapport signal sur bruit optimum est de 41,7 dB à une amplitude $OL = 0,52\,V_{ref}$. Le maintient de la stabilité entraîne une réduction de l'amplitude maximum. De plus, l'optimum du rapport signal sur bruit est à peine plus élevé qu'avec l'ordre 2. Ceci est dû à la réduction des gains des intégrateurs qui, d'après la formule 6.9 s'accompagne d'une augmentation du gain sur la fonction de transfert du bruit. De ce fait, pour un faible rapport de suréchantillonnage, le gain théorique apporté par l'augmentation de l'ordre du modulateur ne se vérifie pas en pratique. Ceci apparaît clairement dans le tableau 6.2 qui résume les caractéristiques obtenues pour les modulateurs un bit d'ordre 2 et 3.

$SINAD_p$ (dB)	R					OL
	8	16	32	64	128	
ordre 2	25,2	41,6	54,6	70,7	86	0,7
ordre 3	20,8	41,7	62,4	83,7	104,7	0,5

TAB. 6.2 – Caractéristiques des modulateurs un bit.

Bien qu'il soit possible d'exploiter des modulateurs 1 bit d'ordre supérieure à trois, le maintien de la stabilité impose une réduction importante de l'amplitude d'entrée.

[1] On se reportera à l'annexe C pour la définition des coefficients

De plus le contrôle des amplitudes en sorties des intégrateurs devient délicat et les performances dynamiques sont bien inférieures aux valeurs théoriques [42]. On préférera dans ce cas utiliser une structure cascade qui permet de s'affranchir du problème de stabilité et qui fait l'objet de la section suivante.

Lorsque l'on augmente le nombre de niveaux du quantificateur, le gain qui lui est associé est très proche de l'unité pour une large plage du signal à son entrée. Le système est alors bien défini par son modèle linéaire avec source de bruit de quantification additive. La stabilité est beaucoup plus facile à garantir et le niveau d'entrée peut être proche de la tension de référence. Le tableau 6.3 donne les performances obtenues pour des modulateurs du deuxième et du troisième ordre avec un quantificateur 4 bit.

$SINAD_p$ (dB)	R					OL
	8	16	32	64	128	
ordre 2	52,5	67,6	82,6	97,7	112,7	0,98
ordre 3	58,9	80,1	101,2	122,3	143,3	0,92

TAB. 6.3 – Caractéristiques des modulateurs avec un quantificateur 4 bit.

Etant donné la nature quasi-linéaire du modulateur, la synthèse peut être faite à partir d'un modèle de filtre pour la mise en forme du bruit. Les coefficients sont obtenus par identifications des fonctions de transfert C.6 et C.7 ($\alpha_i = \beta_i$) avec un prototype de filtre Butterworth passe-haut. Ceux-ci sont reportés dans le tableau 6.4.

ordre	α_1	β_1	α_2	β_2	α_3	β_3
2	3/7	3/7	4/3	4/3		
3	2/7	2/7	3/4	3/4	2	2

TAB. 6.4 – Coefficients des intégrateurs avec quantificateur 4 bit

L'amplitude maximum utilisable est très proche de la tension de référence. Cette méthode exploite également bien la dynamique des différentes sorties d'intégrateurs comme on peut le voir sur la figure 6.17 qui donne l'histogramme des états du modulateur du troisième ordre à l'optimum du rapport signal sur bruit.

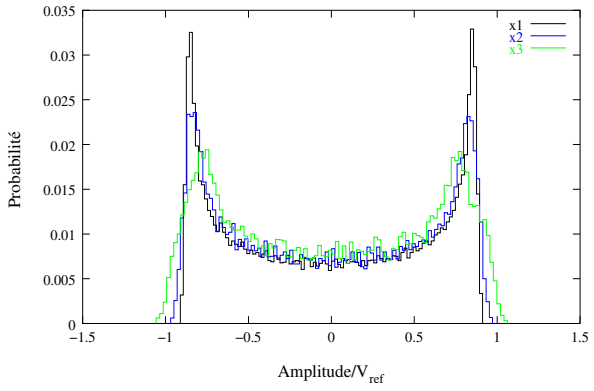


FIG. 6.17 – Histogramme des états du modulateur d'ordre 3 avec un quantificateur 4 bit.

L'utilisation d'un quantificateur multi-bit présente le double avantage d'améliorer la stabilité du modulateur et d'effectuer une mise en forme plus efficace du bruit de

quantification. La contrepartie, comme il a été noté en 6.4.1, est que les imperfections du CNA, et en particulier sa linéarité, affectent directement les performances du modulateur.

6.6 Architecture cascade

6.6.1 Principe

La structure cascade permet de résoudre les problèmes liés à l'instabilité d'un modulateur d'ordre élevé. Elle utilise K modulateurs d'ordre M_i (en général $M_i \leq 2$, $i = 1, .., K$) associés à K quantificateurs. La figure 6.18 donne le schéma de principe de cette structure. Le bruit de quantification de chaque modulateur noté E_i

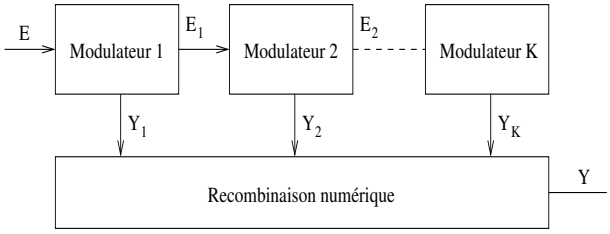


FIG. 6.18 – Principe de la structure cascade.

sur la figure 6.18 est transmis à l'étage suivant. Les sorties numériques Y_i des différents quantificateurs sont recombinées dans un bloc numérique qui réalise l'opération $Y(z) = \sum_{i=1}^{K} H_i(z)\, Y_i(z)$. Les filtres H_i sont liés aux modulateurs utilisés et idéalement cette recombinaison élimine tous les bruits de quantification intermédiaires pour ne conserver que celui (à un coefficient près) du dernier modulateur. Si tous les quantificateurs ont la même résolution, cette réalisation est équivalente à un modulateur d'ordre $M = \sum_{i=1}^{K} M_i$. La stabilité individuelle des modulateurs assure celle de la structure complète. Si le modulateur du premier ordre ne peut pas être utilisé pour le premier étage pour les raisons invoquées au 6.5.2, il constitue cependant un bon choix pour les étages suivant. Celui-ci est en effet inconditionnellement stable et sa dynamique peut être ajustée très simplement. La limitation inhérente à cette structure est liée à la dispersion des gains des intégrateurs. L'élimination des bruits de quantification des étages intermédiaires nécessite une adaptation entre les coefficients du bloc de recombinaison et les coefficients des modulateurs. Du fait des dispersions, ces relations qui font l'objet de la section suivante, ne seront jamais satisfaites parfaitement.

6.6.2 Relations de couplage dans la cascade

La linéarisation du quantificateur permet d'écrire les sorties y_i et s_i de chaque modulateur qui correspondent respectivement à la sortie du modulateur et à la différence entre l'entrée et la sortie du quantificateur (voir figure 6.19). Celui-ci est remplacé par sont modèle linéaire associé de gain K_i et de bruit de quantification n_i :

$$y_i = T_i^s\, s_{i-1} + T_i^n\, n_i \qquad s_i = A_i^n\, n_i + A_i T_i^s\, s_{i-1} \tag{6.11}$$

$$avec \quad A_i = \frac{a_i}{K_i} - b_i \quad et \quad A_i^n = A_i T_i^n - \frac{a_i}{K_i}$$

Les fonctions T_i^s et T_i^n sont déduites du choix du filtre associé au modulateur élémentaire. Celles-ci sont données à l'annexe C pour les modulateurs d'ordre 1 à 4. Le signal

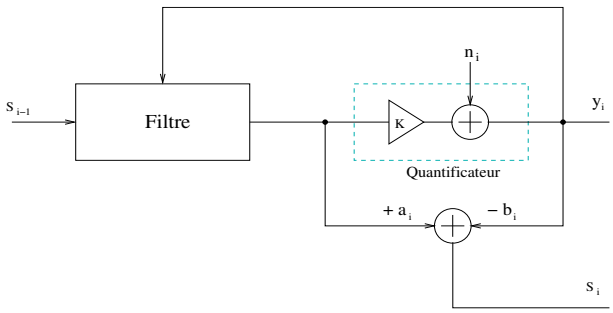


FIG. 6.19 – Couplage dans une cascade de modulateurs

d'entrée e est égal à s_0 et la sortie est obtenue par combinaison linéaire des sorties y_i. Pour m modulateurs on a :

$$y = \sum_{i=1}^{m} h_i y_i = t^s\, e + \sum_{i=1}^{m} t_i^n\, n_i \tag{6.12}$$

On a par exemple pour m=4 :

$$\begin{aligned} t^s &= T_1^s(h_1 + A_1 T_2^s(h_2 + A_2 T_3^s(h_3 + A_3 T_4^s h_4))) \\ t_1^n &= h_1 T_1^n + A_1^n T_2^s(h_2 + A_2 T_3^s(h_3 + A_3 T_4^s h_4)) \\ t_2^n &= h_2 T_2^n + A_2^n T_3^s(h_3 + A_3 T_4^s h_4) \\ t_3^n &= h_3 T_3^n + A_3^n T_4^s h_4 \\ t_4^n &= h_4 T_4^n \end{aligned} \tag{6.13}$$

La recombinaison des différentes sorties doit permettre d'annuler tous les bruits de quantification intermédiaires $\{n_{1,2,3}\}$. Ceci est possible avec :

$$A_1 = A_2 = A_3 = 0 \tag{6.14}$$

$$h_1 T_1^n = h_2 T_2^s b_1 \qquad h_2 T_2^n = h_3 T_3^s b_2 \qquad h_3 T_3^n = h_4 T_4^s b_3 \tag{6.15}$$

A titre d'exemple d'application de ces relations nous utilisons la cascade 2-1-1 constituée d'un modulateur du second ordre suivi de deux modulateur du premier ordre dont le schéma est donné à la figure 6.20.

Exemple : Cascade 2-1-1

A partir des fonctions de transfert de l'annexe C, les relations 6.13 deviennent dans ce cas particulier :

$$\begin{aligned} t^s &\approx \tfrac{k_1\alpha_1\alpha_2 z^{-2}}{D_{10}}(h_1 + A_1\tfrac{k_2 z^{-1}}{D_{20}}(h_2 + A_2\tfrac{k_3 z^{-1}}{D_{30}}h_3)) \approx z^{-2}h_1 \\ t_1^n &\approx h_1\tfrac{(1-z^{-1})^2}{D_{10}} - \tfrac{k_2\alpha_{32}z^{-1}}{D_{20}}(h_2 + A_2\tfrac{k_3 z^{-1}}{D_{30}}h_3) \\ &\approx h_1\tfrac{(1-z^{-1})^2}{D_{10}} - \tfrac{k_2\alpha_{32}z^{-1}}{D_{20}}h_2 \\ t_2^n &\approx h_2\tfrac{(1-z^{-1})}{D_{20}} - \tfrac{k_3\alpha_{42}z^{-1}}{D_{30}}h_3 \\ t_3^n &= h_3\tfrac{(1-z^{-1})}{D_{30}} \end{aligned} \tag{6.16}$$

Dans ces expressions on a noté D_{10}, D_{20}, D_{30} la valeur des dénominateurs des différents modulateurs. Celle-ci sera approximée par la valeur prise pour z=1 pour conduire

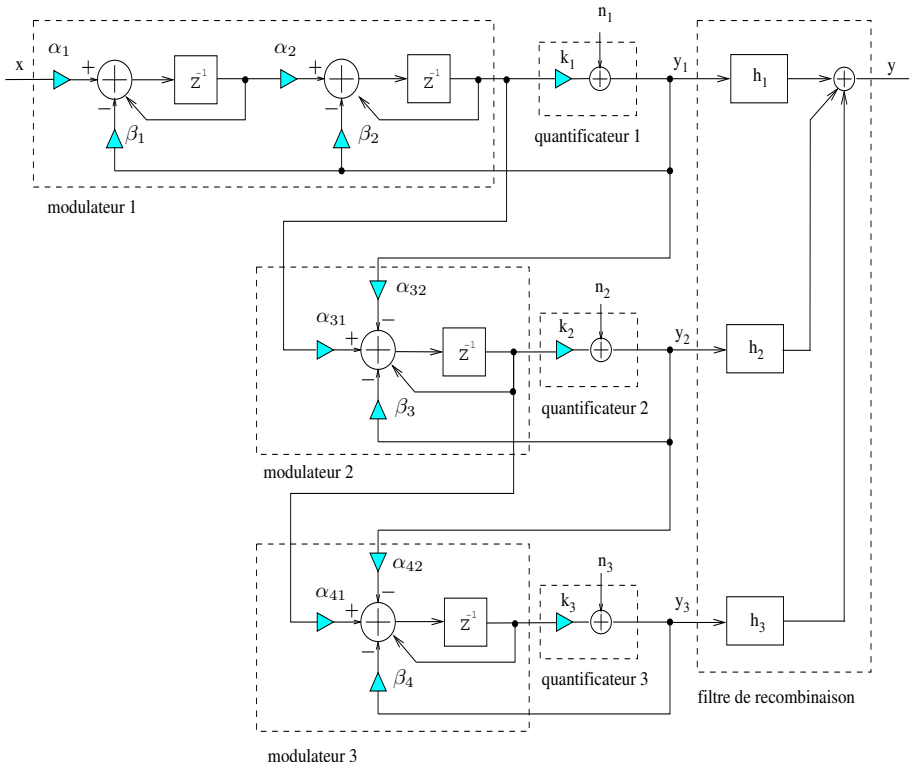


FIG. 6.20 – Cascade 2-1-1

à des fonctions non-récursives simples pour les filtres h_1, h_2, h_3 :

$$\begin{aligned} h_1 &\approx z^{-2} \\ h_2 &\approx \frac{\beta_3}{k_1\alpha_1\alpha_2\alpha_{32}} z^{-1}(1-z^{-1})^2 = d_2\, z^{-1}(1-z^{-1})^2 \\ h_3 &\approx \frac{\beta_4}{k_1 k_2\alpha_1\alpha_2\alpha_{32}\alpha_{42}}(1-z^{-1})^3 = d_3\,(1-z^{-1})^3 \end{aligned} \tag{6.17}$$

La minimisation du bruit de quantification dans le dernier modulateur conduit aux coefficients du tableau 6.5. L'optimum du rapport signal sur bruit avec les filtres

α_1	β_1	α_2	β_2	α_{31}	α_{32}	β_3	α_{41}	α_{42}	β_4
2/7	2/7	3/4	1/2	1	2/5	2/5	1	2/5	2/5

TAB. 6.5 – Coefficients des intégrateurs

précédents et un rapport de sur-échantillonnage de 16 est obtenu pour $d_2 = 1,87$ et $d_3 = 1,84$. Sa valeur est de 64,3 dB à une amplitude d'entrée de 0,7 V_{ref}. La figure 6.21 représente la contribution dans la puissance de bruit en sortie des différents quantificateurs de la cascade ainsi que la puissance de bruit totale. On remarque que la puissance de bruit du premier quantificateur n'a pas été suffisamment éliminée par les filtres de recombinaison et qu'elle représente une part importante de la puissance de bruit dans la bande passante du signal. Ceci est du à la présence du coefficient A_2

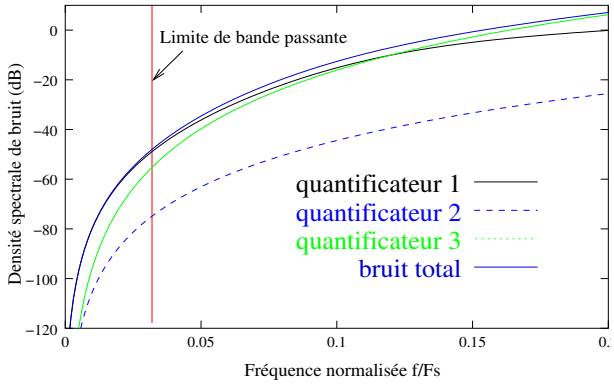


FIG. 6.21 – Contribution des différentes sources de bruit de quantification

dans l'équation 6.16. Du fait de la variation du gain du quantificateur, celui-ci ne peut pas être annulé parfaitement. La prise en compte de ce coefficient conduit à la forme suivante pour h_1 :

$$h1 = z^{-2}\left[1 + d1\,(1-z^{-1})\right] \tag{6.18}$$

L'optimum du rapport signal sur bruit est obtenu pour $d_1 = 0,35$, $d_2 = 1,89$ et $d_3 = 1,83$. Sa valeur est de 71,4 dB à une amplitude d'entrée de 0,7 V_{ref}. Le bruit de quantification est dans ce cas essentiellement celui du dernier quantificateur comme on peut le voir sur la figure 6.22 qui représente la nouvelle contribution du bruit des quantificateurs. Cet exemple montre que le filtre de recombinaison doit être parfaitement adapté à la structure du modulateur. Son efficacité est cependant dépendante de la précision avec laquelle sont réalisés les coefficients. En effet, dans les relations 6.16 de l'exemple précédent, toute variation de ces coefficients va se traduire par une annulation imparfaite des sources de bruit des premiers étages. Cette particularité de

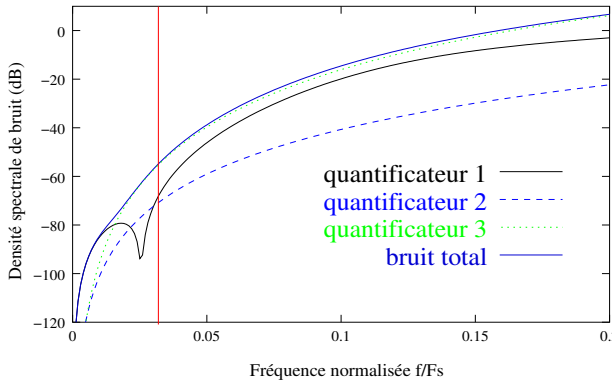


FIG. 6.22 – Nouvelle contribution des bruits de quantification

la structure cascade la rend plus sensible aux variations des coefficients et aux défauts des intégrateurs comparativement à l'utilisation d'un seul quantificateur (voir le chapitre 9).

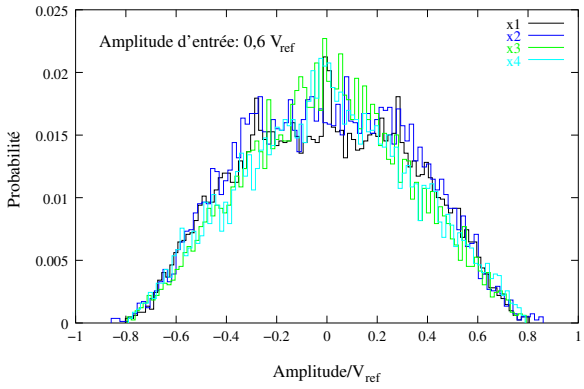


FIG. 6.23 – Histogramme des états pour la cascade 2-1-1

La cascade 2-1-1 constitue un exemple particulièrement intéressant quant aux choix possibles de l'ordre des modulateurs. Nous avons vu à la section 6.5.2 que le choix d'un premier ordre était inadapté pour le premier étage. Celui-ci est par contre bien indiqué pour les étages suivants par sa stabilité inconditionnelle et la possibilité d'exploiter toute l'amplitude permise pour l'évolution des états d'intégrateurs. D'autre part, le choix d'un ordre supérieur à deux pour le premier étage entraîne une limitation sur le niveau maximum d'entrée (cf 6.5.3) et une plus mauvaise utilisation de la dynamique. La figure 6.23 qui représente l'histogramme des états de la cascade 2-1-1 illustre cette utilisation efficace de l'excursion des sorties d'intégrateurs.

6.6.3 Structures cascades du troisième et quatrième ordre

Nous avons vu dans la section précédente que les considérations pratiques sur l'excursion des états et sur la stabilité du modulateur restreint les choix possibles sur l'ordre des filtres utilisés dans la cascade. Le tableau 6.6 donne les caractéristiques obtenues pour les cascades 2-1 et 2-1-1 qui constituent des choix intéressants de ce point de vue. Une comparaison avec le tableau 6.2 montre que l'ordre 3 cascade a des

$SINAD_p$ (dB)	R					OL
	8	16	32	64	128	
cascade 2-1	35	56,1	77	97	115,6	0,75
cascade 2-1-1	43,9	71,4	93,5	108,3	123	0,7

TAB. 6.6 – Caractéristiques des modulateurs cascade 1 bit.

performances environ 14 dB supérieures à la structure utilisant un seul quantificateur. De plus l'amplitude d'entrée est 50% plus élevée. Dans le cas d'un quantificateur 1 bit, la cascade est donc une solution très avantageuse pour augmenter l'ordre du filtrage sans être pénalisé par les contraintes de stabilité du modulateur.

6.6.4 Utilisation de plusieurs niveaux de quantification

Le principe de fonctionnement de la cascade n'est pas limité quant à la résolution des quantificateur utilisés. Le modèle linéaire étudié à la section 6.6.2 est même encore mieux justifié lorsque cette résolution augmente. Cependant, le principe étant basé sur l'élimination du bruit des premiers étages pour ne conserver que celui du dernier, il est particulièrement intéressant d'introduire une quantification à plus de deux niveaux dans le dernier étage. Cette solution présente également un autre avantage. Nous avons vu à la section 6.4.1 que les imperfections du CNA sont équivalentes à une source d'erreur à l'entrée du modulateur (ici le dernier étage). Dans la cascade, cette source d'erreur subit le même traitement que le bruit de quantification du modulateur précédent, c'est à dire un filtrage passe-haut. En reprenant l'exemple de la cascade 2-1-1 le filtrage qui serait ainsi obtenu est donné par le filtre h_2 de l'expression 6.17. Le bruit lié au CNA est ainsi rejeté hors de la bande utile du signal sans qu'il soit nécessaire d'appliquer un traitement particulier tel que la sélection cyclique (DWA) vue en 6.4.1. Le gain apporté par une quantification fine sur le dernier étage est cependant limité par le fait que le bruit des premiers étages ne peut pas être parfaitement annulé par le filtre de recombinaison. Reprenons l'exemple de la cascade 2-1-1 avec une quantification de 4 bit sur le dernier étage dont les coefficients des intégrateurs sont donnés dans le tableau 6.7. Les gains du dernier étage ont été modifiés pour prendre

α_1	β_1	α_2	β_2	α_{31}	α_{32}	β_3	α_{41}	α_{42}	β_4
2/7	2/7	3/4	1/2	1	2/5	2/5	2	4/5	4/5

TAB. 6.7 – Coefficients des intégrateurs

en compte la variation du gain effectif du quantificateur liée à la quantification sur 4 bit. Le $SINAD_p$ est dans ce cas de 81,4 dB pour un suréchantillonnage R de 16, soit un gain de seulement 10 dB alors que le gain attendu par l'accroissement de résolution est de $6\,(n-1) = 18\,dB$. Cette perte de résolution est liée au fait que le bruit de quantification lié au premier étage et représenté sur la figure 6.22 est maintenant le bruit dominant qui limite les performances de l'ensemble de la cascade. Une solution à ce problème est de réduire également le bruit de quantification des premiers étages en introduisant une quantification plus fine sur tous les modulateurs de la cascade [42]. L'inconvénient de cette solution, comme dans le cas d'un seul quantificateur, est que le bruit associé aux imperfections du CNA ne subit plus le filtrage numérique passe-haut et influence directement les performances de l'ensemble du convertisseur.

6.7 Choix des paramètres d'un modulateur

Il existe trois degrés de libertés essentiels dans le choix d'une architecture de modulateur qui sont :
- Le rapport R de suréchantillonnage.
- L'ordre M du filtrage associé à la mise en forme du bruit de quantification.
- Le nombre de niveaux de quantification (ou le nombre n de bit).

Nous considérons successivement ces différents paramètres dans l'ordre précédent qui représente un niveau de complexité croissant de la réalisation.

Rapport de suréchantillonnage Augmenter la valeur de R constitue la solution la plus simple pour augmenter la résolution. Le gain théorique (formule 6.6) est de $(6\,M + 3)\,dB$ pour chaque octave supplémentaire de R. Cette solution permet également une grande souplesse d'utilisation, un seul modulateur pouvant réaliser différents couples (résolution, bande-passante). La limite est imposée par la fréquence maximum d'échantillonnage liée au temps d'établissement fini des amplificateurs et à la bande passante désirée. La recherche de bandes passantes plus importantes a conduit à une réduction progressive de ce paramètre, des valeurs de R de 16 voire de 8 sont courantes aujourd'hui dans les modulateurs large bande. Pour ces valeurs, l'efficacité de la mise en forme du bruit est réduite. D'autre part, le gain réduit des intégrateurs rend la structure moins robuste vis à vis des imperfections technologiques (le gain d'un intégrateur de la forme $\frac{1}{1-z^{-1}}$ est de l'ordre de 2,5 en limite de bande pour R=8).

Ordre du filtrage Le rapport R étant fixé par la technologie, il est possible d'augmenter la résolution en accroissant l'ordre M du filtrage. Le gain est théoriquement donné par la formule 6.6 et la figure 6.7. En pratique, nous avons vu que pour une quantification un bit et une faible valeur de R, le gain apporté par l'ordre du filtrage est bien inférieur à cette valeur théorique. Ceci est dû au choix restreint des coefficients pour maintenir la stabilité. La structure cascade permet de résoudre ce problème au prix d'une sensibilité accrue aux imperfections des intégrateurs et à la variation du gain effectif des quantificateurs.

Résolution des quantificateurs Une autre manière d'accroître l'efficacité du filtrage est d'augmenter la résolution des quantificateurs. La stabilité est alors beaucoup plus facile à garantir et les coefficients des intégrateurs peuvent être choisis pour effectuer une mise en forme efficace du bruit. Cette réduction du pas de quantification procure également un gain de $20\,\log_{10}(2^n - 1)\,dB$ pour une résolution de n bit. Mise à part la complexité accrue des quantificateurs, le gain apporté par une résolution supérieure à un bit est cependant assujetti aux imperfections de la conversion numérique-analogique (cf 6.4.1). Une solution efficace est alors d'utiliser une technique de sélection (DEM) des éléments du CNA procurant une réduction du bruit qui lui est associé.

6.8 Convertisseurs ΣΔ CMOS récents

Le tableau 6.8 donne quelques exemples de caractéristiques de convertisseurs ΣΔ CMOS large bande publiées durant ces 5 dernières années. Dans ce tableau, F_{data} représente la fréquence d'échantillonnage finale après décimation, celle du modulateur étant R fois plus élevée. Le rapport de suréchantillonnage R varie de 8 à 96, la moyenne étant proche de 24. La résolution effective (ENOB) moyenne est de 12,5 bit. La puissance indiquée correspond à la puissance consommée par le modulateur seul, la puissance

totale incluant le filtre de décimation étant rarement fournie. Lorsqu'elle est indiquée, cette dernière est approximativement 25% plus élevée mais la puissance associée au filtre de décimation tend à diminuer progressivement avec l'évolution technologique.

La figure 6.24 représente l'évolution de l'énergie par conversion E_c (cf 4.15) en fonction de la longueur de canal à partir des données du tableau 6.8. Une régression linéaire sur ces données conduit à une réduction similaire au convertisseurs flash et pipeline d'un facteur 5 environ, la valeur actuelle étant comprise entre 1 et 10 pJ.

On constate sur la figure 6.25 que la plupart des réalisations sont dédiées à une bande de 1 à 2 MHz avec une résolution de 12 à 15 bit. Ce domaine étant principalement lié aux applications de type modems pour les transmissions avec ou sans fil. La réalisation [41] (point (a) sur la figure 6.25) marque une évolution importante de la bande passante accessible par le modulateur ΣΔ. Ces performances sont obtenues avec une architecture à un seul quantificateur de 4 bit et un filtrage du troisième ordre. La fréquence d'échantillonnage du modulateur est de 100MHz. Pour réduire les contraintes de linéarités du CNA une selection cyclique (DWA) est utilisée.

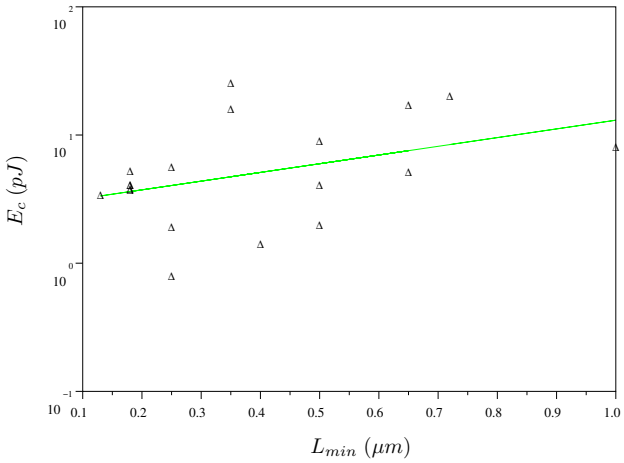


FIG. 6.24 – Evolution de l'énergie par conversion en fonction de la longueur du canal.

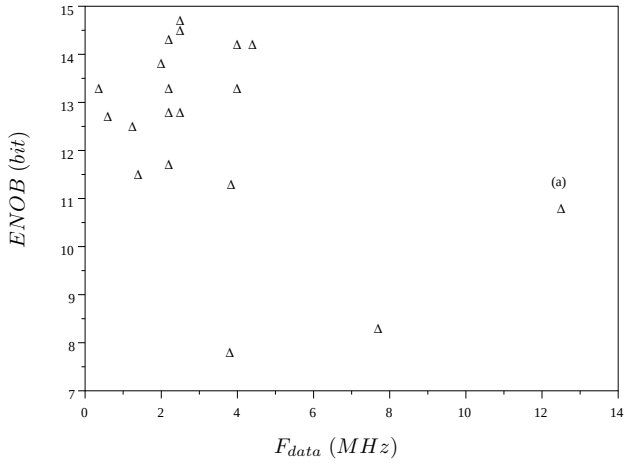


FIG. 6.25 – Résolution effective en fonction de la fréquence d'échantillonnage après décimation.

F_{data} (Ms/s)	R	P (mW)	L_{min} (μm)	V_{dd} (V)	$SNDR$ (dB)	$ENOB$ (bit)	E_c (pJ)	Ref
2	24	230	1	5	85	13,8	8,1	[76]
1,4	16	81	0,72	3,3	71	11,5	20	[34]
2,5	24	295	0,65	5	89	14,5	5,1	[40]
12,5	8	380	0,65	5	67	10,8	17,1	[41]
2,2	24	200	0,5	3,3	82	13,3	9	[39]
2,5	8	270	0,5	5	90	14,7	4,1	[36]
4	16	150	0,5	2,5	87	14,2	2	[128]
0,36	36	5	0,4	1,8	82	13,3	1,4	[91]
2,2	24	187	0,35	3,3	72	11,7	25,5	[86]
2,2	24	248	0,35	3,3	79	12,8	15,8	[86]
7,7	24	13,5	0,25	2,5	52	8,3	5,6	[15]
2,5	32	33	0,25	2,5	79	12,8	1,9	[106]
4,4	16	66	0,25	2,5	87	14,2	0,8	[30]
0,6	96	15	0,18	1,8	78	12,7	3,8	[37]
4	8	150	0,18	1,8	82	13,3	3,7	[64]
2,2	29	180	0,18	1,8	88	14,3	4,1	[49]
3,84	12	50	0,18	2,7	70	11,3	5,2	[82]
1,25	18,4	30	0,18	2,7	77	12,5	4,1	[82]
3,8	12	2,9	0,13	1,5	49	7,8	3,4	[45]

TAB. 6.8 – Convertisseurs ΣΔ CMOS récents

Chapitre 7

Contraintes technologiques

7.1 Introduction

La réduction des dimensions des transistors s'accompagne de nombreux effets bénéfiques. Un avantage évident est la possibilité d'intégrer des systèmes de plus en plus complexes sur une même puce (*System On Chip*), ce qui permet la fabrication de systèmes de taille de plus en plus réduite et à faible coût. La réduction de la longueur du canal a également permis d'accroître la vitesse de traitement. Ceci peut être mis en évidence par le calcul de la fréquence de transition pour laquelle le gain en courant du transistor est égal à l'unité [73]. On a, à partir du modèle quadratique du courant drain $I_{ds} = \frac{\mu C_{ox}}{2}\frac{W}{L}(V_{gs} - V_T)^2$, la valeur suivante pour cette pulsation limite :

$$\omega_t = \frac{g_m}{C_{gs} + C_{gd}} \approx \frac{\mu C_{ox}\frac{W}{L}(V_{gs} - V_T)}{C_{ox}\,W\,L} = \mu\,\frac{V_{gs} - V_T}{L^2} \tag{7.1}$$

Ce qui met clairement en évidence le rôle prépondérant de la longueur du canal (cette tendance doit toutefois être nuancée par les effets de canal court qui conduisent à une transconductance g_m plus faible et à l'importance grandissante des capacités de connexion).

La réduction des dimensions implique cependant de fortes contraintes sur la conception des blocs analogiques. Les deux conséquences majeures étant la *réduction de l'excursion des signaux* et la *dispersion des valeurs des composants*. La première résulte de la réduction de la tension d'alimentation. Cette réduction est nécessaire pour garantir la fiabilité du composant par le maintien d'un champ électrique maximum. Elle est également très bénéfique à la consommation des circuits numériques dont la composante dynamique varie de manière quadratique avec la tension d'alimentation. La dispersion sur les valeurs des composants résulte quant à elle des imprécisions géométriques qui sont plus importantes pour les petits dispositifs. Ces deux effets sont particulièrement préjudiciables à certains types de convertisseurs. Prenons par exemple le cas d'un flash : la réduction des dimensions implique une dispersion plus grande des seuils de décisions et une pleine échelle réduite, ce qui entraîne une perte de résolution. La technologie VLSI favorise la vitesse au détriment de la résolution et ceci explique le succès des techniques utilisant le suréchantillonnage du signal et des quantificateurs de faibles résolution.

Les circuits à capacités commutées constituent un choix privilégié pour le traitement analogique des signaux. Cette technique présente de nombreux avantages :

– capacité à être intégrée dans des technologies CMOS standards

- constantes de temps précises sans réglage
- bonne linéarité

Ce chapitre est consacré à l'étude des conséquences que représente l'évolution de la technologie sur les CAN étudiés dans les chapitres précédents et en particulier sur les composants des circuits à capacités commutées qui sont aujourd'hui fréquemment utilisés pour leur conception.

7.2 Réduction de la tension d'alimentation

Le tableau 7.1, extrait de la référence [14], donne quelques exemples de générations technologiques avec la longueur de canal L_{min}, la tension d'alimentation V_{dd}, l'épaisseur d'oxyde t_{ox}, la tension de seuil V_T et le paramètre de dispersion de la tension de seuil A_{VT}.

L_{min} (μm)	V_{dd} (V)	t_{ox} (nm)	V_T (V)	A_{VT} $(mV.\mu m)$
3	5	70	1,5	35
2,5	5	60	1,2	30
2	5	40	1,1	25
1,5	5	25	1	22
1,2	5	25	1	21
1	5	25	0,95	20
0,8	5	20	0,85	13
0,5	3,3	13,5	0,73	11
0,35	3,3	10	0,59	9
0,25	2,5	6	0,52	6
0,18	1,8	5	0,42	4,2
0,12	1,2	4,2	0,32	3,8
0,1	1,2	3,6	0,31	3,2
0,07	0,9	3	0,3	2,5

TAB. 7.1 – Exemples de technologies [14].

Ce tableau fait apparaître la réduction rapide de la tension d'alimentation pour les dimensions submicroniques après un grand palier à la tension $V_{dd} = 5\,V$. Cette tendance apparue dans le milieu des années 90 résulte des contraintes de fiabilité et de dissipation de puissance occasionnée par la forte densité d'intégration des circuits numériques. Afin de déterminer l'impact de cette évolution sur les circuits analogiques, nous nous appuierons sur les dernières données du tableau 7.1 pour lesquelles l'effet de la réduction des dimensions est particulièrement important.

Impact sur le convertisseur flash A partir de la formule 4.12, on peut déterminer la résolution maximale d'un flash connaissant la surface A occupée par les transistors d'entrée d'un amplificateur, la pleine échelle FSR et la dispersion A_{VT} de la tension de seuil. En considèrant une surface arbitraire $A_o = 50\mu m^2$ pour la longueur de grille $L_o = 0,8\mu m$ du tableau 7.1, une pleine échelle égale à 0,9 V_{dd} et en appliquant un facteur d'échelle $\alpha = \left(\frac{L}{L_o}\right)^2$ pour les technologies suivantes on obtient les résolutions de la ligne (a) du tableau 7.2.

	$L_{min}(\mu m)$	0,8	0,5	0,35	0,25	0,18	0,12	0,1	0,07
(a)	$n_{max}(bit)$	7,9	6,8	6,6	6,3	5,9	4,9	4,8	4,3
(b)	$C_e(fF)$	6,4	15,5	14	18,1	20,5	45	37,2	48,5

TAB. 7.2 – Impact de la technologie sur les performances d'un flash

Ceci montre bien qu'une réduction des dimensions s'accompagne d'une perte de résolution. Le remède utilisé fréquemment consiste à conserver une surface suffisante pour garantir la résolution attendue. L'inconvénient de cette solution est de limiter les performances en vitesse ou en consommation. Pour s'en convaincre, considérons cette fois un convertisseur flash de 6 bit de résolution avec les technologies submicroniques du tableau 7.1. La capacité d'entrée d'un amplificateur peut être obtenue par $C_e = \frac{\epsilon_{ox} \cdot A}{t_{ox}}$, la surface A étant calculée par l'expression 4.12. Cette capacité est reportée en fonction de la longueur de canal à la ligne (b) du tableau 7.2. On note un accroissement notable de la capacité d'entrée pour les faibles longueurs de canal. Cette évolution est défavorable pour la consommation et le comportement dynamique du convertisseur flash.

Contraintes faible tension et $\Sigma\Delta$ La réduction de la tension d'alimentation impose de fortes contraintes sur la conception d'un modulateur $\Sigma\Delta$, l'excursion en sortie de chaque intégrateur étant de plus en plus réduite. Dans un modulateur mono-bit une excursion des intégrateurs de 1,5 fois la tension de référence est fréquemment utilisée pour éviter une dégradation importante du rapport signal sur bruit. Pour des tensions d'alimentation de l'ordre du volt la tension de référence permise devient alors très faible. Nous verrons dans le chapitre 9 que l'on peut déterminer efficacement les coefficients d'un modulateur pour garantir une excursion minimale des sorties d'intégrateurs tout en conservant de bonnes performances dynamiques.

7.3 Circuits à capacités commutées

7.3.1 Commutateur CMOS

L'efficacité du commutateur est liée à la résistance du canal qui doit être suffisamment faible pour que la charge des capacités soit quasi-complète. D'autre part sa taille doit être minimum pour limiter le phénomène d'injection de charges. Celui-ci étant lié au fait qu'à la coupure du canal, une partie de la charge qui y était stockée est accumulée sur la capacité d'échantillonnage. La conductance du canal d'un transistor NMOS dans le mode linéaire étant donnée par :

$$g_{on} = \frac{1}{r_{on}} = \frac{\mu\, C_{ox}\, W\, (V_{gs} - V_T)}{L} \tag{7.2}$$

ces deux exigences conduisent à maintenir une tension de contrôle $V_{gs} - V_T$ suffisante. Ceci n'est possible, pour un transistor NMOS, que dans la partie basse de la tension d'alimentation. Pour garantir une conduction sur toute la tension d'alimentation, la solution traditionnelle consiste à utiliser un transistor PMOS en parallèle avec une commande complémentaire. La conductance globale est alors la somme des conductances des deux transistors comme indiqué à la fugure 7.1.

On remarque qu'il existe une valeur minimum de la tension d'alimentation pour qu'un tel dispositif soit fonctionnel. En effet, pour $V_{dd} = V_{TN} + |V_{TP}|$, la conductance

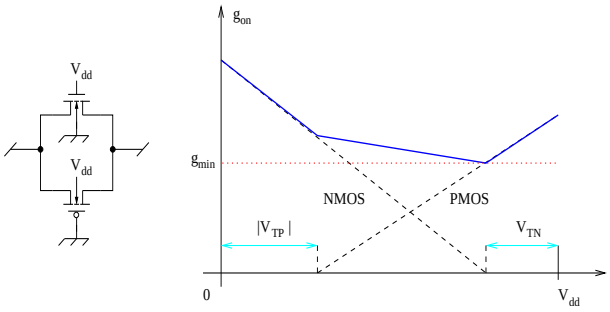


FIG. 7.1 – Conductance d'un commutateur CMOS.

s'annule au voisinage du centre de l'alimentation où la densité de probabilité du signal est souvent maximale dans un dispositif analogique. Le rapport entre la tension d'alimentation et la tension de seuil est donc très important pour garantir une dynamique maximum du signal. En considérant les tensions de seuil identiques pour les deux types de transistors, on obtient une valeur minimum de deux pour ce rapport. La figure 7.2 montre sont évolution avec la technologie à partir des données du tableau 7.1 pour les faibles longueurs de canal.

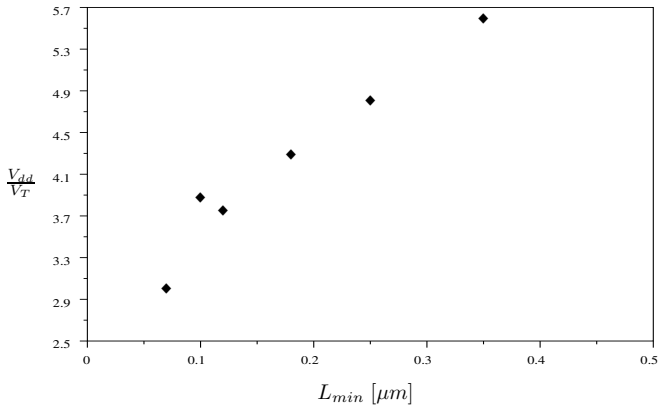


FIG. 7.2 – Evolution du rapport entre V_{dd} et V_T en fonction de L_{min}.

On remarque une nette diminution de ce rapport pour les faibles longueurs de canal qui est très défavorable à l'utilisation du commutateur CMOS. Cette situation est la conséquence d'une contrainte liée à la conduction des transistors sous le seuil. Celle-ci accroît la consommation statique des circuits numériques et impose une valeur minimale de la tension de seuil. Diverses solutions ont été envisagées pour palier à cette déficience du commutateur CMOS pour les faibles tensions d'alimentation [14] :

Transistor à faible V_T La disponibilité de transistors avec une faible tension de seuil résout de manière simple le problème du commutateur. Elle introduit, certes, un coût supplémentaire pour la technologie du fait de l'introduction d'étapes supplémentaires. Elle est cependant également intéressante pour les circuits numériques. En effet, pour

ces circuits, une faible tension de seuil réduit les temps de propagation au détriment des courants de fuite. L'utilisation de faibles tensions de seuils uniquement dans les chemins critiques peut permettre un bon compromis entre vitesse et consommation.

Elévation de la tension de contrôle Dans cette technique, on utilise un circuit spécifique à chaque commutateur dont le rôle est d'augmenter la tension de contrôle grille-source du transistor. Certaines configurations permettent également de maintenir cette tension de contrôle constante indépendamment de la valeur du signal. Cette technique peut cependant poser des problèmes de fiabilité. Si la plupart des configurations garantissent une différence de potentiels inférieure à la tension d'alimentation V_{dd} pour V_{gs},V_{gd} et V_{ds}, pour certains transistors, la tension V_{db} peut être supérieure à 2Vdd.

Amplificateur commuté Les commutateurs les plus critiques quant à l'excursion nécessaire du signal sont généralement ceux qui sont situés en sortie d'un amplificateur. Il est alors possible de supprimer ces commutateurs et d'activer ou non l'étage de sortie de l'amplificateur. En raison du temps d'établissement de cet étage, le temps de commutation avec cette technique est cependant plus important que celui d'un simple commutateur CMOS.

Amplificateur CMOS

Les circuits à capacités commutées utilisent généralement une version particulière d'amplificateur, de type transconductance (*Operational Transconductance Amplifier*). Celui-ci est en fait une version simplifiée de l'amplificateur opérationnel générique. Les performances dynamiques et la consommation sont en particulier étroitement liées aux éléments externes à l'amplificateur. Au prix d'une complexité plus grande de conception du fait de cette interaction, on peut obtenir de meilleurs performances en termes de rapidité, de surface et de consommation. Parmi les critères importants dans le choix d'un amplificateur on peut citer :

- Faible tension de décalage (en particulier si cet amplificateur est utilisé dans un circuit de comparaison)
- Marge de phase importante pour garantir la stabilité
- Excursion maximum de sortie pour un gain donné
- Relative indépendances des performances vis à vis de la technologie
- Consommation minimale

Tension de décalage

Considérons la relation quadratique pour le courant drain :

$$I_{ds} = \frac{\beta}{2}(V_{gs} - V_T)^2 \tag{7.3}$$

Les paramètres V_T et β sont supposés suivre une loi gaussienne d'écart type respectif σ_{V_T} et σ_β avec [1] :

$$\sigma_{V_T} = \frac{A_{VT}}{\sqrt{W \cdot L}} \qquad \frac{\sigma_\beta}{\beta} = \frac{A_\beta}{\sqrt{W \cdot L}} \tag{7.4}$$

[1] Seuls les effets à courte distance sont ici considérés [97]

En absence de corrélation entre ces deux grandeurs, on peut facilement obtenir la tension de décalage $\sigma_{V_{gs}}$ qui résulte de la variation des paramètres V_T et β lorsque le courant drain est gardé constant. On obtient à partir de l'expression 7.3 :

$$\sigma^2_{V_{gs}} = \sigma^2_{V_T} + \left(\frac{V_{gs} - V_T)}{2} \cdot \frac{\sigma_\beta}{\beta}\right)^2 = \frac{1}{W \cdot L}\left(A^2_{VT} + \frac{A^2_\beta}{4}(V_{gs} - V_T)^2\right) \tag{7.5}$$

Le paramètre A_{VT} varie proportionnellement à l'épaisseur d'oxyde t_{ox} alors que le paramètre A_β reste sensiblement constant avec un ordre de grandeur de $2\% \cdot \mu m$ [125]. L'importance relative de ce paramètre peut être évaluée en égalant les deux derniers termes de l'expression 7.5. Ceci correspond à $V_{gs} - V_T = 2\,\frac{A_{VT}}{A_\beta}$. Cette valeur est de l'ordre de 500mV pour la dernière technologie du tableau 7.1. On peut donc considérer que pour toutes ces technologies, le paramètre A_{VT} est prépondérant mais que cette situation va changer pour les technologies futures avec une importance croissante du paramètre A_β. La dispersion sur la tension de seuil est souvent traitée comme une tension d'erreur qui va se superposer au signal. On peut, par cette analogie, faire intervenir une énergie associée à cette erreur [99] :

$$E_d = \frac{1}{2}\,C_{gs} \cdot \sigma^2_{V_{gs}} \approx \frac{1}{2}\,C_{ox} \cdot A^2_{VT} \tag{7.6}$$

Le paramètre A_β a été négligé dans cette expression. Cette energie est indépendante des dimensions et caractérise bien une technologie vis à vis de la dispersion sur les composants. Le tableau 7.3 donne quelques exemples de valeurs normalisées par rapport à l'énergie thermique $E_{th} = \frac{1}{2}\,k_B \cdot T_a$ à la température absolue $T_a = 300K$ (les paramètres technologiques sont ceux du tableau 7.1).

$L_{min}(\mu m)$	0,8	0,5	0,35	0,25	0,18	0,12	0,1	0,07
E_d/E_{th}	70,5	74,7	67,5	50	29,4	28,7	23,7	17,4

TAB. 7.3 – Energie associée à la dispersion de V_T

On constate que cette erreur est toujours supérieure au bruit thermique et constitue donc une limite importante pour la précision. Elle décroit cependant avec L_{min}, ce qui représente une évolution favorable de la technologie quant à la dispersion de la tension seuil.

Consommation dynamique

La figure 7.3 représente un amplificateur associé à une capacité C qui est censée représenter la charge totale de sortie (une partie de celle-ci étant liée à la capacité d'entrée de l'étage suivant).

On suppose une efficacité totale de l'élément amplificateur telle que seul le courant fourni à la capacité intervienne dans la consommation. La tension de sortie est supposée varier de V_{pp} pendant une période T du signal. Durant cette période, la charge fournie par l'alimentation est donc égale à $C \cdot V_{pp}$ et l'énergie consommée est égale à $C \cdot V_{pp} \cdot V_{dd}$. On a donc une puissance moyenne :

$$P_{dyn} = \frac{C \cdot V_{pp} \cdot V_{dd}}{T} = f\,C\,V^2_{pp}\,\frac{V_{dd}}{V_{pp}} \tag{7.7}$$

La tension V_{pp} doit être suffisante pour garantir un rapport signal sur bruit minimum SNR_{min}. Considérons le cas d'un signal sinusoïdal de fréquence f en présence du seul

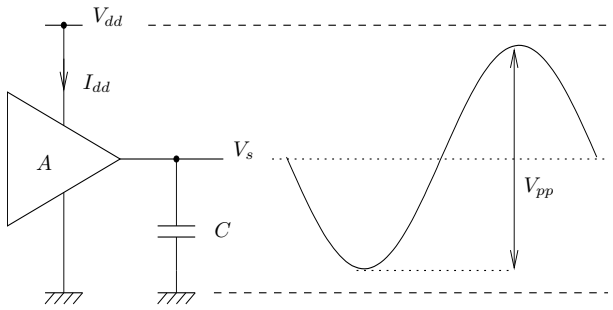


FIG. 7.3 – Modèle pour la consommation.

bruit thermique à une température absolue T_a. On a dans ce cas :

$$P_{dyn} = 8\, k_B\, T_a \cdot f \cdot SNR_{min} \cdot \frac{V_{dd}}{V_{dd} - V_{sat}} \tag{7.8}$$

où l'on a fait apparaître la tension $V_{sat} = V_{dd} - V_{pp}$ nécessaire en pratique pour la polarisation correcte des transistors de l'amplificateur. On voit que pour minimiser la puissance dynamique sous la contrainte d'un certain rapport signal sur bruit on doit avoir une tension V_{sat} la plus faible possible par rapport à la tension d'alimentation. Cette exigence est contradictoire avec les performances en gain et bande passante de l'amplificateur qui imposent une valeur minimum à la tension de saturation. Avec cette contrainte, on remarque que la réduction de la tension d'alimentation s'accompagne d'une augmentation de la puissance dynamique.

La puissance obtenue par l'expression 7.8 est cependant généralement plusieurs ordres de grandeur inférieure à la puissance consommée en pratique. Plusieurs raisons expliquent cet écart. L'expression 7.8 ne considère que le seul bruit thermique. L'erreur associée à la dispersion de la tension de décalage peut être assimilée à un bruit qui va également limiter la dynamique du signal. Dans ce cas, la référence [125] établit une formule similaire à 7.8 mais où l'energie thermique est remplacée par l'energie E_d définie par 7.6. Or, comme l'indique le tableau 7.3, cette energie est actuellement bien supérieure à l'énergie thermique. D'autre part, une consommation existe également indépendamment de l'activité du signal. C'est la puissance statique, nécessaire à la polarisation des transistors. Celle-ci dépend de la topologie de l'amplificateur, de la technologie et des performances dynamiques qui lui sont demandées. La formule 7.8 représente donc une borne inférieure de la consommation que l'on cherchera à approcher au plus près.

Types d'amplificateurs :

La figure 7.4 représente deux architectures classiques d'amplificateurs. Le cascode simple (a) est un montage efficace qui minimise le nombre de noeuds critiques internes. La marge de phase est ainsi facilement garantie et la consommation statique est réduite. Le maintien d'une tension V_{ds} minimale garantissant le fonctionnement saturé de chaque transistor limite cependant l'excursion possible du signal de sortie qui, commme nous l'avons vu dans la section précédente, entraîne une augmentation de la consommation dynamique. Une alternative très utilisée est le cascode replié (b) qui procure une plus grande excursion de la tension de sortie (qui est de plus indépendante du mode commun de l'entrée). Ceci est obtenu au prix d'une consommation qui est

pratiquement le double de celle du cascode simple. L'avantage de cette architecture est de présenter de bonnes performances avec une complexité réduite. La réduction progressive de la tension d'alimentation associée à une tension de seuil relativement importante (figure 7.2) conduit cependant à une excursion de plus en plus faible des signaux. Une alternative au cascode est d'utiliser plusieurs étages de gain inférieur. La marge de phase est dans ce cas plus difficile à satisfaire, du fait de l'introduction de noeuds internes supplémentaires et la conception est plus complexe.

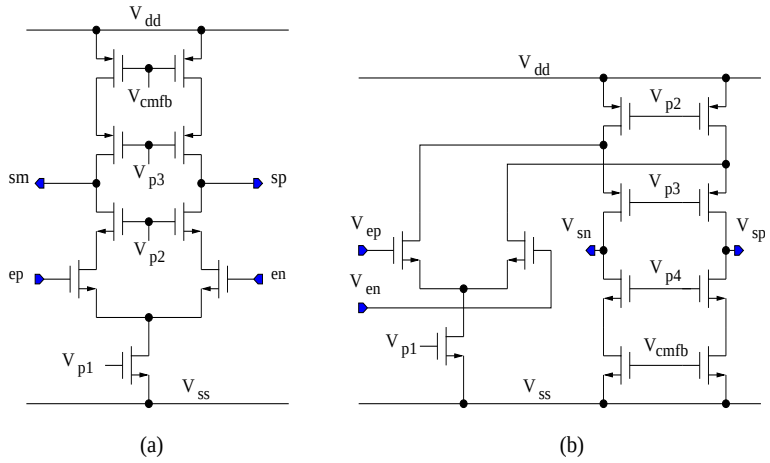


FIG. 7.4 – Type d'amplificateur : (a) Cascode simple (b) Cascode replié .

Pour les structures différentielles représentées à la figure 7.4, la tension de mode commun doit être définie par un circuit séparé. Un exemple de tel circuit est donné à la figure 7.5. C'est un des avantages de la technique des capacités commutées de pouvoir réaliser cette fonction avec un circuit simple qui n'ajoute pratiquement pas de consommation supplémentaire.

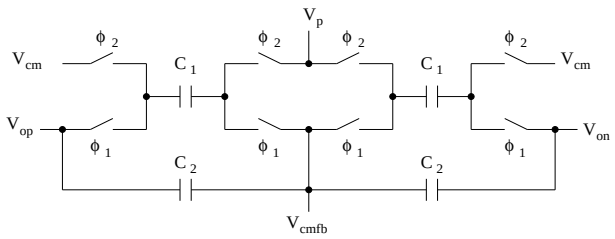


FIG. 7.5 – Exemple de circuit de contrôle de mode commun.

7.4 Conclusions

Nous avons examiné dans ce chapitre l'impact de l'évolution technologique sur les composants des CAN étudiés dans les chapitres précédents. L'étude est restreinte aux circuits à capacités commutées en raison de leur utilisation fréquente aujourd'hui dans les convertisseurs. L'implication la plus importante de la réduction des dimensions est

une diminution de la tension d'alimentation. Nous avons vu quel était son impact sur les blocs de base des circuits à capacités commutées. L'efficacité réduite du commutateur et la réduction de l'excursion en sortie des amplificateurs étant les deux limites les plus importantes pour la conception de ces circuits.

Troisième partie

Simulation comportementale des CAN

Chapitre 8

Simulation comportementale

8.1 Introduction

Dans le cadre de la conception des circuits mixtes, l'exploration de l'espace de conception est souvent faite au niveau transistor, ce qui nécessite des temps de simulation très importants pour évaluer une solution. Les choix au niveau système sont essentiellement basés sur des heuristiques qui permettent de retenir une solution potentiellement valide. A partir de ce choix, l'approche descendante traditionnelle est très coûteuse, car elle nécessite un grand nombre d'itérations entre les différents niveaux d'abstraction comme indiqué à la figure 8.1. Afin de permettre une exploration plus

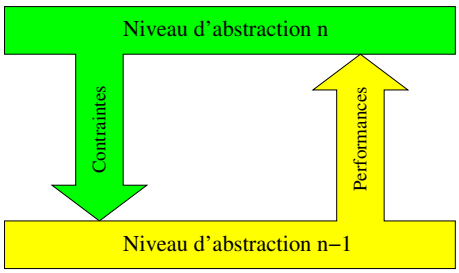


FIG. 8.1 – Interaction entre les niveaux d'abstraction.

efficace de l'espace de conception dans un temps restreint, L. P. Carloni et al. [17] proposent une étape intermédiaire entre le système et l'architecture appelée plate-forme. Celle-ci est composée d'une bibliothèque de cellules paramétrisées et hiérarchisées par leur niveau d'abstraction. Ce point de vue est illustré à la figure 8.2. Ce niveau intermédiaire est défini par une relation sur les performances qui seront utilisées au niveau système et au niveau architectural. Par exemple, pour un convertisseur analogique-numérique de fréquence d'échantillonnage F_s consommant une puissance P pour un $SINAD$ donné, on aura une relation du type :

$$\Psi(F_s, P, SINAD) \tag{8.1}$$

qui définit un domaine sur ces paramètres. Un point dans ce domaine correspond à des performances données au niveau système (par exemple le taux d'erreur binaire d'un récepteur). Il est également en correspondance avec une ou plusieurs architectures

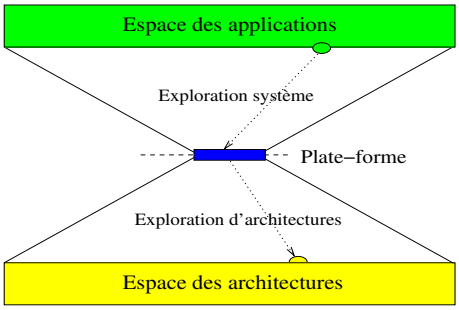


FIG. 8.2 – Plate-forme.

pouvant garantir ce niveau de performances. Une relation du type 8.1 est très précieuse au niveau système pour un CAN car elle peut conditionner un choix particulier d'architecture.

Nous proposons d'utiliser la simulation comportementale pour guider les premières phases de la synthèse. Ces différentes étapes dans la conception d'un circuit mixte sont présentées à la section suivante. Nous traitons ensuite plus spécifiquement le problème du choix d'une architecture de CAN. Le principe de la simulation comportementale utilisée pour ces composants est ensuite décrit. Son application à la synthèse fera l'objet des chapitres 9 et 10.

8.2 Généralités sur la synthèse

La synthèse de circuits mixtes comprend généralement les étapes suivantes [107] :

1. l'expression des spécifications
2. la sélection de topologie
3. Le dimensionnement
4. la génération des masques

Les spécifications sont établies à partir de l'étude au niveau système. Pour un CAN les spécifications générales peuvent être par exemple : la fréquence d'échantillonnage, la résolution (ENOB), la linéarité (SFDR) et la puissance consommée.

La sélection de topologie consiste à choisir dans une bibliothèque de blocs ou de cellules les composants appropriés pour le problème à résoudre. L'existence d'un grand nombre de topologies candidates pour une même fonction rend cette tâche particulièrement délicate pour les circuits mixtes.

Le dimensionnement peut être défini comme la méthode permettant de déterminer l'ensemble des paramètres de conception $\mathcal{I}$ permettant de satisfaire aux mesures de performances $\mathcal{O}$. Au niveau d'une cellule, l'ensemble $\mathcal{I}$ peut être constitué, par exemple, des dimensions des transistors, des courants de polarisation. Au niveau d'un bloc, ce peut être la valeur des coefficients ou l'ordre d'un modulateur $\Sigma\Delta$. Le passage des paramètres de conception aux mesures de performances peut s'effectuer de différentes manières :

- Expression analytique
- Simulation basée sur un modèle comportemental simplifié
- Simulation électrique utilisant un modèle élaboré de transistor (BSIM3, EKV,...).

Le tableau 8.1 présente quelques avantages et inconvénients entre les approches basées sur des expressions analytiques et celles utilisant la simulation. L'avantage en terme de précision attribué à la simulation doit être nuancé par le degré de précision des modèles qu'elle utilise. En fonction de ce choix, un grand nombre de situations intermédiaires sont possibles, où rapidité d'execution et précision sont nécessairement incompatibles.

	Equations	Simulation
Avantages	- rapidité - interactivité	- précision - généralité
Inconvénients	- dérivation des équations - approximations	- temps de simulation - peu prédictif

TAB. 8.1 – Comparaison des méthodes de synthèse

Il existe deux catégories d'outils pour le dimensionnement. Ceux-ci peuvent utiliser une base de connaissance ou une optimisation. Dans le premier cas, les équations de base nécessaires au dimensionnement sont intégrées dans l'outil. Dans le second, un optimiseur ajuste les variables de conception afin de minimiser une fonction objectif sous certaines contraintes.

Enfin, la génération des masques doit fournir une représentation fidèle du circuit dimensionné avec la prise en compte des problèmes géométriques (respect des règles technologiques, prise en compte des dispersions et des perturbations de voisinage des objets). Pour un circuit analogique, le nombres de composants élémentaires (transistors, capacités, résistances,...) est généralement de quelques milliers pour un sous-système mais les interactions entre ces composants sont complexes. Celles-ci sont très dépendantes de la technologie et également de l'environnement (température, Vdd, couplage substrat,...). Les problèmes de couplages et d'appariement de composants influencent fortement la disposition géométrique et les distances entre les différents éléments du circuit. Il faut noter que les étapes précédentes ne sont pas forcément aussi distinctes. Par exemple la synthèse au niveau cellule peut imbriquer fortement le dimensionnement et la synthèse des masques [31].

Méthodes d'optimisation Toutes les phases précédentes sont susceptibles d'utiliser l'optimisation lorsqu'il n'existe pas de solutions directes pour passer d'une étape à l'autre. Le tableau 8.3 présente une classification des principaux algorithmes qui sont généralement envisagés pour l'optimisation dans la conception de circuits. La distinction essentielle tient au caractère local ou global de l'optimisation. Il existe, pour les méthodes locales, des algorithmes efficaces qui convergent rapidement. Malheureusement, selon le point de départ de l'optimisation, ils peuvent fournir un minimum local de la fonction de coût très éloigné de l'optimum global. Les méthodes globales permettent de palier à cet inconvénient en explorant plus efficacement l'espace des variables. Ceci est généralement obtenu au prix d'un temps de convergence plus important. Parmi les problèmes d'optimisation numérique, le cas convexe est particulier. Celui-ci n'apparaît pas dans le tableau 8.3 car c'est à la fois une technique locale et globale. Cette technique s'avère très intéressante pour sa capacité à résoudre efficacement des systèmes de grande taille [12]. Son inconvénient est de nécessiter une formulation particulière du problème qui peut demander un effort important. Nous donnons un bref apercu de cette formulation à l'annexe D.

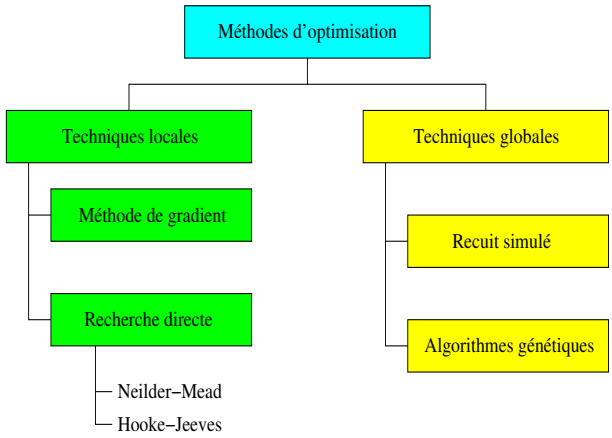


FIG. 8.3 – Classification des méthodes d'optimisation.

8.3 Sélection d'une architecture de CAN

Comme il a été précisé dans l'introduction, le choix d'une architecture est souvent basé sur l'experience. Une représentation fréquente pour les CAN consiste en une partition du plan (résolution, bande-passante) comme indiqué à la figure 8.4. Nous avons

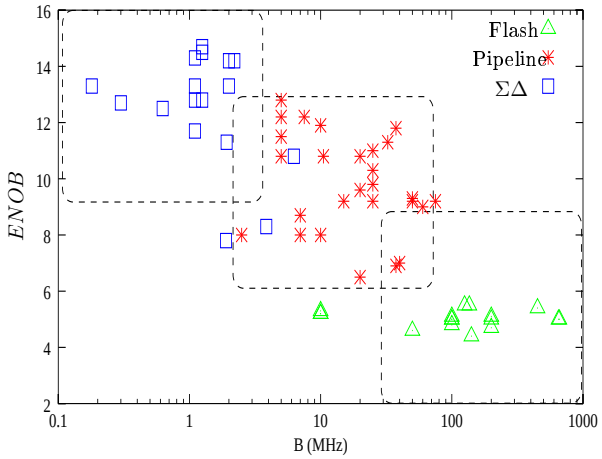


FIG. 8.4 – espace résolution - bande passante pour les différents convertisseurs

reporté sur la figure les réalisations des convertisseurs qui ont été présentés dans la partie 2 (tableaux 4.1, 5.4 et 6.8). Celles-ci couvrent des régions assez distinctes du plan ce qui permet, à partir des spécifications de résolution et de bande passante, de faire un choix particulier d'architecture. Ce mode de sélection présente plusieurs inconvénients. Il est assujetti au choix d'un nombre limité de réalisations propres à un domaine d'application. Certaines région du plan sont ainsi moins explorées que d'autres et le choix à la frontière de deux domaines est délicat. D'autre part, cette représentation ne donne aucune indication sur l'évolution de ces frontières avec la technologie, ce qui peut compromettre la pérennité du choix. Un critère de sélection plus quantitatif

est nécessaire pour réduire la part d'arbitraire dans le choix d'un convertisseur. La référence [129] donne un exemple de facteur de mérite pour la consommation que l'on peut introduire à cette fin. Celui-ci combine les spécifications et les données technologiques pour les CAN flash, pipeline et ΣΔ. L'erreur sur l'estimation de ce facteur pour un nombre réduit de réalisations peut cependant être importante. La figure 8.5 donne un exemple d'ajustement de l'energie de conversion en fonction de la longueur du canal pour les CAN CMOS présentés dans la partie 2. Nous avons également fait

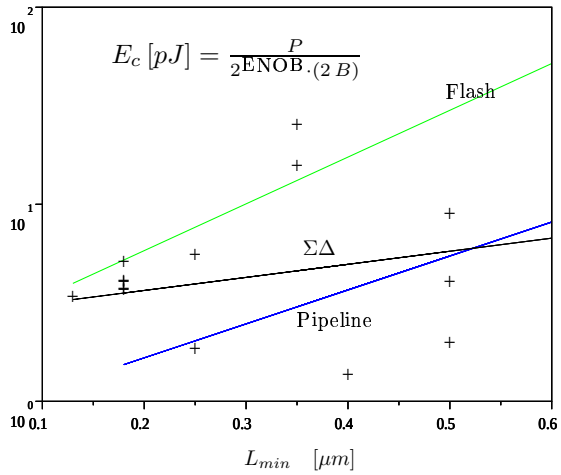


FIG. 8.5 – Energie de conversion en fonction de L_{min}

apparaître les données d'ajustement pour le modulateur ΣΔ pour mettre en évidence la dispersion importante des points (elle est du même ordre pour les autres convertisseurs). L'estimation de l'énergie de conversion faite à partir de cet ajustement est donc très grossière et le prolongement de celui-ci pour des longueurs plus faibles est délicat. Nous verrons dans le chapitre 10 une méthode pour estimer ce paramètre directement à partir de la simulation. La suite de ce chapitre est consacrée à la description de cette dernière à partir de classes C++.

8.4 Classes C++ pour la simulation de CAN

Pour décrire un système constitué de composants discrets, il existe différents formalismes qui sont liés à nature des variables internes et des instants d'observation. Le tableau 8.2 donne une classification des possibilités rencontrées en fonction de la nature continue ou discrète de ces grandeurs.

		Variables descriptives	
		Continues	Discrètes
Temps	Continu	*Equations différentielles*	*Evenements discrets*
	Discret	*Equations aux différences*	*Machine à états*

TAB. 8.2 – Classification des formalismes de modélisation

Pour un système logique ce formalisme est directement lié à son niveau d'abstraction. Une description de type *Machine à états* peut ainsi être utilisée pour une

structure à base de registres et de logique combinatoire (description dite Transfert de Registre ou RTL). Si l'on considère le temps de propagation des signaux dans les portes qui vont constituer cette dernière on devra alors faire appel à la notion d'*Evenement discret* pour décrire les changements d'états. Des langages tel que VHDL ou Verilog peuvent être utilisés pour la description et la simulation à ce niveau d'abstraction (Event-driven simulation). Enfin si ces portes sont décrites au niveau transistor, on sera amené à considérer l'aspect continu des signaux à partir des *Equations différentielles algébriques* déduites des modèles et des lois de conservations physiques (lois de Kirchoff). A ce niveau la simulation est effectuée par un simulateur de circuit de type SPICE. Cette simulation, qui nécessite l'intégration des équations différentielles, est bien sûr la plus coûteuse en temps de calcul.

La simulation d'un système mixte analogique-numérique nécessite généralement tous les niveaux de représentation précédemment décrits. Afin de combiner efficacement ces différents formalismes des langages de description matérielle tels que VHDL-AMS ou Verilog-A ont été développés. La méthode de simulation la plus appropriée à chaque niveau d'abstraction est utilisée (type SPICE pour le temps continu, événementielle pour les états discrets,...). Dans le cas particulier des CAN, les circuits échantillonnés constitués d'interrupteurs contrôlés par une horloge tels que les circuits à capacités commutées ou les circuits à courants commutés sont très fréquemment utilisés. Un tel système est décrit par des *Equations aux différences* où seuls les variables aux instants $n\,T$ avec n entier sont à considérer. Ainsi si $\mathbf{x}(t)$ représente l'état du composant à l'instant $t \in [n\,T, (n+1)T)$ et que celui-ci est décrit par l'équation différentielle :

$$\frac{d}{dt}\mathbf{x} = g(\mathbf{x}, \mathbf{u}_n) \tag{8.2}$$

où $\mathbf{u}_n$ est l'entrée échantillonnée sur l'intervalle $t \in [n\,T, (n+1)T)$, l'état futur est donné par :

$$\mathbf{x}((n+1)T) = \mathbf{x}(nT) + \int_{nT}^{(n+1)T} \phi(\tau, \mathbf{u}_n)d\tau = f(\mathbf{x}(nT), \mathbf{u}_n) \tag{8.3}$$

où $\phi(t, \mathbf{u}_n)$ est la solution de 8.2 telle que $\phi(t, \mathbf{u}_n) = \mathbf{x}(nT)$ pour $t = nT$.

Dans un simulateur mixte, l'intégration numérique de 8.2 conduit à un temps de calcul important. En effet, pour des raisons de stabilité, les méthodes d'intégration sont généralement d'ordre réduit et le seul moyen pour augmenter la précision est de réduire le pas d'intégration. Pour un circuit échantillonné, un nombre de points de calcul important est ainsi dédié aux régimes transitoires [135]. Nous proposons d'utiliser les spécificités de ce type de circuits pour réduire au maximum le temps de calcul, à savoir :

- Tous les composants sont "actifs" à des instants réguliers de l'horloge.
- Si la solution de 8.2 peut être préalablement établie, la sortie d'un composant peut être déterminée directement à partir de 8.3.

La suite de ce chapitre est consacrée à la description structurelle des composants à partir de classes C++ pour la simulation de ce type de circuits. Le choix de ce langage est lié d'une part aux impératifs de rapidité et d'autre part à la possibilité de décrire les composants de manière intuitive et hiérarchique. Les classes de base utilisées pour la description structurelle des convertisseurs sont inspirées de [25] et sont représentées graphiquement sur la figure 8.6. La classe `Component` définit le corps du composant. Elle contient la liste des ports (`Input` et `Output`). Les connexions entre ports sont effectuées par la classe `Wire`. Toutes ces classes (`Input`,`Output`,`Wire`) dérivent d'une même classe `Connector`.

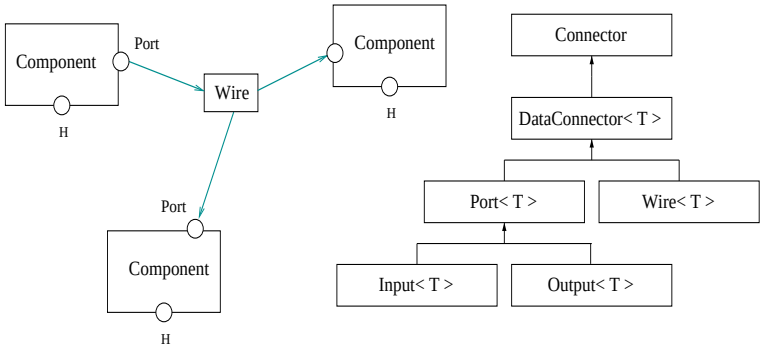


FIG. 8.6 – Classes de base.

La simulation de cette structure est simplement effectuée par une activation simultanée des différents composants à partir d'une horloge. Tous les composants ont une entrée d'horloge (H) et une méthode `update()` qui définit l'action du composant lors de l'activation. Cette méthode calcule la ou les sorties à partir des signaux présents sur les ports d'entrée. Afin d'illustrer le fonctionnement de ces classes, nous utiliserons comme exemple un modulateur ΣΔ du premier ordre représenté à la figure 8.7. Cet

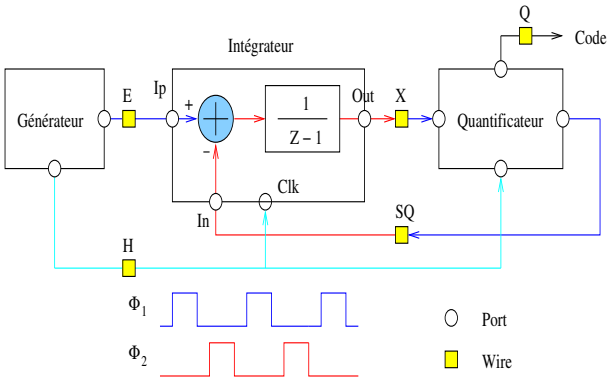


FIG. 8.7 – ΣΔ du premier ordre.

exemple utilise trois composants et cinq connecteurs (`Wire`). Le générateur fournit le signal d'entrée du modulateur et les deux phases de l'horloge qui sont également représentées sur la figure 8.7. Ces deux phases sont représentées par un entier qui est transmis sur le connecteur d'horloge. Voici le code C++ correspondant à la description et à la simulation du modulateur :

```
Wire<Analog> E,X,SQ;
Wire<Binary> Q;
Wire<Clock> H;
INT_BHV Int(E,SQ,X,H);
QTZ_BHV Qtz(X,SQ,Q,H);
Generator G(E,H);
G.set_sin(0.5,1);
G.update();
```

Les types définis pour ces connecteurs (`Analog, Binary, Clock`) sont ici équivalent aux types `flottants, booléens` et `entiers` du langage. Un type quelconque peut cependant être attribué à un connecteur. Le graphe d'héritage des composants correspondant à cet exemple, en supposant deux types de description pour l'intégrateur [1], est donné à la figure 8.8. L'intérêt de cet héritage est de pouvoir réutiliser la description structurelle et de ne redéfinir que les méthodes spécifiques au comportement du composant. Ceci est similaire à la distinction entité-architecture pour la partie structurelle des langages de simulation et de description matérielle tels que Verilog ou VHDL. Dans ces langages, la simulation gère une liste d'événements discrets pour décider de

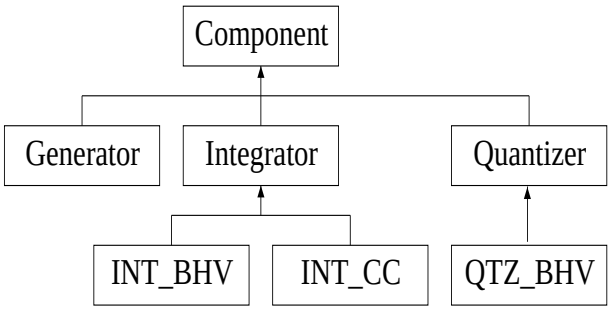


FIG. 8.8 – Hiérarchie des composants pour le modulateur

l'activation des processus qui définissent l'état futur du composant. Ici, l'activation du composant est systématique sur tout changement d'état de l'horloge. Nous donnons à titre d'exemple la méthode d'activation de l'intégrateur `INT_BHV` défini avec les ports `Ip, In, Out, Clk` :

```
void INT_BHV::update()
{
   switch(Clk.get()) {
     case 0: {
        En=In.get();
        State+=Ep-En;
        Out.put(State);
        }; break;
     case 1: {
        Ep=Ip.get();
        }; break;
   }
}
```

Comme nous le voyons sur cet exemple, chaque connecteur dispose de deux méthodes `put()` et `get()` pour écrire et lire des données sur les ports. La description avec un modèle plus complexe tel que celui décrit à l'annexe E est tout à fait similaire. Dans ce cas, l'état de l'intégrateur est définit par le vecteur de charges. La méthode `update()` calcule alors ce vecteur de charge à partir de son état antérieur à chaque période d'horloge.

La hiérarchie simplifiée des classes pour les CAN pipeline et $\Sigma\Delta$ est représentée à la figure 8.9. Elles dérivent toutes les deux de la classe `Component` et toutes les méthodes définies dans cette classe sont applicables aux deux architectures. En particulier, la

[1] BHV : Modèle comportemantal simple, CC : Modèle pour les capacités commutées

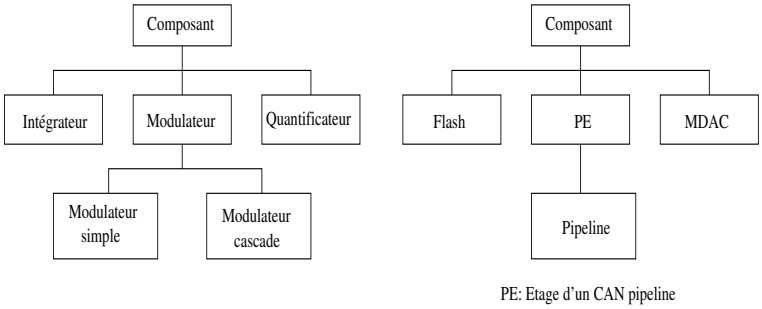


FIG. 8.9 – Hiérarchie simplifiée des classes pipeline et $\Sigma\Delta$.

classe `Component` définit une méthode virtuelle `sample()` implantée dans toute les classes dérivées et qui permet l'échantillonnage statistique. Cette méthode effectue un tirage aléatoire de tous les paramètres définis par une loi statistique donnée.

Un intérêt du language C++ est également de pouvoir intégrer facilement l'environnement de test et le composant dans une même entité. La génération du signal de test et les méthodes d'analyses décrites au chapitre 3 sont ainsi regroupées dans une classe générique (`TestBench`).

L'application de cette description matérielle à la simulation dépend du niveau de modélisation utilisé pour chaque composant. Les deux prochains chapitres présentent différents niveaux de modélisation. Le premier est particulier au modulateur $\Sigma\Delta$. Il permet une évalution rapide de ces performances pour la synthèse des coefficients. Le dernier chapitre est consacré à la modélisation spécifique des circuits à capacités commutées. Nous y décrirons en particulier une méthode d'exploration des performances dynamiques pour les convertisseurs pipeline et $\Sigma\Delta$.

Chapitre 9

Application à la modélisation linéaire du modulateur $\Sigma\Delta$

9.1 Introduction

Du fait de leur caractère non linéaire, certains circuits sont très difficiles à décrire analytiquement. C'est par exemple le cas du modulateur $\Sigma\Delta$ où, à l'heure actuelle, le recours à la simulation est très fréquent pour déterminer ces performances. Pour en simplifier l'analyse, ce circuit est très souvent linéarisé en utilisant le modèle de quantificateur étudié en 1.3.2. L'approximation consiste à remplacer cet élément non linéaire par un gain K et un bruit additif uniforme et blanc de variance P. Un problème délicat est de déterminer analytiquement ces paramètres qui dépendent de la distribution statistique du signal à l'entrée du quantificateur. Ardalan et Paulos [3] ont déterminé ces paramètres dans le cas d'un modulateur du second ordre sous l'hypothèse d'un signal gaussien à l'entrée du quantificateur. Cette hypothèse étant difficile à vérifier dans le cas général, nous proposons de déterminer K et P par une simulation du modulateur sur une courte séquence de N échantillons [100]. A partir des paramètres linéaires du quantificateur, un modèle linéaire complet pour le modulateur est alors déterminé. Celui-ci peut être utilisé pour évaluer rapidement le rapport signal sur bruit ou pour déterminer le degré de stabilité par la localisation des pôles. Nous exploiterons cette possibilité pour déterminer les coefficients d'un modulateur sous des contraintes d'excursion réduites des états d'intégrateurs. Cette possibilité de réduire l'amplitude des états tout en conservant des caractéristiques dynamiques acceptables est particulièrement intéressante pour une réalisation du modulateur avec une faible tension d'alimentation. Un autre intérêt de ce modèle est de permettre une analyse précise et rapide de tous les effets qui conservent le caractère linéaire du filtrage. Il peut ainsi être utilisé pour l'analyse de la sensibilité à la variation des coefficients du filtre ou à tout défaut de l'intégrateur qui admet une modélisation linéaire.

9.2 Estimation des paramètres du quantificateur

Le gain moyen K du quantificateur est donné par 1.26 et sera estimé à partir d'une simulation du modulateur sur N échantillons :

$$\hat{K} = \frac{\sum_{i=1}^{N} e_n \cdot s_n}{\sum_{i=1}^{N} e_n^2} \tag{9.1}$$

Dans cette expression, e_n et s_n sont respectivement les séquences d'entrée et de sortie du quantificateur (figure 9.1). Ces signaux sont assimilés à des variables aléatoires centrées de sorte que l'expression précédente représente une estimation du coefficient de régression linéaire :

$$K = \frac{Cov(e,s)}{Var(e)} \tag{9.2}$$

Cette expression du gain minimise la variance $P = Var(b)$ du bruit de quantification

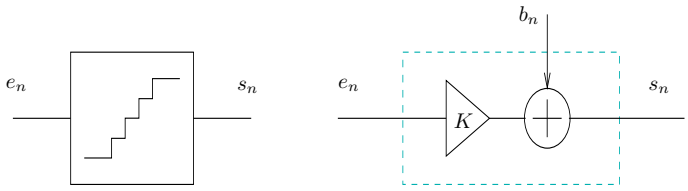


FIG. 9.1 – Modèle linéaire du quantificateur

$b_n = s_n - K \cdot e_n$ qui sera estimée par :

$$\hat{P} = \frac{1}{N}\sum_{i=1}^{N}(s_n - K \cdot e_n)^2 \tag{9.3}$$

Le signal d'entrée et le bruit de quantification sont tels que (d'après 9.2) :

$$Cov(e,b) = Cov(e, s - K \cdot e) = Cov(e,s) - K \cdot Var(e) = 0 \tag{9.4}$$

Le signal et le bruit sont ainsi décorrélés, ce qui justifie le traitement séparé de ces deux signaux.

Le nombre d'échantillons nécessaire au calcul du bruit par 9.3 peut être évalué à partir de la variance de l'erreur $\epsilon_P^2 = E\{(\hat{P} - P)^2\}$ sur cet estimateur. Celle-ci est donnée par l'expression suivante [18] :

$$\epsilon_P^2 = \frac{2}{N}\sum_{n=-(N-1)}^{N-1}\left(1 - \frac{|n|}{N}\right) R_{bb}^2(n) \tag{9.5}$$

où R_{bb} est la fonction d'autocorrélation du bruit de quantification b_n. Sous l'hypothèse d'un bruit de quantification blanc (cf 1.3.2) de variance σ^2, l'expression précédente se réduit à $\epsilon_P^2 = \frac{2\sigma^4}{N}$ soit une erreur relative de l'estimation égale à :

$$\frac{\epsilon_P}{P} = \sqrt{\frac{2}{N}} \tag{9.6}$$

Le tableau 9.1 donne un exemple de valeurs pour les paramètres $\{\hat{K}, \hat{P}\}$ obtenus à partir de 9.1 et 9.3 pour un modulateur monobit du second ordre en fonction de la longueur de la séquence. Les coefficients du modulateur sont ceux du tableau 6.1 et le signal d'excitation est une sinusoïde d'amplitude égale à $0{,}6V_{ref}$ avec $V_{ref} = \frac{\Delta}{2} = 1V$. La dispersion sur les valeurs de la puissance de bruit est conforme à celle donnée par 9.6.

N	$\hat{K}$	$\hat{P}$ (V^2)	ϵ_P (V^2)
256	2.486	0.2785	0.025
512	2.449	0.2873	0.018
1024	2.510	0.3041	0.013
2048	2.729	0.2689	0.008
4096	2.618	0.2835	0.006
8192	2.492	0.2812	0.004

TAB. 9.1 – Paramètres K et P en fonction de la longueur de la séquence

9.3 Calcul de la puissance du bruit de quantification

La connaissance du gain K associé au quantificateur permet de déterminer complètement les fonctions de transfert du signal et du bruit pour un modulateur donné. Ainsi, en reprenant l'exemple du modulateur du second ordre précédent, les fonctions de transfert pour le signal et pour le bruit peuvent être calculées à partir des expressions de l'annexe C :

$$NTF = \frac{(1-z^{-1})^2}{D} \qquad STF = \frac{0{,}5394}{D} \tag{9.7}$$
$$D = 1 - 0{,}7412\, z^{-1} + 0{,}2806\, z^{-2}$$

Celles-ci sont représentées sur la figure 9.2. Nous avons également fait apparaître ces fonctions pour une amplitude du signal égale à $0{,}1V_{ref}$ et la limite de bande pour un rapport de suréchantillonnage R de 16. Nous remarquons la sensibilité de ces fonctions

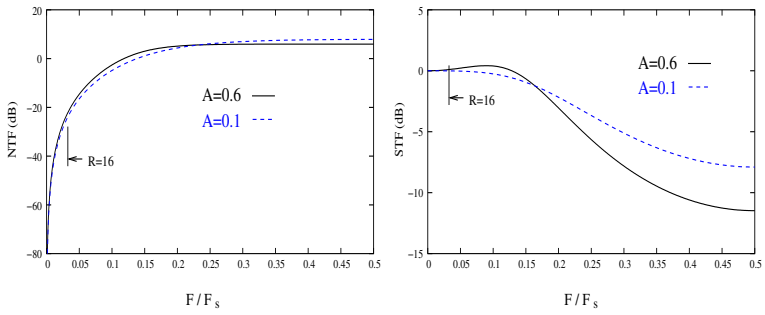


FIG. 9.2 – Fonctions de transfert d'un modulateur du second ordre

à la valeur du gain K du quantificateur qui est égal à 3 pour cette seconde amplitude. Le choix d'une fonction de filtrage particulière est important de ce point de vue et la forme 6.5 est bien adaptée si le rapport de suréchantillonnage est suffisant.

Pour un modulateur avec un seul quantificateur la puissance du bruit de quantification dans une bande B est obtenue à partir de sa fonction de transfert de bruit. En

effet, avec l'hypothèse d'un bruit de quantification blanc de densité spectrale $S_b = \frac{P}{F_s}$, cette puissance est égale à :

$$P_B = \frac{P}{F_s} \int_{-B}^{B} |NTF|^2 df \tag{9.8}$$

Pour une cascade de modulateurs, chaque quantificateur va contribuer au bruit total par l'intermédiaire de son bruit de quantification propre P_i et de la fonction de transfert NTF_i associée à ce bruit. En supposant toutes ces sources indépendantes (elles sont au moins décorrélées d'après 9.4), on a la puissance totale de bruit :

$$P_B = \sum_{i=1}^{M} \frac{P_i}{F_s} \int_{-B}^{B} |NTF_i|^2 df \tag{9.9}$$

Le rapport signal sur bruit est alors calculé à partir de l'intégration numérique de de ces fonctions de transfert dans la bande du signal. Nous avons vu à la section 9.2 que l'estimation de ce rapport necessite un nombre réduit d'échantillons liés au calcul des paramètres K et P. Ceci permet de réduire considérablement le temps de simulation par rapport à ce que nécessiterait une transformée de Fourier, en particulier lorsque le rapport de suréchantillonnage est important. Le calcul du bruit par la FFT n'utilise en effet que N/R points pour une séquence de longueur N. Ceci est mis en évidence sur la figure 9.3 qui montre le rapport signal sur bruit obtenu par les deux méthodes pour la structure cascade de la figure 6.20 et une séquence de 4096 points. On notera

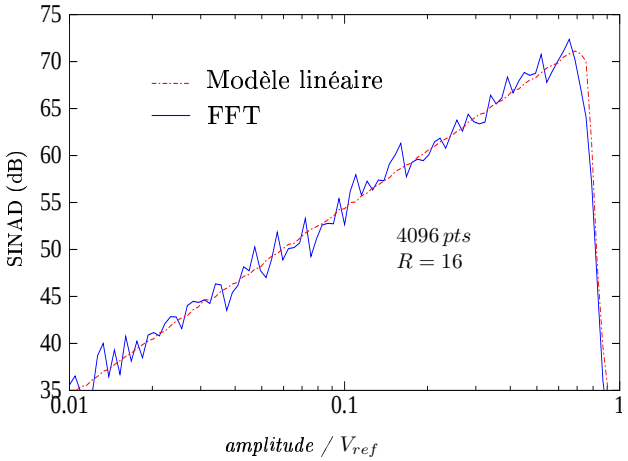


FIG. 9.3 – Rapport signal sur bruit en fonction de l'amplitude

que le modèle linéaire prend bien en compte la saturation du quantificateur [3], celle-ci étant due à une chute rapide du gain K qui entraîne à son tour une modification de la fonction de transfert de bruit. Ceci permet de déterminer rapidement le maximum du SNR par une méthode itérative.

Un autre intérêt du modèle linéaire est de pouvoir juger du degré de stabilité du modulateur par la localisation des pôles. La section suivante décrit une méthode d'optimisation des coefficients du filtre exploitant les avantages de cette modélisation.

9.4 Optimisation des coefficients d'un modulateur

Comme nous l'avons vu dans le chapitre 7, une excursion réduite des états d'intégrateurs est imposée par la diminution de la tension d'alimentation. Le contrôle de l'amplitude des états est cependant délicat lorsque la résolution du quantificateur est faible et en particulier pour le cas monobit. Nous proposons d'utiliser le modèle linéaire précédent pour optimiser les coefficients du modulateur répondant à cette contrainte tout en garantissant la stabilité du modulateur. Pour cela, à chaque simulation du modulateur, le modèle linéaire est calculé ainsi que les pôles. La stabilité est alors imposée par une borne maximale β sur le module des pôles. D'autre part, la fonction de transfert STF pour le signal doit garantir un écart maximal α en limite de bande passante B par rapport au gain unité. Enfin, nous imposons une valeur minimum aux coefficients g_k du filtre pour garantir une meilleur précision (cf 5.13). En notant z_i et x_j respectivement les pôles et les sorties d'intégrateurs, nous pouvons donc formaliser le problème de la manière suivante :

Minimiser la variance P du bruit de quantification sur l'ensemble des coefficients sous les contraintes :

$$\begin{array}{rcl} \max(|1 - STF(B)|) & < & \alpha \\ \max(|z_i|) & < & \beta \\ \max(|x_j|) & < & \gamma \cdot V_{ref} \\ \min(|g_k|) & > & g_{min} \end{array} \tag{9.10}$$

Toutes ces contraintes sont formulées à l'aide d'une fonction objectif :

$$f = \log(P) \cdot \prod_i f_i \tag{9.11}$$

où l'indice i se réfère aux contraintes 9.10. Chaque fonction f_i est définie par deux coefficients. Un des coefficients fixe une limite stricte et l'autre attribue un poids spécifique à la contrainte. En considérant par exemple la stabilité, si le pôle $|z_j|$ de plus grand module est supérieur à la borne β, la fonction associée aura la forme $f_j = \exp[-w_j \cdot (|z_j| - \beta)]$ où w_j est le poids associé à cette contrainte. Dans le cas où tous les pôles sont en module inférieur à β, on imposera $f_j = 1$. Avec cette formulation, le problème se résume à minimiser (sans contrainte) la fonction f précédente sur le vecteur des coefficients du modulateur. L'algorithme de recherche directe de Hooke et Jeeves [54] a été utilisé pour effectuer cette optimisation. Celui-ci s'avère robuste pour les problèmes utilisant la simulation où l'estimation, liée au nombre limité d'échantillons, introduit nécessairement une certaine incertitude. Le point initial doit cependant être correctement choisi pour garantir la convergence de l'algorithme. Une solution simple pour fixer les coefficients initiaux est de choisir un prototype de filtrage particulier pour la fonction de transfert de bruit. Si l'on choisit par exemple un prototype de Butterworth passe-haut pour cette fonction il ne reste qu'une inconnue qui est la fréquence de coupure F_c de ce filtre. Pour le choix de cette fréquence, nous pouvons utiliser la relation 6.10 qui devient :

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} \|NTF_{proto}\|^2 df = P_o \in [1,5:3] \tag{9.12}$$

où NTF_{proto} est dans ce cas la fonction Butterworth passe-haut. La valeur de P_o en fonction de F_c est représentée sur la figure 9.4 pour un modulateur du second ordre, ainsi que les limites données par 9.12. La valeur de F_c peut alors être choisie dans la

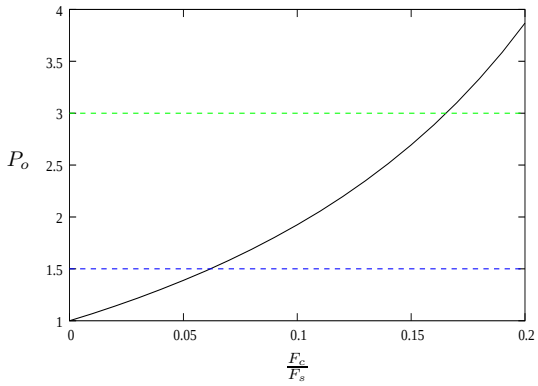


FIG. 9.4 – Valeur de P_o en fonction de F_c pour un modulateur du second ordre

plage des valeurs autorisées par cette relation. Nous avons, par exemple, la fonction prototype suivante, obtenue pour $F_c = 0,125$:

$$NTF_{proto} = \frac{(1 - z^{-1})^2}{1 - 0.9428z^{-1} + 0.3333z^{-2}} \tag{9.13}$$

En choisissant les coefficients α et β identiques, la fonction de transfert de bruit d'un modulateur du second ordre devient (cf annexe C) :

$$NTF = \frac{(1 - z^{-1})^2}{1 + (K\beta_2 - 2)\, z^{-1} + (K\beta_2\,(\beta_1 - 1) + 1)z^{-2}} \tag{9.14}$$

L'identification de 9.13 avec 9.14 pour déterminer les coefficients β_1 et β_2 nécessite encore un choix particulier du gain K. Celui-ci n'intervient qu'au travers du terme $K\beta_2$, produit du gain du quantificateur et du gain du dernier intégrateur (bien que la séquence de sortie du modulateur soit indépendante de ce dernier pour un quantificateur monobit [42], l'amplitude du dernier état dépend de sa valeur). Un gain intermédiaire $K = 2$ a été utilisé pour déterminer la valeur initiale des coefficients. La recherche de l'optimum du rapport signal sur bruit pour ce point initial permet également de fixer l'amplitude du signal pour la phase d'optimisation. Celle-ci nécessite alors les étapes suivantes pour l'évaluation de la fonction objectif :

1. Simulation du modulateur
2. Calcul des paramètres K et P (formules 9.1 et 9.3)
3. Calcul des fonctions de transfert (cf annexe C)
4. Calcul de la puissance de bruit dans la bande du signal (équation 9.8)
5. Calcul du gain sur le signal en limite de bande
6. Calcul des pôles
7. Test de dépassement des états d'intégrateurs
8. Calcul de la fonction objectif (formule 9.11)

Les coefficients des modulateurs 1 bit d'ordre 2 et 3 qui ont été présentés au chapitre 6 ont été optimisés à partir de cette méthode avec une contrainte forte sur l'amplitude des états d'intégrateurs ($\gamma = 0,8$). Pour ces modulateurs, la perte en rapport signal sur bruit introduite par le contrôle des états d'intégrateurs est réduite. La stabilité

représente cependant une limite à cette contrainte. Considérons par exemple à nouveau un modulateur du troisième ordre avec deux bornes différentes $\gamma = 0,6$ et $\gamma = 0,8$ sur l'amplitude de ces états. La figure 9.5 montre l'histogramme obtenu à l'optimum du SNR (R=16).

Les valeurs de SNR obtenues sont très voisines ainsi que les amplitudes correspondantes. La valeur des coefficients pour ces deux cas est reportée dans le tableau 9.2. [1] La stabilité de ces deux solutions n'est cependant pas équivalente. Les pôles à l'is-

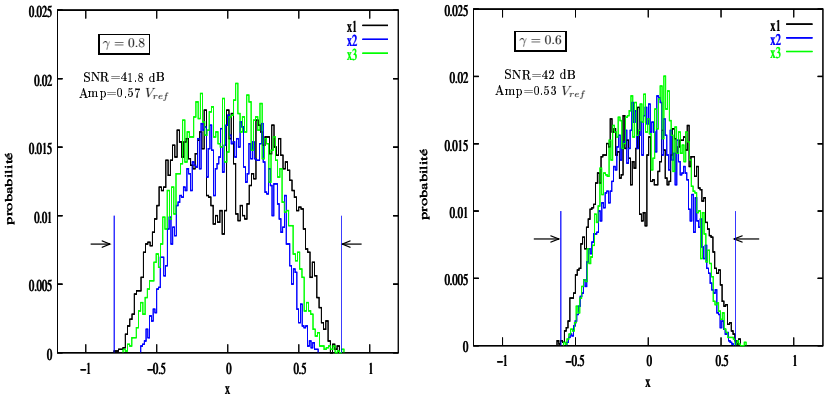


FIG. 9.5 – Excursion des états pour un modulateur d'ordre 3

γ	α_1	β_1	α_2	β_2	α_3	β_3
0,6	1/6	1/6	1/3	2/9	3/4	3/8
0,8	1/5	1/5	1/3	1/3	2/3	4/9

TAB. 9.2 – Coefficients des intégrateurs

sue de l'optimisation sont représentés sur la figure 9.6. La restriction sur leur module est représentée par le cercle de rayon r=0,75. On remarque que cette contrainte n'est plus respectée pour $\gamma = 0,6$. Même si le module reste à l'intérieur du cercle unité ($r_{max} = 0,85$) la stabilité du modulateur sur une longue séquence n'est pas garantie. Les valeurs obtenues pour les coefficients des modulateurs d'ordre 2 et 3 du chapitre 6 résultent de ce compromis entre stabilité et amplitude des états. Comme il a été signalé au chapitre 6 le maintien de la stabilité pour les modulateurs d'ordre supérieur à trois est délicat et impose une réduction importante de la tension d'entrée. L'application de la procédure précédente à l'ordre 4 conduit ainsi à une solution très proche du point initial qui, bien que fonctionnel, ne satisfait pas aux contraintes précédentes sur la stabilité et l'amplitude des états (celle-ci étant près de deux fois supérieure à la tension d'entrée). Il semble donc préférable de limiter l'ordre à trois pour une réalisation du modulateur avec une contrainte de faible tension d'alimentation.

[1] On se reportera à l'annexe C pour la définition des coefficients

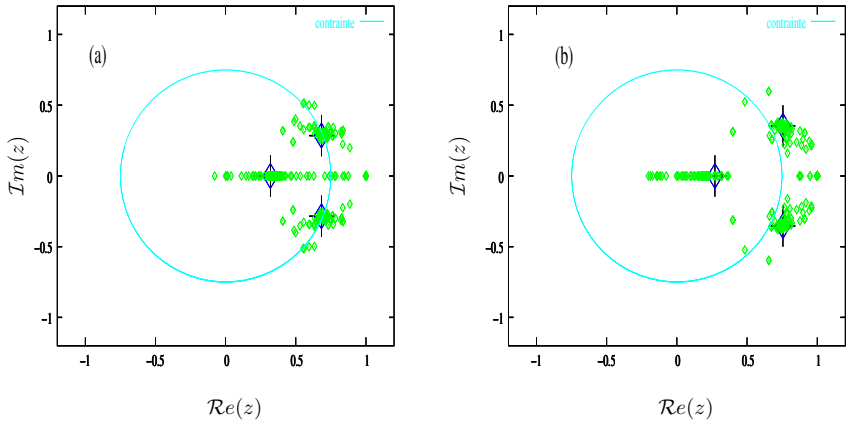


FIG. 9.6 – Lieu des pôles pour (a) $\gamma = 0,8$ et (b) $\gamma = 0,6$

9.5 Prise en compte des effets linéaires

9.5.1 Effet du gain fini dans un intégrateur

Tout effet qui conserve le caractère linéaire du filtrage dans le modulateur peut être efficacement déterminé à partir du modèle précédent. C'est le cas, par exemple, si l'on considère l'effet d'un gain fini constant A pour un amplificateur dans le bloc intégrateur de la figure 9.7. La conservation des charges lors du transfert (fermeture

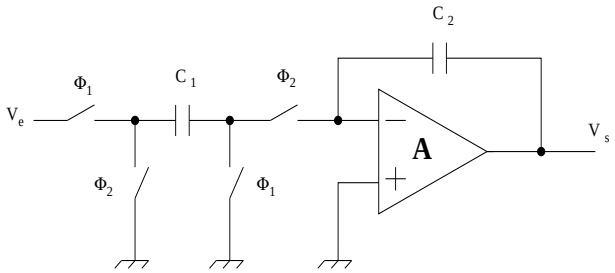


FIG. 9.7 – Intégrateur à capacités commutées

de Φ_2) s'écrit :

$$C_2\ (1 + A^{-1})\ [V_s(n) - V_s(n-1)] + C_1\ [A^{-1}\ V_s(n) - V_e(n-1)] = 0$$

Ceci conduit à une nouvelle relation de récurrence linéaire :

$$V_s(n) = b\ V_s(n-1) + a\ V_e(n-1)$$

$$\text{avec}\quad b = \frac{1 + A^{-1}}{1 + A^{-1}\ (1 + K)}\quad \text{et}\quad a = \frac{K}{1 + A^{-1}\ (1 + K)}$$

et la fonction de transfert devient :

$$T(z) = \frac{V_s(z)}{V_e(z)} = \frac{a \cdot z^{-1}}{1 - b \cdot z^{-1}} \tag{9.15}$$

La figure 9.8 montre l'effet du gain fini de l'amplificateur pour un modulateur du second ordre (tableau 6.1) et la structure cascade étudiée à la section 6.6.2. Dans le cas idéal, le SNR est sensiblement identique pour ces deux architectures avec respectivement R=64 et R=16. La structure cascade s'avère beaucoup plus sensible à l'effet du gain fini. Celui-ci modifie les fonctions de transfert de bruit, ce qui conduit à une recombinaison inefficace des différentes sorties de quantificateurs. Le modèle li-

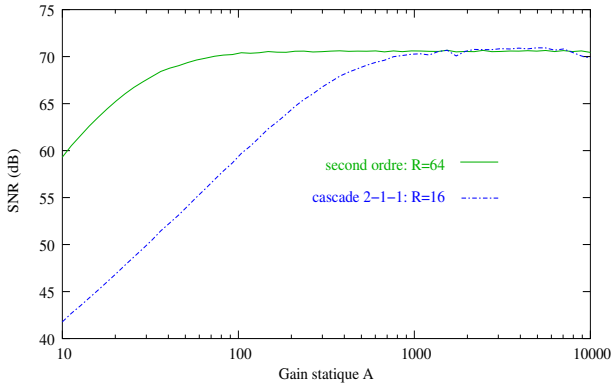


FIG. 9.8 – Effet du gain fini de l'amplificateur.

néaire permet d'effectuer rapidement ce type d'analyse étant donné le faible nombre de points nécessaires à la simulation [2]. Il ne pose aucune restriction sur le rapport de suréchantillonnage qui est souvent supposé suffisamment grand dans les formules approximatives qui prennent en compte cet effet [78].

9.5.2 Analyse statistique

Le modèle linéaire associé au modulateur est également bien adapté pour effectuer une analyse statistique sur les coefficients du filtre. Il évite en particulier le recours à la transformée de Fourier qui est particulièrement coûteuse en raison du suréchantillonnage. La figure 9.9 représente les histogrammes du SNR pour le modulateur du second ordre ($R = 64$) et la structure cascade 2-1-1 ($R = 16$) de la section 9.5. Ceux-ci sont obtenus après 200 simulations de chaque architecture et un tirage aléatoire de tous les coefficients suivant une loi normale d'écart type σ=1% autour de la valeur nominale.

On note la grande robustesse du modulateur du second ordre vis-à-vis de la variation des coefficients [11]. La dispersion plus importante du SNR pour la structure cascade est quant à elle toujours liée à la recombinaison imparfaite des sources de bruit issues des différents quantificateurs.

9.6 Classes C++ associées

Le modèle linéaire est décrit par un ensemble de classes (fonction objet) qui sont associées aux architectures simples et cascade de la figure 9.10. A chaque modulateurs simples d'ordre 1 à 4 (SL1 à SL4) est associée une fonction de transfert pour le bruit et le signal. Les étages, autres que le premier sont constitués par les modulateurs SL1A et SL2A d'ordre 1 et 2 et comportant une entrée supplémentaire pour la connexion en cascade qui peut elle même comporter deux (CXY) ou trois (CXYZ) étages. Un

[2] Chaque point de la figure 9.8 est obtenu avec une séquence de 4096 échantillons

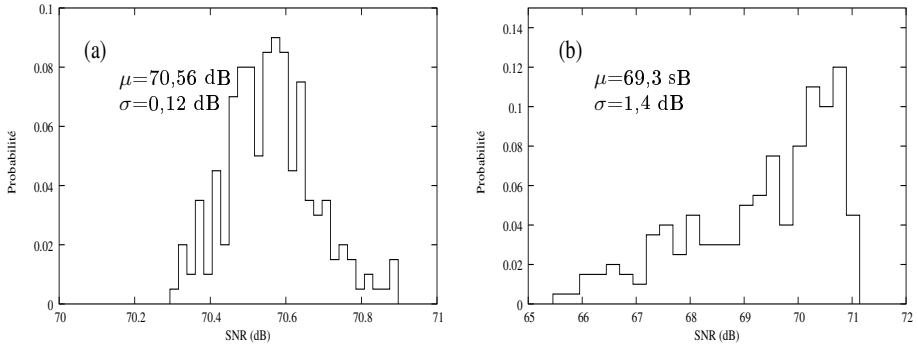


FIG. 9.9 – Analyse statistique : (a) second ordre (b) cascade

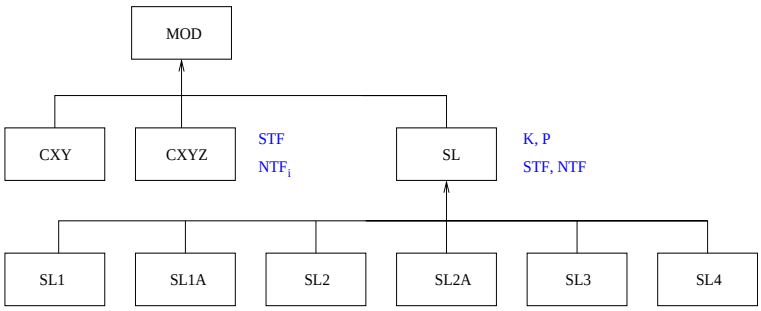


FIG. 9.10 – Classes de modulateurs

ensemble de classes spécifiques pour la cascade permet le calcul des fonctions de transfert en prenant en compte les relations de couplages décrites à la section 6.6.2. Ces fonctions sont mises à jour à l'issue de chaque simulation après le calcul des paramètres K et P associés à chaque quantificateur.

Afin de faciliter l'utilisation interactive du modèle, une interface écrite en Tcl/Tk (figure 9.11) permet d'effectuer les opérations suivantes :

1. Choix d'une architecture de modulateur
2. Configuration (Paramètres de simulation, du modulateur...)
3. Simulation (SNR maximum, Dynamique, Statistique,...)
4. Analyse des résultats (Histogramme des états, Fonctions de transfert,FFT,...)

9.7 Conclusions

Nous avons présenté dans ce chapitre une modélisation linéaire pour l'analyse des performances du modulateur ΣΔ à partir de la simulation. L'utilisation d'un modèle linéaire pour le quantificateur n'est pas nouvelle mais le choix de son gain est souvent fait de manière arbitraire, ce qui conduit à des écarts importants entre les performances théoriques et pratiques. La simulation du modulateur sur une courte séquence du signal permet de déterminer efficacement les paramètres de ce modèle. Celui-ci fournit par ailleurs les fonctions de transfert associées au signal et au bruit de quantification. Il renseigne également sur le degré de stabilité du modulateur par la localisation des

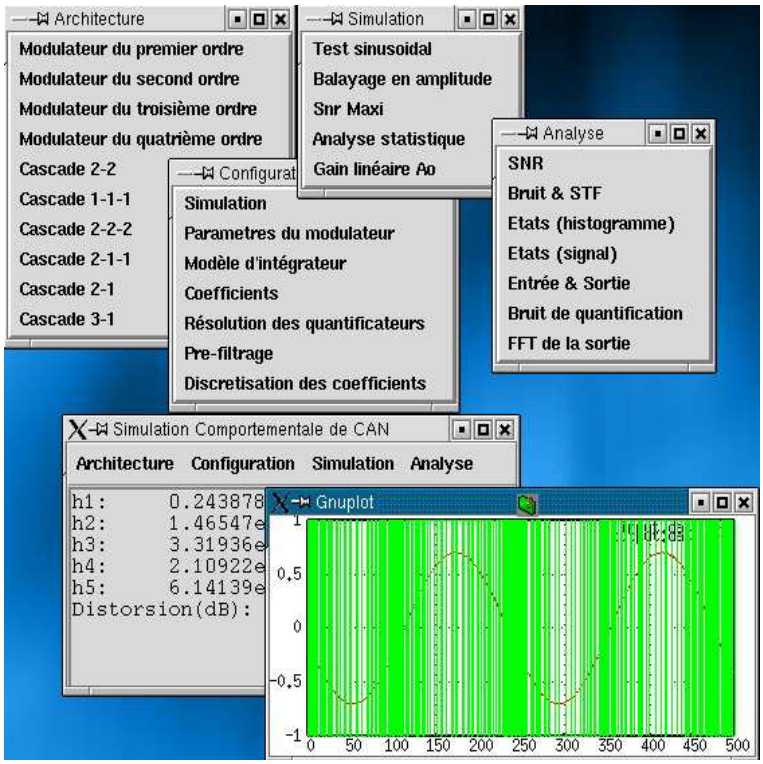


FIG. 9.11 – Interface Tcl/Tk pour l'analyse des modulateurs

pôles. Nous avons exploité cette propriété pour optimiser les coefficients afin d'avoir une excursion contrôlée des sorties d'intégrateurs. Enfin, ce modèle permet une analyse rapide de l'effet des perturbations qui conservent le caractère linéaire du filtrage telles que la modification de la fonction de transfert d'un intégrateur ou la variation des coefficients du filtre. Il n'est cependant d'aucun secours pour mettre en évidence des non linéarités dans le filtrage ou dans la caractéristique d'un quantificateur multibit. Pour l'étude de tels phénomènes, le recours à l'analyse spectrale de la sortie du modulateur est inévitable.

Chapitre 10

Application aux circuits à capacités commutées

10.1 Introduction

Ce chapitre est consacré à l'application de la modélisation comportementale aux convertisseurs utilisant la technique des capacités commutées. Cette technique est largement utilisée aujourd'hui pour ces composants du fait de sa compatibilité avec la technologie CMOS. Nous y mettrons en évidence les effets non linéaires de ces circuits qui fixent des limites importantes aux performances. Ceci est particulièrement vrai pour les effets dynamiques tel que le transfert incomplet des charges qui impose une limite basse à la consommation du convertisseur. Celle-ci sera analysée pour différentes configurations de convertisseurs $\Sigma\Delta$ et pipeline à partir d'un modèle simplifié d'amplificateur associé à une technologie.

10.2 Blocs de base des CAN à capacités commutées

Dans les convertisseurs à capacités commutées, les deux blocs essentiels autres que le CAN flash pour les conversions analogique-numérique intermédiaires sont le MDAC pour la structure pipeline et l'intégrateur pour le modulateur $\Sigma\Delta$. La conversion numérique-analogique nécessaire aux quantifications intermédiaires est avantageusement réalisée par un réseau de capacités élémentaires à l'intérieur de ces blocs, ce qui permet une intégration avec un coût réduit. Un exemple de réalisation du MDAC utilisant cette technique est donné par la figure 5.6 du chapitre 5. Pour ces circuits une structure différentielle est généralement employée. Celle-ci procure le double avantage de réduire l'effet des perturbations présentes sur la tension d'alimentation et d'augmenter l'excursion des signaux (théoriquement dans un facteur 2). Un exemple de structure différentielle pour l'intégrateur est donné à la figure 10.1. Une des deux entrées (phase ϕ_2) réalise la conversion numérique-analogique associée à une quantification 1 bit. Ce circuit réalise idéalement la fonction (en supposant que les signaux V_e, V_s et b changent sur la phase ϕ_2) :

$$V_s(n) = V_s(n-1) + \alpha(V_e(n-1) - q \cdot V_{ref}) \tag{10.1}$$

avec

$$\alpha = \frac{C_1}{C_2} \quad , \quad q = \begin{cases} +1 & \text{si } b = \text{“1”} \\ -1 & \text{si } b = \text{“0”} \end{cases}$$

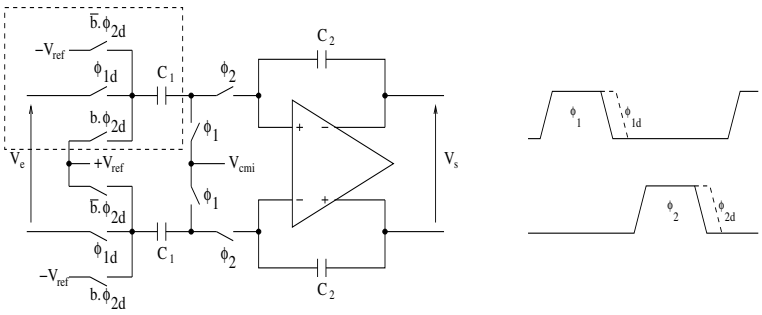


FIG. 10.1 – Intégrateur différentiel.

Les phases de commande des interrupteurs sont également représentées sur la figure 10.1. Elles font intervenir deux signaux décalés pour les interrupteurs d'entrée qui ont pour but de réduire l'effet de l'injection de charges parasites lors de la commutation de l'horloge. L'utilisation de phases retardées, sans supprimer cette erreur, la rend indépendante du signal ce qui améliore nettement la linéarité du circuit [66]. Comme pour le MDAC l'utilisation d'une quantification sur n bit est aisément réalisée par le partage de la capacité d'entrée C_1 en $m = 2^n$ capacités C_u comme indiqué à la figure 10.2.

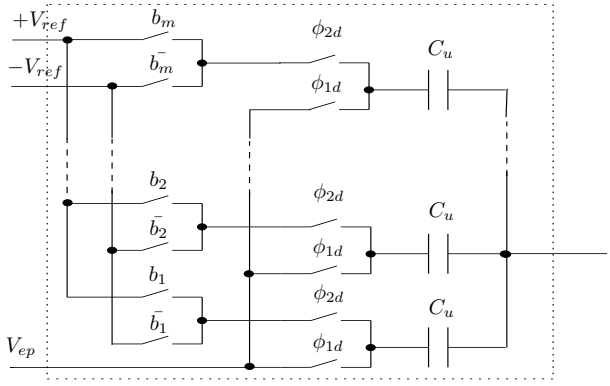


FIG. 10.2 – Entrée d'un intégrateur multibit

10.2.1 Sources d'erreur

Différentes sources d'erreurs, autres que l'injection de charge précédemment mentionnée, vont altérer le comportement des blocs à capacités commutées. On peut classer ces erreurs de la façon suivante :

1. Erreurs conservant le caractère linéaire du circuit
   - Gain fini (constant) de l'amplificateur
   - Dispersions sur les valeurs des capacités
2. Erreurs introduisant un comportement non linéaire
   - Comportement dynamique de l'amplificateur

- Non linéarité du gain statique
- Non linéarités des capacités
- Injection de charges d'horloge
- Incertitude d'échantillonnage

3. Erreurs pouvant être représentées par une source de bruit additif
   - Le bruit thermique lié aux commutateurs
   - Le bruit de l'amplificateur

La modélisation de toutes ces erreurs est complexe et généralement, seule une simulation électrique complète peut rendre compte de tous ces phénomènes. Celle-ci exige cependant un temps de calcul important et ne peut être appliquée que pour une vérification finale ou des tests partiels lors de la conception. Nous présentons dans la suite quelques exemples d'utilisation de la simulation comportementale pour analyser les performances des CAN en présence des défauts précédents. Parmi ceux-ci, les effets non linéaires sont les plus délicats à évaluer. Ils seront mis en évidence dans deux cas particuliers :

les non linéarités statiques liées à la charge de capacités imparfaites.

les non linéarités dynamiques liées aux transferts incomplets des charges.

10.3 Charge non linéaire des capacités

Dans les circuits à transfert de charge, les performances ultimes sont limitées par la qualité des capacités. Leur valeur est généralement dépendante de la fréquence d'utilisation, de la tension appliquée et de la température. Nous ne considérons ici que la variation de la capacité en fonction de la tension appliquée. Celle-ci peut être modélisée par l'équation suivante [136] :

$$C(V) \cdot V = [C\,(1 + \beta\,V + \alpha\,V^2)] \cdot V = C \cdot f_c(V) \tag{10.2}$$

Les coefficients α et β sont très dépendant de la technologie utilisée et plus précisément de la nature de l'oxyde qui constitue le diélectrique. Un exemple de valeurs est donné dans le tableau 10.1 pour l'oxyde de silicium (SiO_2) et pour le nitrure de silicium (Si_3N_4) [136].

	$\alpha\ (V^{-2})$	$\beta\ (V^{-1})$
SiO_2	$-9,4\ 10^{-6}$	$-2,2\ 10^{-6}$
Si_3N_4	$35\ 10^{-6}$	$-130\ 10^{-6}$

TAB. 10.1 – Coefficients α et β en fonction de l'oxyde

Nous considérons successivement l'effet de cette non linéarité dans les convertisseurs pipeline et $\Sigma\Delta$.

convertisseur pipeline : La relation donnant le résidu d'un étage pipeline est idéalement :

$$V_{r+1} = b\,V_r - q\,V_{ref} \tag{10.3}$$

où b est la base associée à l'étage et q est le code issu du CAN flash. Dans le MDAC à capacités commutées, cette opération est effectuée par un échantillonnage de la tension V_r sur b capacités C, puis un échantillonnage de $\{+V_{ref}\}$, $\{0\}$ ou $\{-V_{ref}\}$ sur $b-1$

capacités en fonction du code numérique. Autrement dit, la relation 10.3 se traduit en pratique par :

$$C\,V_{r+1} = b\,C\,V_r - C\sum_{j=0}^{b} q_j\,V_{ref} \tag{10.4}$$

A partir de 10.2 et 10.4 on a :

$$f_c(V_{r+1}) = b\,f_c(V_r) - \sum_{j=0}^{b} f_c(q_j \cdot V_{ref}) \tag{10.5}$$

Le membre de droite peut être aisément calculé. Le résidu V_{r+1} est alors la racine réelle de l'équation polynomiale du troisième ordre résultante. La figure 10.3 donne le résultat de simulation de la non linéarité d'un CAN pipeline utilisant l'équation 10.5 pour le calcul du résidu et les coefficients α et β correspondant à l'oxyde Si_3N_4 du tableau 10.1. Le CAN simulé a une résolution théorique de 14 bit et est constitué de 4 étages avec redondance de résolution respective 4, 3, 3 et 4 bit. La non linéarité intégrale

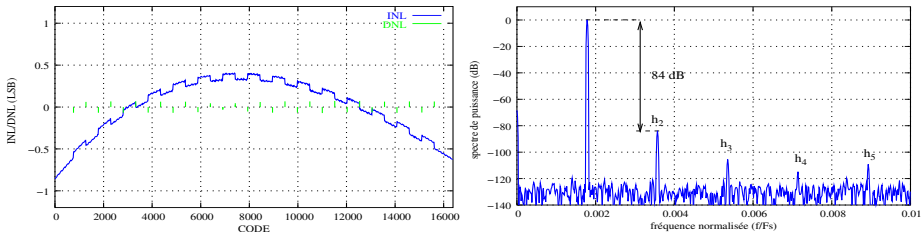


FIG. 10.3 – Effet de la non linéarité des capacités

est pratiquement une fonction paire. Ceci se traduit par la présence de l'harmonique 2 qui est 84 dB en dessous du fondamental. Dans le cas de l'oxyde de silicium, les coefficients α et β sont beaucoup plus faibles et conduisent à une erreur imperceptible pour cet exemple.

modulateur $\Sigma\Delta$: La relation de conservation de charge pour un intégrateur à M entrées est idéalement :

$$C_I\,V_s(n) = C_I\,V_s(n-1) + \sum_{j=0}^{M} m_j\,C_{Sj}\,V_{ej} \tag{10.6}$$

où C_I est la capacité d'intégration et C_{Sj} est la capacité d'échantillonnage pour l'entrée j. Le coefficient $m_j = \pm 1$ détermine le signe de l'entrée et V_{ej} est la tension appliquée sur cette entrée. La prise en compte de la charge non linéaire (équation 10.2) conduit à l'équation fonctionnelle :

$$f_c(V_s(n)) = f_c(V_s(n-1)) + \sum_{j=0}^{M} m_j\,\frac{C_{Sj}}{C_I} f_c(V_{ej}) \tag{10.7}$$

Cette équation peut être résolue de la même manière que pour le pipeline en calculant la racine réelle de l'équation polynomiale obtenue en remplaçant $f_c(\bullet)$ par l'expression 10.2. La figure 10.4 donne le résultat d'une simulation comportementale pour une

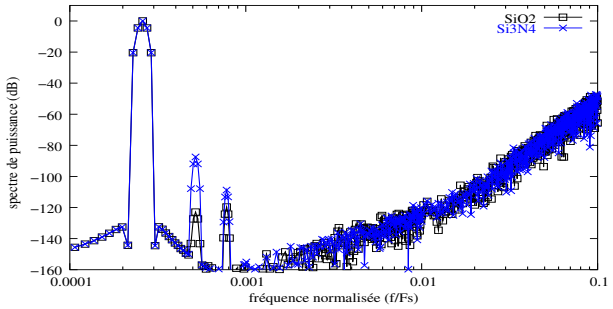


FIG. 10.4 – Effet de la non linéarité des capacités dans le modulateur ΣΔ

architecture cascade 2-1-1 monobit en calculant la sortie des intégrateurs à partir de l'équation 10.7. La distorsion introduite est légèrement inférieure à celle qui est observée dans le convertisseur pipeline pour l'oxyde Si_3N_4. Elle se traduit toujours par une harmonique 2 prépondérante qui est dans ce cas 87,5 dB en dessous du fondamental (l'amplitude du signal est égale à 0,65 V_{ref}). Le résultat de la simulation avec des coefficients de tension correspondant à l'oxyde SiO_2 est également reporté sur la figure 10.4. Les harmoniques deux et trois sont 120 dB en dessous du fondamental.

Ces exemples illustrent la limitation en résolution liée à la qualité des éléments passifs dans ce type de circuits. Le choix de ces composants est crucial lorsque la résolution attendue est importante. La simulation comportementale est un outil efficace pour analyser ce type d'erreur. Il est en effet généralement difficile d'obtenir une formulation analytique de la distorsion pour de tels circuits. La situation est encore plus complexe lorsque les non linéarités sont associées à un comportement dynamique [1]. C'est le cas du transfert de charge et du comportement dynamique de l'amplificateur que nous examinons dans la section suivante.

10.4 Exploration des performances dynamiques

Parmi les erreurs précédentes, le transfert incomplet des charges constitue une limite importante. Il conditionne directement la fréquence maximale d'échantillonnage et la consommation du circuit. Nous décrivons dans cette partie une méthode d'analyse pour le comportement dynamique des CAN à capacités commutées. Celle-ci nécessite un modèle simple pour la prise en compte du comportement dynamique de l'amplificateur. Celui-ci est développé pour le circuit intégrateur et sa présentation détaillée est reportée à l'annexe E. Nous en donnons ici uniquement le principe.

10.4.1 Modèle dynamique pour le transfert de charge

Le principe de cette modélisation est de résoudre préalablement le système d'équations différentielles décrivant le circuit. L'amplificateur est modélisé par un système du premier ordre avec une source linéaire par partie (figure 10.5). La conservation de la charge est garantie à toutes les étapes du calcul. Ce point est très important ici, car quel que soit le comportement de l'amplificateur, la conservation de la charge est une caractéristique du circuit de la figure 10.1 (et des circuits à capacités commutées en général). Seuls quatre points sont effectivement calculés par période du signal (3 points

[1] C'est également le cas si l'on considère le phénomène de relaxation pour la capacité [136]

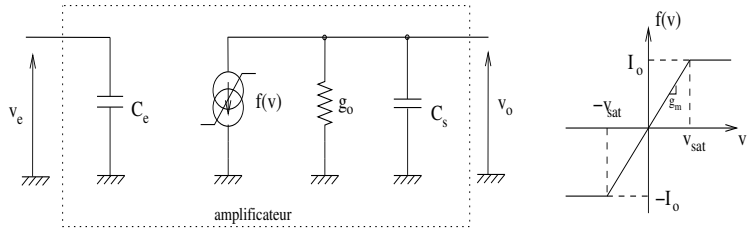


FIG. 10.5 – Modélisation de l'amplificateur

pendant $\phi 2$ et un point pendant $\phi 1$). La figure 10.6 montre la tension de sortie lors du

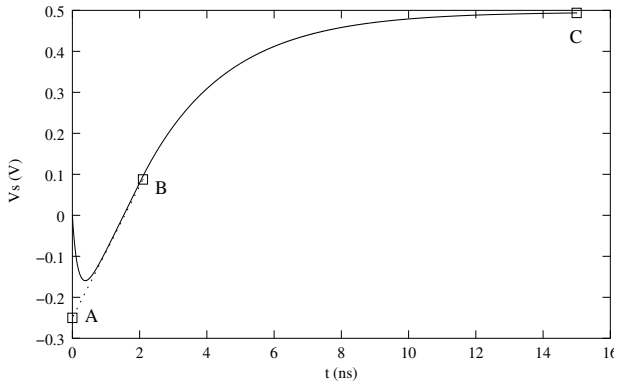


FIG. 10.6 – Tension de sortie de l'intégrateur et points de calcul (□)

transfert de charge (phase ϕ_2 du circuit de la figure 10.1). La courbe correspond à une intégration numérique et les points (A,B,C) sont déterminés par le modèle simplifié de l'annexe E. L'écart entre le point calculé (A) et la sortie est important car celui-ci suppose une redistribution instantanée des charges. Les points (B) et le point (C) correspondent respectivement au début et à la fin du régime linéaire. Dans la mesure où ce régime représente une part importante du transfert, le point (C) constitue une bonne approximation de la valeur finale.

Nous proposons d'utiliser ce modèle pour analyser les performances dynamiques des CAN utilisant les capacités commutées. Ce modèle pour le transfert de charge peut en effet être appliqué également à l'amplification dans un étage pipeline qui est similaire au problème précédent si l'on néglige l'effet de la résistance du commutateur en série avec la capacité de contre-réaction (capacité C_0 de la figure 5.6 du chapitre 5). Celle-ci est la capacité unitaire du MDAC et la constante de temps associée est généralement faible. Dans ce cas, seule la phase nommée "phase de transfert" de l'annexe E est considérée. On doit cependant prendre en compte le couplage entre étages pour déterminer les caractéristiques de l'amplificateur. Ce problème est traité à l'annexe F.

10.4.2 Principe de la méthode

La modélisation de l'amplificateur nécessite la connaissance des différents paramètres du circuit de la figure 10.5. Ceux-ci sont déterminés à partir du point de repos

d'un seul transistor NMOS et de ces paramètres linéaires et statiques :

$$I_o = 2 * I_{ds} = g_m V_{sat} \quad g_o = \alpha\, g_{ds} \quad C_e = C_{gs} \quad C_s = \beta\, C_{jd} \tag{10.8}$$

Si l'on considère un amplificateur du type de la figure 7.4 (a) du chapitre 7, ces paramètres sont ceux d'un transistor de la paire différentielle d'entrée et le courant I_o est son courant de polarisation. Les coefficients α et β sont introduits pour prendre en compte la diminution de la conductance de sortie par l'étage cascode et l'accroissement de la capacité de jonction qui résulte de cette association de transistors. Bien que ce modèle puisse sembler très simple, il donne de bons résultats lorsque ces paramètres sont directement déduits du point de repos de l'amplificateur [101]. Afin d'explorer les limites de performances de différentes architectures de CAN nous fixerons arbitrairement $\alpha = 0,01$ et $\beta = 5$ sachant que ces coefficients sont dépendants de la topologie et du dimensionnement exact de l'amplificateur. La connaissance du courant de polarisation I_{ds} et de la géométrie (W, L) du transistor NMOS fournit ainsi tous les paramètres du modèle. Afin de garantir un gain en tension suffisant de l'amplificateur nous imposerons de plus une longueur L du transistor 5 fois plus grande que le minimum technologique [2]. Nous pouvons ainsi déterminer les paramètres du modèle sous la forme d'une table obtenue par simulation électrique d'un seul transistor. Les entrées de cette table étant la largeur W et le courant de polarisation I_{ds}. Une exploration des performances dynamiques est alors effectuée à partir du modèle comportemental du CAN comme indiqué à la figure 10.7. Pour chaque point de la table, la simulation

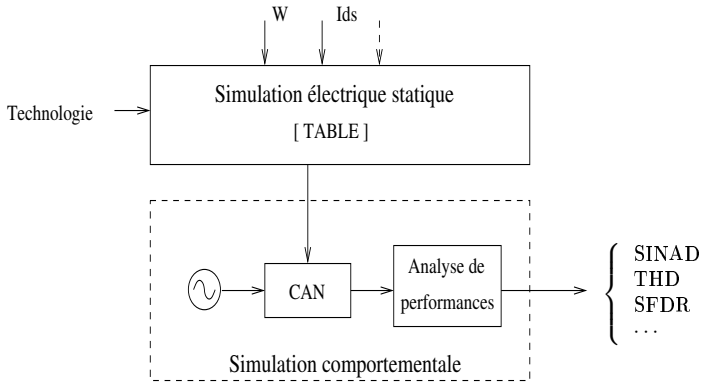


FIG. 10.7 – Exploration des performances dynamiques.

comportementale du CAN est effectuée et le SINAD est calculé à partir de la FFT de sa sortie. Un exemple de domaine pour la variation des paramètres $I = I_{ds}$ et W est donné sur la figure 10.8 ainsi que quelques caractéristiques du transistor pour une technologie CMOS $L_{min} = 0,13\mu m$. Les caractéristiques du modèle d'amplificateur étant connues, il nous reste à faire le choix de la capacité minimum utilisée pour le CAN. Celui-ci conditionne la consommation, le bruit thermique et la dispersion sur les rapports de capacités (équation 5.13). Nous supposerons dans les exemples suivants qu'une capacité minimum $C_u = 0,1pF$ est suffisante pour garantir la précision du convertisseur ou qu'un calibrage est effectué. Pour garantir un niveau de bruit thermique et de quantification comparable pour l'ensemble du CAN, on pourra cependant être amené à choisir localement une valeur de capacité supérieure. La tension de référence conditionne également l'amplitude du bruit thermique. Une valeur de 1 V a

[2] [59] : *Analog and Mixed-signal Devices Technology Requirements-Near–term*

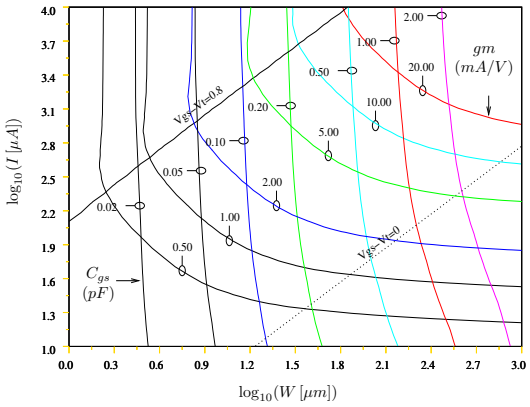


FIG. 10.8 – Caractéristiques technologiques CMOS $L_{min} = 0,13\mu m$

été retenue pour les simulations envisagées dans la suite de ce chapitre. Par ailleurs, la résistance des commutateurs est choisie suffisamment faible pour que son influence soit négligeable dans toutes ces simulations.

La figure 10.9 montre les contours à SINAD constant obtenus pour un modulateur du second ordre avec une quantification sur 4 bit. Les paramètres (I,W) sont identiques pour les deux intégrateurs. Le domaine de la figure 10.8 est couvert par une table de 20x20=400 points et le temps de simulation est de l'ordre d'une heure pour un ordinateur Intel Pentium III 750 MHz et 8192 échantillons par point pour le calcul du SINAD. Cette figure montre qu'il existe un couple (I,W) qui minimise

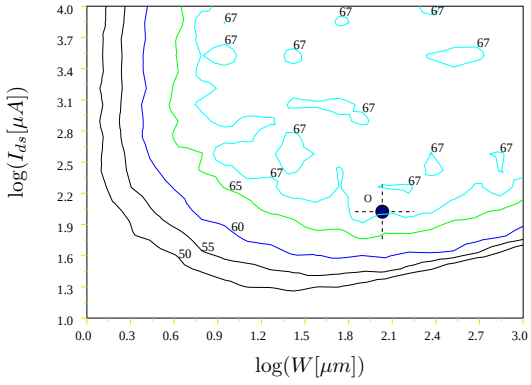


FIG. 10.9 – Contours à SINAD (dB) constant pour un modulateur du second ordre 4 bit

la consommation des amplificateurs pour un SINAD donné. Cette consommation est en effet proportionnelle à la somme des courants $I_{tot} = \sum_{i=1}^{m} I_{dsi}$. Les paramètres I et W peuvent être obtenus par une optimisation du SINAD sous la contrainte d'une consommation minimum. La solution envisagée consiste à minimiser (sans contraintes)

la fonction objectif :

$$f_{obj} = \lambda \sum_i I_i - SINAD \tag{10.9}$$

Le paramètre λ est choisi pour que les deux termes de la fonction objectif soient d'amplitudes comparables. Dans l'exemple précédent, on obtient ainsi le point **O** de la figure10.9.

Les sections suivantes sont consacrées à l'application de cette méthode pour des architectures de CAN ΣΔ et pipeline afin de déterminer les configurations les plus intéressantes du point de vue de la consommation.

10.4.3 Application au CAN ΣΔ

Nous avons regroupé dans le tableau 10.2 les structures les plus performantes des modulateurs considérés au chapitre 6 avec, pour chacun d'eux, le nombre de bit effectifs et le niveau maximum de l'entrée. L'exploration sera limitée à l'étude des configura-

ENOB	R					OL
	8	16	32	64	128	
ordre 2 (1 bit)	3,9	6,6	8,8	11,5	14	0,7
cascade 2-1 (1 bit)	5,5	9	12,5 (d)	15,8	18,9	0,75
cascade 2-1-1 (1 bit)	7	11,6 (a)	15,2	17,7	20,1	0,7
ordre 2 (4 bit)	8,4	10,9 (b)	13,4 (e)	15,9	18,4	0,98
ordre 3 (4 bit)	9,5	13 (c)	16,5	20	23,5	0,92

TAB. 10.2 – Performances des modulateurs ΣΔ

tions notées (a),(b),(c),(d) et (e) du tableau 10.2 qui correspondent a des résolutions intermédiaires de 10-13 bit avec un faible rapport de suréchantillonnage. Nous choisissons une fréquence d'échantillonnage F_s=100 MHz pour les configurations (a) à (c) et F_s=200 MHz pour (d) et (e). Ce choix permet de comparer ces solutions avec une bande passante identique $B = \frac{F_s}{2\,R} \approx 3,1\,MHz$. Pour chaque configuration les paramètres (I,W) associés au modèle de l'intégrateur sont déterminés afin de minimiser 10.9. Le courant total $I_{tot} = \sum_i I_i$ pour l'ensemble des intégrateurs est reporté dans le tableau 10.3 pour deux technologies. Le nombre effectif de bit obtenu à l'optimum est également indiqué entre parenthèses. Celui-ci est proche de la valeur théorique pour les

Configuration	Technologie	
	(1) L_{min}=0,13 μm	(2) L_{min}=0,25 μm
(a) F_s=100 MHz	0,3 (11,3)	0,4 (11,3)
(b) F_s=100 MHz	0,28 (10,8)	0,5 (10,8)
(c) F_s=100 MHz	0,24 (12,1)	0,37 (12)
(d) F_s=200 MHz	0,38 (12,4)	0,6 (12,2)
(e) F_s=200 MHz	0,3 (12,9)	0,5 (12,9)

TAB. 10.3 – I_{tot} (mA) et ENOB pour les modulateurs retenus

configurations (a),(b),(d) et (e). L'écart est cependant important pour la configuration (c) correspondant au modulateur du troisième ordre. Pour celui-ci, l'erreur introduite par le transfert incomplet des charges peut d'ailleurs conduire à une instabilité pour certains couples de paramètres (I,W). L'augmentation des performances dynamiques,

liées à la diminution de la longueur du transistor [3] conduit à une consommation plus faible pour la technologie (1). La réduction du courant total est de l'ordre de 40% par rapport à la technologie (2). La puissance consommée par les amplificateurs se déduit de ce courant total par :

$$P_{amp} = \eta \cdot V_{dd} \cdot I_{tot} \tag{10.10}$$

où η est un facteur qui dépend de l'architecture choisie pour l'amplificateur. Ce facteur est par exemple au minimum égal à quatre pour une structure telle que celle de la figure 7.4 (b). Avec cette valeur de η, une tension d'alimentation de 1 V et la technologie (1), on obtient la puissance donnée dans le tableau 10.4 pour les amplificateurs.

(a)	(b)	(c)	(d)	(e)
1,2	1,1	1	1,5	1,2

TAB. 10.4 – Puissance (mW) consommée par les amplificateurs

Dans les structures à capacités commutées, la puissance consommée par les quantificateurs est essentiellement liée à celle du CAN flash. Pour les modulateurs monobit le CAN flash se résume à un simple comparateur dynamique dont la consommation peut être négligée par rapport à celle des amplificateurs. Ce n'est pas le cas des architectures (b),(c) et (e) qui emploient une quantification de 4 bit. On a vu au chapitre 4 que l'énergie de conversion pour ces convertisseurs est actuellement de l'ordre de 1 pJ. Pour une fréquence d'échantillonnage de 100 MHz et une résolution de 4 bit, on a ainsi (équation 4.15) :

$$P_{flash} = E_c \cdot 2^{ENOB} \cdot F_{clk} = 1,6mW \tag{10.11}$$

En incluant la puissance liée au CAN flash dans le tableau 10.4, on obtient, pour le modulateur, la puissance totale $P_{tot} = P_{amp} + P_{flash}$ et l'énergie de conversion du tableau 10.5. Cette valeur de l'énergie de conversion est optimiste. D'autres sources de

	(a)	(b)	(c)	(d)	(e)
P_{tot} (mW)	1,2	2,7	2,6	1,5	4,4
E_c (pJ)	0.08	0.24	0.09	0.04	0.09

TAB. 10.5 – Puissance totale (mW) et énergie de conversion

consommation doivent en effet être considérées, telles que la consommation dynamique liée à la commande des interrupteurs ou la consommation statique nécessaire à la polarisation des amplificateurs. Les structures monobit apparaissent particulièrement intéressantes du point de vue de la consommation. Ce fait est essentiellement lié a la présence du CAN flash dont la fréquence d'horloge est imposée par celle du modulateur. L'efficacité de ce dernier, pour un ordre faible, est d'autant plus grande que R est élevé. On retrouve cette caractéristique avec la solution (d) qui conduit à l'énergie de conversion la plus faible.

Nous n'avons précédemment considéré que la puissance relative au modulateur. Une partie de la consommation est également associée au filtrage numérique en sortie du modulateur. Celle-ci peut être exprimée de la manière suivante [8] :

$$P_{dec} = K_t \cdot F_d \left[\sum_i nc_i \cdot \frac{Fs_i}{F_d} \cdot lw_i \right] \tag{10.12}$$

[3] On rappelle que celle-ci est prise égale à 5 L_{min}

Dans cette expression K_t est un paramètre technologique et F_d est la cadence des données en sortie du filtre. La somme entre crochets est étendue à tous les étages du filtre et dépend de la fréquence d'échantillonnage de l'étage Fs_i, du nombre de coefficients (nc_i) et de la longueur du mot de sortie (lw_i). Cette somme caractérise la complexité du filtre. Celle-ci dépend beaucoup du traitement qui est généralement effectué simultanément avec la décimation tel que le filtrage canal dans un récepteur radio [8]. Bien que la consommation liée au filtrage numérique soit actuellement comparable à celle du modulateur, l'évolution technologique conduit à une réduction croissante de cette dernière par l'intermédiaire du paramètre K_t qui est lié à l'énergie de transition d'un état logique.

10.4.4 Application au CAN pipeline

Il existe plusieurs topologies candidates pour une résolution donnée du pipeline. En effet, pour un convertisseur de n bit on doit simplement avoir $n = \sum_{i=1}^{m} n_i$ pour un convertisseur de m étages avec une résolution par étage n_i. Pour réduire le nombre de configurations possibles nous envisageons le cas où tous les étages ont la même résolution. On impose de plus que celle-ci soit inférieure ou égale à 4 pour limiter la complexité de l'étage.

Nous considérons successivement des convertisseurs pipelines de 10, 12 et 14 bit [4] de résolution avec des fréquences d'échantillonnage respectives de 100, 50 et 20 MHz. Ces caractéristiques sont choisies pour garder des paramètres (I,W) des transistors dans le domaine définit par la figure 10.8. Le calcul de la capacité de charge pour la construction du modèle est effectué comme indiqué à l'annexe F. Pour réduire le nombre de variables à optimiser, lorsque le nombre d'étages du pipeline est supérieur à 6, le courant de polarisation n'est plus distribué arbitrairement entre les étages mais est pondéré par le vecteur k_c utilisé pour les capacités (voir la section F.3).

Résolution de 10 bit : Pour cette résolution une valeur de capacité unitaire de 0,1pF peut être retenue pour tous les étages sans pondération. Dans la configuration la plus défavorable d'une résolution de 1 bit par étage, le bruit thermique donné par F.7 et F.8 est en effet inférieur d'environ 10 dB au bruit de quantification. Le courant total $I_{tot} = \sum_i I_i$ est reporté dans le tableau 10.6 pour deux technologies. Le nombre de bit effectifs obtenu à l'optimum est également indiqué entre parenthèses. Pour cette résolution, la configuration de 1 bit par étage est la plus performante en

Configuration	Technologie	
	(1) L_{min}=0,13 μm	(2) L_{min}=0,25 μm
1 x 10	0,9 (9,8)	3,6 (9,9)
2 x 5	1,2 (9,9)	3,8 (9.8)
2 x 3 + 4	3,6 (9,8)	-

TAB. 10.6 – I_{tot} (mA) et SINAD (dB) pour un pipeline 10 bit à F_s=100 MHz

terme de consommation des amplificateurs. On note une réduction d'un facteur 4 de la consommation avec la technologie pour cette architecture.

Résolution de 12 bit : Lorsque la résolution par étage est faible, la capacité unitaire $C_u = 0,1\,pF$ utilisée conduit à un bruit thermique important. En effet d'après

[4] Pour toutes ces configuration, nous utiliserons la redondance dans le codage

la formule F.8, l'utilisation de la capacité minimum pour tous les étages conduit aux rapports donnés dans la ligne (a) du tableau 10.7 entre le bruit thermique et le bruit de quantification. Pour réduire l'influence du bruit thermique pour les configurations

bit/étage	1	2	3	4
(a)	1,39	0,55	0,26	0,13
(b)	0,35	0,29	0,26	0,13

TAB. 10.7 – Influence du bruit thermique en fonction de la résolution

à 1 et 2 bit par étage, deux vecteurs de pondération kc (formule F.8) égaux respectivement à $kc_1 = 8, 4, 2, 1, 1, .., 1$ et $kc_2 = 2, 1, 1, .., 1$ sont utilisés. Le bruit thermique correspondant (normalisé par rapport au bruit de quantification) est donné à la ligne (b) du tableau 10.7. Le courant total ainsi que le ENOB obtenus après optimisation sont donnés dans le tableau 10.8. On remarque que c'est la configuration à deux bit

Configuration	Technologie	
	(1) L_{min}=0,13 μm	(2) L_{min}=0,25 μm
1 x 12	1,3 (11,9)	3 (11,8)
2 x 6	1 (11,8)	2,4 (11,6)
3 x 4	2 (11,7)	5,5 (11,5)
4 x 3	7,1 (11,5)	-

TAB. 10.8 – Courant total (mA) pour un pipeline 12 bit à F_s=50 MHz

par étage qui conduit cette foi à la consommation la plus faible pour les amplificateurs. Etant donné la faible résolution du CAN flash associé à chaque étage du pipeline, on peut considérer que la consommation des amplificateurs est prépondérante devant celle des CAN flash.

Résolution de 14 bit : Pour les mêmes raisons qu'avec la résolution précédente, la réduction du bruit thermique nécessite l'application d'une pondération sur les capacités. Les vecteurs utilisés pour les configurations à 1, 2 et 3 bit par étage sont les suivants :

$$kc_1 = \{64, 32, 16, 8, 4, 2, 1, .., 1\}\ , \quad kc_2 = \{16, 8, 4, 2, 1, 1\}\ , \quad kc_3 = \{8, 4, 2\}$$

Le courant total pour ces configurations est reporté dans le tableau 10.9. La dernière architecture fait exception au choix d'une résolution identique par étage qui n'est pas applicable dans ce cas particulier. Pour cette résolution c'est encore la configuration

Configuration	Technologie	
	(1) L_{min}=0,13 μm	(2) L_{min}=0,25 μm
1 x 14	3 (13,5)	7,3 (13,5)
2 x 7	1,6 (13,5)	3,3 (13,6)
4-3-3-4	3 (13,6)	9,8 (13,6)

TAB. 10.9 – Courant total (mA) pour un pipeline 14 bit à F_s=20 MHz

à deux bit par étage qui conduit à la consommation la plus faible. La réduction de la consommation avec la technologie n'est plus que d'un facteur 2. Ceci est justifié par

l'utilisation de capacités plus importantes pour réduire le bruit thermique. Celles-ci masquent en partie l'amélioration des performances liées à la réduction des dimensions.

Si nous regroupons les configurations qui conduisent à la consommation minimale, nous obtenons pour la technologie (1) et V_{dd}=1 V le tableau 10.10. La puissance

n	F_s (MHz)	n_{pe}	ENOB	P (mW)	E_c (pJ)
10	100	1	9,8	3,6	0,04
12	50	2	11,8	4	0,02
14	20	2	13,5	6,4	0,03

TAB. 10.10 – Caractéristiques des configurations pipeline

consommée par les amplificateurs a été déterminée à partir de la formule 10.10 avec η=4. n_{pe} est le nombre de bit par étage qui conduit à la consommation minimale.

Nous obtenons une énergie de conversion très faible en regard des réalisations actuelles. Nous avons cependant négligé la consommation liée aux CAN flash et à la logique nécessaire pour la constitution du code de sortie à partir des codes intermédiaires. Celle-ci pénalise légèrement les solutions dont la résolution est plus élevée. L'évolution technologique entraîne par ailleurs une réduction croissante de la consommation des circuits logiques et des CAN flash de faible résolution (figure 4.15). Le tableau 10.6 montre également une réduction importante de la consommation avec la technologie pour les CAN pipeline de faible résolution. Cette réduction est bien moindre pour le modulateur $\Sigma\Delta$ avec une résolution équivalente. Elle fait du CAN pipeline un excellent candidat pour les applications de résolution moyenne et qui exigent une bande passante importante.

Outre l'exploration des performances dynamiques de différentes architectures, la méthode précédente fournit un premier dimensionnement de l'amplificateur. Certes, celui-ci est très sommaire puisqu'il se résume a une estimation des dimensions des transitors de l'étage d'entrée et à leur courant de polarisation. Il faut en effet prendre en compte le fait que le modèle utilisé pour l'amplificateur est simplifié. Ainsi, une simulation électrique pour le pipeline 10 bit précédent où seule l'étage d'entrée de l'amplificateur est modélisé au niveau transistor conduit à un SINAD de 59,6 dB et une distorsion de 67,3 dB. La simulation comportementale conduit quant à elle aux valeurs respectives de 61 dB et 71 dB. Cet écart est le prix à payer pour un temps de simulation raisonnable. Ce dernier est en effet de quelques secondes pour la simulation comportementale. Il est plus de 100 fois supérieur pour la simulation électrique, ce qui la rend inutilisable pour l'optimisation précédente. Elle devient exploitable à partir de ce premier dimensionnement de l'amplificateur pour en préciser les spécifications et permettre sa synthèse.

Conclusions et perspectives

Conclusions

La place des CAN dans les circuits mixtes est particulièrement importante car elle fixe la frontière entre le traitement analogique et le traitement numérique. L'efficacité toujours plus grande de ce mode de traitement, liée à la réduction des dimensions, tend à repousser cette frontière le plus possible en plaçant des contraintes très fortes sur les CAN en termes de vitesse, de résolution et de consommation. La réduction de la tension d'alimentation et la perte de précision occasionnées par la réduction des dimensions ne sont pas sans conséquences sur la conception des blocs analogiques. Dans ce contexte, les architectures de CAN étudiés dans la seconde partie de cette thèse se sont avérées robustes. Le filtrage spatial pour le flash, la redondance dans le codage pour le pipeline et la modulation du bruit pour le $\Sigma\Delta$ sont des exemples de techniques très efficaces pour obtenir de bonnes performances dans cet environnement.

Ces dernières années, de nombreuses recherches ont été menées pour la synthèse analogique au niveau des blocs. Dans le domaine des CAN, des outils de synthèse dédiés a une architecture particulière de convertisseur ont également été développés[5]. Le choix d'une architecture reste cependant guidé par une partition de l'espace résolution-vitesse dont les frontières évoluent très rapidement avec la technologie. L'exemple le plus marquant de cette évolution est sans doute le convertisseur $\Sigma\Delta$ qui couvre un domaine de plus en plus grand de cet espace. Il est bien sûr possible de guider son choix à partir de l'expérience acquise, en nous appuyant par exemple sur les réalisations récentes de CAN CMOS présentés dans la seconde partie. Ce choix risque néanmoins d'être très conservatif dans un environnement technologique qui évolue rapidement. L'exploration de l'espace de conception au niveau transistor, qui est efficace au niveau d'une cellule (amplificateur, comparateur,...) n'est pas envisageable au niveau d'un bloc tel qu'un CAN étant donné le temps de simulation très important qu'il nécessiterait. Nous avons pour cela introduit une méthode de simulation comportementale à partir de classes C++ pour les CAN échantillonnés. Celle-ci utilise différents niveaux d'abstraction.

Le premier niveau de modélisation introduit, spécifique du modulateur $\Sigma\Delta$, est lié au quantificateur. Nous avons montré que l'on peut extraire efficacement, par simulation, un modèle linéaire de cet élément. Celui-ci a été utilisé pour l'optimisation des coefficients du modulateur sous des contraintes d'excursion réduite des états d'intégrateur. L'utilisation de tensions d'alimentation toujours plus faibles impose en effet de limiter au maximum cette excursion. Ce modèle fournit également un moyen d'analyse rapide pour les effets linéaires et l'étude statistique. Cette dernière étant particulièrement importante pour le choix d'une architecture.

Les autres niveaux de modélisation introduits sont propres aux circuits à capacités commutées. Il sont particulièrement intéressants pour l'étude des effets non linéaires

[5] On trouvera un tableau récapitulatif de ces outils dans [78]

qui n'admettent pas de modélisation analytique simple. Les effets non linéaires dynamiques sont les plus délicats à évaluer étant donné leur lien étroit avec la technologie. Ils conditionnent pourtant des paramètres fondamentaux tel que la fréquence d'échantillonnage ou la consommation. un modèle simple basé sur la conservation de la charge a été introduit. Il permet un gain très important dans le temps de simulation. Il a été appliqué à différentes configurations de convertisseurs $\Sigma\Delta$ et pipeline pour mettre en évidence les configurations les plus intéressantes du point de vue de la consommation. Pour le modulateur $\Sigma\Delta$, les structures monobit s'avèrent plus intéressantes de ce point de vue. Ceci est lié à l'utilisation du CAN flash dont la fréquence d'horloge est imposée par celle du modulateur. Pour le CAN pipeline, les solutions utilisant une résolution réduite par étage (1 à 2 bit) combinent une grande bande passante et une faible consommation. Cette structure bénéficie beaucoup de la réduction de la longueur de canal du transistor MOS. L'estimation de la consommation laisse espérer une énergie de conversion E_c inférieure à 1 pJ avec les technologies actuelles. De tels niveaux de consommation sont déjà accessibles pour de faibles fréquences d'échantillonnage [112]. Il reste à montrer la possibilité d'obtenir ce niveau de consommation dans le domaine des applications de la figure 1.

Perspectives

Un prolongement naturel de ce travail est de définir une méthode de synthèse complète d'une architecture de CAN à partir du modèle comportemental obtenu. La spécification des sous blocs nécessite d'introduire des niveaux de modélisation supplémentaires pour prendre en compte des effets secondaires ignorés par la simulation comportementale, dont le but initial est de fournir une méthode d'évaluation rapide des performances qui permettre l'optimisation de paramètres. Dans cette voie, des outils tels que VHDL-AMS ou Verilog-AMS sont mieux adaptés et plus généraux. La production automatique d'une description structurelle dans ces langages avec un premier niveau de modélisation est préférable. Cette description pouvant ensuite être affinée pour permettre la vérification, lors des différentes étapes de la synthèse.

La technique de simulation proposée est propre aux circuits échantillonnés. Elle a été exploitée pour les circuits à capacités commutées qui constituent la grande majorité des réalisations actuelles. Elle peut être étendue à d'autres techniques de traitement échantillonné tel que celle des courants commutés par l'utilisation d'une modélisation comportementale spécifique. Ce type de traitement est bien adapté aux technologies CMOS VLSI puisqu'il n'utilise que des transistors et tolère une plus faible tension d'alimentation.

Annexes

Annexe A

Couplage dans le convertisseur flash

Les capacités d'entrée des amplificateurs introduisent un couplage capacitif entre le signal d'entrée et le réseau de référence. Ces capacités étant dépendantes des potentiels d'entrée des amplificateurs, le circuit résultant est en fait non-linéaire. Afin d'estimer l'effet du couplage nous supposerons néanmoins le circuit linéaire constitué des capacités nominales d'entrée des amplificateurs et du réseau de résistances. On peut ainsi étudier seul l'effet de ce couplage et ensuite ajouter ces perturbations aux potentiels de référence (figure A.1).

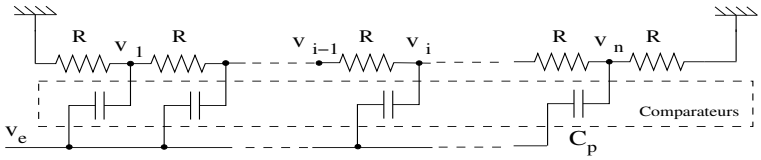


FIG. A.1 – Couplage capacitif dans le flash.

A partir du schéma de la figure A.1 et en notant v_i une perturbation particulière en un point de l'échelle résistive on a :

$$C_p \frac{d(v_i - v_e)}{dt} = \frac{v_{i-1} - vi}{R} + \frac{v_{i+1} - vi}{R} \tag{A.1}$$

En normalisant le temps avec $t^{'} = RC_p t$ on a :

$$\frac{dv_i}{dt'} = v_{i-1} + v_{i+1} - 2v_i + \frac{dv_e}{dt'} \tag{A.2}$$

L'équation d'état du système s'écrit :

$$\begin{bmatrix} \frac{dv_1}{dt'} \\ \vdots \\ \frac{dv_n}{dt'} \end{bmatrix} = A \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} + \begin{bmatrix} \frac{dv_e}{dt'} \\ \vdots \\ \frac{dv_e}{dt'} \end{bmatrix} \tag{A.3}$$

avec

$$A = \begin{bmatrix} -2 & 1 & 0 & \cdots & 0 \\ 1 & -2 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & -2 & 1 \\ 0 & \cdots & 0 & 1 & -2 \end{bmatrix}$$

La figure A.2 montre l'évolution des perturbations sur les potentiels de référence lorsque l'entrée est un échelon de 1 V pour un convertisseur de 4 bit.

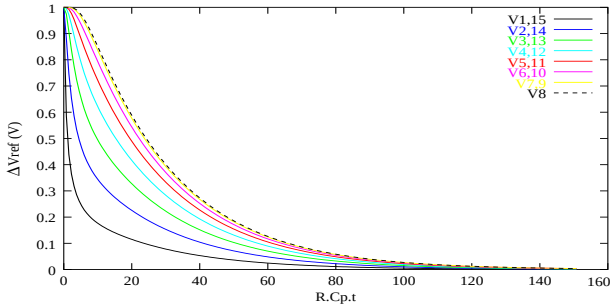


FIG. A.2 – Perturbations sur les tensions de référence.

En notant $\{\lambda_k\}$ l'ensemble des valeurs propres de la matrice A, la réponse libre du réseau précédent s'écrit :

$$v(t^{'}) = \sum_{k=1}^{m} v_k.exp(\lambda_k.t^{'}) = \sum_{k=1}^{m} v_k.exp(\frac{t^{'}}{\tau_k}) \tag{A.4}$$

m étant l'ordre de la matrice.

La constante de temps maximum du circuit est obtenue à partir de la valeur absolue minimale $|\lambda_{min}|$ des valeurs propres. Celle-ci donne une indication sur le temps nécessaire pour que les perturbations sur les potentiels de référence puissent être considérées comme négligeables.

Les valeurs propres de la matrice A sont données par [70] :

$$\lambda_k = -4\,sin^2\frac{k\pi}{2(m+1)} \qquad avec \qquad k = 1, ..., m \tag{A.5}$$

d'où la valeur $\lambda_{min} = -4\,sin^2\frac{\pi}{2(m+1)}$. On peut exprimer cette valeur en fonction du nombre de bit n du convertisseur sachant que dans ce cas on a $m = 2^n - 1$ potentiels de référence on a $\lambda_{min} = -4\,sin^2\frac{\pi}{2^{n+1}}$. A partir de n=4 bit cette valeur peut être approchée par $\lambda_{min} \simeq \frac{\pi^2}{2^{2n}}$. La constante de temps maximum (normalisée par rapport à RC_p) correspondante est alors :

$$\tau_{max} = \frac{2^{2n}}{\pi^2} \tag{A.6}$$

Cette valeur est reportée dans le tableau A.1 pour des résolutions entre 4 et 8 bit.

Pour qu'une perturbation égale à l'amplitude pleine échelle du convertisseur ($A = \frac{2^n q}{2}$) donne une erreur inférieure au quantum on doit avoir :

$$e^{-\frac{T_s}{\tau}} < 2^{-n} \tag{A.7}$$

n	4	5	6	7	8
τ_{max}	25,9	103,8	415	1660	6640

TAB. A.1 – Constante de temps dominante

D'où la contrainte sur la constante de temps RC_p par rapport à la période T_s de conversion :

$$\frac{RC_p}{T_s} < \frac{\pi^2}{n2^{2n}ln(2)} \tag{A.8}$$

Annexe B

Codage redondant des nombres

B.1 Ecriture des nombres avec chiffres signés

Ce système de représentation des nombres a été introduit par Avizienis [4] pour accélérer les opérations arithmétiques dans les calculateurs. La redondance introduite par cette notation permet par exemple de limiter la propagation de la retenue lors de l'addition de deux nombres, le temps nécessaire à cette opération étant ainsi indépendant de la longueur des mots. Nous introduisons ici cette représentation qui constitue une extension de la numération simple de position.

B.1.1 Numération simple de position

C'est la manière habituelle de représenter un nombre avec une suite de symboles étant donné une base. Si $b \geq 2$ et $n \geq 1$ (n entier) alors tout nombre entier $x \in [0, b^n - 1]$ peut s'écrire de façon unique sous la forme :

$$x = \sum_{i=0}^{n-1} d_i b^i \tag{B.1}$$

$(d_0, d_1, ..., d_{n-1})$ est la décomposition de x en base b.
Elle est notée : $d_{n-1}d_{n-2}...d_1d_0$.

B.1.2 Extension à des chiffres signés

On considère le système de nombres $S\{b, \alpha, \beta, n\}$ de base b constitués de n chiffres pris dans l'ensemble $[\alpha, \alpha+1, ..., \alpha+\beta]$. α est un entier relatif quelconque alors que β est strictement positif. L'écriture précédente (B.1) est un cas particulier avec $\alpha = 0$ et $\beta = b - 1$.

Soit $N\{b, \alpha, \beta, n\}$ l'ensemble des nombres que l'on peut écrire dans ce système. On peut distinguer trois cas selon la valeur de β :

$\boxed{\beta < b - 1}$ On a des "trous" dans le système de numération : entre deux éléments de N, on peut trouver un entier qui n'appartient pas à N. En effet

$$N \in [\alpha(b^n - 1)/(b - 1), (\alpha + \beta)(b^n - 1)/(b - 1)]$$

et contient au plus $(\beta + 1)^n$ éléments. Or entre ces deux bornes on a

$$\beta(b^n - 1)/(b - 1) + 1$$

éléments, nombre strictement supérieur à $(\beta+1)^n$ pour $\beta < b-1$.
Ce système de numération est donc peu satisfaisant.

$\boxed{\beta = b-1}$ N est égal à l'ensemble des entiers appartenant à

$$[\alpha(b^n-1)/(b-1), (\alpha+b-1)(b^n-1)/(b-1)]$$

et l'écriture de chacun de ces entiers est unique. On a dans ce cas une bijection entre un nombre x dans $S\{b, \alpha, b-1, n\}$ et un nombre

$$x - \alpha(b^n-1)/(b-1)$$

écrit dans $S\{b, 0, b-1, n\}$.

$\boxed{\beta > b-1}$ N est égal à l'ensemble des entiers appartenant à

$$[\alpha(b^n-1)/(b-1), (\alpha+\beta)(b^n-1)/(b-1)]$$

mais l'écriture de ces entiers n'est plus unique : le système est dit redondant.

Ceci nous montre que pour écrire un nombre en base b il faut utiliser au moins b chiffres consécutifs. Une écriture unique des nombres nécessite exactement b chiffres consécutifs.

symétrie : Pour coder les nombres négatifs on doit avoir $\alpha < 0$. Pour obtenir facilement l'opposé d'un nombre on peut prendre l'opposé de chaque chiffre ce qui nécessite une symétrie de l'ensemble des chiffres :

$$-m, -m+1, -m+2, ..., 0, 1, ..., m-1, m$$

Comparaison : On doit pouvoir facilement décider du signe d'un entier dans le système $S\{b, -m, 2m, n\}$. Soit un entier $x = (d_{n-1}, d_{n-2}, ...d_0)$ avec d_k premier chiffre non nul en partant de la gauche. Pour que le signe de x soit celui de d_k, on doit avoir :

$$|d_k|b^k > \sum_{i=0}^{k-1} |d_i|b^i$$

Ceci est vrai quels que soient les d_j pour :

$$b^k > \sum_{i=0}^{k-1} mb^i = m\frac{b^k-1}{b-1}$$

Une condition nécessaire et suffisante de comparaison en utilisant le premier chiffre non nul est obtenue pour :

$$m \leq b-1$$

En conclusion, les exigences de symétrie et de comparaison imposent les restrictions suivantes sur l'écriture des nombres avec des chiffres signés :

$$d_i \in \{-m, -m+1, -m+2, ..., 0, 1, ..., m-1, m\} \qquad \text{(B.2)}$$

avec

$$\tfrac{b-1}{2} \leq m \leq b-1$$

Le rapport $\rho = \frac{m}{b-1}$ est appelé facteur de redondance. Avec les restrictions précédentes, on a :

$$\frac{1}{2} \leq \rho = \frac{m}{b-1} \leq 1 \qquad \text{(B.3)}$$

Si $\rho = \frac{1}{2}$ le système est dit non redondant. Pour $\rho = 1$, la redondance est maximale.

B.2 Algorithme de division et chiffres signés

La relation générale de récurrence pour obtenir les différents éléments d'un quotient de m chiffres est donnée par :

$$r_{j+1} = b \cdot r_j - q_{j+1} \cdot d \tag{B.4}$$

j	indice de la récurrence : $j = 0, 1, ..., m-1$
r_j	reste partiel ou résidu au cycle j
r_0	dividende
r_n	reste
q_j	$j^{ème}$ chiffre du quotient : $q = q_0.q_1q_2...q_m$
d	diviseur
b	base

Dans le cas pratique de représentation binaire, on a $b = 2^n$ et le diviseur est donné sous forme normalisée : $\frac{1}{2} \leq d < 1$. Après m itérations on obtient la valeur approchée du quotient :

$$q = \sum_{j=1}^{n} q_j \cdot b^{-j} \tag{B.5}$$

Nous nous intéressons ici au cas où les chiffres du quotient sont signés et aux conditions qui font que cette récurrence converge, c'est à dire que le reste partiel reste borné indépendamment de l'indice j de l'itération.

B.2.1 Sélection d'un élément du quotient

On suppose qu'un élément de quotient q_j est dans l'ensemble D_m défini par :

$$D_m = \{-m, -m+1, ..., -1, 0, 1, ..., m-1, m\}$$

La sélection de q_{j+1} en fonction du résidu r_j et du diviseur d doit satisfaire aux deux propriétés [33] :

- Le résidu doit rester borné indépendamment de j :

$$\underline{R} \leq r_j \leq \bar{R}$$

- Pour chaque valeur de $b \cdot r_j$ on doit pouvoir sélectionner un élément de quotient dans D_m qui respecte la borne précédente.

Bornes sur le résidu et intervalle de sélection

On note $[L_k, U_k]$ l'intervalle de $b \cdot r_j$ tel que $q_{j+1} = k$ avec un résidu borné :

$$L_k \leq b \cdot r_j \leq U_k \quad \Rightarrow \quad \underline{R} \leq r_{j+1} = b \cdot r_j - k \cdot d \leq \bar{R} \tag{B.6}$$

Cette relation est satisfaite avec :

$$\bar{R} = U_k - k \cdot d \qquad \underline{R} = L_k - k \cdot d \tag{B.7}$$

D'autre part, les bornes extrêmes du domaine D_m imposent :

$$U_m = b \cdot \bar{R} \quad \text{et} \quad L_m = b \cdot \underline{R} \tag{B.8}$$

Les relations B.7 et B.8 conduisent aux bornes suivantes pour le résidu :

$$\bar{R} = \frac{m}{b-1}d = \rho \cdot d \qquad \underline{R} = \frac{-m}{b-1}d = -\rho \cdot d \tag{B.9}$$

En résumé, l'intervalle de sélection pour $q_{j+1} = k$ qui donne un résidu borné est tel que :

$$L_k = (-\rho + k) \cdot d \leq b \cdot r_j \leq (\rho + k) \cdot d = U_k \tag{B.10}$$

La représentation graphique de la fonction de sélection d'un élément de quotient est donnée à la figure B.1. Cette représentation porte le nom de diagramme de Robertson. Elle permet de déterminer à chaque itération de la division la valeur du résidu à l'étape suivante. On a également fait apparaître sur ce diagramme les bornes de sélection $[L_k, U_k]$ qui correspondent au choix de l'élément de quotient $q_{j+1} = k$.

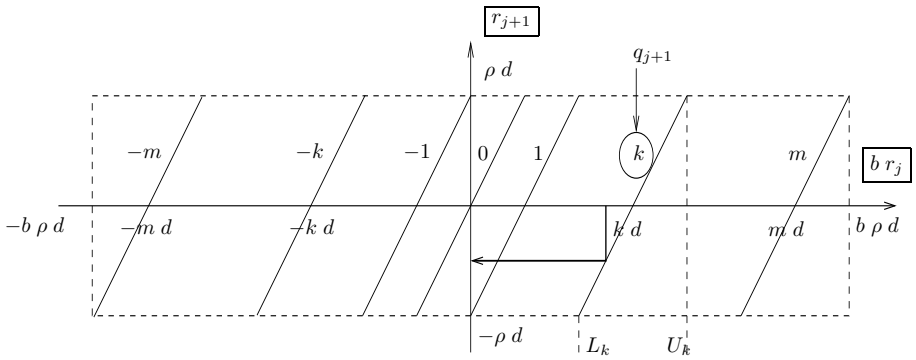


FIG. B.1 – Diagramme de Robertson.

Existence d'un élément de quotient et facteur de redondance

Il est évident, d'après cette figure, que pour garantir l'existence d'un élément de quotient pour toute valeur de $b \cdot r_j$ on doit avoir :

$$U_{k-1} \geq L_k \tag{B.11}$$

D'après la relation B.10, cette condition est satisfaite pour $\rho \geq \frac{1}{2}$ qui correspond à la limite d'un système de numération avec redondance. On obtient également la zone de recouvrement qui autorise le choix de deux valeurs possibles pour un élément du quotient :

$$U_{k-1} - L_k = (k - 1 + \rho)d - (k - \rho)d = (2\rho - 1)d \tag{B.12}$$

L'importance de cette zone est directement liée au facteur de redondance. Elle est nulle pour un système non redondant et égale à d lorsque la redondance est maximale.

Annexe C

Fonctions de transfert des modulateurs $\Sigma\Delta$

C.1 Fonctions de transfert des modulateurs à un seul quantificateur

C.1.1 Fonctions de transfert complètes

Le schéma général est donné à la figure C.1 :

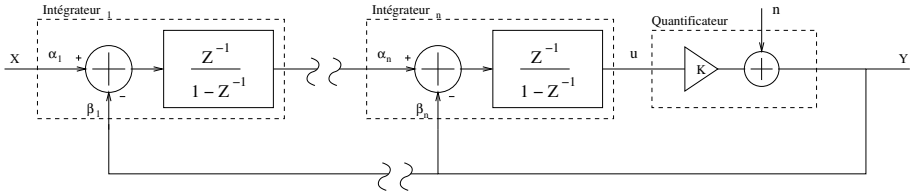


FIG. C.1 – Modulateur simple d'ordre n.

Fonction de transfert de l'intégrateur

$$X_i = \frac{a_i z^{-1}}{1 - b_i z^{-1}}[\alpha_i e^+ - \beta_i e^-]$$

Fonction de transfert du quantificateur (modèle linéaire)

$$y = Ku + n$$

Forme générale des fonction de transfert pour le signal et pour le bruit

$$STF = \frac{N_s}{D} \qquad NTF = \frac{N_n}{D}$$

Pour tous les modulateurs :

$$N_s = K \prod_{i=1}^{L} a_i \alpha_i$$

où L est le nombre d'intégrateurs

Modulateur du premier ordre

$$N_n = 1 - b1z^{-1} \qquad D = 1 + (Ka_1\beta_1 - b_1)z^{-1} \tag{C.1}$$

Modulateur du second ordre

$$\begin{gathered} N_n = 1 - (\textstyle\sum_{i=1}^{2} b_i)z^{-1} + (\prod_{i=1}^{2} b_i)z^{-2} \\ D = 1 + [Ka_2\beta_2 - \textstyle\sum_{i=1}^{2} b_i]z^{-1} + (Ka_2W + \prod_{i=1}^{2} b_i)z^{-2} \\ W = a_1\alpha_2\beta_1 - b_1\beta_2 \end{gathered} \tag{C.2}$$

Modulateur du troisième ordre

$$\begin{gathered} N_n = 1 - (\textstyle\sum_{i=1}^{3} b_i)z^{-1} + \theta z^{-2} - (\prod_{i=1}^{3} b_i)z^{-3} \\ \theta = b_1b_2 + b_2b_3 + b_1b_3 \\ D = 1 + [Ka_3\beta_3 - \textstyle\sum_{i=1}^{3} b_i]z^{-1} + (Ka_3R + \theta)z^{-2} \\ +(Ka_3U - \textstyle\prod_{i=1}^{3} b_i)z^{-3} \\ R = a_2\alpha_3\beta_2 - (b_1 + b_2)\beta_3 \qquad U = a_2\alpha_3W + b_1b_2\beta_3 \end{gathered} \tag{C.3}$$

Modulateur du quatrième ordre

$$\begin{gathered} N_n = 1 - (\textstyle\sum_{i=1}^{4} b_i)z^{-1} + \gamma z^{-2} - \delta z^{-3} + (\prod_{i=1}^{4} b_i)z^{-4} \\ \gamma = (b_1 + b_2 + b_3)b_4 + \theta \qquad \delta = b_4\theta + b_1b_2b_3 \\ D = 1 + [Ka_4\beta_4 - \textstyle\sum_{i=1}^{4} b_i]z^{-1} + (Ka_4T + \gamma)z^{-2} \\ +[Ka_4(\beta_4\theta + a_3\alpha_4R) - \delta]z^{-3} \\ +(Ka_4V + \textstyle\prod_{i=1}^{4} b_i)z^{-4} \\ V = a_3\alpha_4U - b_1b_2b_3\beta_4 \qquad T = a_3\alpha_4\beta_3 - \beta_4(b_1 + b_2 + b_3) \end{gathered} \tag{C.4}$$

C.1.2 Fonctions de transfert simplifiées

Les expressions suivantes donnent les fonctions de transfert précédentes lorsque les intégrateurs sont considérés comme parfaits ($a_i = b_i = 1$).

Modulateur du premier ordre

$$NTF = \frac{(1 - z^{-1})}{1 + (K\beta_1 - 1)z^{-1}} \tag{C.5}$$

Modulateur du second ordre

$$NTF = \frac{(1-z^{-1})^2}{1+(K\beta_2-2)\,z^{-1}+(K\,(\alpha_2\beta_1-\beta_2)+1)z^{-2}} \tag{C.6}$$

Modulateur du troisième ordre

$$NTF = \frac{(1-z^{-1})^3}{1+(K\beta_3-3)z^{-1}+(K\,r+3)z^{-2}+(K\,s-1)z^{-3}} \tag{C.7}$$

$$r = \alpha_3\beta_2 - 2\beta_3 \qquad s = \alpha_3(\alpha_2\beta_1-\beta_2)+\beta_3$$

Modulateur du quatrième ordre

$$NTF = \frac{(1-z^{-1})^4}{1+(K\beta_4-4)\,z^{-1}+(K\,t+6)\,z^{-2}+(K\,u-4)\,z^{-3}+(K\,v+1)z^{-4}} \tag{C.8}$$

$$t = \alpha_4\beta_3 - 3\,\beta_4 \qquad u = 3\,\beta_4 + \alpha_4\,(\alpha_3\beta_2 - 2\,\beta_3)$$

$$v = \alpha_4\,(\alpha_3\,(\alpha_2\beta_1-\beta_2)+\beta_3)-\beta_4$$

Annexe D

Optimisation convexe

Un grand nombre de mesures de performances peuvent s'exprimer sous la forme de fonctions particulières des variables de conception dites *posynômiales* [29]. Si x est un vecteur positifs de dimension n ($x \in \mathcal{R}_n^+$), une telle fonction s'exprime de la manière suivante :

$$f(x) = \sum_k \left(c_k \prod_n x_n^{a_{nk}} \right) \tag{D.1}$$

où $c_k \in \mathcal{R}^+$ et $a_{nk} \in \mathcal{R}$. Ces fonctions sont fermées pour l'addition et la multipliction. Le problème d'optimisation peut alors s'exprimer par la minimisation de la fonction objectif $f_0(x)$ sous les contraintes :

$$f_i(x) = \sum_k \left(c_{ik} \prod_n x_n^{a_{nk}} \right) \leq 1 \qquad g_j(x) = c_j \prod_m x_m^{a_{mj}} = 1 \tag{D.2}$$

où la fonction objectif et les contraintes sont sous forme posynômiale. La contrainte égalité est particulière : elle ne possède qu'un seul terme de la somme D.1. Ce cas particulier de fonction est également fermé pour la division.

Le problème ainsi définit n'est pas de type convexe mais on peut facilement s'y ramener par le changement de variable $y_i = \log x_i$ et $b_{ik} = \log c_{ik}$ qui conduit à :

$$\begin{aligned} \text{minimiser} \quad & p_0(y) = \log \sum_k \exp(a_{0k}^T y + b_{0k}) && \\ \text{sous les contraintes} \quad & p_i(y) = \log \sum_k \exp(a_{ik}^T y + b_{ik}) && \leq 0 \\ & a_j^T y + b_j && = 0 \end{aligned} \tag{D.3}$$

Comme le logarithme d'une somme d'exponentielles est une fonction convexe, la fonction objectif et les contraintes le sont également, ainsi que l'ensemble des points qui satisfont D.3.

Une propriété essentielle de ce type de problème et que lorsqu'un point x^* constitue un minimim local, il est également un minimum global. Ceci a comme conséquences que la résolution du système fournit le point optimum s'il existe et ceci indépendamment du point de départ. Il existe de plus des méthodes récentes très performantes pour sa résolution sous la forme D.2 qui permettent de traiter plusieurs centaines de variables [29].

Annexe E

Modèle dynamique de l'intégrateur à capacités commutées

Bien qu'une structure différentielle soit généralement adoptée, la modélisation sera basée sur le schéma de principe non différentiel de la figure E.1. Les conséquences de cette simplification sont que les grandeurs relatives à l'amplificateur différentiel qui interviendront dans ce modèle doivent être considérées sur le demi circuit associé à l'amplificateur. Durant la phase ϕ_1, la capacité d'échantillonnage C_1 est chargée

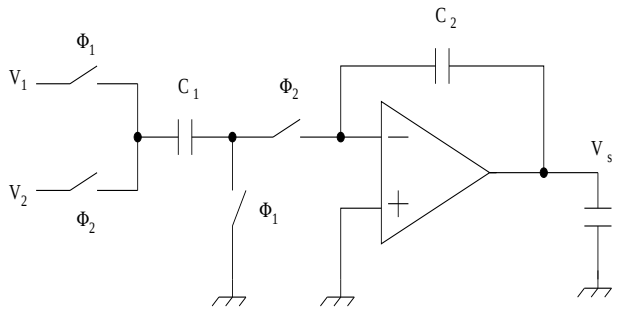


FIG. E.1 – Intégrateur à capacités commutées

à la tension V_1 et l'amplificateur maintient à sa sortie l'échantillon précédent. Nous appellerons cette phase la phase de maintien. Durant la phase ϕ_2, que nous appellerons phase de transfert, la charge $C_1(V_1 - V_2)$ est transférée dans la capacité C_2 ce qui conduit, dans le cas idéal et en supposant que les tensions V_1 et V_2 ne changent que sur ϕ_2, à la fonction de transfert :

$$V_s(z) = K \cdot \frac{z^{-1}}{1 - z^{-1}}[V_1(z) - V_2(z)] \quad \text{avec} \quad K = \frac{C_1}{C_2} \tag{E.1}$$

E.1 Intégrateur en phase de transfert

L'analyse de l'intégrateur dans la phase de transfert peut être effectuée à partir du schéma de la figure E.2. L'amplificateur est modélisé par une capacité d'entrée C_P,

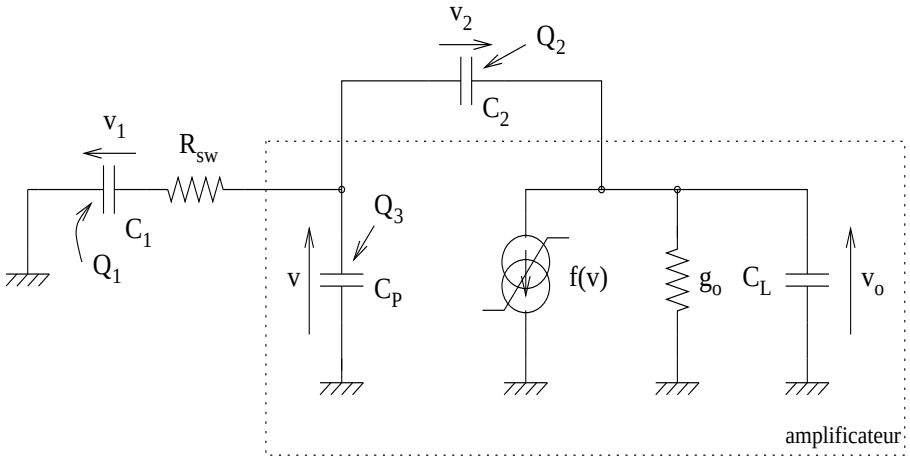


FIG. E.2 – Schéma équivalent de l'intégrateur en phase de transfert

une conductance de sortie g_o, une capacité de sortie C_L et un générateur de courant contrôlé dont la caractéristique non linéaire par parties est donnée à la figure E.3.

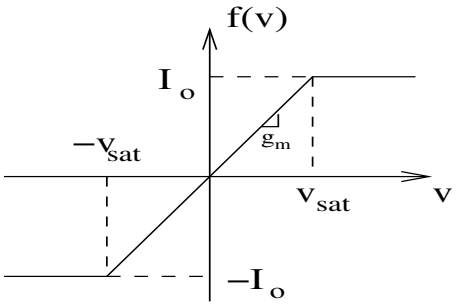


FIG. E.3 – modèle de la source de courant

E.1.1 Equations de conservations des charges

Les équations du circuit de la figure E.2 sont établies à partir des charges Q_1, Q_2 et Q_3 qui décrivent entièrement l'état de ce système du troisième ordre :

$$v_1 = \frac{Q_1}{C_1} \qquad v = \frac{Q_3}{C_P} \qquad v_o = \frac{Q_2}{C_2} + \frac{Q_3}{C_P}$$

$$\frac{dQ_1}{dt} = -\frac{Q_1}{R_{sw}C_1} - \frac{Q_3}{R_{sw}C_P} \qquad \frac{dQ_3}{dt} = \frac{dQ_1}{dt} + \frac{dQ_2}{dt}$$

$$f(v) + g_o v_o + C_L \frac{dv_o}{dt} + \frac{dQ_2}{dt} = 0 \tag{E.2}$$

A partir de ces relations de base on peut déterminer l'évolution du vecteur d'état $Q = [Q_1, Q_2, Q_3]^T$:

$$\frac{d}{dt}\begin{bmatrix} Q_1 \\ Q_2 \\ Q_3 \end{bmatrix} = M \begin{bmatrix} Q_1 \\ Q_2 \\ Q_3 \end{bmatrix} - \alpha f(\frac{Q_3}{C_P}) \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \tag{E.3}$$

avec

$$M = \begin{bmatrix} -\frac{1}{R_{sw}C_1} & 0 & -\frac{1}{R_{sw}C_P} \\ \frac{C_L C_2}{R_{sw}C_1 \Pi} & -\frac{g_o C_P}{\Pi} & \frac{-C_2(g_o R_{sw} C_P - C_L)}{R_{sw} C_P \Pi} \\ \frac{-C_P(C_2+C_L)}{R_{sw}C_1 \Pi} & -\frac{g_o C_P}{\Pi} & -\frac{C_2+C_L+g_o R_{sw} C_2}{R_{sw}\Pi} \end{bmatrix}$$

et

$$\Pi = C_2C_P + C_LC_P + C_LC_2 \qquad \alpha = \frac{C_2C_P}{\Pi}$$

E.1.2 Influence des paramètres R_{sw} et I_{sat}

Afin de déterminer l'influence de ces paramètres, la figure E.4 représente l'évolution de la tension de sortie pour une tension initiale de 1V et un gain fonctionnel K de 0,5 pour différentes valeurs de R_{sw} et avec les valeurs nominales données dans le tableau E.1.

R_{sw}	C_1	C_2	C_P	C_L	I_{sat}	V_{sat}	g_o
1 $k\Omega$	0,3 pF	0,6 pF	0,3 pF	0,3 pF	100 μA	0.2 V	2 μS

TAB. E.1 – Paramètres du circuit.

On note qu'une variation importante de la résistance du commutateur a peu d'effet sur la partie finale de la réponse. A g_m constant, la valeur finale est, par contre, fortement dépendante du courant de saturation (ou de la tension de saturation) comme on peut le voir sur la figure E.5.

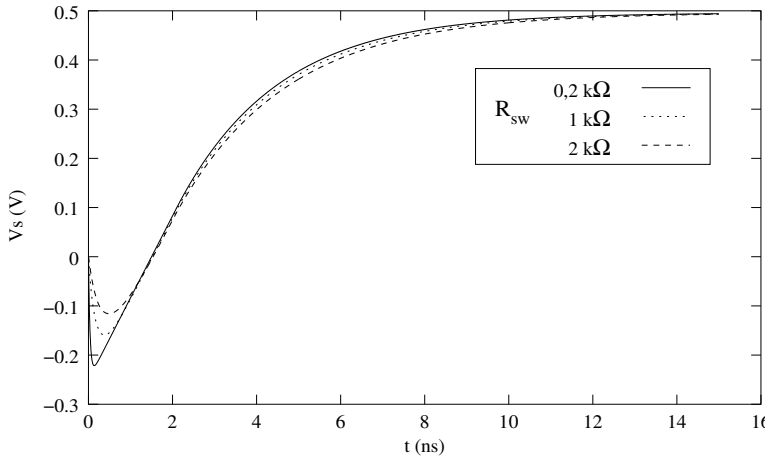


FIG. E.4 – Effet de la résistance du commutateur sur la réponse temporelle

E.1.3 Analyse simplifiée

Cette analyse va chercher à approcher le comportement non linéaire d'ensemble du transfert de charge en une succession de régimes élémentaires plus simples à modéliser. Le point central de cette analyse est la conservation de la charge dans la surface isolée

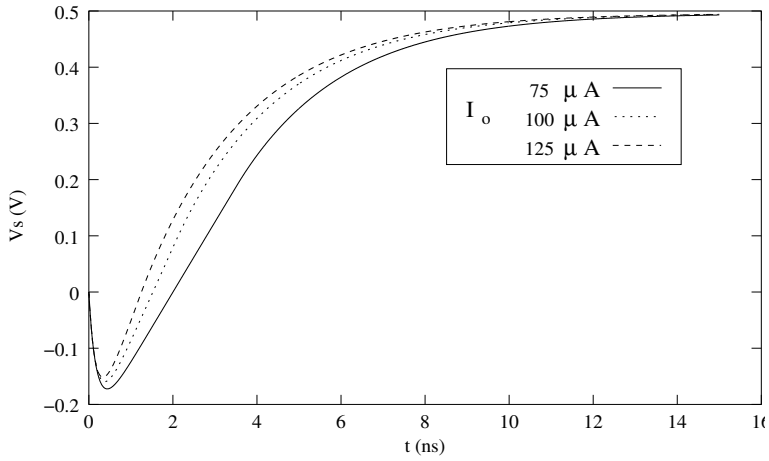


FIG. E.5 – Effet du courant de polarisation sur la réponse temporelle

qui englobe l'entrée de l'amplificateur de façon à garantir que le régime asymptotique (lorsque t tend vers l'infini) soit conservé.

On remarque que la résistance du commutateur a une influence prépondérante dans le début du transfert mais que l'ensemble des réponses tend asymptotiquement vers une même limite lorsque la résistance du commutateur tend vers zéro. On peut simplifier l'analyse en supposant celle-ci négligeable et considérer que le transfert des charges est effectué de manière instantanée. La configuration du circuit juste après le transfert est alors celle de la figure E.6.

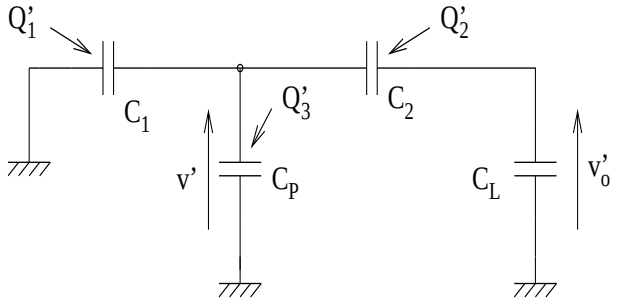


FIG. E.6 – Schéma équivalent simplifié au début du transfert de charge

La conservation de la charge dans la surface isolée permet d'écrire :

$$\begin{aligned}(C_1 + C_P)v^{'} + C_2(v^{'} - v_o^{'}) &= -C_1.v_1 + C_P.v + C_2(v - v_o) \\ C_2(v_o^{'} - v^{'}) + C_L.v_o^{'} &= C_2(v_o - v) + C_L.v_o \end{aligned} \qquad \text{(E.4)}$$

ce qui nous donne pour les tensions d'entrée et de sortie juste après le transfert :

$$v^{'} = v + \Delta v \qquad v_o^{'} = v_o + \Delta v_o \qquad \text{(E.5)}$$

avec

$$\Delta v = -\frac{C_1(v + v_1)}{C_1 + C_P + \frac{C_L C_2}{C_L + C_2}} \qquad \Delta v_o = \frac{C_2 \Delta v}{C_L + C_2}$$

A partir de ces relations on peut connaître le nouvel état des charges au début de la phase de transfert. Deux situations sont alors possibles selon la valeur obtenue pour la tension d'entrée de l'amplificateur. Celle-ci peut en effet appartenir au domaine linéaire de l'amplificateur ou bien correspondre au régime de saturation à courant constant. Dans le second cas, nous devons déterminer la réponse correspondante jusqu'à ce que la tension d'entrée soit inférieure à la tension de saturation. Pour ce faire, nous ferons l'hypothèse que la tension d'entrée de l'amplificateur varie linéairement avec le temps. Nous avons alors pour les charges Q_1 et Q_3 les équations suivantes :

$$\frac{dQ_3}{dt} = S \qquad \frac{d^2Q_1}{dt^2} + \frac{1}{\tau_1}\frac{dQ_1}{dt} + \frac{S}{\tau_3} = 0 \tag{E.6}$$

avec

$$\tau_1 = R_{sw}C_1 \qquad \tau_3 = R_{sw}C_P$$

L'intégration de la seconde équation nous donne la charge Q_1 :

$$Q_1(t) = Q_1(0)e^{-\frac{t}{\tau_1}} + \frac{C_1}{C_P}[(S.\tau_1 - Q_3(0))(1 - e^{-\frac{t}{\tau_1}}) - S.t] \tag{E.7}$$

Pour évaluer la pente S, nous négligeons la résistance du commutateur et celle de sortie de l'amplificateur. Le courant de sortie I_o charge alors une capacité équivalente C_T donnée par :

$$C_T = C_L + \frac{C_2(C_1 + C_P)}{C_1 + C_2 + C_P} \tag{E.8}$$

La pente du signal de sortie est alors égale à $SR = \frac{I_o}{C_T}$ et celle du signal d'entrée s'obtient à partir du diviseur capacitif :

$$S = C_P\frac{dv}{dt} = \frac{C_PC_2}{C_1 + C_2 + C_P}\frac{I_o}{C_T} \tag{E.9}$$

A partir de la pente S, on peut déterminer le temps de saturation t_{sat} de l'amplificateur :

$$t_{sat} = \frac{C_P(v' - V_{sat})}{S} \tag{E.10}$$

Il est alors facile de déterminer les charges à l'issue du régime de saturation. La charge Q_3 croît ou décroît linéairement (selon le signe de la pente S) : $Q_3(t_{sat}) = Q_3(0) + S.t_{sat}$ et la charge Q_1 est obtenue à partir de l'équation E.7. La conservation de la charge totale $Q_T = Q_3 - Q_1 - Q_2$ permet alors de déterminer Q_2.

L'état des charges à la fin de la saturation peut ensuite être utilisé pour le calcul du régime linéaire qui va s'établir durant le temps $t_{lin} = t_{int} - t_{sat}$ où t_{int} est le temps total de la phase de transfert. Ce calcul est simple puisque l'équation E.3 devient une équation linéaire homogène :

$$\frac{d}{dt}\begin{bmatrix} Q_1 \\ Q_2 \\ Q_3 \end{bmatrix} = A\begin{bmatrix} Q_1 \\ Q_2 \\ Q_3 \end{bmatrix} \quad \text{avec} \quad A = M - \alpha\frac{g_m}{C_P}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \tag{E.11}$$

L'intégration sur le temps t_{lin} donne alors simplement $Q(t_{int}) = Q(t_{sat}).e^{A.t_{lin}}$ et la seule difficulté réside alors dans l'évaluation de l'exponentiel d'une matrice. La figure E.7 montre l'évolution de la tension de sortie ainsi que les points effectivement calculés à partir de l'analyse simplifiée précédente.

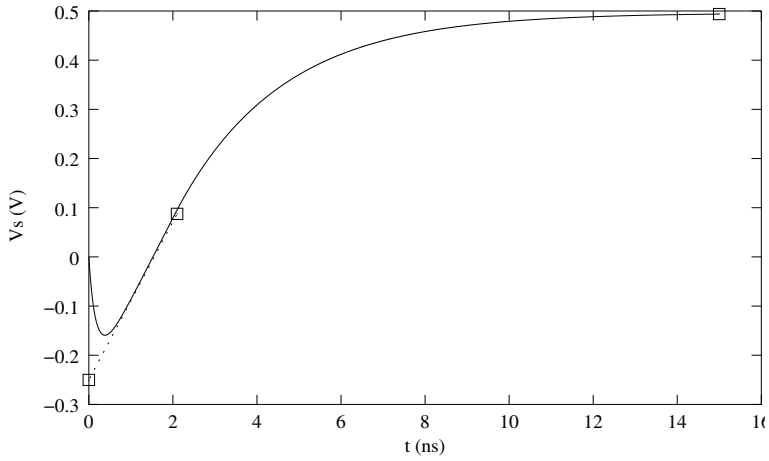


FIG. E.7 – Tension de sortie de l'intégrateur et points de calcul (□)

Dans le cas où deux capacités C_{1a} et C_{1b} sont connectées à l'entrée de l'intégrateur on peut reconduire l'analyse simplifiée précédente en supposant une redistribution quasi instantanée des charges sur les deux capacités :

$$C_{1a}[v+\Delta v-(-v_{1a})]+C_{1b}[v+\Delta v-(-v_{1b})]+C_P\Delta v+\frac{C_L.C_2}{C_L+C_2}\Delta v=0$$

$$\Delta v=\frac{C_{1a}(v+v_{1a})+C_{1b}(v+v_{1b})}{C_{1a}+C_{1b}+C_P+\frac{C_L.C_2}{C_L+C_2}} \qquad \Delta v_o=\frac{C_2\Delta v}{C_L+C_2} \tag{E.12}$$

On peut étendre ce résultat à n capacités C_{1i} avec une tension initiale V_{1i}. L'analyse précédente reste valable en considérant une charge Q_1 initiale et une capacité C_1 telle que :

$$Q_1=\sum_{i=1}^{n}C_{1i}V_{1i} \qquad C_1=\sum_{i=1}^{n}C_{1i} \tag{E.13}$$

Prise en compte du bruit thermique échantillonné Le bruit thermique peut être considéré en introduisant une perturbation sur la charge initiale Q_1. Cette perturbation est simulée par l'ajout d'une variable aléatoire gaussienne d'écart type $\sigma_Q=\sqrt{k_B\cdot T_a\cdot C_1}$ préalablement au transfert décrit ci-dessus.

E.2 Intégrateur dans la phase de maintien

Le schéma correspondant à l'intégrateur dans la phase de maintien est représenté à la figure E.8. Dans cette configuration où la tension de sortie est exploitée, la capacité C1 est déconnectée et une capacité CU (capacité d'échantillonnage de l'étage suivant) est connectée sur la sortie via un commutateur de résistance R_sw. Durant cette phase, c'est la charge $Q_o=Q_3-Q_2$ qui est conservée. D'autre part, dans cette phase, seul le régime linéaire de l'amplificateur est exploité. L'état du circuit est entièrement décrit par les charges Q_2 et Q_x :

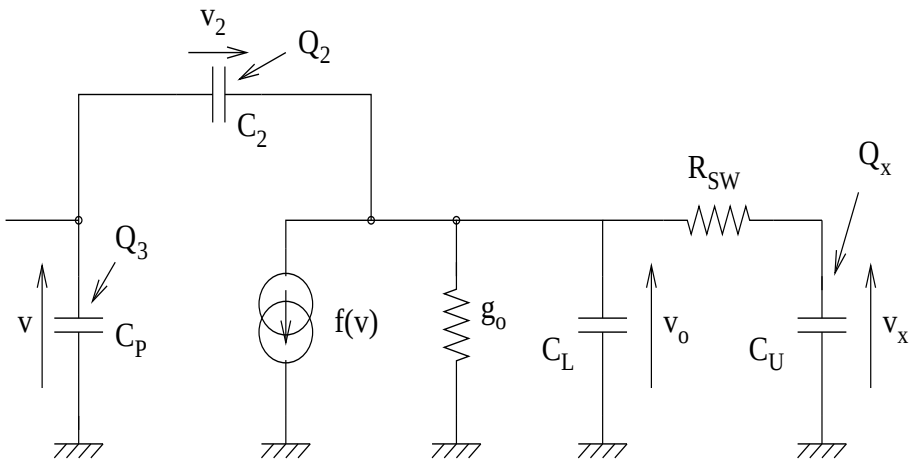


FIG. E.8 – Intégrateur dans la phase de maintien

$$\frac{d}{dt}\begin{bmatrix} Q_2 \\ Q_x \end{bmatrix} = M \cdot \begin{bmatrix} Q_2 \\ Q_x \end{bmatrix} + B \cdot Q_o \tag{E.14}$$

avec

$$M = \begin{bmatrix} -\beta^{-1}\left(\frac{g_m+g_o}{C_P} + \frac{g_o}{C_2} + \frac{1}{R_{sw}C_P} + \frac{1}{R_{sw}C_2}\right) & \frac{\beta^{-1}}{R_{sw}C_U} \\ \frac{1}{R_{sw}C_P} + \frac{1}{R_{sw}C_2} & -1\frac{1}{R_{sw}C_U} \end{bmatrix}$$

et

$$B = \begin{bmatrix} -\beta^{-1}\left(\frac{1}{R_{sw}C_P} + \frac{g_m+g_o}{C_P}\right) \\ \frac{1}{R_{sw}C_U} \end{bmatrix} \quad , \quad \beta = 1 + \frac{C_L}{C_P} + \frac{C_L}{C_2}$$

Ce système linéaire peut facilement être intégré pour obtenir la valeur des charges Q_2 et Q_x à la fin de la phase de maintient. Sachant que la charge initial $Q_o = Q_3 - Q_2$ au début de cette phase a été conservée, on en déduit la valeur de la charge Q_3.

E.3 Réponse de l'intégrateur sur une période complète

La figure E.9 représente la tension de sortie de l'intégrateur sur une période complète de l'horloge avec $T_s = 10\,ns$ et une tension initiale de 1 V sur la capacité C_1. Les phases de transfert et de maintien sont égales à $\frac{T_s}{2}$.

Les paramètres du circuit sont reportés dans le tableau E.2. Le gain théorique de l'intégrateur est $0,5 = \frac{C_1}{C_2}$ et la valeur finale de la tension de sortie dans le cas idéal est égale à 0,5 V.

Les courbes RK (a) et (b) correspondent à une intégration numériques des équations E.3 et E.14 pour le calcul des charges [1]. Dans le cas (a) la source contrôlée est linéaire

[1] Méthode de Runge-Kutta d'ordre 2 à pas fixe (5 ps)

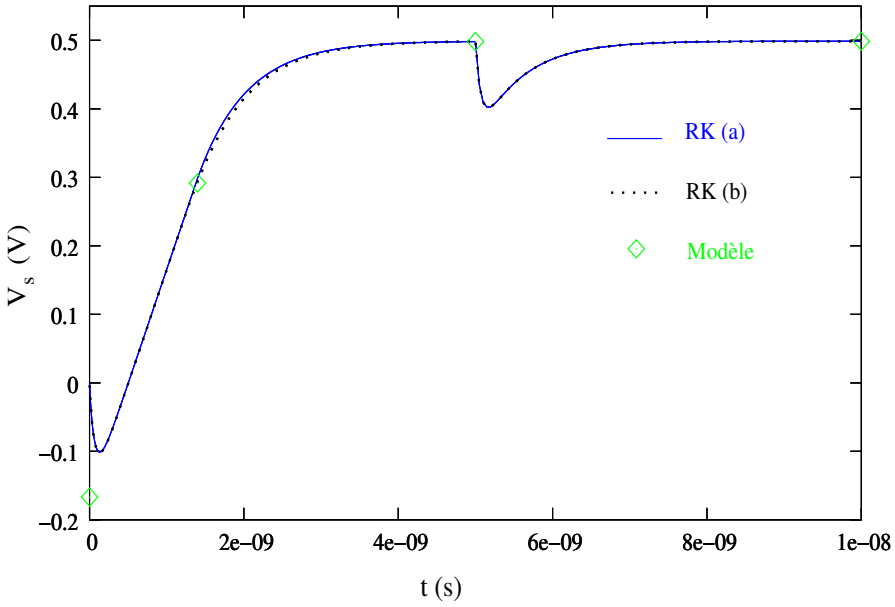


FIG. E.9 – Tension de sortie de l'intégrateur

R_{sw}	C_1	C_2	C_P	C_L	C_U	I_{sat}	V_{sat}	g_o
1 $k\Omega$	$0,1\ pF$	$0,2\ pF$	$0,1\ pF$	$0,2\ pF$	$0,1\ pF$	100 μA	0.1 V	1 μS

TAB. E.2 – Paramètres du modèle.

par parties comme indiqué sur la figure E.3. Dans le cas (b), un modèle quadratique avec la même pente à l'origine est utilé :

$$i = f(v) = I_{sat} \cdot x \cdot \sqrt{1 - \frac{x^2}{4}} \quad \text{avec} \quad -\sqrt{2} < x = \frac{v}{V_{sat}} < \sqrt{2}$$

Celui-ci est proche du comportement d'un amplificateur avec une paire différentielle MOS et montre un très faible écart avec le modèle linéaire par parties.

Les 4 points correspondent aux valeurs calculées par le modèle simplifié. L'écart initial est important du fait des hypothèses de transfert instantané des charges. La précision obtenue sur la valeur finale est cependant très bonne comme l'indique le tableau E.3.

RK (a)	RK (b)	modèle
0,498766	0,498755	0,498773

TAB. E.3 – Valeur de la tension de sortie (V) en fin de période.

Annexe F

Caractéristiques d'un étage amplificateur pipeline

En pratique, l'activation des étages d'un CAN pipeline s'effectue à partir de deux phases complémentaires ϕ_1 et ϕ_2 comme indiqué sur la figure F.1.

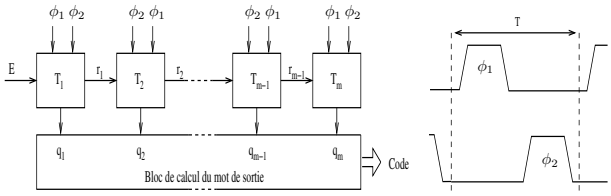


FIG. F.1 – structure du convertisseur pipeline

En une période T de l'horloge, un échantillon du signal traverse ainsi deux étages et le temps de latence pour la génération du code de sortie est de $\frac{m}{2} \cdot T$ pour un pipeline de m étages. Durant une phase un étage réalise l'une des deux opérations suivantes :

1. Echantillonnage du résidu de l'étage précédent et génération du code intermédiaire
2. Génération du résidu par le MDAC à partir du code intermédiaire

Ces opérations sont alternées comme le montre la figure F.2 où l'étage $k-1$ réalise l'opération 2 et l'étage k l'opération 1. Nous nous intéressons ici aux performances de l'amplificateur dans la configuration de l'étage $k-1$. Sa capacité de charge est constituée de celle du MDAC de l'étage k en parallèle avec celle du CAN flash. La capacité de charge totale est donc :

$$C_L = b_k\, C_u + C_{flash} = 2^{n_k}\, C_u + C_{flash} \tag{F.1}$$

où n_k est le nombre de bit de l'étage k et C_u est la capacité élémentaire du MDAC. Pour déterminer les performances de l'amplificateur, il nous faut donc au préalable évaluer la capacité d'entrée du CAN flash.

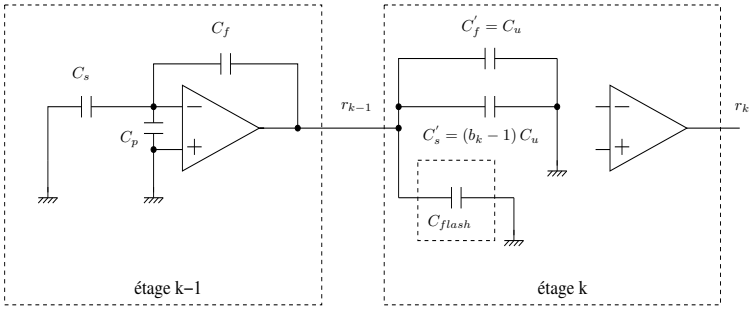


FIG. F.2 – Couplage entre deux étages d'un CAN pipeline

F.1 Calcul de la capacité d'entrée du flash

L'erreur maximale autorisée par la redondance sur un seuil de comparateur est (cf 5.11) :

$$\delta = \frac{V_{ref}}{2^{n_k+1}} \tag{F.2}$$

Pour satisfaire cette contrainte, prenons un rapport η entre l'écart type de la tension de seuil des transistors d'entrée du comparateur et cette erreur :

$$\sigma_{VT} = \frac{A_{VT}}{\sqrt{A}} = \frac{V_{ref}}{\eta \cdot 2^{n_k+1}} \tag{F.3}$$

On peut alors facilement en déduire la capacité d'entrée du flash à partir de celle du transistor ($C_{ox} = \frac{\epsilon_{ox} \cdot A}{t_{ox}}$) et du nombre de comparateurs constituant le flash ($n_c = 2^{n_k+1}$).

Exemple :
$n = 4 \quad \eta = 6 \quad V_{ref} = 1V \quad t_{ox} = 5nm \quad A_{VT} = 5mV.\mu m \quad C_{flash} = 0,2pF$

F.2 Bande passante

La bande passante d'un étage amplificateur du pipeline peut être obtenue à partir du schéma de la figure F.2 et un modèle simple d'amplificateur. Nous supposerons que celui-ci est caractérisé par une transconductance g_m et une capacité d'entrée C_p. Le coefficient de contre-réaction β de l'étage $k-1$ est donné par :

$$\beta = \frac{C_f}{C_s + C_p + C_f} = \frac{C_u}{b_{k-1}\,C_u + C_p} = \frac{1}{b_{k-1} + \gamma} \tag{F.4}$$

où C_u est la capacité unitaire du MDAC et $\gamma = \frac{C_p}{C_u}$.

La capacité totale à la sortie de l'amplificateur est égale à :

$$C_o = C_L + \frac{C_f\,(C_s + C_p)}{C_f + C_s + C_p} = C_L + C_u\,(1 - \beta) \tag{F.5}$$

où C_L est la capacité de charge donnée par l'équation F.1.

On déduit facilement la fréquence de coupure de l'amplificateur :

$$\omega_c = \beta \frac{g_m}{C_o} = \frac{\beta\, g_m}{C_L + C_u\,(1 - \beta)} \tag{F.6}$$

En supposant γ suffisamment faible, le coefficient β est complètement déterminé par le choix de la base. Il est égale à $\frac{1}{2}$ en base 2 et tend vers 0 lorsque la base augmente. On voit ainsi, d'après la formule F.6, que la bande passante est d'autant plus réduite que la base est élevée. Ceci résulte de l'effet combiné de la réduction du facteur β et de l'augmentation de la capacité de charge C_L (formule F.1). Ceci est illustré par la figure F.3 qui donne la fréquence de coupure d'un étage pipeline pour une technologie CMOS avec $L_{min} = 0,13\,\mu m$ et pour deux résolutions différentes de 1 bit et 4 bit. L'amplificateur est supposé constitué d'un seul transistor avec un courant de polarisation I et une largeur W, la longueur de grille L étant prise égale à $5\,L_{min}$ pour garantir un gain suffisant de l'amplificateur [1]. La capacité minimum utilisée pour le MDAC est $C_u = 0.1pF$. Le choix de la capacité est important car il conditionne la consommation de l'étage pour une bande passante donnée (équation F.6), le bruit thermique et la dispersion sur les rapport de capacités (équation 5.13). La figure F.3 montre également qu'il existe un optimum de consommation pour une bande passante donnée de l'étage.

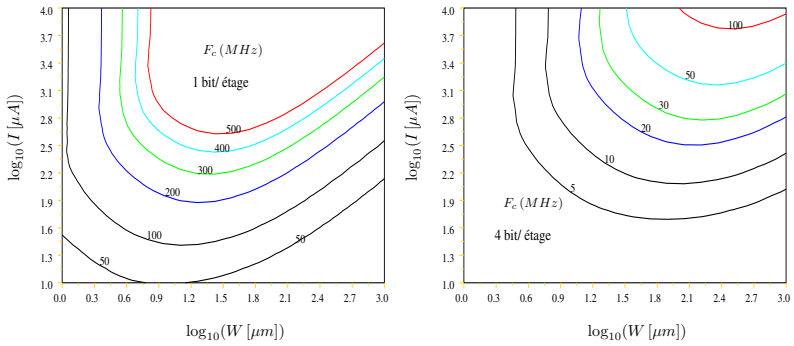


FIG. F.3 – Bande passante d'un étage pipeline

F.3 Bruit thermique échantillonné

Le bruit thermique échantillonné à l'entrée de l'amplificateur correspond à l'échantillonnage d'une capacité $C_s = b\,C_u$ où C_u est la capacité élémentaire du MDAC et b est la base de l'étage considéré (figure F.2). La variance de cette tension de bruit est donnée par (formule 2.27) :

$$\sigma_{pe}^2 = \frac{k_B \cdot T}{b \cdot C_u} \tag{F.7}$$

On remarque que le bruit thermique est inversement proportionnel à la base. Ce fait est encore accru par la structure pipeline où le bruit ramené à l'entrée d'un étage particulier est inversement proportionnel au carré du gain total lié aux étages précédents. On a ainsi, pour un convertisseur de m étages de base b, le bruit total à l'entrée (le gain étant égal à la base b) :

$$\sigma^2 = \sigma_{pe}^2 \sum_{i=0}^{m-1} \frac{1}{kc_i \cdot b^{2i}} \tag{F.8}$$

[1] [59] : *Analog and Mixed-signal Devices Technology Requirements-Near–term*

On a introduit un coefficient $kc_i = C_u^{'}/C_u$ pour prendre en compte une pondération éventuelle des capacités dans la structure pipeline (cf section 5.4). L'expression complète du bruit d'un étage doit également inclure les bruit liées à la phase de transfert (configuration de l'étage $k-1$ de la figure F.2). Dans cette phase, le bruit de l'amplificateur doit également être considéré. On trouvera une expression complète du bruit lié au MDAC dans [5].

Bibliographie

[1] A. M. Abo et P. R. Gray : A 1.5-V, 10-bit, 14.3-MS/s CMOS Pipeline Analog-to-Digital Converter. *IEEE Journal of Solid-State Circuits*, 1999.

[2] H. Aboushady et M.-M. Louerat : Systematic approach for discrete-time to continuous-time transformation of Sigma Delta modulators. Dans *Circuits and Systems, 2002. ISCAS 2002. IEEE International Symposium on*, volume 4, pages 229–232, 2002. TY - CONF.

[3] S.H. Ardalan et J.J. Paulos : An analysis of nonlinear behavior in delta-sigma modulators. *IEEE Transactions on Circuits and Systems*, juin 1987.

[4] A. Avizienis : Signed-digit number representation for fast parallel arithmetic. *IRE Transactions on Electronic Computers*, pages 389–400, 1961.

[5] M. Y. Azizi, A Saeedfar, H. Z. Hoseini et O. Shoaei : Thermal noise analysis of multi-bit SC gain-stages for low-voltage high-resolution pipeline ADC design. *International Symposium on Signals, Circuits and Systems*, juillet 2003.

[6] R.T. Baird et T.S. Fiez : Stability analysis of high-order delta-sigma modulator for ADC's. *IEEE Transactions on Circuits and Systems-II : Analog and Digital Signal Processing*, 1994.

[7] R.T. Baird et T.S. Fiez : Linearity enhancement of multibit $\Sigma\Delta$ A/D and D/A converters using Data Weighted Averaging. *IEEE Transactions on Circuits and Systems-II : Analog and Digital Signal Processing*, décembre 1995.

[8] C. Barrett : *Low-Power Decimation Filter Design for Multi-Standard Transceiver Applications*. Thèse de doctorat, université de Californie, Berkeley, 1997.

[9] W.R. Bennet : Spectra of quantized signals. *Bell Systems Technical Journal*, 27, juillet 1948.

[10] A.S. Blum, B.H. Engl, H.P. Bichfield, R. Hagelauer et A.A. Abidi : A 1.2 V 10-b 100-MSamples/s A/D Converter in 0.12 μm CMOS. *Symposium on VLSI circuits*, 2002.

[11] B.E. Boser et B.A. Wooley : The design of sigma-delta modulation analog-to-digital converters. *IEEE Journal of Solid-State Circuits*, décembre 1988.

[12] S. Boyd : . *http ://www.stanford.edu/∼boyd*, .

[13] W. Bright : 8b 75MSample/s 70mW Parallel Pipelined ADC Incorporating Double Sampling. *IEEE International Solid-State Circuits Conference*, 1998.

[14] K. Bult : Analog Design in Deep Submicron CMOS. *Proceedings of European Solid State Circuits Conference*, septembre 2000.

[15] T. Burger et Q. Huang : A 13.5-mW 185-Msample/s $\Sigma\Delta$ Modulator for UMTS/GSM Dual-Standard IF Reception. *IEEE Journal of Solid-State Circuits*, décembre 2001.

[16] P. Carbone : Effect of additive Dither on the Resolution of Ideal Quantizers. *IEEE Transactions on Instrumentation and Measurement*, juin 1994.

[17] L.P. Carloni, F. De Bernardinis, A. Sangiovanni-Vincentelli et M. Sgroi : The Art and Science of Integrated Systems Design. *Proceedings of European Solid State Circuits Conference*, 2002.

[18] M. Charbit : *Eléments de théorie du signal : Les signaux aléatoires.* Ellipses, collection pédagogique de Télécommunications, 1990.

[19] H-S Chen, K. Bacrania et B-S Song : A 14b 20Msample/s CMOS pipelined ADC. *IEEE International Solid-State Circuits Conference*, 2000.

[20] J.A. Cherry et W.M. Snelgrove : *Continuous-time delta-sigma modulator for high-speed A/D conversion.* Kluwer Academic Publishers, 2000.

[21] M-J Choe, B-S Song et K Bacrania : A 13b 40Msample/s CMOS pipelined folding ADC with background offset timming. *IEEE International Solid-State Circuits Conference*, 2000.

[22] H.C. Choi, H-J Park, S-K Bae, J-W Kim et P. Chung : A 1.4V 10-bit 20Msps pipelined A/D converter. *IEEE International Symposium on Circuits and Systems*, 2000.

[23] M. Choi et A.A. Abidi : A 6b 1.3 GSample/s A/D Converter in 0.35 μm CMOS. *IEEE Journal of Solid-State Circuits*, décembre 2001.

[24] F.A. Collins et C.J. Sicking : Properties of Low Precision Analog-to-Digital Converters. *IEEE Transactions on Aerospace and Electronic Systems*, septembre 1976.

[25] D.C. Craig : Extensible Hierachical Object-Oriented Logic Simulation with an Adaptable Graphical User Interface. Mémoire de D.E.A., Memorial University of Newfoundland, 1996.

[26] N. Da Dalt, M. Harteneck, C. Sander et A. Wiesbauer : On jitter Requirement of the sampling Clock for Analog-to-Digital Converters. *IEEE Transactions on Circuits and Systems-I : Fundamental Theory and Applications*, septembre 2002.

[27] N. Da Dalt et C. Sandner : A Subpicosecond Jitter PLL for Clock Generation in 0.12- μm Digital CMOS. *IEEE Journal of Solid-State Circuits*, juillet 2003.

[28] D. Dalton et al : A 200-MSPS 6 bit flash ADC in 0.6-μm CMOS. *IEEE Trans. on circuits and Systems*, 45(11):1433–1444, novembre 1998.

[29] M. del Mar Hershenson : Design of pipeline analog-to-digital converters via geometric programming. *International Conference on Computer Aided Design*, 2002.

[30] R. del Rio, F. Medeiro, J.M. de la Rosa, B. Perez-Verdu et A. Rodriguez-Vazquez : A 2.5-V $\Sigma\Delta$ modulator in 0.25-μm CMOS for ADSL. *IEEE International Symposium on Circuits and Systems*, 2002.

[31] M. Dessouky : *Conception en vue de la réutilisation de circuits analogiques. Application : Modulateur Delta-Sigma à très faible tension.* Thèse de doctorat, Université Paris VI, 2001.

[32] C. Donovan et M.P. Flynn : A "digital" 6-bit ADC in 0.25 μm CMOS. *IEEE Journal of Solid-State Circuits*, mars 2002.

[33] M.D. Ercegovac : *Division and square root : digit-recurrence algorithms and implementations.* Kluwer Academic Publishers, 1994.

[34] A.R. Feldman, B.E. Boser et P.R. Gray : A 13-Bit, 1.4-MS/s sigma-delta modulator for RF baseband channel applications. *IEEE Journal of Solid-State Circuits*, octobre 1998.

[35] M. Flynn et B. Sheahan : A 400-MSample/s 6-b CMOS Folding and Interpolating ADC. *IEEE J. of Solid-State Circuits*, 33(12):1932–1938, décembre 1998.

[36] I. Fujimori, L. Longo, A. Hairapetian, K. Seiyama, S. Kosic, J. Cao et S-L Chan : A 90-dB SNR 2.5-MHz Output-Rate ADC Using Cascaded Multibit Delta-Sigma Modulation at 8 Oversampling Ratio. *IEEE Journal of Solid-State Circuits*, décembre 2000.

[37] R. Gaggl, A. Wiesbauer, C. Schranz et P. Pessl : A 14-bit Delta-Sigma Modulator for ADSL-CO Applications in 0.18 μm CMOS. *Proceedings of European Solid State Circuits Conference*, 2002.

[38] G. Geelen : A 6b 1.1 Gsample/s CMOS A/D converter. *Int. Solid-State Circ. Conf., Digest of Technical Papers*, pages 128–129, février 2001.

[39] Y. Geerts, A. Marques, M. Steyaert et W. Sansen : A 3.3-V, 15-bit, delta-sigma ADC with a signal bandwidth of 1.1 MHz for ADSL applications. *IEEE Journal of Solid-State Circuits*, juillet 1999.

[40] Y. Geerts, M. Steyaert et W. Sansen : A 2.5 Msample/s multi-bit delta-sigma CMOS ADC with 95 dB SNR. *IEEE International Solid-State Circuits Conference*, 2000.

[41] Y. Geerts, M. Steyaert et W. Sansen : A High-Performance Multibit $\Sigma\Delta$ CMOS ADC. *IEEE Journal of Solid-State Circuits*, décembre 2000.

[42] Y. Geerts, M. Steyaert et W. Sansen : *Design of multi-bit delta-sigma A/D converters*. Kluwer Academic Publishers, 2002.

[43] B. Ginetti et P. Gespers : Reliability of code density test for high resolution ADCs. *Electronic letter*, novembre 1991.

[44] J. Goes, J. C. Vital et J. Franca : *Systematic design for optimisation of pipelined ADCs*. Kluwer Academic Publishers, 2001.

[45] G. Gomez et B. Haroun : A 1.5 V 2.4/2.9 mW 79/50 dB DR $\Sigma\Delta$ modulator for GSM/WCDMA in a 0.13 μm digital process. *IEEE International Solid-State Circuits Conference*, 2002.

[46] R.M. Gray : Oversampled Sigma-Delta Modulation. *IEEE Trans. Commun.*, mai 1987.

[47] R.M. Gray : Quantization noise spectra. *IEEE Trans. on Information Theory.*, 36, novembre 1990.

[48] J. Guilherme, P. Figueiredo, P. Azevedo, G. Minderico, A. Leal, J. Vital et J. Franca : A pipeline 15-b 10-Msample/s Analog-to-digital converter for ADSL applications. *IEEE International Symposium on Circuits and Systems*, 2001.

[49] S.K. Gupta, T.L. Brooks et V. Fong : A 64 MHz $\Sigma\Delta$ ADC with 105 dB IM3 distortion using a linearized replica sampling network. *IEEE International Solid-State Circuits Conference*, 2002.

[50] S.K. Gupta, T.L. Brooks et V. Fong : A 64 MHz $\Sigma\Delta$ ADC with 105 dB IM3 distortion using a linearized replica sampling network. *IEEE International Solid-State Circuits Conference*, 2002.

[51] C.S. Güntürk, J.C. Lagarias et V.A. Vaishampayan : On the robustness of Single-Loop Sigma-Delta Modulation. *IEEE Trans. on Information Theory.*, juillet 2001.

[52] S. Hamedi-Hagh et C.A.T. Salama : A 10 bit, 50Msample/s, low power pipelined A/D converter for cable modem application. *IEEE International Symposium on Circuits and Systems*, 2001.

[53] S. Hein et A. Zakhor : *Sigma Delta Modulators : Nonlinear Decoding Algorithms and Stability Analysis*. Kluwer Academic Publishers, 1993.

[54] R. Hooke et T.A. Jeeves : Direct search solution of numerical and statistical problems. *J. Ass. Comput. Mach.*, 1961.

[55] HYPRES, http ://www.hypres.com. .

[56] IEEE-STD-1241 : Standard for Terminology and Test Methods for Analog-to-Digital Converters. , décembre 2000.

[57] J.M. Ingino et B.A. Wooley : A Continuously Calibrated 12-b, 10-MS/s, 3.3-V A/D Converter. *IEEE Journal of Solid-State Circuits*, 1998.

[58] H. Inose, Y. Yasuda et J. Murakami : A telemetering system by code modulation-$\Sigma\Delta$ modulation. *IRE trans. Space Electron. Telemetry*, 1962.

[59] ITRS, http ://public.itrs.net. .

[60] F. Jager : Delta Modulation - a method of PCM transmission using the one unit code. *Philips Res. Rep.*, 1952.

[61] S.M. Jamal, Fu Daihong, P.J. Hurst et S.H. Lewis : A 10b 120MSample/s time-interleaved analog-to-digital converter with digital background calibration. *IEEE International Solid-State Circuits Conference*, 2002.

[62] Y. Jenq : Digital spectra of nonuniformly sampled signal : fundamentals and high-speed waveform digitizers. *IEEE Transactions on Instrumentation and Measurement*, juin 1988.

[63] Y-D Jeon, B-L Jeon, S-C Lee, S-M Yoo et S-H Lee : A 12b 50MHz 3.3V CMOS acquisition time minimized A/D converter. *Asia and South Pacific Design Automation Conference*, 2000.

[64] R. Jiang et T. Fiez : A 1.8V 14b Sigma-Delta A/D Converter with 4MSample/s Conversion. *IEEE International Solid-State Circuits Conference*, février 2002.

[65] White J.L. et Abidi A.A. : Active resistor network as 2D sampled data filters. *IEEE Trans. Circuits and Systems I : Fondamental Theory and Application*, 39:724–733, septembre 1992.

[66] D. Johns et K. Martin : *Analog Integrated Circuit Design*. Wiley, 1997.

[67] A.N. Karanicolas, H-S Lee et K.L. Bacrania : A 15-b 1-Msample/s Digitally Self-Calibrated Pipeline ADC. *IEEE Journal of Solid-State Circuits*, 1993.

[68] K. Kattmann et J. Barrow : A Technique for reducing differential non-linearity errors in flash A/D converters. *Int. Solid-State Circ. Conf., Digest of Technical Papers*, pages 170–171, février 1991.

[69] D. Kelly, W. Yang, I. Mehr, M. Sayuk et L. Singer : A 3 V 340 mW 14 b 75 MSPS CMOS ADC with 85 dB SFDR at Nyquist. *IEEE International Solid-State Circuits Conference*, 2001.

[70] R. Kress : *Numerical Analysis*. Springer, 1998.

[71] S. Kuboki, K. Kato, N. Miyakawa et K. Matsubara : Nonlinearity analysis of resistor string A/D converter. *IEEE Trans. on Circuits and Systems*, 29(6), juin 1982.

[72] S. Kulhali, V. Penkota et R. Asv : A 30mW 12b 21MSample/s pipelined CMOS ADC. *IEEE International Solid-State Circuits Conference*, 2002.

[73] T.H. Lee : *The Design of CMOS Radio-Frequency Integrated Circuits*, chapitre 3. cambrige, 1998.

[74] T.C. Leslie et B. Singh : Sigma-delta modulators with multibit quantising elements and single-bit feedback. *Circuits, Devices and Systems, IEE Proceedings G*, 1992.

[75] J.H-C Lin et B. Haroun : An Embedded 0.8V/480 μW 6b/22 MHz Flash ADC in 0.13μm Digital CMOS Process using Nonlinear Double-Interpolation Technique. *IEEE Journal of Solid-State Circuits*, décembre 2002.

[76] A. Marques, V. Peluso, M. Steyaert et W. Sansen : A 15-b Resolution 2-MHz Nyquist Rate Sigma Delta ADC in a 1 μm CMOS Technology. *IEEE Journal of Solid-State Circuits*, juillet 1998.

[77] S. Mathur, M. Das, P. Tadeparthy, S. Ray, S. Mukherjee et B. L. Dinakaran : A 115mW 12-bit 50 MSPS pipelined ADC. *IEEE International Symposium on Circuits and Systems*, 2002.

[78] F. Medeiro, A. Rodriguez-Vazquez et B. Perez-Verdu : *Top-Down Design of High-Performance Sigma-Delta Modulators*. Kluwer Academic Publishers, 1999.

[79] I. Mehr et D. Dalton : A 500-Msample/s, 6-Bit Nyquist-Rate ADC for Disk-Drive Read-Channel Applications. *IEEE J. of Solid-State Circ.*, 34(7):912–920, juillet 1999.

[80] I. Mehr et L. Singer : A 55-mW, 10-bit, 40-Msample/s Nyquist-rate CMOS ADC. *IEEE Journal of Solid-State Circuits*, 2000.

[81] D.G. Messerschmitt : Quantizing for maximum output entropy. *IEEE Trans. on Information Theory.*, septembre 1971.

[82] M.R. Miller et C.S. Petrie : A Multibit Sigma-Delta ADC for Multimode Receivers for Multimode Receivers. *IEEE Journal of Solid-State Circuits*, mars 2003.

[83] J. Ming : An 8-bit 80-Msample/s Pipelined Analog-to-Digital Converter With Background Calibration. *IEEE Journal of Solid-State Circuits*, 2001.

[84] D. Miyazaki, M. Furuta et S. Kawahito : A 16 mW 30 MSample/s 10 b pipelined A/D converter using a pseudo-differential architecture. *IEEE International Solid-State Circuits Conference*, 2002.

[85] N. Moreau : *Techniques de compression des signaux*. Masson, 1995.

[86] J.C. Morizio et al : 14-bit 2.2-MS/s Sigma-Delta ADC's. *IEEE Journal of Solid-State Circuits*, juillet 2000.

[87] K. Nagaraj et al : Efficient 6-Bit A/D Converter Using a 1-Bit Folding Front End. *IEEE Journal of Solid-State Circuits*, 34(8):1056–1062, août 1999.

[88] K. Nagaraj et al : A dual-mode 700-Msamples/s 6-bit 200-Msamples/s 7-bit A/D converter in a 0.25-μm digital CMOS. *IEEE Journal of Solid-State Circuits*, 2000.

[89] D.G. Nairn : A 10-bit, 3V, 100MS/s pipelined ADC. *IEEE Custom Integrated Circuits Conference*, 2000.

[90] O.J.P. Nys et R.K. Henderson : An analysis of dynamic element matching techniques in sigma-delta modulation. *IEEE International Symposium on Circuits and Systems*, 1996.

[91] O. Oliaei, P. Clement et P. Gorisse : A 5-mW sigma-delta modulator with 84-dB dynamic range for GSM/EDGE. *IEEE Journal of Solid-State Circuits*, 2002.

[92] K. Ono, H. Shimizu, J. Ogawa, M. Takeda et M. Yano : A 6bit 400Msps 70mW ADC using interpolated parallel scheme. *Symposium on VLSI circuits*, 2002.

[93] P.E. Pace, D. Styer et I.A. Akin : A Folding ADC Preprocessing Architecture Employing a Robust Symmetrical Number System With Gray-Code Properties. *IEEE Transactions on Circuits and Systems-II : Analog and Digital Signal Processing*, 47(5), mai 2000.

[94] H. Pan : *A 3.3-V 12-b 50-MS/s A/D converter in 0.6-μm CMOS with 80-dB SFDR.* Thèse de doctorat, University of California, Los Angeles, 1999.

[95] H. Pan, M. Segami, M. Choi, J. Cao et F. Iatori : A 3.3V, 12b, 50Msample/s A/D converter in 0.6 um CMOS with over 80dB SFDR. *IEEE International Solid-State Circuits Conference*, 2000.

[96] Y-I Park, S. Karthikeyan, F. Tsay et E. Bartolome : A 10 b 100 MSample/s CMOS pipelined ADC with 1.8 V power supply. *IEEE International Solid-State Circuits Conference*, 2001.

[97] M. J. M. Pelgrom, A. C. J. Duinmaijer et A. P. G. Welbers : Matching properties of MOS transistors. *IEEE Solid-State Circuits*, 1989.

[98] M.J.M Pelgrom, A.C.J Rens, M. Vertregt et M.B. Dijkstra : A 25-Ms/s 8-bit CMOS A/D converter for embedded application. *IEEE Journal of Solid-State Circuits*, 29(8):879–886, août 1994.

[99] M.J.M. Pelgrom, H.P. Tuinhout et M. Vertregt : Transistor matching in analog CMOS applications. *International Electron Devices Meeting*, 1998.

[100] H. Petit et J-F. Naviner : Modèle linéaire de quantificateur pour la synthèse de modulateur sigma-delta cascade. *Conférence TAISA, Bordeaux*, 2001.

[101] H. Petit et J-F. Naviner : Modèle non linéaire d'intégrateur à capacités commutées pour la simulation de convertisseurs sigma-delta. *Conférence TAISA, Louvain La Neuve*, 2003.

[102] A. Petraglia et K. Mitra : Analysis of mismatch effects among A/D Converters in a time-interleaved waveform digitizer. *IEEE Transactions on Instrumentation and Measurement*, octobre 1991.

[103] C.L. Portmann et T.H.Y. Meng : Power-efficient metastability error reduction in CMOS Flash A/D converters. *IEEE Journal of Solid-State Circuits*, 31:1132–1140, 1996.

[104] K. Poulton et al : A 20Gs/s 8b ADC with a 1Mb memory in 0.18 μm CMOS. *IEEE International Solid-State Circuits Conference*, 2003.

[105] K. Poulton, R. Nett, A. Muto, W Liu, A. Burstein et M. Heshami : A 4 GSample/s 8b ADC in 0.35 μm CMOS. *IEEE International Solid-State Circuits Conference*, 2002.

[106] R. Reutemann, P. Balmelli et Q. Huang : A 33mW 14b 2.5$MSample/$s $\Sigma\Delta$ A/D converter in 0.25 μm digital CMOS. *IEEE International Solid-State Circuits Conference*, 2002.

[107] R.A. Rutenbar, G.G.E. Gielen et B.A. Antao : *Computer-Aided Design of Analog Integrated Circuits and Systems.* Wiley-IEEE Press, 2002.

[108] P. Scholtens : A 2.5 Volt 6 bit 600Msps Flash ADC in 0.25 um CMOS. *Proceedings of European Solid State Circuits Conference*, 2000.

[109] P. Scholtens et M. Vertregt : A 6b 1.6GSample/s Flash ADC in 0.18μm CMOS using Averaging Termination. *IEEE Journal of Solid-State Circuits*, décembre 2002.

[110] R. Schreier : An empirical study of high-order single-bit delta-sigma modulators. *IEEE Transactions on Circuits and Systems-II : Analog and Digital Signal Processing*, août 1993.

[111] R. Schreier, M.V. Goodson et Bo Zhang : An algorithm for computing convex positively invariant sets for delta-sigma modulators. *IEEE Transactions on Circuits and Systems I : Fundamental Theory and Applications.*, janvier 1997.

[112] M.D. Scott, B.E. Boser et K.S.J. Pister : An ultralow-energy ADC for Smart Dust. *IEEE Journal of Solid-State Circuits*, juillet 2003.

[113] C. Shi, J. Wilson et M. Ismail : Design techniques for improving intrinsic accuracy of resistor string DAC's. *IEEE International Symposium on Circuits and Systems*, 2001.

[114] L. Singer, S. Ho, M. Timko et D. Kelly : A 12b 65Msample/s CMOS ADC with 82dB SFDR at 120MHz. *IEEE International Solid-State Circuits Conference*, 2000.

[115] O.M. Solomon : The use of DFT windows in signal-to-noise ratio and harmonic distorsion computations. *IEEE Transactions on Instrumentation and Measurement*, avril 1994.

[116] H.A. Spang et P.M. Schultheiss : Reduction of Quantization Noise by use of Feedback. *IRE Trans. Commun*, 1962.

[117] L. Sumanen, M. Waltari et K. Halonen : A pipeline A/D converter for WCDMA applications. *IEEE International Conference on Electronics, Circuits and Systems*, 1999.

[118] L. Sumanen, M. Waltari et K. A. I. Halonen : A 10-bit 200-MS/s CMOS Parallel Pipeline A/D Converter. *IEEE Journal of Solid-State Circuits*, juillet 2001.

[119] K. Sushihara et al : A 6b 800 MSample/s CMOS A/D Converter. *IEEE Int. Solid-State Circ. Conf., Dig. Tech. Papers*, pages 428–429, février 2000.

[120] J. Talebzadeh, M.R. Hasanzadeh, M. Yavari et O. Shoaei : A 10-bit 150-MS/s, Parallel Pipeline A/D Converter in 0.6 μm CMOS. *IEEE International Symposium on Circuits and Systems*, 2002.

[121] Y. Tamba et K. Yamakido : A CMOS 6b 500MSample/s ADC for a Hard Disk Drive Read Channel. *IEEE Int. Solid-State Circ. Conf., Dig. Tech. Papers*, pages 324–325, février 1999.

[122] S.K. Tewksbury et R.W. Hallok : Oversampling, Linear Predictive and Noise-Shaping Coders of Order $N > 1$. *IEEE trans. on Circuits and Systems*, 1978.

[123] S. Tsukamoto : A CMOS 6-b, 400-MSamples/s ADC with error correction. *IEEE J. of Solid-State Circuits*, 33(12):1939–1947, décembre 1998.

[124] K. Uyttenhove, A. Marques et M. Steyaert : A 6-bit 1 GHz acquisition speed CMOS flash ADC with digital error correction. *Proceedings of the Custom Integrated Circuits Conference*, 2000.

[125] K. Uyttenhove et M.S.J. Steyaert : Speed-Power-Accuracy Tradeoff in High-Speed CMOS ADCs. *IEEE Transactions on Circuits and Systems-II : Analog and Digital Signal Processing*, avril 2002.

[126] P. van Zeijl et al : A Bluetooth radio in 0.18 μm CMOS. *IEEE International Solid-State Circuits Conference*, février 2002.

[127] R.G. Vaughan, N.L. Scott et D.R. White : The theory of Bandpass Sampling. *IEEE Trans. on Signal Processing.*, septembre 1991.

[128] K. Vleugels, S. Rabii et B.A. Wooley : A 2.5-V Sigma-Delta Modulator for Broadband Communications Applications. *IEEE Journal of Solid-State Circuits*, décembre 2001.

[129] M. Vogels et G. Gielen : Figure of merit selection of A/D converters. *DATE Conference*, 2003.

[130] R.H. Walden : Analog-to-Digital converter survey and analysis. *IEEE Journal on Selected Areas in Communications*, avril 1999.

[131] M. Waltari et K. Halonen : 1-V 9-Bit Pipelined Switched-Opamp ADC. *IEEE Journal of Solid-State Circuits*, 2001.

[132] B. Widrow, I. Kollar et M-C Liu : Statistical Theory of Quantization. *IEEE Transactions on Instrumentation and Measurement*, avril 1996.

[133] K. Yoon et al : A 6b 500MSample/s CMOS Flash ADC with a Background Interpolated Auto-Zero Technique. *IEEE Int. Solid-State Circ. Conf., Dig. Tech. Papers*, pages 326–327, février 1999.

[134] S-B You, K-W Lee, H. C. Choi, H-J Park, J-W Kim et P. Chung : A 3.3V 14-bit 10Msps calibration-free CMOS pipelined A/D converter. *IEEE International Symposium on Circuits and Systems*, 2000.

[135] F. Yuan et A. Opal : Computer methods for switched circuits. *IEEE Transactions on Circuits and Systems-I : Fundamental Theory and Applications*, août 2003.

[136] A. Zanchi, F. Tsay et I. Papantonopoulos : Impact of Capacitor Dielectric Relaxation on a 14-bit 70-Ms/s Pipeline ADC in 3-V BiCMOS. *IEEE Journal of Solid-State Circuits*, décembre 2003.

[137] Y. Zhang, E. Hayahara, S. Hirano et N. Sakakibara : An optimal design consideration for higher-order delta-sigma A/D converter. *Asia-Pacific Conference on Circuits and Systems, 2000. IEEE APCCAS*, 2000.