Dept GEII IUT Bordeaux I

ECHANTILLONNAGE, QUANTIFICATION

CONVERSION ANALOGIQUE-NUMERIQUE

et

NUMERIQUE-ANALOGIQUE

(Vol. 3)

G. Couturier

Tel : 05 56 84 57 58

email : couturier@elec.iuta.u-bordeaux.fr

Sommaire

Echantillonnage, quantification, conversion analogique-numérique et numérique-analogique

Dans cette partie on se propose d'étudier le spectre des signaux échantillonnés et la reconstruction analogique des signaux échantillonnés. La quantification des signaux et le bruit de quantification sont également traités.

La théorie de l'échantillonnage est à la base des transmissions numériques, la transmission numérique des signaux permet :

- d'améliorer la qualité de la transmission par une meilleure immunité aux bruits, les niveaux transmis sont des niveaux logiques '0' ou '1'.
- de réaliser des traitements mathématiques sur les signaux (filtrage, codage,...).
- de procéder à un multiplexage temporel, c'est à dire transmettre plusieurs signaux sur une même voie de transmission, cas du téléphone numérique.

Dans une première étape, on aborde l'échantillonnage et la quantification sous l'aspect traitement du signal, c'est à dire sans se soucier des moyens pratiques à mettre en œuvre pour arriver aux résultats. Les techniques de conversion analogique-numérique (CAN) sont étudiées en TP électronique, la conversion numérique analogique (CNA) est étudiée en TD électronique.

I- Spectre d'un signal échantillonné : approche simplifiée

Effectuer un échantillonnage sur un signal continu e(t), c'est d'un point de vue mathématique fabriquer un nouveau signal e*(t) nul partout sauf aux instants d'échantillonnage T_e, $2T_e$, ... , nT_e, ... où e*(t) prend respectivement les valeurs $e(T_e)$, $e(2T_e)$, ..., $e(nT_e)$, ... La distribution peigne de Dirac et la transformée de Fourier sont les outils mathématiques parfaitement adaptés pour traiter le problème d'échantillonnage. Compte tenu de leurs difficultés respectives nous feront une approche un peu moins rigoureuse mais aussi plus proche de la réalité.

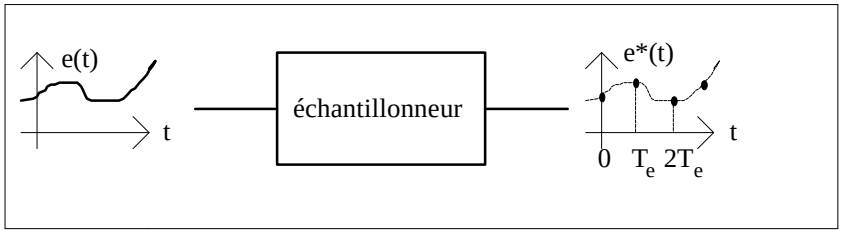


Fig. 1 opération d'échantillonnage

En pratique l'opération d'échantillonnage est réalisée par un simple interrupteur ouvert pendant une durée θ avec une période $T_e=1/F_e$; F_e est appelé la fréquence d'échantillonnage. Le signal échantillonné est noté $e^*_m(t)$ pour le distinguer du signal échantillonné théorique e*(t) mentionné précédemment. D'un point de vue mathématique, l'opération est équivalente à la multiplication du signal d'entrée e(t) par un signal h(t) égal à 1 pendant la durée θ et égal à zéro le reste du temps comme le montre la Fig. 2.

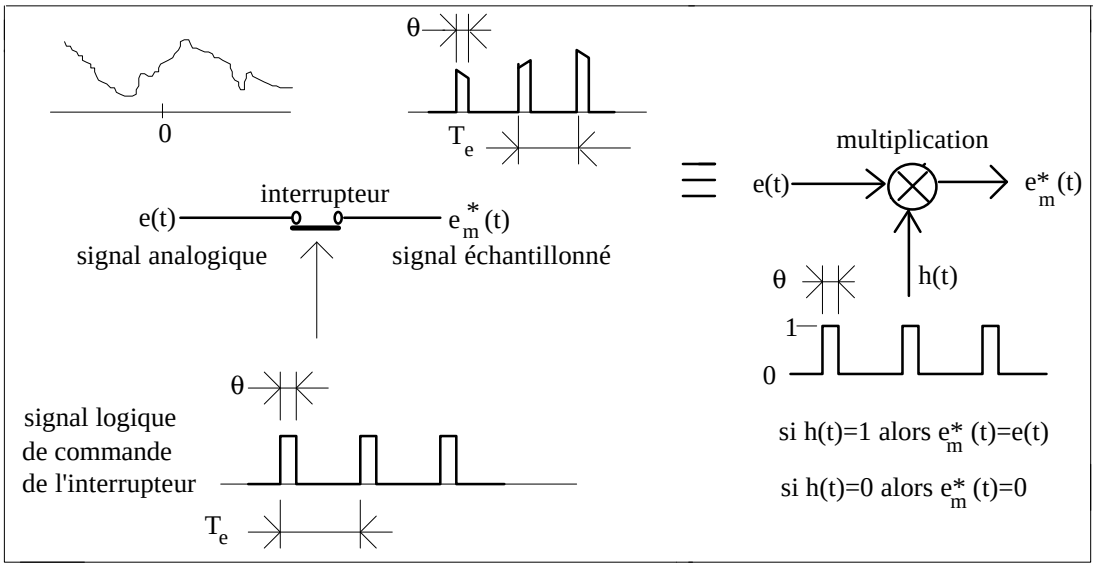


Fig. 2 réalisation de la fonction d'échantillonnage

Pour étudier le spectre du signal échantillonné nous allons supposer un signal test e(t) très simple, constitué par exemple par la somme de trois cosinusoïdes de pulsation ω, 2ω et 3ω (nous aurions pu en prendre deux, quatre, cinq, etc).

$$e(t) = \sum_{n=1}^{n=3} s_n \cos(n\omega t) \qquad \text{avec } \omega=2\pi/T \tag{1}$$

Le signal h(t), de période T_e, accepte quant à lui la série de Fourier suivante :

$$h(t) = \frac{q}{T_e} + \frac{2q}{T_e}\sum_{k=1}^{k\to\infty} \frac{\sin(pkq/T_e)}{(pkq/T_e)}\cos(kw_e t) = \sum_{k=0}^{k\to\infty} p_k \cos(kw_e t - j_k) \quad \text{avec } \omega_e=2\pi/T_e \tag{2}$$

Nous avons volontairement choisi une origine des temps au milieu de l'impulsion de largeur θ afin d'obtenir des expressions simples pour les coefficients de la série de Fourier.

Le signal échantilloné $e_m^*(t)$ s'écrit donc :

$$e_m^*(t)=e(t)h(t) = \sum_{k=0}^{k\to\infty}\sum_{n=1}^{n=3} \frac{p_k s_n}{2}\left\{\cos\left[(kw_e - nw)t - j_k\right] + \cos\left[(kw_e + nw)t - j_k\right]\right\} \tag{3}$$

Le spectre de $e_m^*(t)$ fait donc apparaître les raies suivantes :

- des composantes en ω, 2ω, 3ω correspondant respectivement aux couples (k=0 et n=1), (k=0 et n=2), (k=0 et n=3).

- des composantes de fréquence en $\omega_e \pm \omega$, $\omega_e \pm 2\omega$, $\omega_e \pm 3\omega$ correspondant respectivement aux couples (k=1 et n=1), (k=1 et n=2), (k=1 et n=3).

- des composantes de fréquence en $2\omega_e \pm \omega$, $2\omega_e \pm 2\omega$, $2\omega_e \pm 3\omega$ correspondant respectivement aux couples (k=2 et n=1), (k=2 et n=2), (k=2 et n=3).

- .....

.

- des composantes de fréquence en $q\omega_e \pm \omega$, $q\omega_e \pm 2\omega$, $q\omega_e \pm 3\omega$ correspondant respectivement aux couples (k=q et n=1), (k=q et n=2), (k=q et n=3).

- etc....

application numérique :

Soit le signal $e(t)=1\cos\left(2p\,10^3 t\right)+2\cos\left(2p\,2x10^3 t\right)+3\cos\left(2p\,3x10^3 t\right)$, les coefficients s_n valent respectivement: $s_1=1$; $s_2=2$; $s_3=3$.

D'après la relation (2), les coefficients p_k s'écrivent : $p_0 = \frac{\theta}{T_e}$ et $p_k = \frac{2q}{T_e}\left|\frac{\sin\left(kpq / T_e\right)}{\left(kpq / T_e\right)}\right|$

pour k=1, 2, 3, ..., par ailleurs $\varphi_0 = 0$ et $\varphi_k = 0$ ou π suivant le signe de $\sin(\pi k\theta / T_e)$.

Traçons les spectres de $e_m^*(t)$ pour les trois cas suivants :

α) F_e=20kHz et θ=1μs
β) F_e=20kHz et θ=10μs
γ) F_e=5.5kHz et θ=1μs

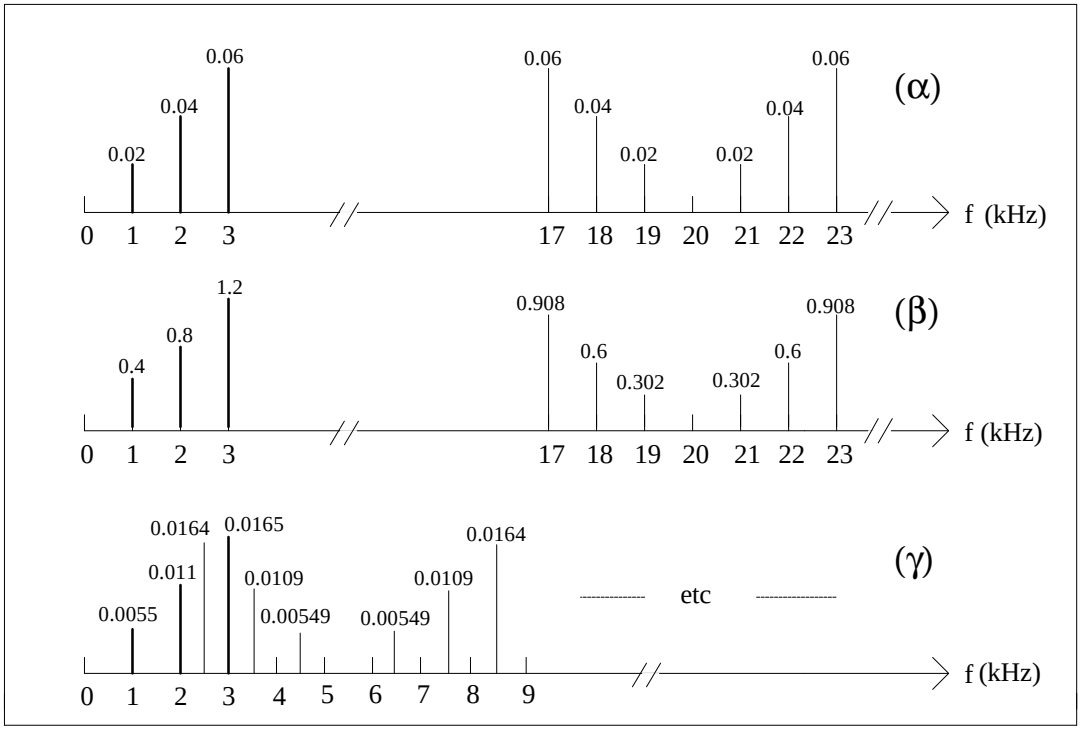


Fig. 3 Amplitudes des composantes de fréquence pour les cas α, β, et γ

Le spectre du signal échantillonné est donc constitué, dans le cas présent, d'une infinité de raies dont les amplitudes tendent vers zéro pour $f \to \infty$. L'échantillonnage est une des étapes dans une transmission numérique, la finalité évidemment est de récupérer au niveau du

récepteur le signal de départ c'est à dire e(t), une question vient naturellement à l'esprit : est-il possible de reconstruire le signal e(t) à partir du signal $e_m^*(t)$?

Compte tenu des spectres de la Fig. 3, on voit de suite que pour le cas (γ) la reconstruction de e(t) est impossible ; en effet il y a chevauchement des raies du *lobe* principal (1, 2 et 3kHz) avec les raies du deuxième *lobe*. On parle de repliement de spectre (aliasing en anglais). Il est donc impossible d'extraire par filtrage les trois composantes à 1, 2, et 3kHz.

Pour les cas (α) et (β) il est à priori possible de reconstruire e(t), il suffit de disposer d'un filtre passe-bas de reconstruction ayant une fréquence de coupure basse F_c telle que : $3kHz<F_c<17kHz$. La largeur θ de l'impulsion a pour seul effet de modifier l'amplitude des raies. Dans le cas général, la reconstruction est donc possible si :

1) le signal e(t) ne contient aucune énergie au-delà d'une certaine fréquence notée B_{max}.

2) le signal est échantillonné à une fréquence $F_e \geq 2B_{max}$ pour éviter les repliements de spectre, c'est le théorème connu sous le nom de théorème de Shannon (1916 - ?). La fréquence $F_e/2$ est appelée fréquence de *Nyquist*.

3) on dispose d'un filtre passe-bas de reconstruction idéal ayant une fréquence de coupure basse F_c telle que : $B_{max}<F_c<F_e-B_{max}$.

II - <u>Aspects pratiques de l'échantillonnage</u>

α) filtre de reconstruction

En supposant que le signal e(t) ne contient aucune composante de fréquence au-delà de B_{max}, il est effectivement possible de satisfaire le théorème de Shannon. Est-il possible de réaliser le filtre passe-bas de fréquence de coupure F_c ?

La réponse en fréquence H(f) du filtre doit satisfaire :

$$\begin{aligned} H(f) &= T_e/\theta \quad \text{pour} \quad -F_c < f < F_c \\ &= 0 \qquad \text{ailleurs} \end{aligned} \tag{4}$$

Intéressons nous à la réponse impulsionnelle h(t) de ce filtre. Elle est donnée par la transformée de Fourier inverse de H(f) :

$$h(t) = \int_{-\infty}^{\infty} H(f) e^{jwt} df = \int_{-F_c}^{F_c} \frac{T_e}{q} e^{jwt} df = \frac{T_e}{q} \frac{e^{j2pF_c t} - e^{-j2pF_c t}}{j2pt} = 2F_c \frac{T_e}{q} \frac{\sin(2pF_c t)}{2pF_c t} \tag{5}$$

Ce filtre est non *causal*, en effet sa réponse impulsionnelle est différente de zéro pour t<0, il n'est donc pas réalisable en temps réel. On ne peut procéder qu'à une reconstruction approchée de e(t). La reconstruction sera d'autant meilleure que :

- la courbe de réponse sera plate entre 0 et B_{max}
- l'atténuation sera grande au-delà de B_{max}

En pratique on utilise, souvent un boqueur d'ordre zéro (un convertisseur analogique numérique est équivalent à un bloqueur d'ordre zéro). Il s'agit d'un filtre qui assure un niveau constant pendant une durée T_e comme le montre la Fig. 4.

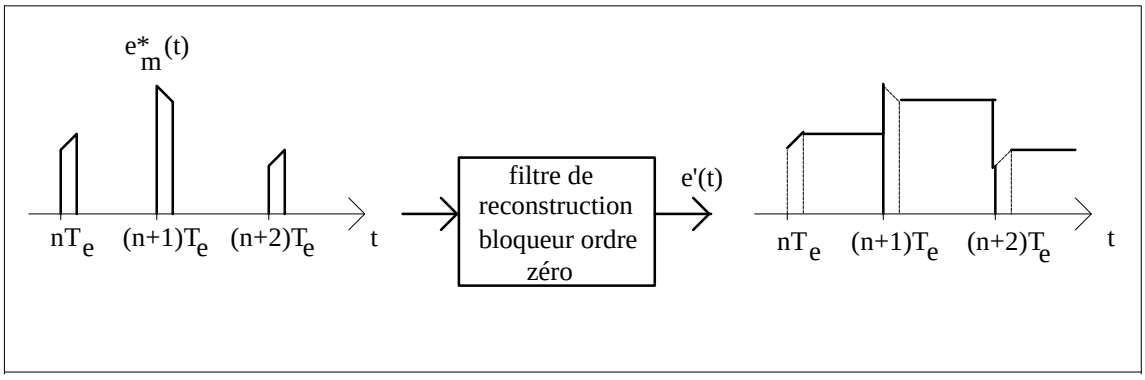


Fig. 4 Reconstruction par bloqueur d'ordre zéro

Un bloqueur d'ordre zéro idéal est un filtre dont la réponse impulsionnelle se met sous la forme : $h(t) = U(t) - U(t-T_e)$

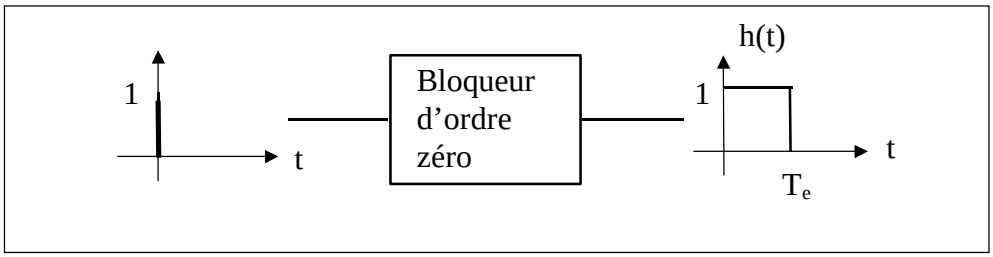


Fig. 5 Bloqueur d'ordre zéro

Cherchons la réponse en fréquence H(f) du bloqueur d'ordre zéro : elle est donnée par la transformée de Fourier de h(t), soit :

$$H(f) = \int_0^{T_e} e^{-jwt} dt = \left[\frac{e^{-jwt}}{-jw} \right]_0^{T_e} = T_e e^{-jwT_e/2} \frac{\sin(wT_e/2)}{wT_e/2} \qquad (6)$$

Sur la Fig. 6, nous avons représenté l'allure du module de H(f).

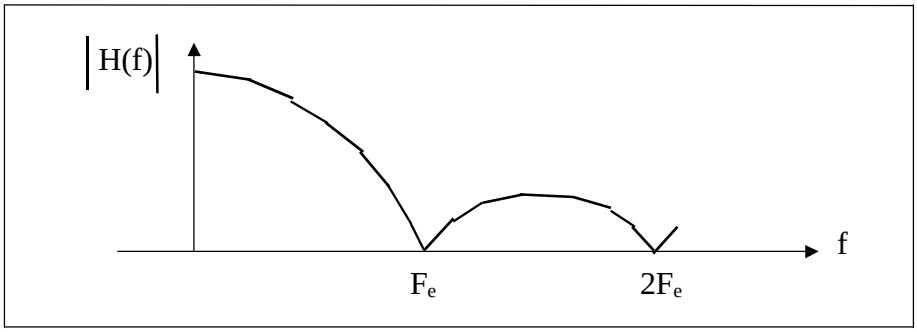


Fig. 6 Réponse en fréquence du bloqueur d'ordre zéro

L'utilisation d'un bloqueur d'ordre zéro dans le but de récupérer le signal e(t) après échantillonnage conduit à la présence de raies autres que celles du spectre de e(t) comme le montre la Fig. 7.

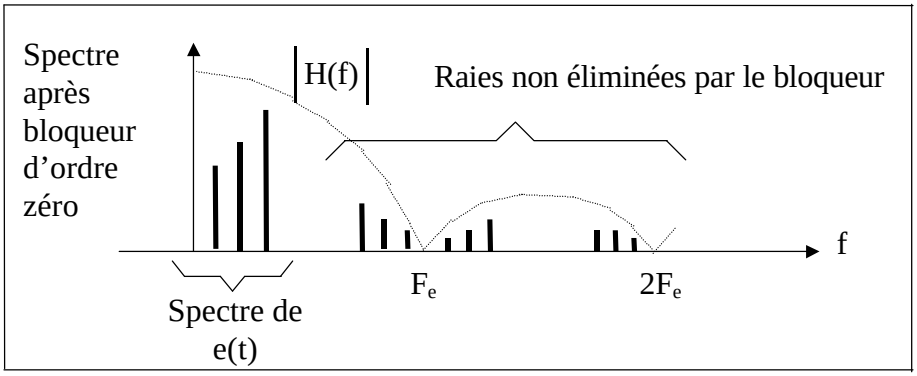


Fig. 7 Spectre du signal après passage dans le bloqueur d'ordre zéro

Après passage dans le bloqueur d'ordre zéro le signal présente des ''marches d'escalier'', celles-ci se traduisent dans le domaine des fréquences par les raies supplémentaires de la Fig. 7.

β) filtre antirepliement (antialiasing filter) : son rôle

Dans le cas (γ) de la Fig. 3 nous avons vu qu'il était impossible de reconstruire correctement le signal, car il y a un chevauchement entre le lobe principal et les lobes secondaires. Dans la plupart des situations, le spectre du signal à échantillonner s'étale dans le domaine des fréquences tout en diminuant du coté des hautes fréquences, mais il n'existe pas une fréquence B_{max} au-delà de laquelle l'énergie est nulle. Il y a donc un problème pour choisir la fréquence d'échantillonnage ; c'est par exemple le cas de la parole.

Grosso modo le spectre de la parole s'étend jusqu'à environ quinze-vingt kHz (en toute rigueur il faut parler de densité spectrale car la parole est un signal aléatoire et dans ce cas on ne peut calculer ni série de Fourier, ni transformée de Fourier, par contre on peut définir une densité spectrale : c'est la valeur quadratique moyenne par unité de fréquence). Dans le cas des CDROM le signal de parole est échantillonné à 44.1 kHz, dans le cas du téléphone numérique le signal est échantillonné à 8 kHz seulement. La Fig. 8 représente très schématiquement le spectre du signal avant et après échantillonnage dans le cas du téléphone numérique.

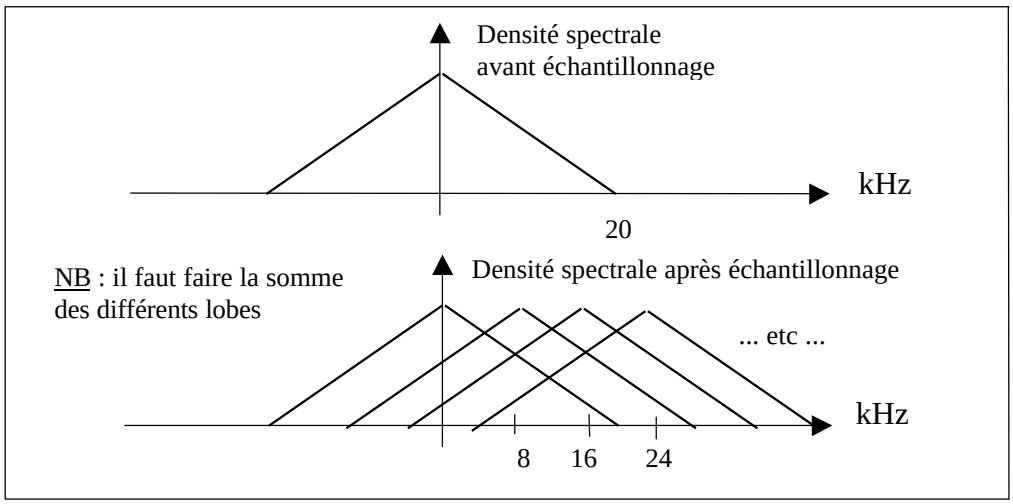


Fig. 8 Densités spectrales avant et après échantillonnage dans le cas d'une fréquence d'échantillonnage à 8 kHz

Il est clair que la récupération est impossible, même le spectre des basses fréquences est perturbé. En effet, prenons par exemple le cas de la composante à 1 kHz, elle provient :

1) du 1 kHz du lobe principal, c'est la seule intéressante
2) du 8kHz-7kHz : 1er lobe secondaire (repliement autour de 8 kHz)
3) du |8kHz-9kHz| : 1er lobe secondaire (repliement autour de 8 kHz)
4) du 16kHz-15kHz : 2ème lobe secondaire (repliement autour de 16 kHz)
5) du |16kHz-17kHz| : 2ème lobe secondaire (repliement autour de 16 kHz)

En téléphonie, on estime que le message est compréhensible pourvu que les composantes basses fréquences soient transmises correctement. Pour remédier aux repliements de spectre qui modifient les basses fréquences, on place avant l'échantillonneur un filtre passe-bas, dit filtre antirepliement (*antialiasing filter*).

Les spectres des signaux avant et après filtrage puis après échantillonnage sont représentés sur la Fig. 9. En téléphonie numérique, la fréquence de coupure du filtre antirepliement est de 3.4kHz. Il est donc possible de récupérer le spectre des basses fréquences par un filtre de reconstruction, évidemment celui-ci doit présenter une pente d'atténuation élevée car la fenêtre spectrale est étroite (entre 3.4kHz et 4.6kHz). Ce filtre est en général réalisé en utilisant la technique des capacités commutées.

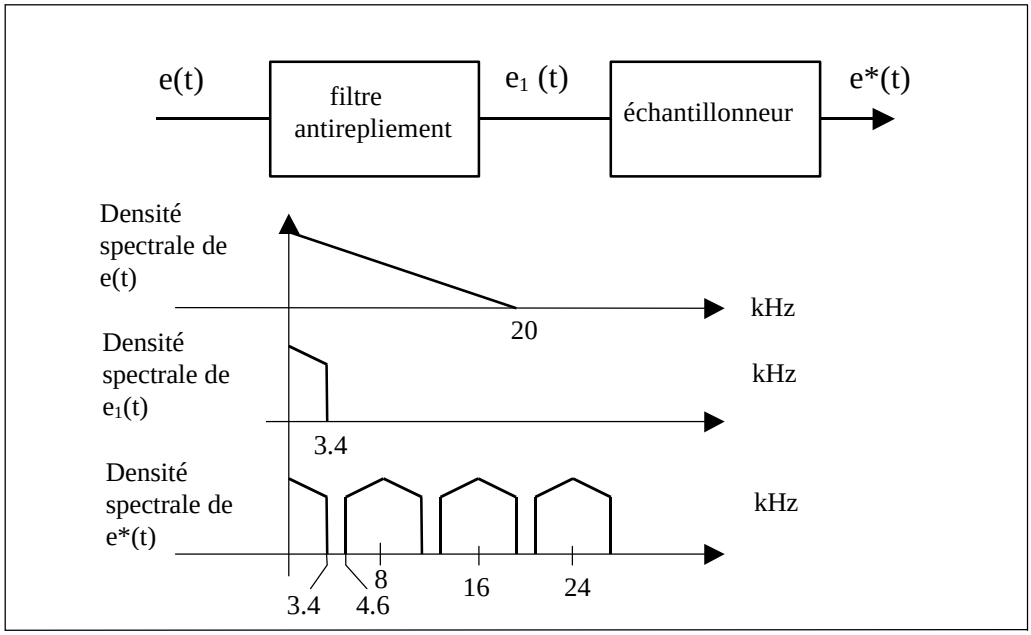


Fig. 9 Filtre antirepliement en téléphonie numérique

III- Quantification

En pratique, on n'échantillonne pas un signal pour le reconstruire juste après. L'échantillonnage est utilisé pour prélever le signal à des instants multiples de T_e et ensuite convertir les échantillons sous forme d'un code binaire (8, 12, 16 bits, ...). En général, juste derrière l'échantillonneur on place un bloqueur pour maintenir le signal constant à l'entrée du convertisseur analogique-numérique (CAN) pendant la durée de conversion, nombre de CAN fonctionnent cependant sans bloqueur. Le schéma de principe d'un échantillonneur-bloqueur (*Sample and Hold*) est donné à la Fig. 10, (exemple de circuit : AD585 de *Analog Devices*, voir annexe I). En téléphonie ou télévision numérique le signal codé module une porteuse en phase, à la réception un démodulateur transforme de nouveau le signal reçu en code et un convertisseur-numérique analogique permet de restituer un signal analogique (en pratique c'est un peu plus compliqué, en effet avant de moduler la porteuse on effectue une compression

d'information pour diminuer le débit binaire, puis un codage de voie pour se prémunir d'éventuelles erreurs de transmission). A la réception, on trouve les opérations inverses comme le montre la Fig. 11.

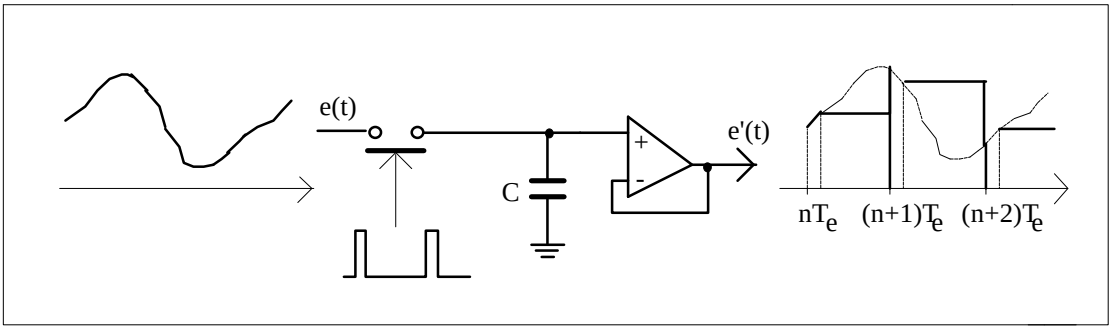


Fig. 10 Schéma de principe d'un échantillonneur-bloqueur

La modulation sert à transposer l'information autour d'une fréquence élévée appelée porteuse, globalement les opérations de modulation et démodulation sont équivalentes à une opération unitaire, il en est de même des opérations de codage de voie et décodage de voie puis de compression et décompression. Finalement l'ensemble émetteur plus récepteur peut se ramener à une chaîne équivalente, représentée en Fig. 11, dans laquelle le CNA reçoit le code du CAN.

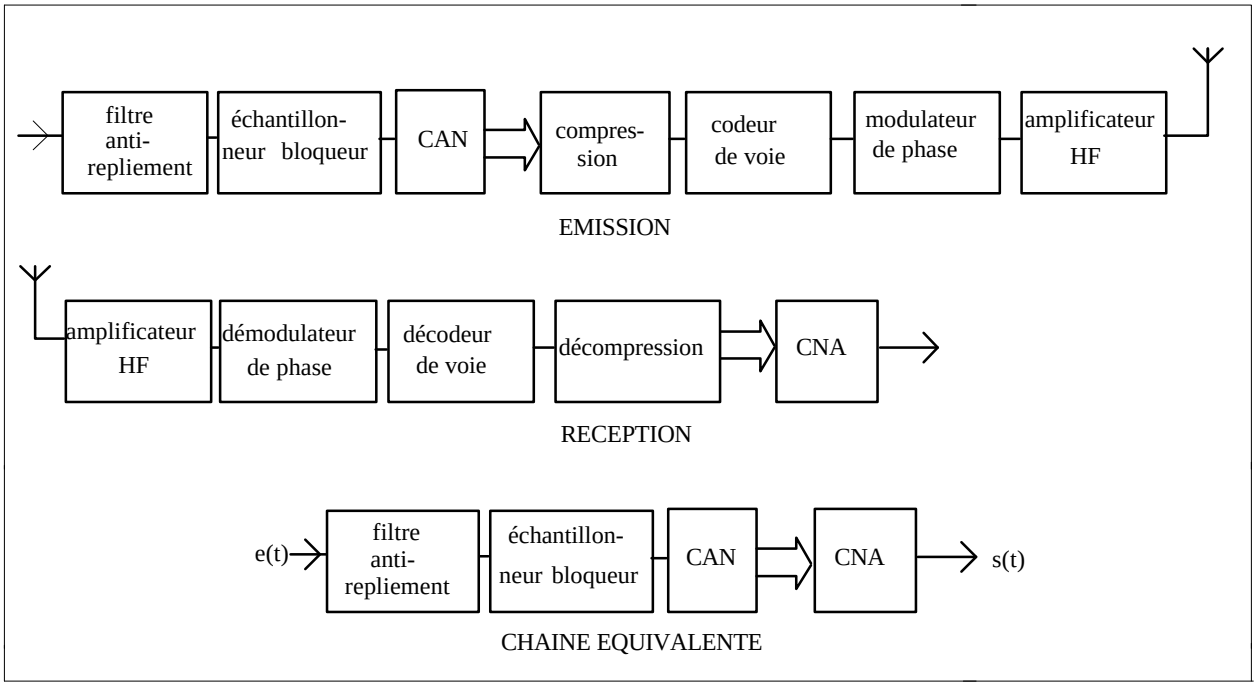


Fig. 11 Synoptique d'un émetteur et d'un récepteur numérique

Chaque échantillon $e(nT_e)$ du signal e(t) est converti sous forme d'un code binaire, en supposant par exemple un codage sur 8 bits, on aura par exemple la suite des codes suivants :

valeurs des échantillons ... e((n-1)Te) / e(nTe) / e((n+1)Te) / ...

codes binaires correspondant ... 00010101 / 00011100 / 00100000 / ...

La Fig. 12 illustre par exemple le cas d'un CAN *unipolaire linéaire* (tension positive seulement et tous les quanta sont égaux) ; on appelle quantum la quantité $V_{max}/2^N$ avec N le nombre de bits utilisés pour la conversion. Il s'agit d'un cas particulier de CAN, on verra par la suite, qu'on a souvent intérêt à choisir des plages q_k d'autant plus petites que le signal e(t) est faible, ceci pour minimiser les erreurs.

A la réception, le CNA délivre une tension analogique, mais la seule connaissance du code ne permet pas de restituer à l'instant $(nT_e+\varepsilon)$ un palier d'amplitude égale à $e(nT_e)$, (ε est un retard introduit par les temps de conversion des CAN et CNA). En effet, le code 00011100 par exemple, veut seulement dire que l'échantillon $e(nT_e)$ vérifie l'inégalité suivante (voir Fig. 12) :

$$V_k < e(nT_e) < V_{k+1} \quad \text{ou encore} \quad V_k < e(nT_e) < V_k + q_k$$

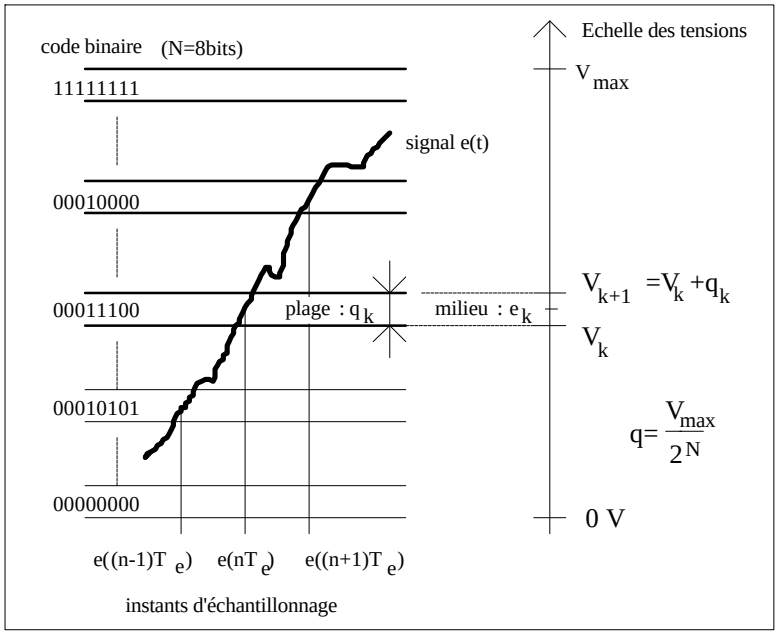


Fig. 12 Codage d'un signal échantillonné sur 8 bits par un CAN

A la réception, il y a donc ambiguité, en effet plusieurs possibilités existent, on peut par exemple attribuer la valeur V_k, V_{k+1}, $V_k+q_k/2$ ou encore à priori toutes autres valeurs entre V_k et V_{k+1}. En choisissant la valeur $V_k+q_k/2$, c'est à dire le milieu de la plage, on minimise l'erreur ; on parle alors de restitution avec demi-échelon. La différence entre le signal e(t) et le signal restitué s(t) en sortie du CNA est appelé bruit de quantification, voir la Fig. 13 ci-dessous.

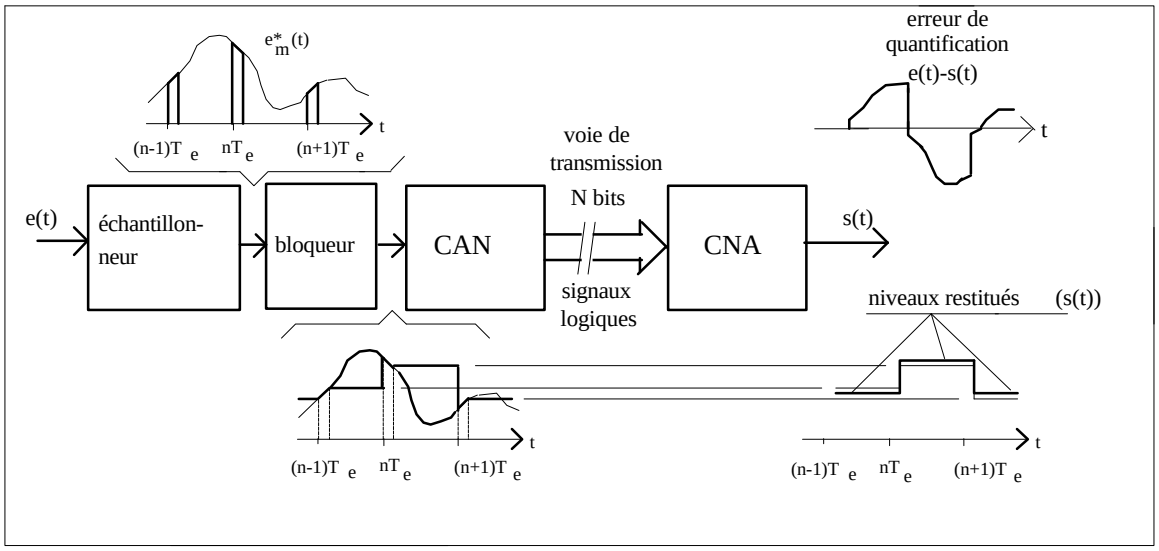


Fig. 13 Illustration du bruit de quantification introduit par les CAN et CNA

La Fig. 14 donne le graphe du signal restitué s(t) en fonction du signal d'entrée e(t), on obtient une loi en marches d'escalier (en téléphonie numérique cette loi est appelée : loi Européenne).

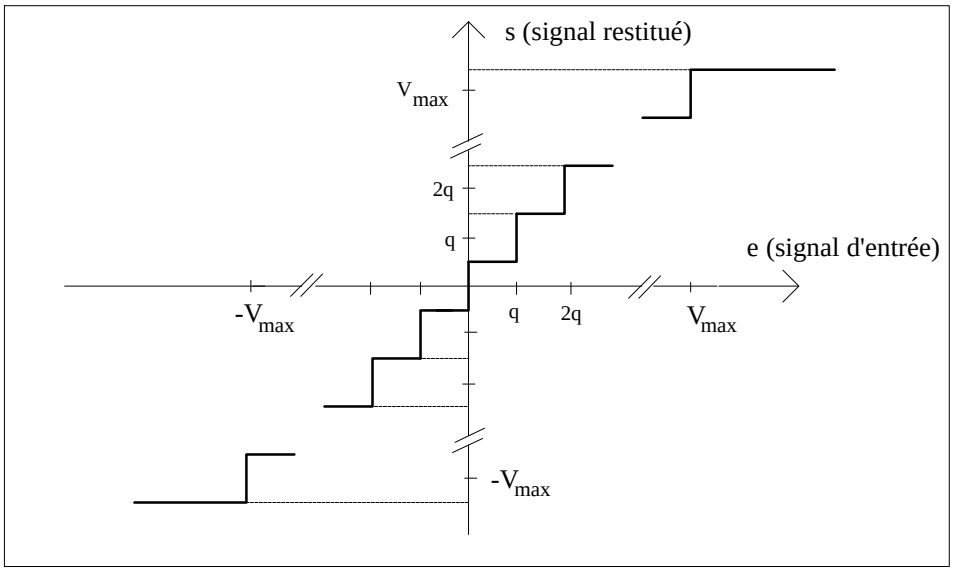


Fig. 14 Graphe représentant le signal restitué en fonction du signal d'entrée

On pourrait également envisager une restitution telle que celle représentée sur la Fig. 15 (en téléphonie numérique, cette loi est appelée : loi Américaine à mi-marche, elle est utilisée aux Etats Unis). Du point de vue du bruit cette loi a un net avantage sur la loi de la Fig. 14. En effet, même en l'absence de signal, il y a toujours du bruit : $e(t)=0+\delta(t)$, avec $\delta(t)$ le bruit. Il s'ensuit que dans le cas de la loi de la Fig. 14 le signal de sortie oscillera entre +q/2 et -q/2 si $-q < \delta(t) < q$, dans le cas de la loi de la Fig. 15 le signal de sortie restera à zéro si $-q/2 < \delta(t) < q/2$.

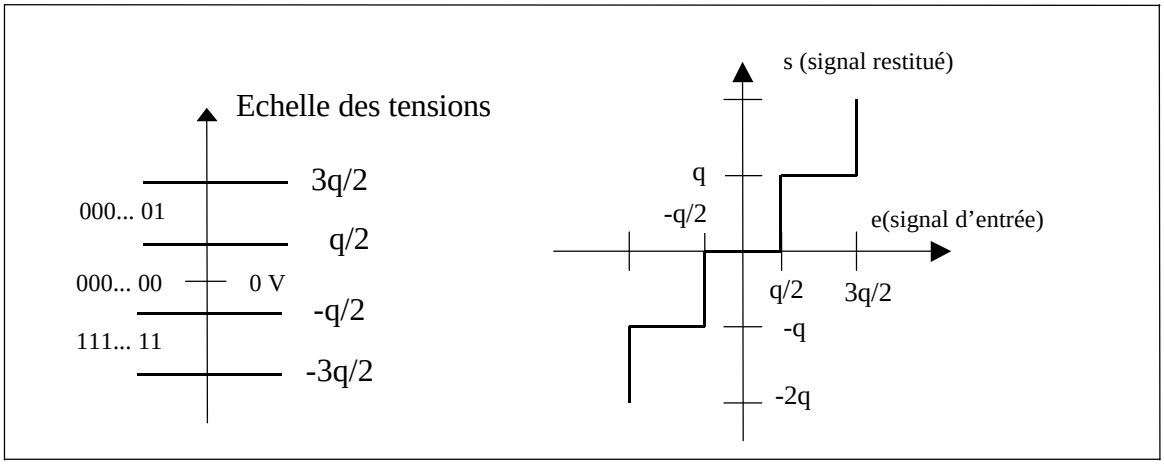


Fig. 15 Autre loi de restitution du signal utilisée en téléphonie aux Etats-Unis

IV- Bruit de quantification et choix du nombre de bits

Il est clair que le bruit de quantification sera d'autant plus génant que le signal e(t) sera de faible amplitude. Pour les fortes valeurs de e(t) le bruit est pratiquement insignifiant. Plus que le bruit, c'est le rapport signal/bruit (S/B) qui est important. Le rapport S/B en dB est donné par :

$$\left(\frac{S}{B}\right)_{dB} = 10\log_{10}\left(\frac{\textit{valeur}\ \text{quadratique moyenne du signal}}{\text{valeur quadratique moyenne de bruit}}\right) \qquad (7)$$

On rappelle que la valeur quadratique moyenne d'un signal est égale à sa valeur efficace au carré. Les signaux e(t) et de bruit ne sont pas des signaux *déterministes* mais des signaux *aléatoires*, on ne dispose pas d'expressions analytiques pour calculer les valeurs quadratiques moyennes. Dans le cas des signaux aléatoires on introduit la notion de *variance* que l'on calcule à partir des lois de probabilité.

$$\left(\frac{S}{B}\right)_{dB} = 10\log_{10}\left(\frac{\text{variance du signal}}{\text{variance du bruit}}\right) \qquad (8)$$

Dans les lignes qui suivent on montre l'équivalence entre valeur quadratique et variance puis on calcule le rapport (S/B). La relation (13) donne le résultat final. Dans une première lecture on peut se reporter directement à la relation (13).

α) *cas d'une variable aléatoire continue*

Prenons le cas particulier du signal f(t) périodique suivant :

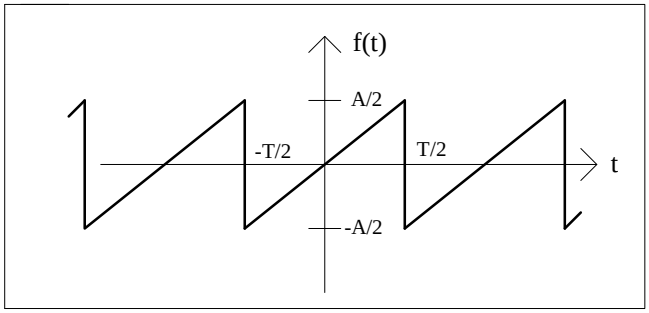


Fig. 16 Signal test pour la définition de la variance d'une variable aléatoire continue

La valeur quadratique moyenne du signal est donnée par :

$$\frac{1}{T}\int_{-T/2}^{T/2} f^2(t)dt = \frac{1}{T}\int_{-T/2}^{T/2}\left(\frac{A}{T}t\right)^2 dt = \frac{A^2}{12}$$

Considérons les valeurs de la fonction f(t) comme les valeurs d'une variable aléatoire X, ces valeurs sont comprises entre -A/2 et A/2 et avec une équiprobabilité d'être entre -A/2 et A/2. Il est alors facile de calculer la probabilité p(x) d'avoir la valeur particulière x de la variable X, p(x) doit en effet vérifier :

$$\int_{-A/2}^{A/2} p(x)dx = 1 \text{ avec p(x) = C}^{te} \Rightarrow p = \frac{1}{A}$$

Dans le cas général d'une variable aléatoire continue la variance est définie par :

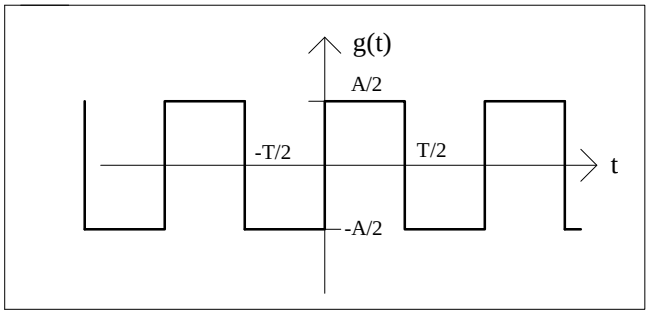
$$s^2 = \int_{x_{min}}^{x_{max}} (x-m)^2 p(x)dx \text{ avec m = valeur moyenne et } x_{min} \le x \le x_{max} \qquad (9)$$

D'après le tracé de la Fig. 16 ; m=0, x_{min}=-A/2 et x_{max}=A/2 d'où :

$$s^2 = \frac{1}{A}\int_{-A/2}^{A/2} x^2 dx = \frac{A^2}{12}.$$

On vérifie bien que variance et valeur quadratique ont même valeur : $\frac{A^2}{12}$.

β) <u>*cas d'une variable aléatoire discontinue*</u>
Prenons maintenant le cas du signal g(t) périodique suivant :



<u>Fig. 17</u> Signal test pour la définition de la variance d'une variable aléatoire discontinue

La valeur quadratique moyenne du signal est donnée par :

$$\frac{1}{T}\int_{-T/2}^{T/2} g^2(t)dt = \frac{1}{T}\left(\int_{-T/2}^{0}\left(-\frac{A}{2}\right)^2 dt + \int_{0}^{T/2}\left(\frac{A}{2}\right)^2 dt\right) = \frac{A^2}{4}$$

Considérons les valeurs de la fonction g(t) comme les valeurs d'une variable aléatoire X. Celle-ci peut prendre deux valeurs distinctes ; x_1=-A/2 avec une probabilité p_1=1/2 ou x_2=A/2 avec une probabilité p_2=1/2. On vérifie aisément que : $p_1+p_2=1$.

Dans le cas général d'une variable aléatoire discrète X pouvant prendre N valeurs différentes x_1, x_2, x_3, ..., la variance est définie par :

$$s^2 = \sum_{i=1}^{N}\left(x_i - m\right)^2 p_i \quad \text{avec} \quad \mathrm{m} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{x_i} = \text{valeur moyenne} \qquad (10)$$

D'après le tracé de la Fig. 17, N=2 et m=0, d'où : $s^2 = \left(-\frac{A}{2}\right)^2\frac{1}{2} + \left(\frac{A}{2}\right)^2\frac{1}{2} = \frac{A^2}{4}$. On vérifie bien que variance et valeur quadratique ont même valeur : $\frac{\mathrm{A}^2}{4}$.

Déterminons dans un premier temps la variance σ_B^2 du bruit de quantification. Pour un signal e(t) tombant par exemple dans la plage q_k (voir Fig. 12), le signal restitué est le milieu de la plage soit e_k, l'erreur ε(t) est donc égal à : $e(t)-e_k$. Pour déterminer la variance, il faut connaître la probabilité p(e) d'avoir la valeur e et sommer sur toutes les plages q_k. D'après les relations (9) et (10), σ_B^2 s'écrit :

$$s_B^2 = \sum_{plages\ k}\left[\int_{e_k - q_k/2}^{e_k + q_k/2}\left(e - e_k\right)^2 p(e)de\right] \qquad (11)$$

Si le nombre de plages est élevé, on peut faire l'hypothèse qu'à l'intérieur d'une plage q_k, on a : $p(e) \approx p(e_k)$. La relation précédente s'écrit alors :

$$s_B^2 = \sum_k\left[p(e_k)\int_{e_k - q_k/2}^{e_k + q_k/2}\left(e - e_k\right)^2 de\right] = \sum_k p(e_k)\frac{q_k^3}{12} \qquad (12)$$

Dans le cas d'une quantification linéaire, toutes les plages ont la même valeur : q_k=q, d'où : $\sigma_B^2 = \frac{\mathrm{q}^3}{12}\sum_{\mathrm{k}} \mathrm{p(e_k)}$. On montre facilement que : $\sum_{\mathrm{k}} \mathrm{p(e_k)} = 1/\mathrm{q}$, en effet si la tension e est comprise dans l'intervalle [$-V_{max}$, V_{max}], alors : $\int_{-V_{max}}^{V_{max}} p(e)de = 1$. Découpons l'intervalle

$[-V_{max}, V_{max}]$ en intervalles de largeur q, il vient :

$$\int_{-V_{max}}^{-V_{max}+q} p(e)de + \int_{-V_{max}+q}^{-V_{max}+2q} p(e)de + \ ... \ + \int_{-V_k}^{-V_{k+1}} p(e)de + \ ... \ + \int_{V_{max}-q}^{V_{max}} p(e)de = 1$$

Avec l'hypothèse $p(e) \approx p(e_k)$, il vient : $\sum_k p(e_k)q = 1$. Il s'ensuit que : $\sigma_B^2 = \frac{q^2}{12}$.

Il nous faut maintenant calculer la variance σ_e^2 du signal e, elle dépend évidemment de la manière dont le signal est réparti dans l'intervalle $[-V_{max}, V_{max}]$. Supposons un signal réparti par exemple entre les valeurs $-V_e$ et V_e, avec $V_e < V_{max}$, il vient :

$$s_e^2 = \int_{-V_e}^{V_e} (e-m)^2 p(e)de \quad \text{avec m = valeur moyenne.}$$

Fin de 🕮

Si on fait par exemple l'hypothèse d'un signal uniformément réparti sur l'intervalle $[-V_e, V_e]$, alors m=0 et $\int_{-V_e}^{V_e} p(e)de = 1$ avec p(e) = p = C^te, d'où $p = 1/2V_e$. Il s'ensuit que : $s_e^2 = \frac{1}{2V_e}\int_{-V_e}^{V_e} e^2 de = \frac{V_e^2}{3}$.

Le rapport $(S/B)_{dB}$ s'écrit : $\left(\frac{S}{B}\right)_{dB} = 10\log_{10}\left(\frac{V_e^2/3}{q^2/12}\right)$ avec $q = \frac{2V_{max}}{2^N}$, N est le nombre de bits. On obtient finalement :

$$\left(\frac{S}{B}\right)_{dB} = 6N + 10\log_{10}\left(\frac{V_e^2}{V_{max}^2}\right) = 6N + 10\log_{10} V_e^2 - 10\log_{10} V_{max}^2 \qquad (13)$$

Dans un système d'axe $\left\{(S/B)_{dB}\ ,\ 10\log_{10} V_e^2\right\}$, on obtient une droite comme le montre la Fig. 18.

Le rapport (S/B) se dégrade au fur et à mesure que le signal diminue. On peut bien sûr augmenter le rapport (S/B) en augmentant le nombre de bits de conversion, mais on augmente du même coup le débit binaire, en conséquence il faut une largeur de canal plus importante pour transmettre l'information. En fait ce sont les faibles niveaux qui sont principalement affectés par le bruit de quantification, il paraît donc intéressant de diminuer la taille des plages pour les faibles niveaux comme le montre la Fig. 19. La quantification n'est alors plus linéaire, on dit qu'il y a compression de la taille des plages.

On peut par exemple chercher les tailles des plages conduisant à un rapport (S/B) constant ; les plages doivent alors vérifier :

$$\left(\frac{S}{B}\right) = C \ \Rightarrow \ \frac{e_k}{q_k/2} = C \quad \forall k \qquad (14)$$

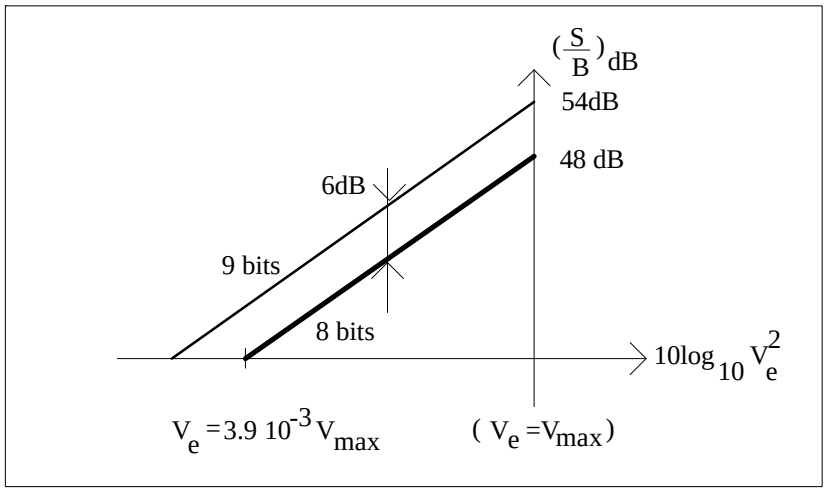


Fig. 18 Rapport $(S/B)_{dB}$ en fonction de $10\log_{10} V_e^2$

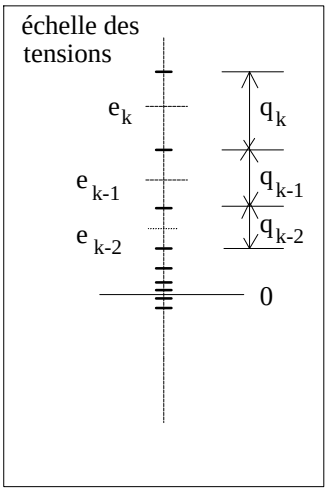


Fig. 19 Quantification non linéaire

En appelant e_k, e_{k-1}, e_{k-2}, ... les milieux des plages successives q_k, q_{k-1}, q_{k-2}, ... on obtient les relations suivantes :

$e_k - q_k/2 = e_{k-1} + q_{k-1}/2$ ou encore : $e_k\left[1-\frac{1}{C}\right] = e_{k-1}\left[1+\frac{1}{C}\right]$

$$e_{k-1}\left[1-\frac{1}{C}\right] = e_{k-2}\left[1+\frac{1}{C}\right]$$

$$e_{k-2}\left[1-\frac{1}{C}\right] = e_{k-3}\left[1+\frac{1}{C}\right]$$

.

.

etc

.

Il s'ensuit la relation de récurrence : $e_k = e_0\left[\frac{C+1}{C-1}\right]^k$ d'où : $k = \frac{\log\left(\frac{e_k}{e_0}\right)}{\log\left(\frac{C+1}{C-1}\right)}$, il s'agit d'une loi de compression logarithmique. Pour des raisons pratiques, il est impossible de diminuer indéfiniment la taille des plages. On définit alors une plage minimum dont le milieu est par exemple la valeur e_0, toutes les plages situées en dessous de cette plage sont identiques, les plages situées au-dessus suivent une loi logarithmique, voir la Fig. 20. Toute loi de compression est composée de deux parties : une partie linéaire jusqu'à une valeur e_0, suivie d'une loi logarithmique.

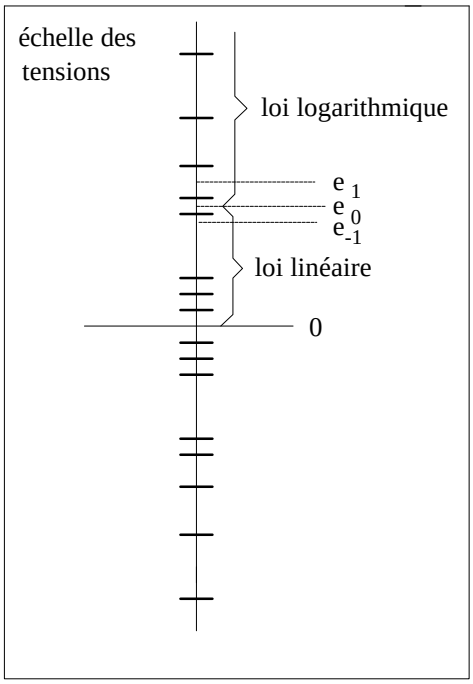


Fig. 20 Loi de compression logarithmique

En Europe, la téléphonie utilise une loi de compression dite loi en **A**, aux Etats Unis on utilise une autre loi appelée loi en **μ**. La loi en **A** est telle que le partage entre loi linéaire et loi logarithmique a lieu pour $e_0=V_{max}/87.6$. Dans la partie linéaire la quantification est équivalente à une quantification sur 12 bits et on obtient encore une relation linéaire entre $20\text{Log}_{10}(S/B)$ et $10\log_{10} V_e^2$. Au-delà de $V_{max}/87.6$ le rapport (S/B) est une constante comme le montre la Fig. 21, (en fait il varie linéairement par tronçon, car c'est le raport $e_k/(q_k/2)$ qui est une constante et non $e/(q_k/2)$) .

En pratique, la loi **A** est approximée par des segments de droite, huit au total, comme le montre la Fig. 22. Pour coder un échantillon sur 8 bits, on procède alors en deux étapes :

1) un codage binaire sur 12 bits à l'aide d'un convertisseur analogique-numérique linéaire.

2) une compression numérique, il s'agit d'une opération purement logique (registre à décalage, compteur, ..)

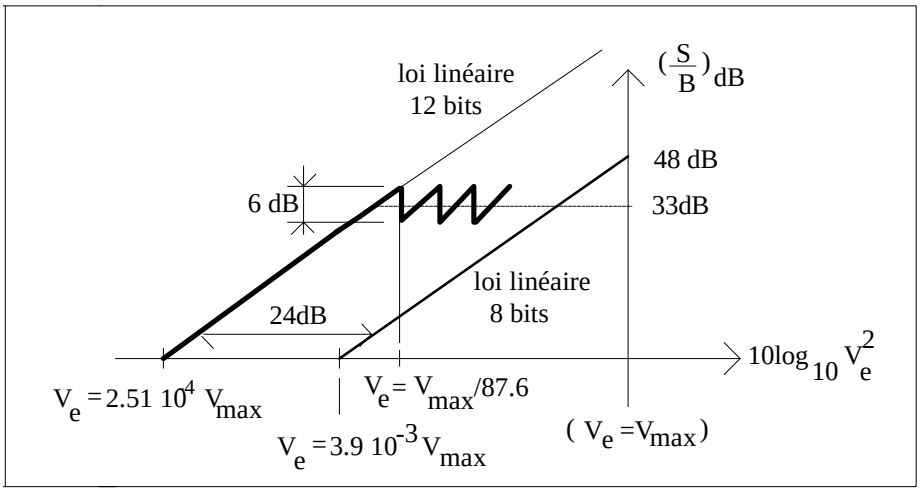


Fig. 21 Rapport $(S/B)_{dB}$ en fonction de $10\log_{10} V_e^2$ pour une compression logarithmique

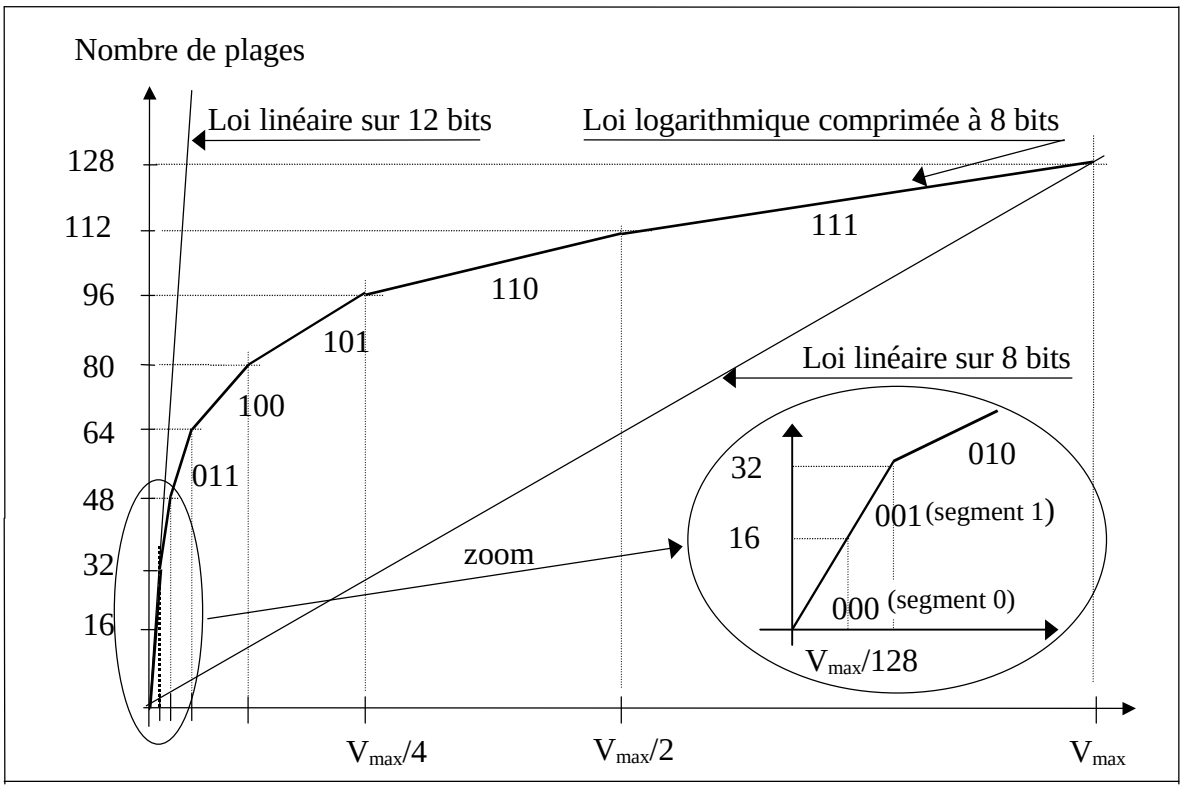


Fig. 22 Approximation de la loi logarithmique par huit segments de droite

On peut établir une correspondance simple entre la quantification sur 12 bits et la compression sur 8 bits.

Niveaux sur 12 bits	Niveaux sur 8 bits	n° du segment	code segment
2047 à 1024	127 à 112	7	111
1023 à 512	111 à 96	6	110
511 à 256	95 à 80	5	101
255 à 128	79 à 64	4	100
127 à 64	63 à 48	3	011
63 à 32	47 à 32	2	010
31 à 16	31 à 16	1	001
15 à 0	15 à 0	0	000

Les échantillons sont codés sous 8 bits de la manière suivante :

- 1 bit de signe
- 3 bits pour le numéro du segment de droite sur la caractéristique
- 4 bits pour la position sur le segment de droite

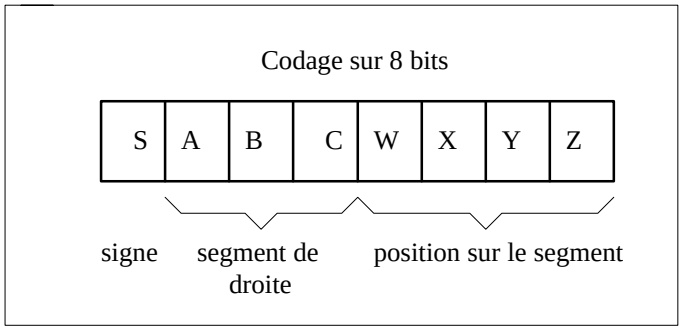


Fig. 23 Codage des échantillons sur 8 bits

Les constructeurs de circuits intégrés commercialisent des circuits intégrant ce type de codage, les circuits portent le nom de CODEC (COder/DECoder), ils sont principalement dédiés à la téléphonie.

(ex : MC14LC5480 - *Motorola* ; annexe IX)

V- Les différents types de CAN et CNA

V- 1- Les convertisseurs analogique-numérique (CAN)

Un CAN est caractérisé par :

- sa **résolution** : elle est fixée par le nombre de bits de conversion
- sa **précision** : elle est fixée par les valeurs des seuils V_k de la Fig. 12. Les seuils déterminent la linéarité du convertisseur. En général, la précision est de (1/2)LSB (Least Significant Bit), la dimension d'une plage est alors comprise entre 1/2 et (3/2)LSB. Un bon CAN n'a pas de code manquant (*no missing code*), c'est à dire que toutes les configurations binaires existent, la Fig. 24 montre un CAN avec et sans code manquant.
- son **temps de conversion** : il est fixée par la structure du CNA
- sa **pleine échelle** (*Full Scale Range* FSR) : c'est la tension maximum acceptable, FSR= $2^N q$ pour un convertisseur linéaire.

Quatre types de CAN sont commercialement disponibles :

(pour une description détaillée des CAN à Approximations Successives, Double Rampes et Flash, on se reportera au cahier de TP électronique 2ème année).

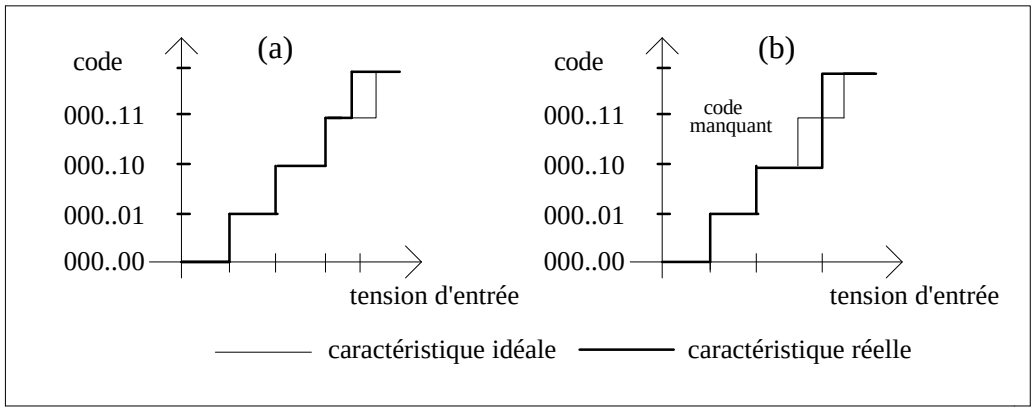


Fig. 24 Caractéristiques de deux CAN, (a) no missing code, (b) missing code

a) Approximations Successives (*Successive Approximation Register ADC* ou SAR)
principe : méthode dichotomique, recherche du bit de poids forts (Most Significant Bit) en divisant la pleine échelle par deux, puis recherche du bit suivant, etc ... . Les CAN de type SAR utilisent en interne un CNA.
avantage : précis et rapide, le temps de conversion dépend du nombre de bits.
(ex : AD670 - *Analog Device* : 8 bits, 10 µs, voir annexe II).

b) Double Rampes (*Dual Slope ADC*)
principe : charge et décharge de condensateur à courant constant et comptage d'impulsions
avantage : très précis et bien adapté à la haute résolution. Ils peuvent présenter une réjection totale du 50 Hz, 100 Hz, ... , si la durée de la première rampe est un multiple de 20ms. Ils sont principalement utilisés dans les multimètres numériques.
inconvénient : temps de conversion important, qq. 10ms.

c) Flash (Flash ADC)
principe : comparaison simultanée de la tension à convertir aux 2^N-1 seuils V_k, puis opération logique pour déterminer le code binaire.
avantage : très rapide
inconvénient : il faut intégrer (2^N-1) comparateurs et 2^N résistances qui sont ajustées par laser (*laser trimmed*). Il est difficile de réaliser des CAN Flash de plus de dix bits. Le temps de conversion est très faible.
(ex : AD9060 - *Analog Device* : 10 bits, 13ns, voir annexe III)

NB : On trouve sur le marché des CAN semi-flash, ces CAN opèrent en deux temps comme le montre le schéma de principe de la Fig. 25. Dans un premier temps, on détermine les N/2 bits de poids forts avec un convertisseur flash, la référence de ce convertisseur est égale à V_{max}. Les N/2 bits sont alors reconvertis en analogique par un CNA. Le résultat du CNA est soustrait au signal à convertir et le résultat de la soustraction est converti en N/2 bits de poids faibles par un convertisseur flash dont la tension de référence est égale à $V_{max}/2^{N/2}$. On peut faire l'économie d'un des deux convertisseurs flash en commutant la tension de référence de V_{max} à $V_{max}/2^{N/2}$ comme le montre la Fig. 26. On utilise ainsi $2^{N/2}$ résistances au lieu de 2^N et ($2^{N/2}$-1) comparateurs au lieu de (2^N-1).
(ex : AD7821 - *Analog Device* : 8 bits, 660 ns, voir annexe IV)

Avec un principe similaire au ½ flash, on trouve des convertisseurs de type pipeline (ex : ADS800 - *Burr-Brown* : 12 bits, 25 ns, voir annexe V).

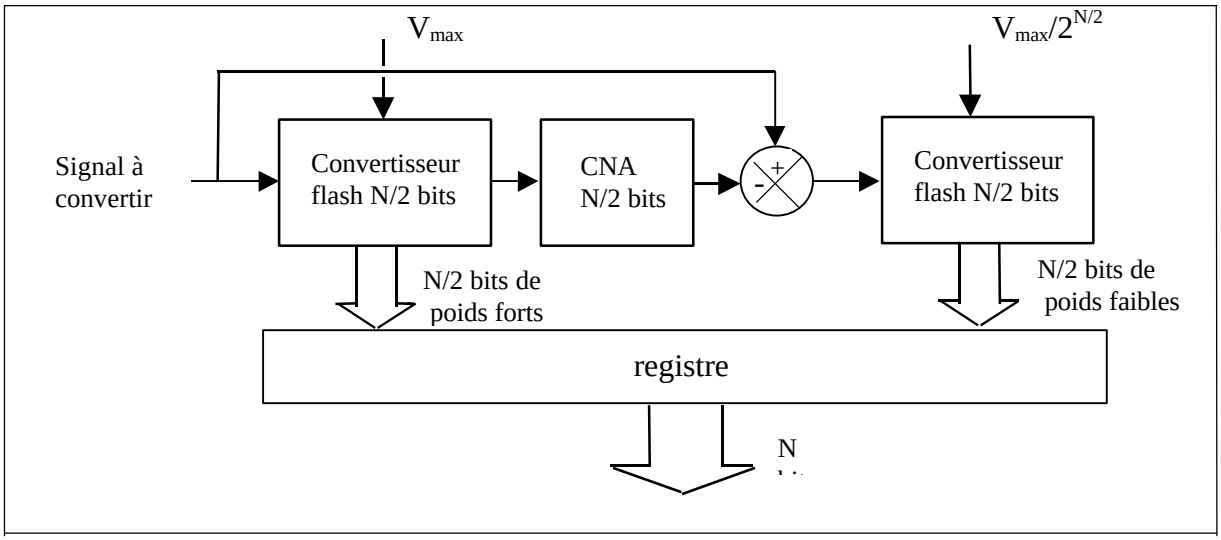


Fig. 25 Schéma de principe d'un convertisseur semi-flash

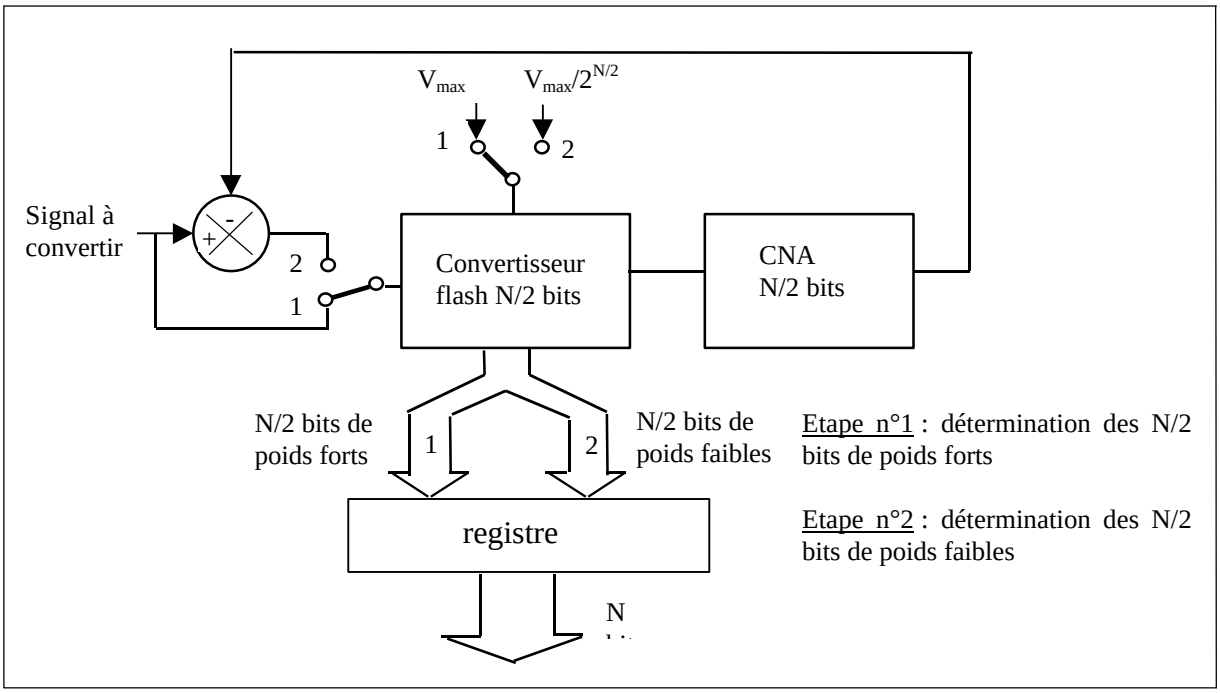


Fig. 26 Réalisation d'un convertisseur semi-flash N bits

d) Delta Sigma (Sigma Delta ADC)

principe : les convertisseurs Delta Sigma sont basés sur le principe de la modulation-démodulation Delta Sigma présentée à la Fig. 27. Le bloc intitulé "quantification un bit" n'est autre qu'un comparateur délivrant un signal A si la sortie du sommateur est positive et -A si la sortie est négative. Les principaux signaux du modulateur sont donnés à la Fig. 28. Le schéma de la Fig. 27 peut se ramener à celui de la Fig. 29.

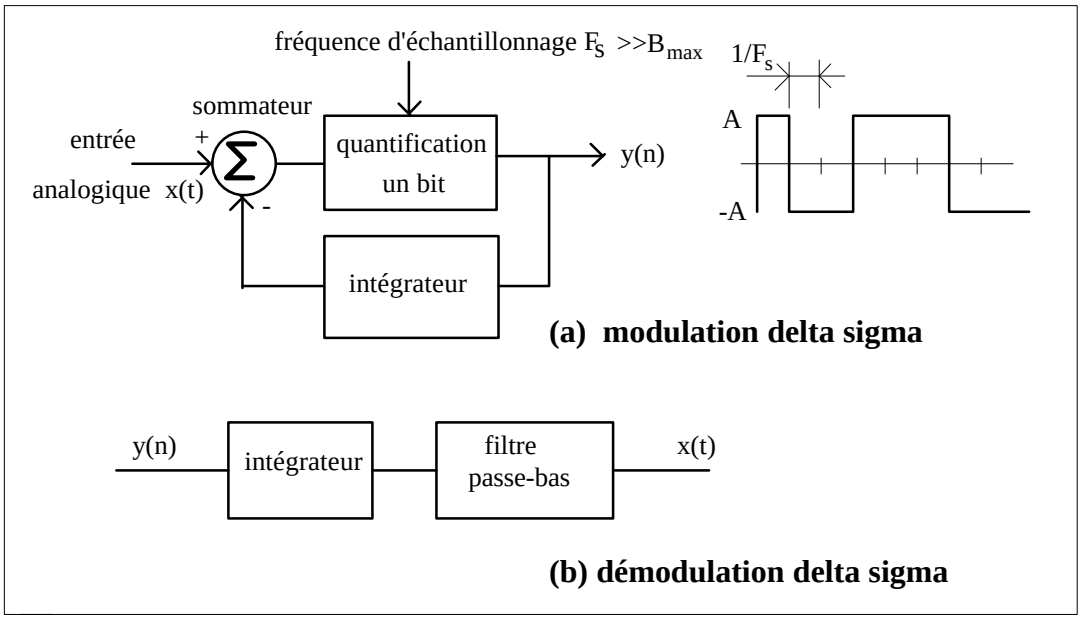


Fig. 27 Principe de la modulation et démodulation delta-sigma

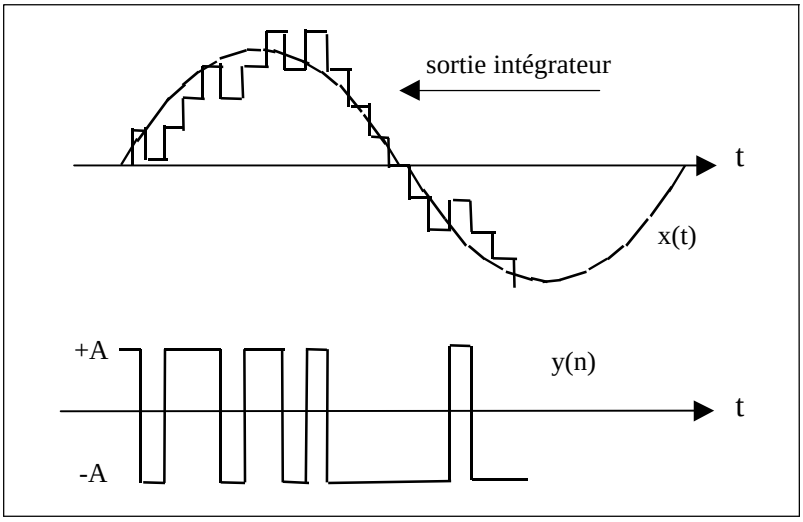


Fig. 28 Principaux signaux du modulateur delta-sigma

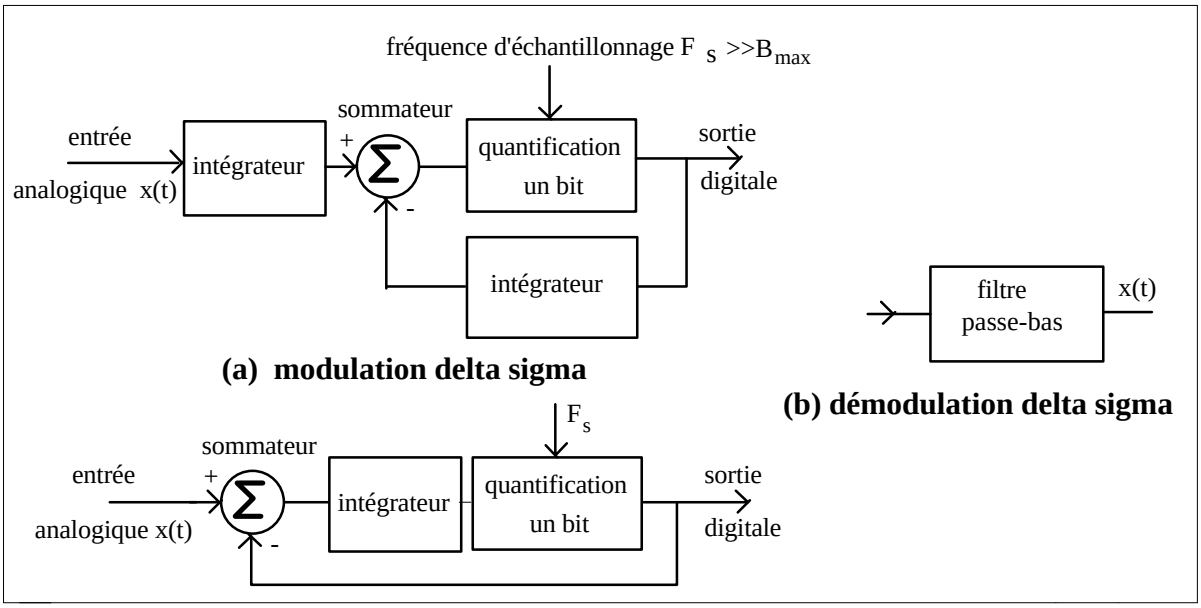


Fig. 29 modulation et démodulation delta sigma modifiée

Dans le cas d'un modulateur delta-sigma, la sortie doit se présenter sous forme de deux niveaux logiques ‘0’ ou ‘1’, il suffit alors de remplacer le quantificateur par un CAN un bit suivi d'un CNA un bit, la sortie étant alors prise sur le CAN comme le montre la Fig. 30. En sortie du CAN le débit est F_s bits/s, ce débit est ramené à (F_s/n) échantillons codés sur N bits après filtrage et décimation. Sur la Fig. 30, on donne un exemple de décimation par 8 (n=8), en sortie on dispose d'échantillons codés sur 3 bits (N=3 et n=8) et la fréquence des échantillons est de F_s/8.

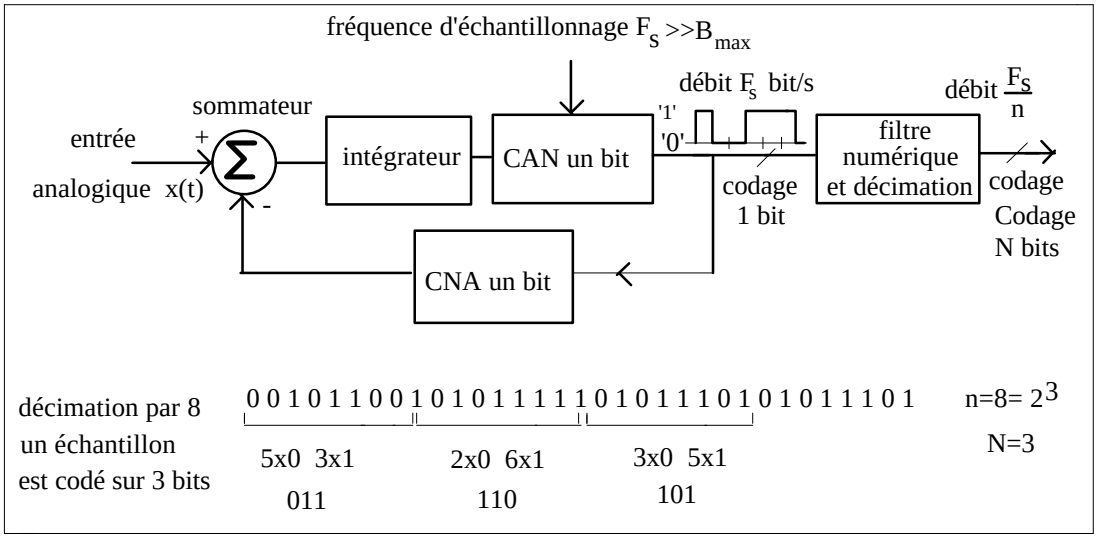


Fig. 30 Convertisseur delta sigma

avantage : Avec la technique de suréchantillonnage (*over sampling*), les contraintes sur le filtre antirepliement sont alors beaucoup moins sévères comme le montre la Fig. 31-a. Un simple filtre R-C peut être utilisé comme filtre antirepliement si la fréquence de suréchantillonnage F_s est très supérieure à la fréquence max. du spectre du signal à convertir. Il n'est donc pas nécessaire de réaliser un filtre avec une pente d'atténuation (*roll-off*) très abrupte, ce type de filtre est par contre indispensable si on utilise un échantillonnage au voisinage de la fréquence de Nyquist ($F_{Nyquist} = 2B_{max}$). Un autre avantage du suréchantillonnage réside dans l'étalement vers les hautes fréquences du bruit de quantification (*noise shaping*) où il est éliminé par le filtre de décimation de sortie, voir Fig. 31-b. Environ 90% du convertisseur est numérique, il s'ensuit une grande stabilité et interchangeabilité. Les convertisseurs delta-sigma et les DSP (Digital Signal Processor) sont les circuits de base pour le traitement des signaux audio-fréquence.

(ex : 1) ADC 16071 de *National Semiconductor*, 16 bits delta-sigma, suréchantillonnage par 64, F_{smax} de 12.288 MHz, soit 192 ks/s.

2) TLC320AD58C de *Texas Instruments*, 18 bits delta-sigma, suréchantillonnage par 64, F_{smax} de 11.29MHz).

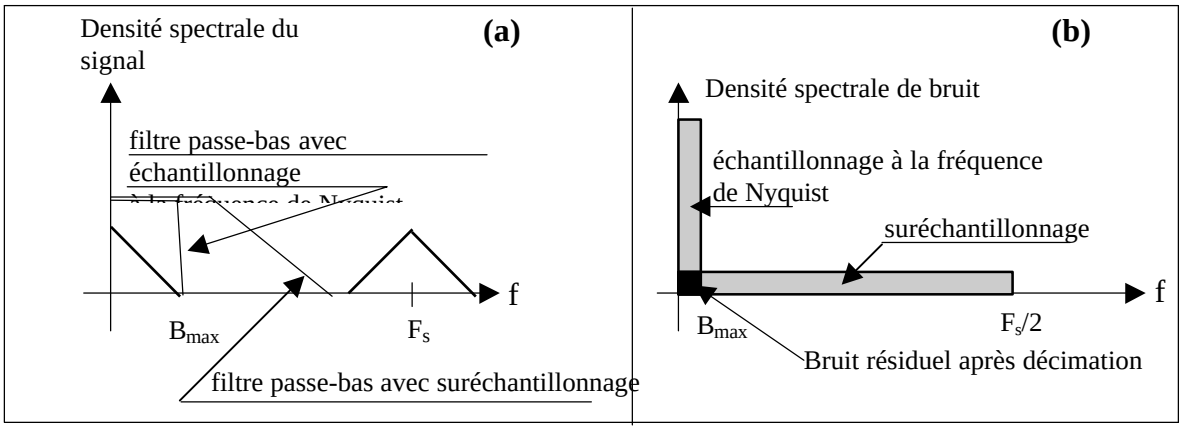


Fig. 31 filtres passe-bas (a) et densités spectrales de bruit (b) avec un échantillonnage à la fréquence de Nyquist ($2B_{max}$) et en présence d'un suréchantillonnage

IV- 2- Les convertisseurs numérique-analogique (CNA)

Le rôle d'un CNA est de convertir une configuration binaire en une grandeur analogique directement proportionnelle à la valeur décimale de la configuration. Les entrées d'un convertisseur N bits sont des niveaux logiques 'a_1', 'a_2', ... 'a_N' prenant les états logiques '0' ou '1'. La sortie est le plus souvent un courant analogique I ; il obéit à l'équation suivante :

$$I = I_0[a_N 2^0 + a_{N-1}2^1 + a_{N-2}2^2 + ... + a_1 2^{N-1}]$$

ou encore : $$I = I_0 2^N \left[a_N \frac{1}{2^N} + a_{N-1} \frac{1}{2^{N-1}} + a_{N-2} \frac{1}{2^{N-2}} + ... + a_2 \frac{1}{2^2} + a_1 \frac{1}{2^1} \right]$$

où I_0 est un courant de référence, les coefficients a_1, a_2, ... , a_N valent 0 ou 1. D'une manière générale le coefficient $a_i = 0$ si le niveau logique 'a_i' = '0', et $a_i = 1$ si 'a_i' = '1'. Les bits 'a_1' et 'a_N' sont respectivement les bits de poids fort et de poids faible.

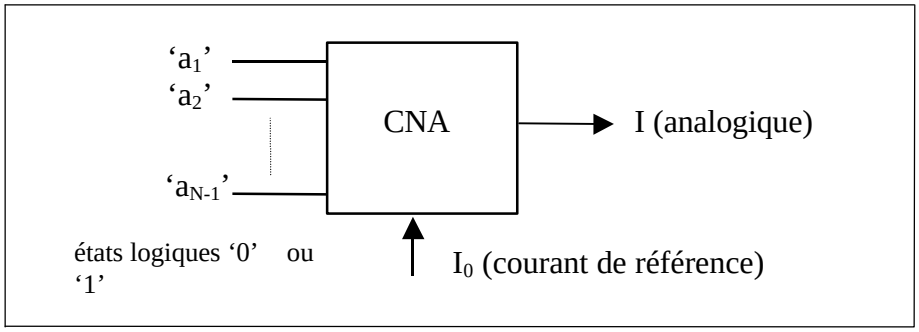


Fig. 32 Synoptique d'un convertisseur numérique-analogique N bits

La réalisation d'un convertisseur de N bits nécessitent donc N générateurs de courant variant dans un rapport 2^{N-1}. Pour générer les différents courants on peut utiliser soit un convertisseur à poids soit un réseau R-2R . Les schémas électriques de principe de ces deux types de convertisseurs sont maintenant décrits.

α) les convertisseurs à poids

Les N générateurs de courant sont réalisés en utilisant N résistances différentes comme le montre le schéma de la Fig. 33.

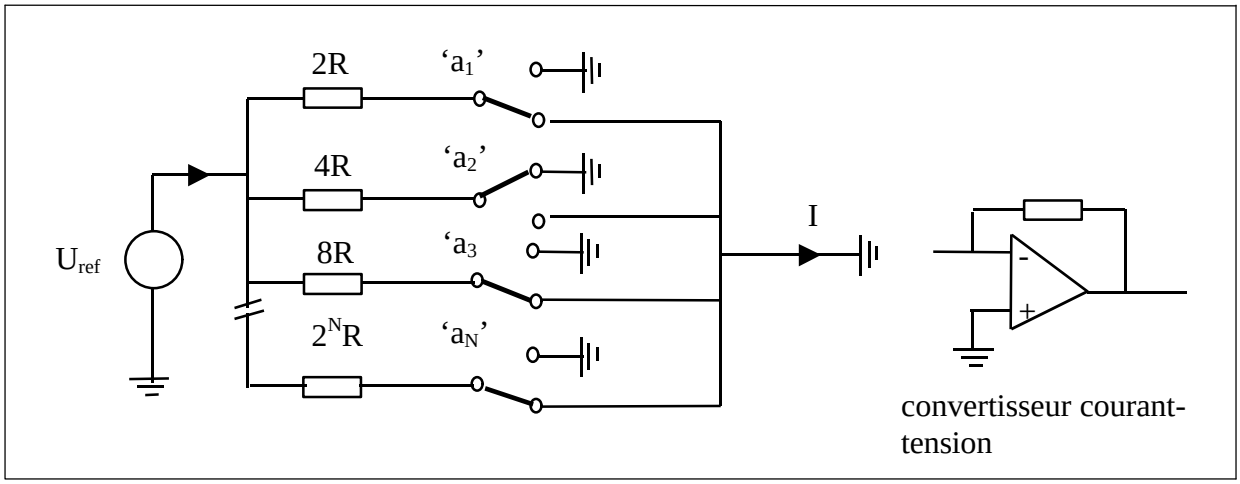


Fig. 33 Schéma électrique de principe d'un convertisseur à poids

On vérifie bien que le courant I se met sous la forme :

$$I = U_{ref}[a_1 \frac{1}{2R} + a_2 \frac{1}{2^2 R} + \quad ... \quad + a_N \frac{1}{2^N R}] = \frac{U_{ref}}{R}[a_1 \frac{1}{2^1} + a_2 \frac{1}{2^2} + \quad ... \quad + a_N \frac{1}{2^N}]$$

On en déduit que $I_0 = U_{ref}/R$. Le courant débité par la source est indépendant des niveaux logiques 'a_1', 'a_2', ... et 'a_N' ; c'est la condition requise pour obtenir un temps de réponse (*settling time*) faible. Les convertisseurs à poids nécessitent la réalisation de N résistances de valeurs différentes.

β) les convertisseurs utilisant un réseau R- 2R

Dans les convertisseurs utilisant un réseau R- 2R, deux valeurs de résistances sont seulement nécessaires. Le schéma de base d'un tel convertisseur est donné à la Fig. 34, le courant $i_1 = U_{ref}/2R$. On peut modifier le schéma de la Fig. 34 pour faire apparaître $i_1/2$ et $i_1/4$ sans modifier le courant total débité par la source, pour cela il suffit de remplacer la dernière résistance de valeur 2R par une résistance R en série avec deux résistances de 2R placées en parallèles comme le montre la Fig. 35.

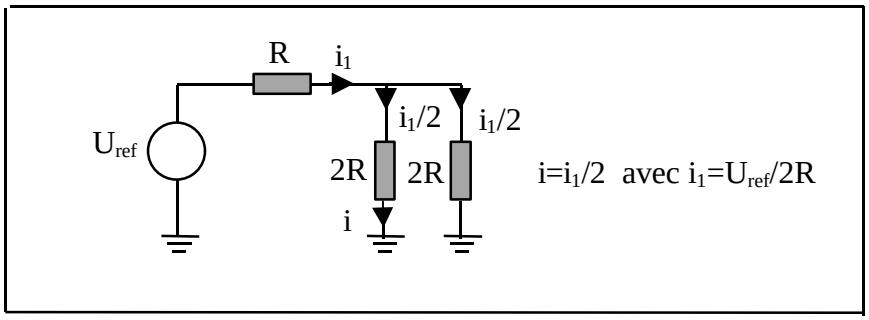


Fig. 34 Schéma électrique de base du réseau R- 2R

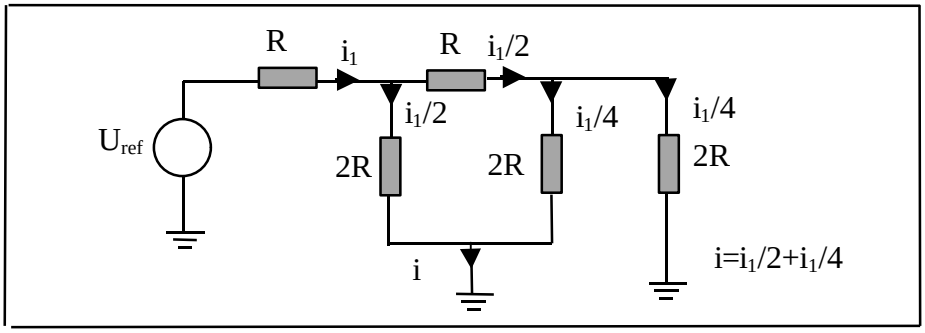


Fig. 35 Réalisation de $i_1/2 + i_1/4$

Le schéma de la Fig. 35 peut être modifier pour faire apparaître les courants $i_1/2$, $i_1/4$ et $i_1/8$, comme le montre la Fig. 36.

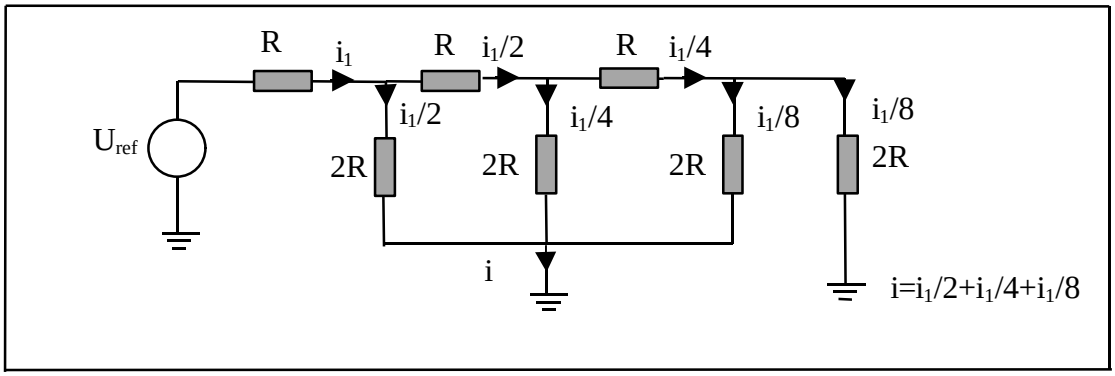


Fig. 36 Réalisation de $i_1/2 + i_1/4 + i_1/8$

Le schéma électrique d'un CNA de type R- 2R de N bits est donné à la Fig. 37, comme dans le CNA à poids il faut réaliser un convertisseur courant-tension pour disposer d'une tension en sortie. Par ailleurs, il faut réaliser des interrupteurs ayant des résistances négligeables devant 2R (voir TD électronique pour la réalisation des interrupteurs).

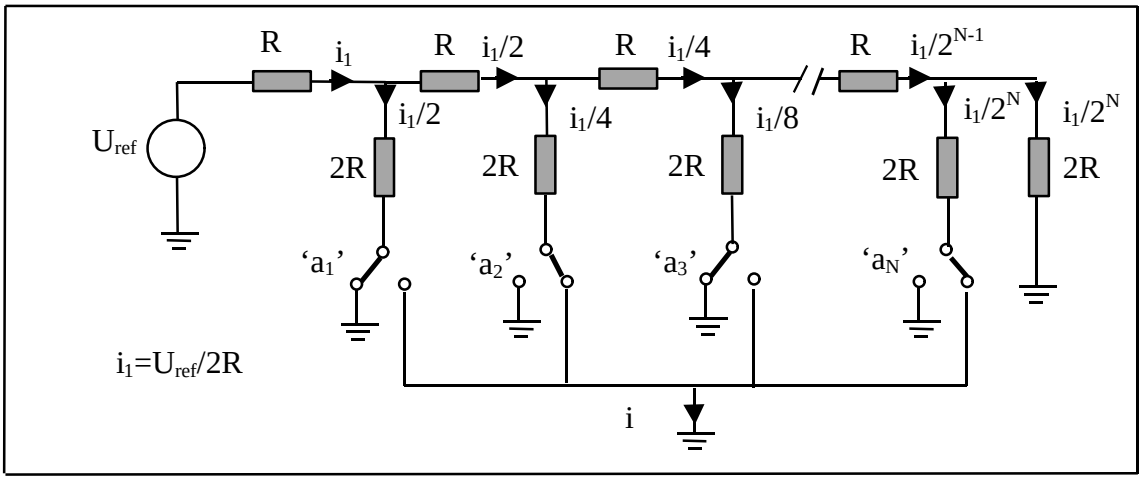


Fig. 37 Réalisation d'un CNA de N bits

(ex : AD568 - *Analog Devices* : 12 bits, 35 ns, voir annexe VIII)

La précision des convertisseurs réside essentiellement dans la précision des résistances. Pour caractériser un CNA on utilise généralement la notion de monotonicité (*monotonocity*) : on dit qu'un convertisseur a une fonction de transfert monotone si la tension de sortie est une fonction croissante du code d'entrée comme le montre la Fig. 38.

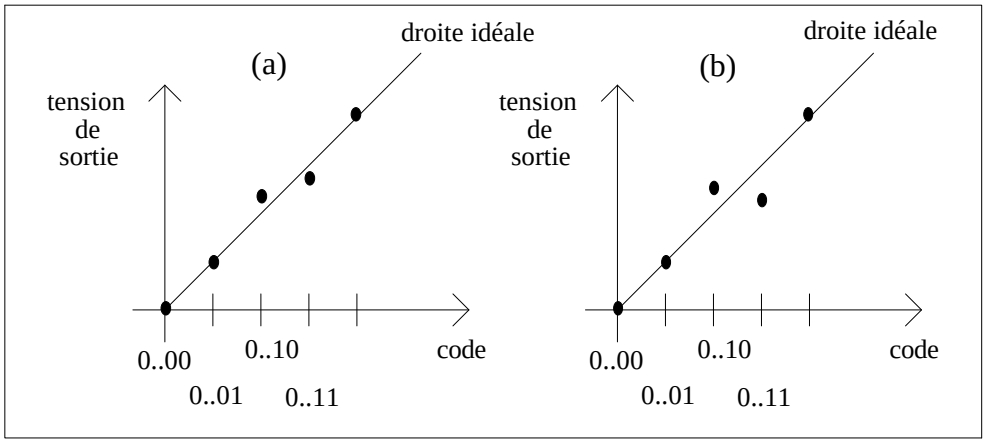


Fig. 38 Caractéristiques de deux CNA, (a) monotone, (b) non monotone

Si la précision d'un CNA est de ½LSB ou meilleure, la caractéristique du CNA est monotone. En pratique, il s'avère difficile de maintenir une précision de ½LSB pour des convertisseurs à nombre de bits élevés. En effet, pour un CNA de 16 bits par exemple, une précision de ½LSB nécessite pour le courant le plus élevé, une précision de 1/65536.

c) les convertisseurs *bit stream*

En audio-fréquence, où le nombre de bits est élevé, on utilise généralement des convertisseurs dit *bit stream*, le concept utilisé est voisin de celui des CAN delta sigma. On décrit ci-dessous le principe du convertisseur *bit stream* développé par la société *Philips*. La conversion nécessite trois étapes comme le montre le schéma de principe de la Fig. 39.

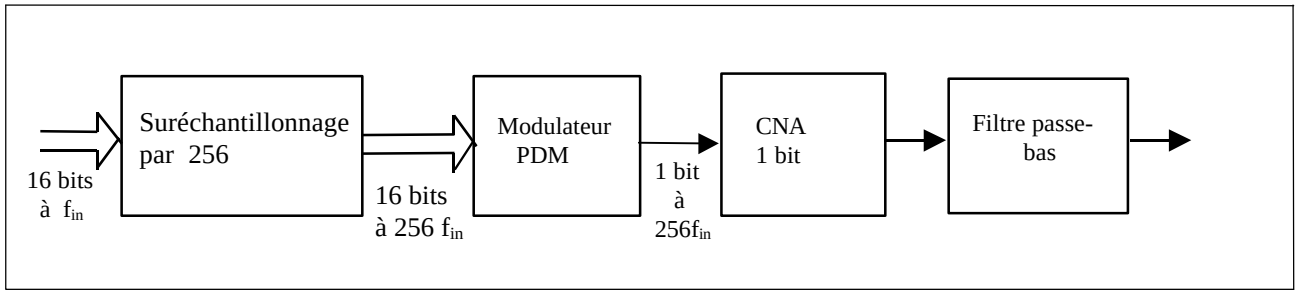


Fig. 39 Principe du CNA *bit stream*

α) la première étape est un suréchantillonnage, ou interpolation, on remplace le train de 16 bits échantillonné à la fréquence F_{in} (44.1 kHz) par 16 bits à la fréquence nF_{in} (n=256), les échantillons sont calculés par interpolation.

β) la seconde étape transforme le train de 16 bits à la fréquence nF_{in} en un train de 1 bit à la fréquence 11.2 MHz (signal PDM pour : *Pulse Density Modulated*). Cette opération est réalisée au moyen d'un modulateur dont le principe est donné à la Fig 40.

γ) la troisième étape consiste à transformer le signal PDM constitué par la suite d'états logiques '0' ou '1' à la fréquence nF_{in} en un signal analogique (+A, -A) par un convertisseur 1 bit, un filtre analogique passe-bas permet ensuite de récupérer le signal analogique. La linéarité du convertisseur est liée à la linéarité du CNA 1 bit, c'est à dire à la précision des niveaux +A et -A.

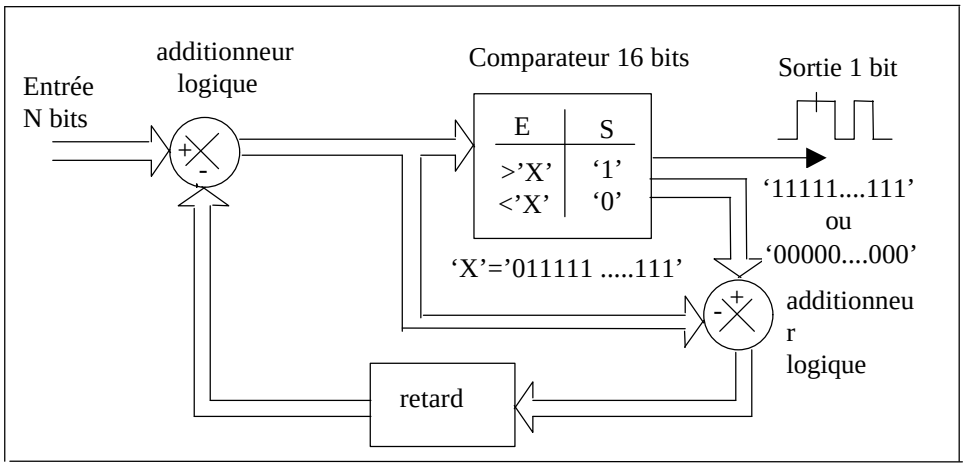


Fig. 40 Transformation du train de 16 bits en un signal PDM