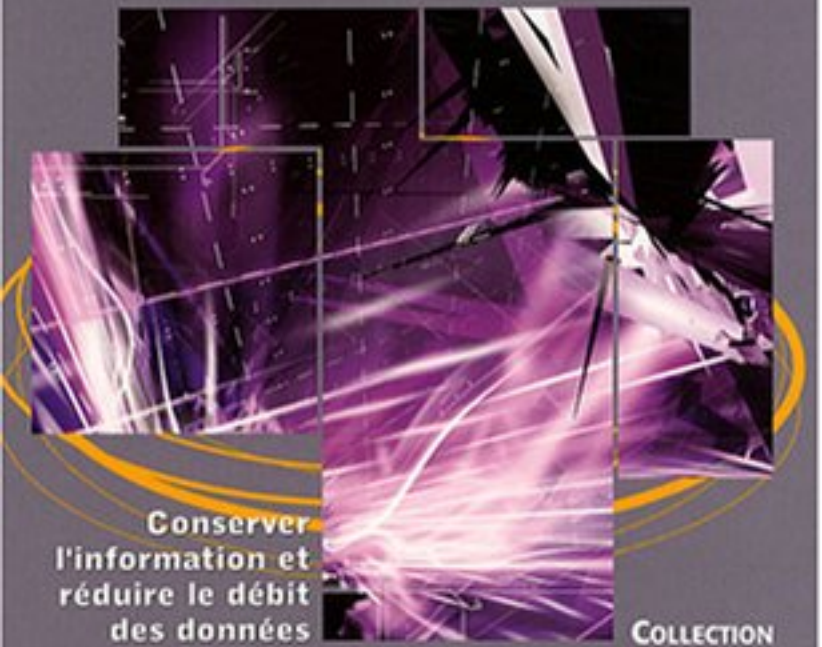
François Luxereau

CST

Compression du signal audiovisuel



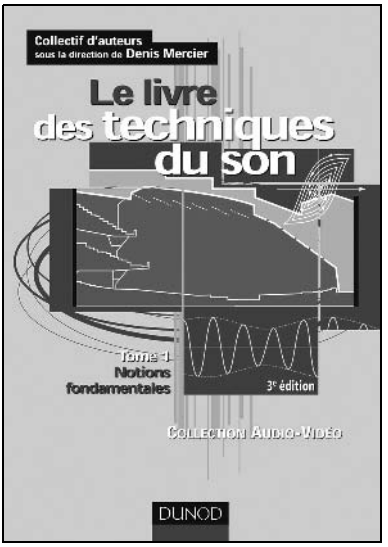
Conserver l'information et réduire le débit des données

Collection Audio-Photo-Vidéo

DUNOD

Compression du signal audiovisuel

CHEZ LE MÊME ÉDITEUR



Le livre des techniques du son
(3e édition)
Sous la direction de Denis Mercier

Traitement du signal audiovisuel
Laurent Millot

Dictionnaire des techniques audiovisuelles et multimédias
Fabien Marguillard

Colorimétrie appliquée à la vidéo
Jacques Gaudin

François Luxereau

Compression du signal audiovisuel

Conserver l'information et réduire le débit des données

DUNOD

Couverture : Rachid Maraï
Illustrations intérieures : Alain et Ursula BOUTEVEILLE

ISBN 978-2-10-053683-2

TABLE DES MATIÈRES

AVANT-PROPOS

La rédaction d'un ouvrage sur la réduction de débit dans l'univers audiovisuel m'a été demandée par le Conseil d'Administration de la CST il y a un peu plus d'une année.

Nous avons alors constitué au sein de la CST un groupe de réflexion afin de définir le niveau technique ainsi que les contours du contenu de ce livre en fonction du public auquel il était principalement destiné : les professionnels de l'audiovisuel ainsi que les étudiants.

Rester simple sans être réducteur ni encyclopédique, fournir aux uns et aux autres les clefs leur permettant ensuite d'approfondir leurs connaissances, tel a été le mot d'ordre.

J'ai accepté ce travail, qui m'a aidé à traverser une sombre période de ma vie, à condition que les différents chapitres soient vérifiés, complétés et modifiés par des spécialistes des différents sujets. Je ne suis quant à moi qu'un généraliste passionné.

Il en a été ainsi grâce aux compétences d'Olivier Amato, Hervé Bernard, Alain Delhaise, Pierre Lavoix, Yves Louchez, Denis Mercier, Yvon Penarguear, Matthieu Sintas, Bernard Tichit, Jean-Baptiste Touchard.

Je veux les remercier pour leur concours compétent, efficace, dévoué et surtout amical.

Merci.

F. L.

CHAPITRE PAGE

1 DU SON À L'IMAGE

1.1 LE MONDE, LES DÉCIBELS ET NOUS

Nous sommes, comme tout organisme vivant, en interaction avec le monde qui nous entoure. Celui-ci nous communique des stimulations par l'intermédiaire de signaux qui sont les variations des paramètres physico-chimiques de notre environnement. Ces signaux physico-chimiques excitent certaines cellules du système nerveux qui sont des récepteurs sensoriels. Il existe différents types de récepteurs sensibles à des phénomènes physico-chimiques spécifiques : la chaleur, la lumière, des variations de la pression de l'air… Ces récepteurs, interfaces entre le monde et nous, transforment les stimulations en sensations qui, interprétées après passage au crible de notre expérience, conduisent à une perception de notre environnement.

1.1.1 De la stimulation à la sensation

La stimulation d'un récepteur se caractérise par sa durée, son intensité et sa localisation.

Il existe des seuils en dessous desquels la stimulation n'entraîne pas de sensation.

On doit distinguer le seuil physiologique, en dessous duquel les récepteurs ne réagiront que dans moins de 50 % des cas, du seuil psychologique, en dessous duquel cette réaction physiologique ne donnera naissance à une perception que dans moins de 50 % des cas.

Le seuil de discrimination spatiale détermine le « pouvoir séparateur » du système perceptif. Il représente la distance minimale qui doit séparer deux sources ponctuelles de stimuli pour qu'elles soient perçues comme séparées.

Le seuil de discrimination temporelle représente l'intervalle de temps minimum qui doit séparer deux stimuli afin qu'ils soient discernables. La fréquence de fusionnement est l'inverse de cet intervalle minimum.

Au-delà de cette fréquence le cerveau interprète cette succession de stimuli comme un phénomène continu. C'est ce qui permet au cinématographe ainsi qu'à la télévision d'exister.

Pour ce qui concerne l'intensité des stimuli, il existe des seuils absolus au-dessous desquels on ne perçoit aucune sensation. Il existe également des seuils différentiels qui caractérisent la variation minimale du stimulus qui puisse être perçue. Ces seuils différentiels sont notés JND dans la littérature anglo-saxonne (pour *Just Noticeable Difference*).

1.1.2 La loi de Weber-Fechner

La sensation croît moins vite que la stimulation. C'est sans doute un moyen de nous protéger contre les agressions de notre environnement.

En s'appuyant sur les travaux de l'anatomo-physiologiste Ernst Weber (1795–1878) Gustav Fechner (1801–1887) énonça sa célèbre loi (dont le domaine de validité reste sujet à débats) selon laquelle « la sensation varie comme le logarithme de l'excitation ».

$$S = k \,.\, \log(E)$$

Cette relation reposait sur les « fractions de Weber ». Il s'agit de la notion de seuil différentiel ΔE, qui représente la plus petite différence d'intensité d'excitation qui puisse être perçue.

Cette différence dépend de l'intensité de l'excitation selon la relation :

$$\Delta E / E = c$$

(ou c est une constante dépendant du système perceptif concerné.)

Il se trouve que cette expression $\Delta E/E$ représente un élément de l'aire comprise entre la courbe représentant la fonction $1/E$ (une branche d'hyperbole) et l'axe hori-

zontal… il se trouve également que c'est cette aire qui permet de définir géométriquement la fonction log(E).

1.1.3 Logarithmes et décibels

Blaise Pascal avait bien raison, l'homme est cerné par deux infinis : le grand et le petit. L'un comme l'autre représentent tellement de zéros (en numérateur ou en dénominateur, avant ou après la virgule) que, lorsqu'on veut les dénombrer, cela en devient impossible.

Prenons par exemple un tout petit infini, celui du nombre très approximatif des habitants de l'Inde ou de la Chine : un milliard ; 1 000 000 000 ou encore 1 suivi de 9 zéros. Ceci traduit le fait qu'on peut trouver ce nombre en portant 10 à la puissance 9 :

$$1\,000\,000\,000 = 10^9$$

Inversement :

$$1/1\,000 = 0{,}0001 \text{ s'écrit } 10^{-3}$$

Les mathématiciens, pour se simplifier la vie ainsi que celle des astronomes, ont eu l'idée d'utiliser couramment cette écriture en puissances de 10. Plus largement ils ont décidé, au début du XVII^{e} siècle, de remplacer les nombres qui avaient tendance au gigantisme, les nombres astronomiques, par un nouvel être mathématique qu'ils ont appelé du nom bizarre de logarithme (du grec logos : rapport et arithmeticos : nombre).

Le logarithme d'un nombre x est la puissance a à laquelle il faut élever une constante b, appelée base, pour obtenir ce nombre :

$$x = b^a$$

a est le logarithme en base b de x :

$$a = \log_b(x)$$

Ainsi, pour la base 2, on trouve $2^1 = 2$; $2^2 = 4$; $2^3 = 8$; $2^4 = 16$; etc.

Et on pourra écrire $\log_2 4 = 2$; $\log_2 16 = 4$; etc.

Une base très utilisée est la base 10. À tout nombre x on fait correspondre son logarithme en base 10 ou logarithme décimal a par la relation :

$$x = 10^a$$

et l'on écrit simplement :

$$a = \log x$$

On trouve immédiatement pour les multiples de 10 que $\log 10 = 1$; $\log 100 = 2$; $\log 10\,000 = 4$; $\log 1/100 = -2$.

Pour ce qui concerne les nombres fractionnaires il faut se rappeler que $1/100$ s'écrit également 10^{-2} :

$1/100 = 10^{-2}$; donc $\log 1/100 = -2 = -\log 100$

Des tables permettent de trouver que, par exemple, $\log 2 = 0{,}30103$; $\log 3 = 0{,}47712$.

On peut aussi dire qu'on fait correspondre à la suite des nombres entiers la suite des nombres d'une progression géométrique de raison 2 ou de raison 10 ou, plus généralement, de raison n.

Ainsi pour une raison $n = 2$ on retrouve les valeurs 2, 4, 8, 16…

Les logarithmes présentent une particularité intéressante : le logarithme d'un produit est égal à la somme des logarithmes des facteurs de ce produit. On peut en effet écrire par exemple :

$\log 10\,000 = 4 = \log 10 + \log 1\,000 = 1 + 3$

$= \log 100 + \log 100 = 2 + 2$;

$\log 100 = \log 1\,000/10 = 3 + (-1) = 2$;

ou encore $\log 200 = \log 2 + \log 100$

$= 0{,}30103 + 2 = 2{,}30103.$

On en déduit immédiatement que :

$$\log x^n = n \,.\, \log x$$

Les puissances acoustiques perceptibles par notre oreille peuvent prendre des valeurs variant dans un rapport supérieur à 1 012 (1 000 000 000 000) depuis le microwatt (le tic-tac d'une montre à plus d'un mètre), jusqu'au mégawatt (le décollage d'une fusée à Kourou).

Les acousticiens ont donc tout naturellement eu recours aux logarithmes pour exprimer ces rapports de puissance, et cela d'autant plus naturellement que notre système de perception acoustique fonctionne de cette manière (loi de Weber-Fechner).

On a défini les rapports de puissance en bel (en hommage à Graham Bell) par l'expression :

$$1\ B = \log P_2 \,/\, P_1$$

Cette unité, trop grande, s'est révélée mal adaptée aux besoins des acousticiens. Ils ont créé le décibel (dB) qui exprime ces rapports par l'expression

$$1\ dB = 10 \log P_2 \,/\, P_1$$

Une dynamique sonore de 60 dB pour un enregistreur signifie qu'il existe un rapport d'un million (10^6) entre la puissance du plus fort et du plus faible des signaux enregistrés.

Une augmentation de 3 dB signifie que le rapport de puissances a doublé puisque $10 \log 2 = 3{,}103$; une baisse de 3 dB signifie qu'il a été divisé par deux.

La notation en décibels a été étendue aux rapports de tensions électriques qui expriment par exemple le gain d'un amplificateur ou encore le fameux rapport « Signal sur Bruit » qui caractérise la pureté d'un signal électrique. Ces rapports de tension (V_2 tension de sortie et V_1 tension d'entrée) s'expriment en décibels par l'expression $20 \log V_2/V_1$.

En effet la puissance électrique est proportionnelle au carré de la tension.

Un gain de 100 en tension correspond donc à 40 décibels. Si ce gain est doublé on atteint une valeur de 46 décibels (20 log 2 = 6).

La notation en décibels, en raison de la loi de Weber-Fechner, est parfaitement appropriée à toutes les études de psychophysiologie. On ne s'étonnera donc pas de la rencontrer bien souvent.

1.2 UN PEU DE THÉORIE DE LA COMMUNICATION

1.2.1 Information et données : le message

La communication est à l'ordre du jour. Son sens profond, étymologique, est de mettre en commun. Communiquer c'est, au sens large, établir des relations avec quelqu'un, partager avec lui des idées, des sentiments.

L'apparition des réseaux de télécommunication ; du téléphone au satellite, a ouvert l'ère des sciences de la communication pour lesquelles communiquer, c'est échanger un message entre un émetteur et un récepteur.

Ce message véhicule un certain nombre d'éléments qui peuvent lui conférer une signification pour le récepteur. Ces éléments constituent l'information du message.

Pour que cette information puisse être utilisable il faut qu'émetteur et récepteur disposent d'un répertoire commun : une langue (un texte en caractères cyrilliques ne me concerne guère), une culture de l'image (la compréhension de la perspective), etc.

La notion d'information reste en général théorique, une quantité indépendante de la signification du message. L'information doit être structurée, organisée pour produire du sens ; elle devient alors des données qui

permettent, par exemple, des actions judicieuses (ou non) ou l'augmentation d'un savoir.

Afin de pouvoir transiter par les canaux de transmission le message a besoin d'être mis sous une forme physique adaptée (en général électrique). Il a souvent besoin d'être codé (tam tam, signaux de fumée, position des bras du télégraphe optique Chappe, code Morse, etc.).

Plus le canal est rudimentaire plus le codage doit être élaboré et le répertoire commun précis.

Le mathématicien Claude Shannon, qui travaillait sur les systèmes de transmission en langage binaire pour les laboratoires Bell, publia en 1948, son ouvrage *A Mathematical Theory of Communication* qui formalisait un ensemble de connaissances plus ou moins implicites à cette époque.

1.2.2 L'information binaire

Ce sont les méthodes de codage binaire qui ont rendu possible cette avancée.

Fonctionner en binaire signifie que le codage devra se contenter de deux briques seulement qu'il conviendra d'assembler pour constituer le message. On note conventionnellement 0 et 1 ces deux possibilités car elles sont presque exclusivement mises en œuvre dans des systèmes numériques. On parle alors de bit pour *binary digit* (chiffre binaire).

On aurait pu également les appeler *oui* et *non* comme dans l'arborescence du jeu des questions : est-ce un être vivant ? oui ; est-ce un végétal ? non ; est-ce que c'est un vertébré ? oui ; etc.

À chaque réponse la généralité se resserre. Les possibilités de choix diminuent, l'imprévu recule.

Si l'on suppose que le *oui* et le *non* sont aussi vraisemblables l'un que l'autre, lorsque le récepteur reçoit la

réponse *oui*, l'une des propositions est validée, le nombre des possibles est divisé par deux. On dit qu'il y a eu réception d'une unité d'information. Dans le quotidien numérique d'aujourd'hui on dira qu'il s'agit d'un bit.

1.2.3 L'entropie

Ce terme d'entropie est né, à la grande époque des machines à vapeur, lors de l'étude du phénomène de l'échange thermique d'énergie entre une source chaude et une source froide. Il s'agit du deuxième principe de la thermodynamique, énoncé par Clausius en 1850 : dans un système énergétiquement isolé, toutes les différences de température tendent spontanément à s'annuler.

Il est bien connu qu'il existe une très forte probabilité pour que les températures des deux sources mises en relation deviennent égales.

Cela signifie que les molécules de ces deux sources qui avaient été initialement rangées, ordonnées suivant leur énergie, leur vitesse, se retrouvent mélangées. L'ordre initial est devenu désordre. L'incertitude quant à cet état final est quasi nulle.

L'entropie, notée S, représente la quantité de chaleur échangée entre les sources, divisée par la température du système :

$$S = \int dQ/T$$

Lorsque l'équilibre a été atteint le système est devenu inerte et son entropie est maximale. L'entropie est donc un marqueur de la dégradation progressive et inéluctable de l'énergie d'un système isolé.

Les réfrigérateurs ainsi que les climatiseurs parviennent à trier les molécules « chaudes » et les molécules « froides », mais c'est au prix d'un apport d'énergie ; c'est EDF qui fournit les moyens ! Le système n'est pas isolé.

Le physicien écossais James Clerk Maxwell inventa en 1871 la fable d'un être « dont les facultés seraient assez

développées pour qu'il puisse suivre chaque molécule dans sa course ; un tel être dont les capacités seraient cependant finies comme les nôtres serait capable de faire ce qui nous est impossible… Les molécules d'un gaz à température uniforme se déplacent à des vitesses qui ne sont pas uniformes, bien que la vitesse moyenne d'un grand nombre quelconque de ces molécules soit exactement uniforme. Imaginons un récipient contenant un tel gaz divisé en deux compartiments A et B par une paroi percée d'un petit trou ; imaginons que notre être surdoué puisse surveiller chaque molécule individuellement et qu'il ouvre ou ferme le trou afin de permettre aux molécules les plus rapides de passer de A vers B et aux plus lentes de passer de B vers A. Les particules rapides s'accumuleront alors dans le compartiment B dont la température s'élèvera ; les lentes dans le compartiment A qui se refroidira ».

La connaissance permet au démon de Maxwell, petit être facétieux, Sysiphe rééduqué par Denis Papin, James Watt et Nicolas Carnot, de faire baisser l'entropie du système en bousculant les étagères de plus en plus paisibles de l'univers, en restaurant l'organisation initiale du système diluée dans un chaos final apathique… il s'agit d'une fable !

En 1877, le physicien autrichien Ludwig Boltzmann tenta de comprendre ce « chaos moléculaire ». Son idée était que l'évolution thermodynamique d'un système vers l'équilibre correspond au passage d'un état initial à un autre état statistiquement plus probable.

Or un état donné, observable à l'échelle macroscopique, peut être réalisé par un grand nombre de configurations microscopiques résultant de l'agitation moléculaire, indiscernables à notre échelle et supposées équiprobables. Il en déduisit une nouvelle définition statistique de l'entropie donnée par la formule, gravée sur sa tombe à Vienne :

$$S = k \cdot \log(\Omega)$$

dans laquelle Ω représente le nombre d'états possibles pour le système.

L'entropie rend ainsi compte de la probabilité d'occurrence d'un état donné pour un système thermique isolé.

Nous avons vu que l'information, de son côté, caractérise la réduction de la probabilité d'occurrence d'un évènement ou d'un état.

Dans l'exemple cité plus haut, l'utilisation d'un seul bit permet de choisir entre deux possibilités, celui de 2 bits ouvre le choix entre 4 possibilités, celui de 8 bits entre 256, etc. On pense bien sûr aux logarithmes de base 2.

On constate que le nombre de bits (unités d'information) nécessaires pour spécifier un état correspond au logarithme en base 2 du nombre total N de possibilités.

On peut donc écrire :

$$I = \log_2 N$$

ou encore :

$$I = \log_2 1/p = -\log_2 p$$

ou p représente la probabilité d'occurrence d'un des états (formule de Hartley 1928).

La formule est identique (à une constante près, la constante de Boltzmann) à celle issue de la thermodynamique. On peut donc parler de l'entropie d'un signal pour définir l'information qu'il transmet.

La situation est un peu plus compliquée si toutes les possibilités de choix ne sont pas égales mais que l'on sait que certaines pourront apparaître plus fréquemment que d'autres.

C'est le cas de l'alphabet où (en français) la fréquence d'occurrence de la lettre E est 15 fois plus élevée que celle du F et celle de la lettre T, 8 fois plus élevée. L'apparition d'un E dans un message apportera beaucoup moins de renseignements (d'information) que

celle d'un W, par exemple, dont la fréquence d'occurrence est environ 50 fois plus faible.

L'entropie totale du message est alors la moyenne de celle de chacun des éléments du message. Si la probabilité d'occurrence de chaque évènement est p_n, la quantité d'information transmise est définie par (formule de Shannon 1948) :

$$I = \sum p_n \cdot \log_2 1/p_n = -\sum p_n \cdot \log_2 p_n$$

L'entropie est maximale quand tous les événements sont équiprobables, c'est-à-dire :

$$p_n = 1/N$$

1.2.4 Le canal de transmission

Communiquer nécessite un canal de transmission entre émetteur et récepteur. Les canaux naturels, ondes sonores ou lumineuses, adaptés à nos sens, sont de faible portée. Les canaux artificiels ont été conçus pour des distances plus grandes. Ils peuvent être de type point à point (le téléphone) ou multi-récepteur (télévision). Ils utilisent des ondes électromagnétiques se propageant dans l'espace, des signaux électriques ou optiques se propageant dans des câbles (fils de cuivre ou fibre optique).

Le message doit être mis en forme afin de s'adapter aux propriétés du canal.

Cette opération peut être réalisée en mode continu. On parlera alors de systèmes analogiques qui mettent en œuvre la modification des paramètres d'une onde porteuse ; c'est la modulation.

Elle peut avoir lieu en mode discontinu. On parlera alors de systèmes échantillonnés ou discrets. Le message sera représenté par des symboles en nombre variable suivant le type de canal : le 0 et le 1 des chiffres binaires, les 32 symboles alphabétiques français (y compris les lettres accentuées, le point et la virgule),

les 92 positions des bras sémaphoriques (régulateur et indicateur) du télégraphe Chappe…

Un code permet de faire correspondre à un groupe de symboles une information (un mot, un ordre ou une commande, un morceau d'image, etc.).

On notera que l'on peut exprimer les symboles d'un code complexe grâce à ceux d'un code plus simple. Par exemple, les 32 symboles de l'alphabet peuvent être exprimés à l'aide de 5 bits ($2^5 = 32$). Le code ASCII (*American Standard Code for Information Interchange*), utilisé notamment pour les claviers d'ordinateurs, permet de coder 128 signes typographiques à l'aide de 7 bits.

Les symboles sont traduits physiquement par des signaux électriques. Le mode binaire est de ce point de vue le plus simple et le plus sûr : il y a une tension électrique ou il n'y en a pas. C'est pourquoi il est universellement utilisé.

La performance essentielle d'un canal est son débit d'information. Le télégraphe Chappe transmettait une page entre Paris et Strasbourg en une demi-heure environ, de jour et par beau temps !

Ce débit est défini par une capacité en bit par seconde (bps ou b/s) pour les canaux numériques et en largeur de bande en Hertz (Hz) pour les systèmes analogiques. Un canal du Réseau téléphonique Numérique à Intégration de Services (RNIS) dispose d'une capacité de 64 kb/s ; les fournisseurs d'accès ADSL annoncent des débits « jusqu'à 20 Mb/s ». La bande passante d'une ligne téléphonique est limitée à 3 kHz, celle d'un canal TV à 6 MHz.

La distance maximale de transmission (10 km entre deux relais pour le télégraphe Chappe) est évidemment une donnée importante.

Il faut également tenir compte de la pertinence économique à plus ou moins long terme (obsolescence).

Il convient notamment de bien gérer la ressource et d'éviter le gaspillage.

Ceci est possible en dimensionnant le canal afin de prendre en compte des situations moyennes grâce au multiplexage statistique. Il s'agit de faire passer dans le même canal plusieurs programmes à débit variable dont on peut penser qu'ils ne manifesteront pas simultanément des exigences extrêmes.

Le transport des données présente des dangers de corruption de l'information.

Certains se souviendront peut être de l'Histoire du Petit Rat Justin dans laquelle la phrase « c'est la girouette du clocher qui grince au vent » était devenue après transmission de proche en proche « c'est le squelette du cocher qui se rince les dents ». Ce risque est particulièrement redoutable en langage binaire.

Claude Shannon a montré que l'on peut obtenir un risque d'erreur aussi faible que souhaité en ajoutant des bits de sécurité au message initial.

On pourrait par exemple transmettre deux fois chaque bit du message original. Le message 0 1 1 0 deviendrait ainsi pendant la transmission 000 111 111 000 ; un message fortement corrompu comme 010 111 110 001 serait facilement interprété à la réception comme 0 1 1 0. La longueur du message est hélas triplée.

Il existe heureusement des codes correcteurs d'erreur beaucoup moins simplistes et qui sont extrêmement efficaces au prix d'un alourdissement du message tolérable.

1.2.5 Redondance, informations pertinentes ou superflues

Ces bits de sécurité indispensables, dont on cherche à limiter le nombre, sont généralement appelés bits de

redondance. La redondance est le fait de répéter une information.

Le phénomène de redondance peut également être observé dans un message.

On se souvient d'Yves Montand dictant dans un télégramme pour Mademoiselle Colette Mercier de Besançon (Doubs) : « Je pense à toi, je t'aime, je t'aime, je t'aime… » et l'opératrice, Simone Signoret, lui répondant « alors trois fois je t'aime ».

La redondance dans les messages humains est souvent nécessaire et signifiante… « trois fois je t'aime » ce n'est évidemment pas la même chose que « je t'aime, je t'aime, je t'aime ».

Tout le problème des systèmes de transmission est de repérer les informations redondantes et de les transmettre dans un « empaquetage » efficace.

Deux images d'une séquence vidéo peuvent être identiques ; faut-il les transmettre toutes les deux ou peut-on simplement, dans le but d'alléger la transmission, indiquer par un code approprié leur identité ?

On se pose également la question de savoir s'il convient de transmettre des informations que le récepteur ne pourra pas exploiter, des information qui sont superflues pour lui (*irrelevant* en anglais) : des détails imperceptibles sur son écran de télévision, des sons trop aigus pour son oreille…

Comment répondre à ces question, c'est tout l'objet des méthodes de réduction de débit.

1.3 UN PEU DE PSYCHOPHYSIOLOGIE DE L'AUDITION ET DE LA VISION

Nos modes de perception du réel sont riches et complexes, astucieux, mais ils ne sont pas encore totalement élucidés.

Ce chapitre sera volontairement rudimentaire. Les lecteurs pourront trouver de nombreux ouvrages de référence sur ce sujet.

Il est indispensable cependant de faire le point sur certaines performances et certaines limites de nos systèmes perceptifs sur lesquelles se sont construits les systèmes de télévision.

1.3.1 L'oreille

L'oreille est l'interface entre l'univers physique des variations de pression de l'air alentour et notre cerveau.

Le pavillon collecte les variations de pression acoustique, les transmet au tympan ; celui-ci met en mouvement trois petits os (marteau, enclume, étrier) qui transmettent ce mouvement à un liquide remplissant un ensemble de conduits.

L'un de ceux-ci dénommé limaçon, en raison de sa forme, contient les cellules ciliées qui produisent des signaux électriques constituant l'influx nerveux auditif informant le cerveau.

Trois autres conduits, les canaux semi-circulaires, servent à donner au cerveau des indications lui permettant de repérer la position de la tête dans l'espace ce qui nous permet, par exemple, de nous tenir debout.

Les cellules ciliées sont sensibles à la fréquence de la vibration transmise ainsi qu'à son amplitude. Ces deux grandeurs correspondent à la hauteur du son et à son intensité.

Rappelons au passage que nous possédons deux oreilles, que le cerveau est capable d'analyser les différences de niveau ou ILD (*Interaural Level Difference*) et de phase ou ITD (*Interaural Time Difference*) entre les signaux reçus par chacune d'elles et provenant de la même source sonore, ce qui lui permet, en prenant également en compte la cohérence entre les deux signaux

ou IC (*Interaural Coherence*) de déterminer assez précisément la localisation spatiale de la source.

1.3.2 Les propriétés de l'audition

Le fonctionnement de l'oreille est limité aussi bien dans le domaine des fréquences que dans celui des puissances.

Dans le domaine des fréquences la perception auditive s'étend, du grave à l'aigu, d'une dizaine de Hertz à 20 kHz (dans le meilleur des cas). Dans le domaine de la puissance il est limité vers le bas par le seuil de perception en dessous duquel le mouvement du tympan devient trop faible et vers le haut par le seuil de la douleur au delà duquel le système auditif risque la destruction.

L'oreille dispose par ailleurs d'une possibilité d'adapter, par le jeu des muscles des osselets, sa sensibilité au niveau sonore prévisible ; c'est ce qui nous permet de « tendre l'oreille » ou de ne pas être assourdi par un son très fort dont on a anticipé la puissance.

Le domaine utile de l'audition est représenté par le diagramme du champ auditif.

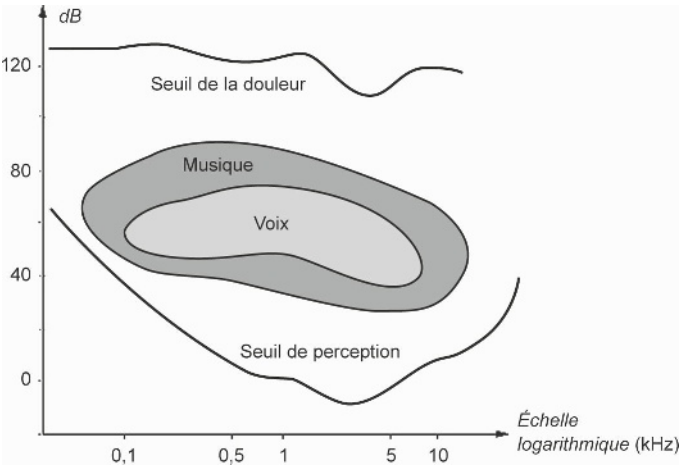


Figure 1.1. Champ auditif humain.

Le seuil de perception peur être momentanément modifié par la réception d'un son fort perturbateur. Cette modification de la perception, dénommée masquage, intervient avant même que l'auditeur ait pu percevoir ce son perturbateur. Ceci résulte du fait que l'analyse des signaux sonores prend un certain temps et qu'elle peut se trouver perturbée, inhibée, plus rapidement.

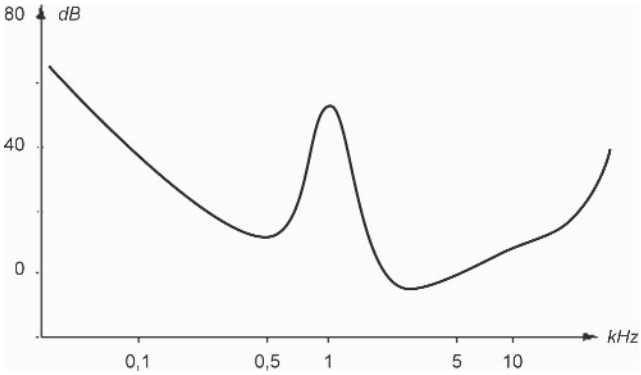


Figure 1.2.
Modification du seuil de perception par un son à 1 kHz.

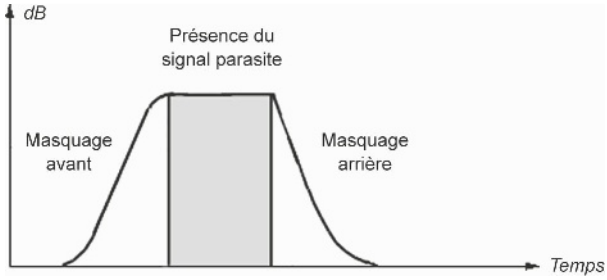


Figure 1.3.
Durée du masquage.

La perception auditive présente également un pouvoir séparateur temporel ; deux impulsions sonores consécutives ne seront perçues comme séparées que si elles sont séparées par un silence de quelques millisecondes ; en deçà, elles apparaîtront comme fusionnées. Ce pouvoir séparateur, mis en évidence par Émile Leipp dépend des individus.

« Tel sujet, disait Leipp, distinguera les impulsions distantes de quelques millisecondes ; tel autre, par

contre, les fusionnera en un magma uniforme. Celui qui possède une finesse de résolution temporelle exceptionnelle s'intéressera, par exemple, aux chants très rapides de certains oiseaux (alouette, rouge-gorge…) dont la « mélodie » le séduira, alors que tel autre trouvera ce chant inintéressant. De même, tel musicien sera attiré par les instruments de musique particuliers, dont il perçoit les transitoires très brefs (clavecin), qu'un autre n'entend même pas parce que son oreille n'est pas assez rapide. »

Ce pouvoir séparateur temporel peut aller de 5 ms à 300 ms, la valeur normale étant de 25 à 50 ms.

La réduction de débit pour les signaux sonores exploite ces propriétés de l'audition suivant la règle « on ne transmet pas ce qui ne peut être perçu ».

1.3.3 Le fonctionnement de l'œil

Notre système de vision, c'est-à-dire le couple œil-cerveau, est extrêmement complexe, notamment pour ce qui concerne l'interprétation par le cerveau de l'influx nerveux : stabilisation de l'image, analyse diagonale, détection de mouvements, codage de l'information… et cette complexité le rend difficile à comprendre.

Le fonctionnement de l'œil, en lui-même, est mieux connu. Ce fonctionnement est analogue à celui d'un appareil photographique (c'est historiquement plutôt l'inverse).

L'œil comporte un système optique formant du réel une image sur une surface sensible : la rétine.

Le système optique est astucieux, car il réalise la mise au point par déformation de l'objectif, mais il est rudimentaire, car il n'est constitué que d'une seule lentille, la cornée et le cristallin. Ceci entraîne des aberrations chromatiques. La mise au point sur la rétine (on parle d'accommodation) ne peut être correctement réalisée simultanément pour l'ensemble des couleurs.

La rétine, surface sensible physiologique, comporte un ensemble de récepteurs : les bâtonnets et les cônes qui sont des cellules spécialisées dans la production, sous l'action de la lumière, d'un signal électrique : l'influx nerveux, que le nerf optique transmet au cerveau pour exploitation.

Les bâtonnets sont sensibles à l'énergie globale de la lumière.

Les cônes sont de trois types qui réagissent à trois gammes de longueurs d'onde de la lumière. Ils sont baptisés S, M, L (pour *Small wave*, *Medium wave* et *Long wave*) selon les longueurs d'onde qui les concernent. On note que les courbes des cônes M et L sont très proches l'une de l'autre.

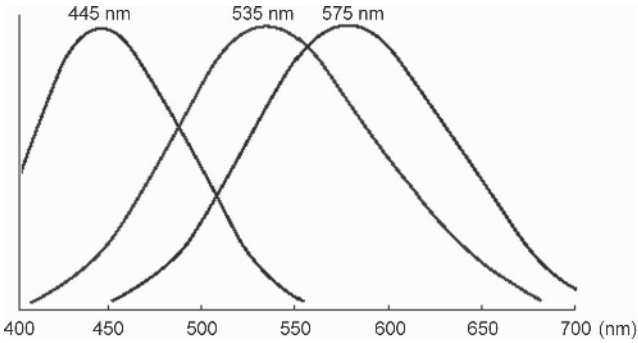


Figure 1.4.
Sensibilité spectrale des trois types de cônes.

À cette propriété physique de longueur d'onde le cerveau fait correspondre une sensation colorée.

La gamme des longueurs d'onde susceptibles d'exciter les récepteurs de la rétine s'étend de 400 à 700 nanomètres (nm). La sensibilité est maximale à 550 nm. Les sensations colorées vont du bleu au rouge, la sensibilité maximale correspond au milieu de la gamme, soit au jaune-vert.

On peut dire que c'est le cerveau qui fabrique les couleurs à partir de trois signaux électriques. C'est l'existence de ces trois types de cônes qui permet de comprendre la notion de synthèse additive.

Les bâtonnets sont très nombreux (plus de 100 millions) et sont répartis sur l'ensemble de la rétine. Ils constituent des grappes dont tous les éléments sont reliées à une même terminaison nerveuse. Les grappes situées en bordure du champ visuel connectent un plus grand nombre de bâtonnets à une fibre nerveuse que les grappes situées dans la partie centrale de notre visuel. L'addition des signaux individuels conduit à une sensibilité élevée (ils demeurent sensibles même à des niveaux d'énergie lumineuse très faibles : vision nocturne ou scotopique) mais donne des informations spatiales moins précises que celles fournies par les cônes.

Les cônes, environ 6 millions, sont reliés individuellement à une terminaison nerveuse. Ils sont responsables de la vision dite photopique (intensité lumineuse supérieure à 3 candelas par mètre carré). Leur densité est maximale dans une étroite zone centrale de la rétine, située sur l'axe optique de l'œil : la fovéa.

C'est cette petite zone qui permet une vision précise dans un angle de champ inférieur à 20°. Le reste de notre champ visuel donne une image imprécise, il s'agit d'un champ de veille.

On a montré en particulier que notre champ visuel binoculaire a une largeur approximative de 160 à 180° (ce qui correspond au champ d'une optique de focale 20 mm pour un appareil 24 × 36) tandis que la région de ce champ accessible par les deux yeux couvre un domaine de 120° de large pour une hauteur de 135°. Sur la largeur du champ visuel de chacun de nos yeux, il reste donc environ 40° qui ne sont pas ou peu sensibles au relief puisqu'il faut deux images pour créer le relief.

Afin d'obtenir une bonne résolution spatiale d'une scène complète, l'œil explore celle-ci grâce à des mouvements rapides : les saccades oculaires. Il ne s'agit pas d'un balayage géométrique régulier mais de déplacements vers des zones considérées comme particuliè-

rement intéressantes, généralement déterminées en fonction de la saccade précédente à moins de surgissement intempestif d'un objet dans le champ visuel.

Le cerveau, informé de la position du regard grâce aux influx nerveux qui lui sont fournis par les muscles supportant l'œil et la tête ainsi que par ceux provenant des canaux semi-circulaires, fonctionne comme un stabilisateur optique afin de rétablir une image « utilisable » à partir des images « chahutées » provenant des différentes saccades oculaires.

Faites l'expérience suivante : déplacez votre globe oculaire avec le doigt ; le cerveau ne sait plus corriger les déplacements du regard. Vous comprendrez alors quelle chance nous avons d'avoir un appareil de la vision bien adapté.

Ce serait une grossière erreur de penser que le cerveau est également capable de stabiliser des images enregistrées par une caméra agitée !

1.3.4 Les propriétés de la vision

L'œil ne peut distinguer deux points proches que si l'angle sous lequel il les perçoit est supérieur à une minute (1/60°). C'est le pouvoir séparateur de l'œil.

Cette propriété a conduit au choix d'environ 600 lignes pour le balayage des images de télévision « standard ». On avait en effet constaté que les spectateurs s'installaient spontanément à une distance du récepteur qui correspond à une angle de vision vertical de l'ordre de 10°, soit environ 3 fois la diagonale de l'image. C'est pour cela que les tailles d'écran sont toujours exprimées par la diagonale de l'image et non par la taille de la base.

L'œil fait spontanément la mise au point pour sa zone de sensibilité maximale aux environs de 550 nanomètres, au milieu du domaine des longueurs d'onde perceptibles. Les aberrations chromatiques de son système

optique fait qu'il distingue mal les détails fins colorés. On se permettra donc de réduire les informations relatives aux couleurs.

On pense qu'il existe un traitement de l'information provenant des trois types de cônes qui conduirait à une information noir/blanc (issue essentiellement des signaux des cônes L et M), une information rouge/vert, une information jaune/bleu.

Ces deux propriétés justifient le choix des systèmes Y,U,V avec réduction d'information pour les signaux de chrominance.

L'œil peut discerner assez peu d'échelons lumineux, on parle de sensibilité différentielle au niveau lumineux $\Delta L/L$ (ΔL plus faible écart discernable) Cette sensibilité dépend de l'éclairement, elle est très faible en vision scotopique. Entre 1 et 100 cd/m^2 (ce dernier niveau lumineux est courant pour des écrans de télévision), l'œil discerne environ 300 niveaux. C'est une donnée intéressante pour le choix de la profondeur de quantification des signaux vidéo.

Dans le domaine temporel on constate que la résolution de l'œil dépend, là encore, du niveau lumineux. Elle est faible aux basses lumières mais, même sous un fort éclairement, elle s'effondre au dessus d'une vingtaine de Hertz. Au-delà de cette fréquence apparaît le phénomène de fusionnement. Les différentes images successives sont interprétées comme un continuum temporel. C'est ce qui a rendu possible le cinématographe.

1.3.5 La synthèse additive

Au XVII^{e} siècle, Isaac Newton avait découvert que la lumière blanche peut se décomposer, à la manière de l'arc-en-ciel, en rayons multicolores susceptibles de se recomposer à nouveau en lumière blanche.

Au tout début du XIX^{e} siècle, Thomas Young, médecin et physicien britannique, constata qu'il n'était pas nécessaire d'utiliser toutes les couleurs du spectre pour reconstituer de la lumière blanche mais que trois d'entre elles, rouge, vert, bleu (les couleurs primaires) suffisaient.

Une cinquantaine d'années plus tard James Clerk Maxwell allait formaliser une représentation graphique de ce phénomène à l'aide d'un triangle dont les sommets correspondent aux couleurs primaires.

La télévision fait appel à cette méthode de la synthèse additive. La captation des images sépare la lumière issue de la scène en trois faisceaux : rouge, vert et bleu. Chacun de ces faisceaux donne naissance à un signal électrique.

À la réception, ces signaux servent à moduler trois sources lumineuses rayonnant dans le rouge, le vert et le bleu. Il peut s'agir d'écrans récepteurs actifs composés de petites sources électroluminescentes accolées dont on peut commander individuellement la luminosité ; il peut également s'agir des trois faisceaux d'un projecteur dont l'intensité est modulée par les signaux électriques correspondant aux trois couleurs.

1.4 TÉLÉVISION : HISTORIQUE DES PROGRÈS ET DES MANQUES

1.4.1 L'image animée : une information critique

La vision est le sens nous fournissant le plus d'informations sur notre environnement. On estime que 80 % de notre connaissance du réel passe par l'œil. Nos autres sens (l'ouie pour les musiciens, l'odorat pour les parfumeurs) nécessitent une véritable éducation qu'on leur accorde rarement.

Très tôt l'homme en a pris conscience. Ainsi, Œdipe voulant se punir de manière exemplaire se prive de la vision.

L'homme a eu beaucoup de mal à « mettre à plat le réel » en tentant de le représenter par des images. L'invention de la perspective au Quattrocento fut une étape décisive.

La représentation du mouvement dut attendre 1832 avec le phénaskistiscope de Joseph Plateau, 1877 avec le praxinoscope d'Émile Reynaud et surtout 1895 avec le cinématographe des frères Lumière… Encore fallut-il patienter quelques dizaines d'années pour admirer des images réalistes en mouvement et en couleur.

Cette lente évolution prouve que l'image animée est une information critique, à manipuler avec précaution.

Tout comme pour le cinématographe, les images de télévision se développent dans trois dimensions : deux dimensions spatiales x et y (le plan de l'écran), une dimension temporelle t.

Toute réflexion sur ces images devra prendre en compte deux exigences : spatiale et temporelle. Puisqu'il s'agit de télévision elle devra également s'accommoder des capacités du canal de transmission.

Pour transmettre ses images le cinématographe disposait d'un canal large bande : le faisceau lumineux du projecteur mais dont la portée était faible et surtout unipolaire.

La télévision ayant d'autres ambitions a dû avoir recours à des signaux électriques (transmission par câbles) ou électromagnétiques (transmission par ondes hertziennes). Les canaux utilisés sont hélas à bande limitée.

Il n'est plus possible de transmettre simultanément les information concernant tous les points de l'image. Il a donc été imaginé de les transmettre séquentiellement en explorant l'image ligne à ligne. On parle de balayage.

Ce terme évoque peut-être le pantélégraphe de Giovanni Caselli qui en 1862 réussit à transmettre grâce

à une ligne téléphonique des textes et des dessins de Paris à Amiens.

L'élément à transmettre était tracé avec une encre isolante sur un plaque de cuivre. Un pendule à l'extrémité conductrice venait balayer la plaque selon l'axe *x*. Celle-ci se déplaçait régulièrement selon l'axe *y*. Ce dispositif permettait d'analyser le dessin par une succession de lignes.

Un dispositif à relais faisait que, lorsque l'extrémité du pendule frottait sur une zone encrée, un courant passait dans la ligne téléphonique, lorsqu'il frottait sur la plaque le courant était interrompu.

À la réception on trouvait un dispositif mécanique analogue. C'était un papier imprégné de cyanure de potassium qui devenait bleu lors du passage du courant dans la ligne.

La grande difficulté résidait dans la parfaite synchronisation des deux mécanismes.

On a là tous les éléments de l'analyse d'une image TV… ou presque, il ne manque que la prise en compte de la variable *t* !

C'est Paul Nipkow qui a maîtrisé cette variable en 1884 grâce à l'utilisation d'un disque percé de trous disposés en spirale. Ces trous analysent l'image selon des portions d'arcs de cercle (18 sur le croquis) assimilables à des lignes.

Ce système amélioré par John Loggie Baird et par Marcel Brillouin, qui inséra des lentilles dans les trous, a été utilisé à partir du milieu des années 20 pour des émissions de télévision sur une trentaine de lignes Il était difficile d'obtenir plus d'une quinzaine d'images par seconde, ce qui exigeait déjà une vitesse de rotation du disque de 900 tours par minute.

Toute cette astucieuse mécanique (il y en eu bien d'autres qui ne dépassèrent pas le stade du projet ou celui du prototype) fut en fin de compte remplacée

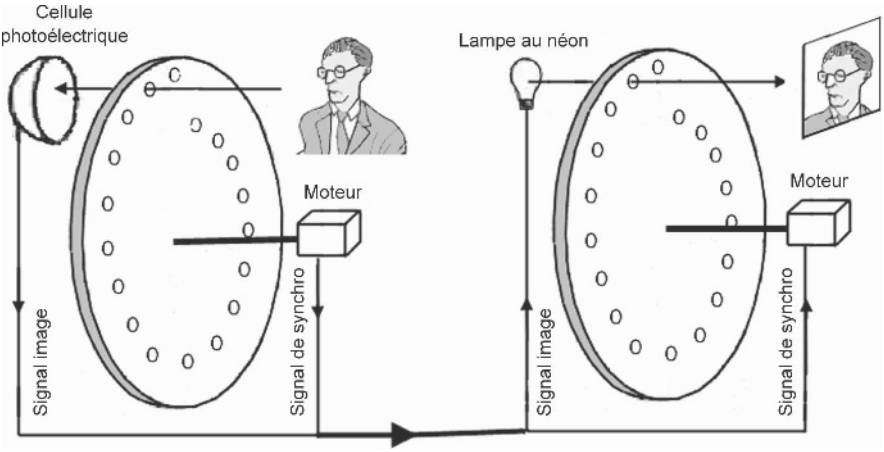


Figure 1.5.
Synoptique d'une transmission d'image grâce au disque de Nipkow.

par l'agilité d'un balayage électronique dans un tube cathodique.

Celui-ci est le fruit des travaux de l'anglais Crookes des allemands Braun, Vichert et Wehnelt (à la fin du dix-neuvième siècle), puis du russe Boris Rossing (entre 1907 et 1911).

C'est vraisemblablement la société Intégra de Marc Chauvière, s'appuyant sur les travaux de René Barthélémy et Henri de France, qui a présenté en 1936 le premier récepteur à tube cathodique. Ce n'est, bien sûr, qu'après la guerre que ce genre de récepteur fut universellement adopté.

Parallèlement le russe Zworykin, émigré aux USA, conçut en 1931 la première caméra utilisant un tube cathodique : l'iconoscope.

Un pinceau d'électrons, convenablement dirigé dans un tube à vide, est en effet tout à fait capable de balayer un petit rectangle avec régularité et vélocité.

Les électrons émis par une cathode en substance émissive chauffée (oxyde de strontium et de baryum) sont accélérés et focalisés en un pinceau étroit par un système d'électrodes portées à un potentiel positif par rapport à la cathode (lentilles électrostatiques).

Le pinceau électronique est ensuite dévié horizontalement et verticalement en passant soit entre des paires de plaques dont l'une est à potentiel positif, l'autre à potentiel négatif, soit entre des bobines parcourues par un courant. La déviation est proportionnelle aux valeurs du potentiel ou du courant dont les variations peuvent être très rapides. La combinaison des deux déviations permet de balayer la totalité de la face avant du tube.

Les signaux de balayage horizontaux et verticaux sont en dents de scie : croissance régulière correspondant à la durée utile du cycle d'analyse, décroissance rapide correspondant au retour du pinceau d'électrons.

La durée d'un cycle ligne est de 64 µs, celle d'un cycle trame (voir plus loin) de 20 ms. Les fréquences des signaux de balayage doivent être précises et absolument asservies l'une à l'autre. Elles sont obtenues à partir du même oscillateur à quartz. En multipliant quatre fois par cinq la fréquence trame de 50 Hz, on obtient 31 250 Hz ; cette fréquence divisée par deux donne 15 625 Hz, soit 625 × 25, qui est la fréquence ligne. Ceci explique pourquoi on a choisi pour les systèmes de télévision actuels 625 lignes plutôt que 611 ou 599…

Dans le tube récepteur le faisceau électronique, convenablement modulé par le signal vidéo grâce à une électrode appelée Wehnelt (du nom de son inventeur), vient transmettre de l'énergie à un revêtement électroluminescent appliqué sur la face interne du tube.

Dans l'iconoscope et sa descendance (Plumbicon, Vidicon, Saticon… tous systèmes dont la désinence ne doit pas laisser à penser qu'ils sont un peu idiots mais qu'ils sont des dispositifs imageurs ; ikon = image en Grec.) le faisceau d'électrons transporte des charges engendrées par la lumière, issue de l'objectif de la caméra, vers une « cible » photo-émissive ou photo-résistante. Il en résulte la circulation d'un courant dont l'intensité dépend de l'insolation locale de la cible.

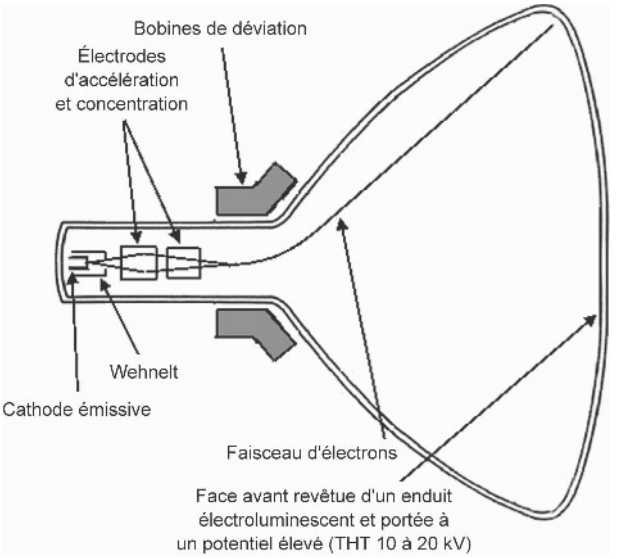


Figure 1.6.
Schéma d'un tube « noir et blanc ».

Dès cette nouvelle ère électronique on a pu transmettre 25 images par seconde (30 dans les deux Amériques en raison de la fréquence du courant industriel) et le nombre de lignes dépassa 400… mais l'image sur « l'étrange lucarne » du tube cathodique de réception ne dépassait pas le format A4, voire A5 (15 cm × 21 cm), comme on dirait aujourd'hui.

C'était alors, et pour quelques temps encore, de la télévision en noir et blanc… ou plutôt en gris clair et gris foncé.

1.4.2 Résolutions spatiale et temporelle. Balayages entrelacé et progressif

Les trois variables *x*, *y*, *t* sont a prendre en compte pour une restitution satisfaisante d'une séquence animée. Il s'agit de faire croire au spectateur qu'il assiste à une scène réelle. On met à profit à cet effet les caractéristiques de la psychophysiologie de la vision humaine.

L'histoire de la peinture montre que nous avons pu maîtriser peu à peu le passage à la représentation 2D d'une scène 3D.

Le cinématographe nous a appris que le phénomène de persistance rétinienne, lié sans doute à des capacités d'interpolation cérébrales, faisait qu'on pouvait prendre une série d'images fixes pour un continuum temporel. Il suffisait pour que ça marche (enfin pour que ça ne marche pas trop mal !) d'enchaîner 24 images fixes par seconde.

Pour que le système demeure crédible, il faut également que chacune de ces images propose un nombre de points suffisants. On parle de résolution temporelle et de résolution spatiale.

Dans le cas de la télévision ces deux résolutions sont liées entre elles par l'intermédiaire de la capacité, toujours limitée, du canal de transmission. On peut caractériser schématiquement celle-ci par un nombre de lignes à la seconde.

Dans les Amériques, les systèmes de télévision « standard » (SDTV, *Standard Definition TV*), antérieurs à la Haute Définition, ont une résolution temporelle de 30 images par seconde (en fait 29,97) pour une résolution spatiale verticale de 480 lignes.

Dans le reste du monde, à l'exception du Japon qui a adopté pour des raisons géopolitiques le système américain, ces résolutions sont respectivement de 25 images par seconde et de 575 lignes. Il s'agit de lignes utiles destinées effectivement à la transmission de l'information « image ».

Ces différences de fréquence de renouvellement des images sont liées, du moins initialement, à la différence de fréquence des courants électriques.

On parle (abusivement) de systèmes à 525 ou 625 lignes parce qu'on trouve, à chaque image 8 % des lignes qui correspondent a la durée du retour du pinceau électronique vers le haut de l'image, qui ne transportent aucune information concernant l'image elle même et qui ont été utilisées pour transmettre des données de service. (ces lignes ne sont pas visibles sur l'écran).

Dans les deux cas, ceci représente au total un peu plus de 15 500 lignes transmises par seconde.

On s'est rendu compte, dès que les écrans eurent des dimensions raisonnables, que le nombre de 25 images par seconde était souvent insuffisant. Mais on le savait déjà, pour le cinématographe pour lequel on avait été amené à introduire un obturateur découpant la projection de chaque image en deux illuminations successives.

Cet effet de papillotement ou de scintillement est encore plus gênant en télévision où les points de l'image sont affichés successivement et temporairement avec un effet de rémanence, tel que lorsque le dernier point d'une image est affiché, le premier s'efface (au moins pour les écrans cathodiques).

On a donc imaginé, en s'appuyant sur les effets de persistance rétinienne, de substituer au balayage dit aujourd'hui progressif, (celui où toutes les lignes de l'image sont affichées dans une même passe), un balayage dit entrelacé (*interlaced*).

On ne transmettait plus 25 images de 625 lignes par seconde, mais 50 trames de 312,5 lignes. La première trame est dite impaire (car elle concerne les lignes impaires), elle se termine par une demi ligne ; la seconde paire commence par une demi ligne (ceci au moins tant que l'on a utilisé des systèmes de prise de vues et d'affichage à tubes cathodiques ; dans les nouveaux systèmes matriciels, capteurs CCD et écrans à cristaux liquides ou à plasma le nombre de lignes utiles est passé de 575 à 576).

Le débit global est le même, mais l'effet de papillotement est diminué.

L'entrelacement des deux trames provoque un nouvel effet dit « scintillement interligne » (*interline flicker*) qui se manifeste particulièrement lors d'un déplacement vertical des images. Il correspond à un détail horizontal qui apparaît sur une trame mais pas sur la suivante.

La résolution verticale, c'est-à-dire le plus petit détail qu'il est possible d'afficher sur l'écran, résulte de l'échantillonnage de l'image en lignes… et de l'application de la règle de Nyquist (se reporter au chapitre concernant les techniques numériques). Si l'échantillonnage comporte 576 lignes on ne pourra, théoriquement, afficher plus de 288 lignes horizontales. En fait; on se donne une marge de tolérance (c'est-à-dire le risque d'artefacts possibles pour les détails fins verticaux) en adoptant une valeur de $576 \times 0{,}7$. Ce coefficient 0,7 est le facteur de Kell.

En 1949, la France se dotait d'un ambitieux système de télévision N & B à 819 lignes (768 lignes utiles). Sans doute un luxe inutile.

1.4.3 La couleur n'est pas un luxe

Toutes les recherches sur la télévision en couleur exploitent les possibilités de l'analyse et de la synthèse trichrome liées au fait que notre rétine possède trois types de récepteurs chromatiques spécialisés pour le bleu, le vert et le rouge.

Dès le fin des années vingt John Loggie Baird expérimentait déjà la transmission d'images en couleurs.

Mais c'est en 1950, aux USA, que la télévision en couleurs à réellement démarré.

Deux procédés étaient en concurrence. Ils exploitaient, chacun à sa manière, les possibilités d'analyse et de synthèse trichromatique ainsi que les propriétés de notre système de la vision.

Le procédé de CBS était séquentiel. Des disques à 3 secteurs colorés (rouge, vert et bleu) tournaient en synchronisme en avant d'une caméra et d'un récepteur N & B.

Celui de RCA faisait appel au merveilleux (et improbable, tant il doit être de construction précise) tube

cathodique à masque (*shadows mask tube*). On peut considérer ce tube à masque comme l'imbrication étroite de trois tubes N & B.

Trois sources d'électrons commandées par les trois signaux de couleurs primaires R, V, B fournissent trois faisceaux d'électrons. La face avant du tube est tapissée d'une mosaïque de petites pastilles électroluminescentes juxtaposées susceptibles de rayonner, sous l'impact du faisceau d'électrons, selon chacune des couleurs primaires. Elles sont disposées soit en points groupés par triplets, soit en bandes verticales adjacentes alternativement R, V, B. (Tubes PIL, *Precision In Line*). Un dispositif (masque, ou grille de focalisation) est placé entre la source et l'écran de manière que les électrons provenant de la source rouge ne puissent viser que les luminophores rayonnant de la lumière rouge, etc.

Un bien beau système qui a fait ses preuves.

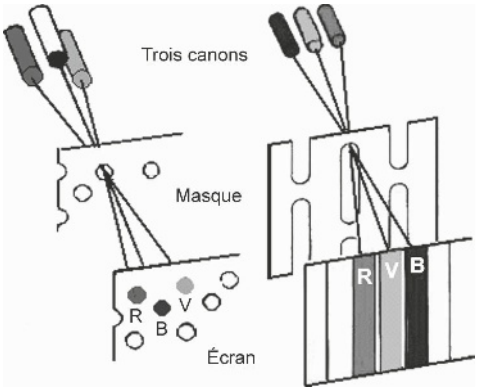


Figure 1.7.
Deux types de tube « couleur ».

La première diffusion par CBS eut lieu sur 5 stations de la Côte Est en juin 1951. La première diffusion par NBC avec le procédé RCA eut lieu dans 22 villes en janvier 1954.

Dès 1952 il fut évident qu'il était indispensable d'avoir un standard commun. Le « National Television Stan-

dard Committee » publia en 1953 les données de ce qui allait être connu sous les initiales NTSC.

Tous les éléments essentiels de ce qui, plus d'un demi siècle après, constitue les fondements de la télévision en couleurs avaient été codifiés.

On analyse, dans la caméra, la scène à transmettre suivant trois couleurs primaires : le rouge, le vert et le bleu, grâce à un jeu de réflexions (totales et partielles) et à trois filtres colorés. On obtient trois signaux électriques R, V, B.

À la réception, notre système de la vision combine les informations optiques affichées sur l'écran (aujourd'hui il s'agit toujours d'une mosaïque de triplets R, V, B juxtaposés) grâce aux trois signaux R, V, B, afin de reconstituer par synthèse additive trichrome les couleurs initiales.

1.4.4 La première réduction du débit vidéo

On commença par transmettre (simultanément ou successivement) les signaux R, V, B. (*R, G, B* en anglais), solution simple et efficace mais qui demande la transmission de trois fois plus d'information (à résolution égale) que le N & B.

La télévision en couleurs s'étant établie à partir d'un existant N & B il lui a fallu se plier à des règles de compatibilité :

- même bande passante du canal de transmission bien qu'on transmette des informations supplémentaires
- un poste N & B doit pouvoir exploiter une émission « couleur ».

Afin de répondre à ces contraintes de débit et de compatibilité on estima plus judicieux de transmettre les trois informations nécessaires sous une forme plus élaborée que celle du système R, V, B et on opta pour :

1) Un signal dit de luminance Y, composé à partir des signaux primaires R, V, B pondérés suivant la courbe de sensibilité chromatique de l'œil :

$$Y = 0{,}3\ R + 0{,}6V + 0{,}1\ B$$

On note que pour R = V = B = 1, ce qui correspond au blanc, Y = 1, et que pour R = V = B = 0, le noir, Y = 0.

Ce signal, qui informe sur l'intensité lumineuse globale de la scène, est transmis comme le signal noir et blanc. Il est donc directement utilisable par les récepteurs N & B (la pondération donne aux plages colorées l'intensité que l'œil leur attribuerait dans la réalité).

2) Deux signaux complémentaires dits signaux de chrominance, (qui auraient pu être, par exemple, R et B), permettant d'obtenir l'information V à partir de l'expression de Y grâce à des opérations arithmétiques simples à la portée d'une électronique banale.

En fait, on a choisi de compliquer un peu l'arithmétique ; les deux signaux de chrominance sont (R - Y) et (B - Y). Ce sont les signaux de différence de couleur notés Dr et Db, ou Cr et Cb, ou encore U et V (en anglais il n'y a pas de risque de confusion avec le G de green). On remarque que ces deux signaux sont nuls pour toute la gamme des gris correspondant à des combinaisons telles que R = V = B.

On prend alors en compte le fait que l'œil, instrument optique non corrigé des aberrations chromatiques (il ne possède qu'une seule lentille), est peu exigeant quant à la couleur des détails fins.

On transmet la luminance Y, avec une large bande afin de maintenir une bonne résolution pour le « dessin » des détails (mais sans qu'on puisse leur attribuer une couleur). En revanche, on ampute largement la bande passante attribuée aux informations chromatiques portées par U et V. Les informations de couleur concernant les détails fins sont ainsi éliminées, mais cette

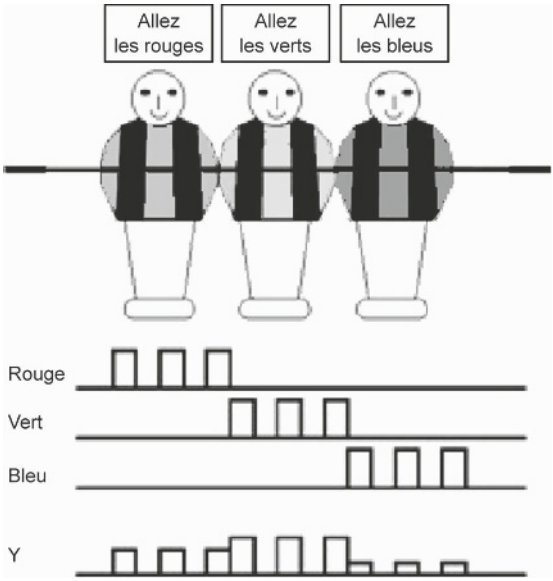


Figure 1.8.
Les signaux primaires rouge, vert, bleu, sont transformés en un signal de luminance Y, et en deux signaux de chrominance U et V.

opération ne pénalise pas le regard du spectateur : « on ne transmet pas ce qui ne peut pas être perçu ».

Les premiers systèmes de télévision NTSC puis PAL et SECAM n'avaient pas hésité à diviser par 4 la bande passante des signaux de chrominance puisque, en tout état de cause, le spectateur ne percevrait pas ces éléments colorés.

Il s'agit en fait de la première réduction de débit avec pertes. Le taux de compression de 2:1 était loin d'être négligeable, la détérioration de l'image était cependant peu apparente. Seuls des traitements tels que l'incrustation en chrominance (*chroma key*) pouvaient en souffrir, mais ceci ne constituait pas un handicap car la post-production vidéo était à cette époque balbutiante et le système était totalement « orienté diffusion ».

La bande passante nécessaire (approximativement 9 MHz) était encore trop importante pour les possibilités de transmission des canaux existants. Il a donc fallu imaginer pour les systèmes dits composites : NTSC, PAL, SECAM, une manipulation supplémen-

taire, l'imbrication des spectres de luminance et de chrominance.

Cette nouvelle étape du « compactage » du signal est, hélas, généralement pénalisante en raison de la difficulté qui existe quant à l'indispensable désimbrication, à coût supportable, des deux spectres dans le récepteur familial. La mise en œuvre de désimbrications bon marché, mais simplistes, conduisit à une perte de résolution ainsi qu'à des artefacts comme le *cross-color* qui fait que les détails fins de l'image (costumes pied-de-poule, chemises à rayures, rideaux…) ont tendance à donner naissance à de magnifiques arcs-en-ciel.

Ces trois systèmes, bien que reposant sur des schémas technico-physiologiques identiques, sont en outre incompatibles entre eux. Il existait déjà les deux lignées 60 et 50 Hz ; les signaux composites scindent cette dernière en deux branches, PAL et SECAM. Un résultat qui pesa lourd sur l'image électronique à partir de la fin des années 60. Ce phénomène a été particulièrement ressenti en France, isolat SECAM au milieu d'une Europe occidentale convertie au PAL.

Les systèmes en composantes YUV utilisent des canaux séparés pour les trois signaux. Ils évitent ce problème d'imbrication et sont plus généreux pour les signaux de chrominance (division par 2 seulement de leur bande passante, taux de compression 1,5:1). En outre ils ont constitué une élégante solution pour sortir du grand schisme PAL/SECAM, dont bénéficia la gamme des magnétoscopes Betacam de Sony.

Ce sont des signaux qui sont destinés au tournage et à la post-production vidéo, la diffusion hertzienne, au moins jusqu'à ces dernières années, continuant très majoritairement sur des canaux composites.

1.4.5 La deuxième réduction du débit vidéo

Dès les années 80, la question d'une télévision, analogique bien entendu, à résolution améliorée était à l'ordre du jour.

En 1986 les industriels japonais présentèrent, lors de la réunion plénière du CCIR, leur norme HiVision (1 125 lignes dont 1 035 utiles, 30i/s et format 16/9, bande passante 30 MHz). En 1988, ils l'avaient rendu compatible avec le NTSC : c'était le système MUSE. La transmission de l'ensemble d'une image était étalée sur 4 trames ; des mémoires de trame intégrées dans le récepteur permettaient de recombiner les quatre flux d'information successifs afin de reconstruire l'image. Le système fonctionnait évidemment très bien pour des images fixes mais se révélait plus délicat lorsque des mouvements importants entraient en jeu. Des mouvements lents et réguliers (panoramiques) pouvaient être identifiés et pris en compte par des « vecteurs mouvement » (*motion vectors*) afin d'éviter des artefacts grossiers.

Les européens réagirent avec le lancement du programme « Euréka 95 » et la création en 1990 d'un GIE (Groupement d'Intérêt Économique) européen « Vision 1250 » au titre explicite.

En 1992 ils proposaient le système HD-MAC (1 250 lignes – le double du PAL et du SECAM afin de permettre une conversion simple vers des récepteurs classiques, dont 1 152 utiles, format 16/9). La réduction de débit était assurée de manière plus soignée que dans le système MUSE.

L'image était décomposée en 10 000 petits rectangles à l'intérieur desquels on analysait l'activité de l'image. Si celle-ci était réputée fixe on transmettait l'information sur 4 trames ce qui conduisait à une faible résolution temporelle. Si cette activité était importante, on rafraîchissait l'information à chaque trame mais avec une résolution spatiale quatre fois plus faible.

Le système fonctionnait parfaitement. Il l'a prouvé pour quelques dizaines de récepteurs (et quelques milliers de spectateurs – dont j'ai eu le plaisir de faire partie) lors de jeux olympiques d'Albertville et de Barcelone en 1992.

L'arrivée rapide des techniques numériques n'a laissé leur chance ni à l'un ni à l'autre de ces deux systèmes.

2.1 PRINCIPES DE BASE DE LA RÉDUCTION DE DÉBIT

Comme nous l'avons vu précédemment il est impératif pour les systèmes de télévision de diminuer « l'encombrement » des séquences d'images pour le stockage et la transmission, sans trop dégrader la qualité.

Les méthodes analogiques y sont parvenues astucieusement mais avec des résultats qui, pour n'être pas négligeables, demeuraient insuffisants.

La mise en œuvre de techniques numériques a totalement modifié la donne.

Les méthodes de réduction du débit numérique (BRR : *Bit Rate Reduction* en anglais) sont aujourd'hui fondamentales dès lors qu'il y transmission ou stockage d'information. On utilise souvent pour les désigner le terme, moins précis, de compression.

La règle de base est « respecter autant que faire se peut l'information tout en réduisant l'encombrement du message ».

Il faut être conscient que la mise en œuvre d'algorithmes de réduction de débit entraîne inévitablement un décalage temporel qui peut devenir gênant voire rédhibitoire dans certaines utilisations « temps réel ».

Le taux de compression s'exprime :

- soit par le rapport entre le volume initial des données et le volume après réduction ; si ce volume est deux fois plus faible alors on écrira qu'il s'agit d'un taux de 2:1 ;
- soit en pourcentage du volume après réduction par rapport au volume initial ; si le volume final représente la moitié du volume initial on écrira qu'il s'agit d'un taux de 50 %.

On groupe ces méthodes en deux catégories : les méthodes sans pertes (*lossless*), dites aussi transparentes, qui ne détruisent aucune information ; les méthodes avec pertes (*lossy*) qui font disparaître une

partie de l'information.. On souhaite que ces dernières soient « virtuellement transparentes », c'est-à-dire que le récepteur ne perçoive pas la déperdition d'information.

La capacité de tolérance du récepteur est donc toujours un paramètre essentiel. Les données scientifiques ou bancaires ne supportent que des réductions absolument sans pertes tandis que les images et les sons peuvent supporter une légère dégradation sans que le spectateur ou l'auditeur ne se sente pénalisé.

Il s'agit toujours de promouvoir des modes plus économiques de description de l'information avec comme mots d'ordre « ne jamais transmettre ce qui a déjà été transmis et que l'on peut réutiliser » et « éliminer le superflu pour ne conserver que l'essentiel ».

L'information superflue est en général énorme mais elle a été longtemps trop complexe à éliminer efficacement.

2.1.1 Le codage à longueur variable

Ce type de codage plus connu sous les initiales RLC (*Run Lenght Coding*) ou RLE (*Run Lenght Encoding*) ou encore VLC (*Variable Lenght Coding*) exploite la redondance entre éléments successifs (des pixels d'une image ou des chiffes d'un tableau ou des lettres d'un texte).

Soit la suite des nombres suivants :

2 1 1 1 5 0 0 0 0 1 4 4 4 4

On pourra l'écrire de la manière suivante :

$2 ; 3 \times 1 ; 5 ; 4 \times 0 ; 1 ; 4 \times 4$

C'est une astuce de codage simple, mais souvent peu efficace.

2.1.2 Les codages entropiques

La notion d'entropie a été évoquée dans un chapitre précédent ; elle caractérise la probabilité d'occurrence d'un élément d'un message ou d'un état d'un système.

Le physicien Américain Samuel Morse, inventeur du télégraphe électromécanique, proposa en 1844 un codage des lettres de l'alphabet qui tenait compte de la fréquence d'occurrence des lettres (en anglais). Le code ne disposait que de deux symboles correspondant à deux durées d'impulsions électriques la ligne téléphonique : brève et longue ou encore, en mode graphique, un point et un trait. Ce code binaire était conçu afin de rendre les transmissions les plus rapides possibles.

Dans ce but la lettre E, la plus fréquente, est représentée par le mot le plus court : une brève ; la lettre T par une longue ; quant au J, assez rare en anglais, il bénéfice d'une brève et de trois longues.

Le fax met également en œuvre un codage entropique Il s'agit, le plus souvent de transmettre des textes. Lors de l'analyse ligne à ligne on a statistiquement plus de chance de rencontrer des groupes de points noirs assez courts (les jambages des lettres) et des groupes de points blancs plus longs (les espaces entre lettres).

On crée une table d'indexation ou LUT (*Look Up Table*) qui fait correspondre un code aux différents motifs.

Points	blanc	noir
1	000111	010
2	0111	11
3	1000	10
4	1011	011
5	1100	0011
6	1110	0010
7	1111	00011
8	10011	000101
9	10100	000100
10	00111	0000100

Un examen de cette table permet de remarquer que l'on a effectivement choisi de représenter les motifs noirs jusqu'à 4 points successifs d'une manière plus économique que les blancs, tandis qu'à partir de 7 points consécutifs c'est l'inverse.

Dans le codage de Huffman (1952), on commence par faire des statistiques de la fréquence d'apparition de chacun des éléments. Ils sont classés dans l'ordre décroissant de probabilité d'occurrence.

On regroupe les deux éléments ayant la probabilité la plus faible pour en faire un nouvel élément dont la probabilité est la somme des probabilités des deux éléments initiaux ; ils peuvent être discriminés en utilisant un seul bit.

On réitère l'opération en créant une arborescence des éléments suivant l'augmentation de la probabilité d'occurrence.

On trouvera ci-dessous un schéma d'arborescence pour 4 éléments.

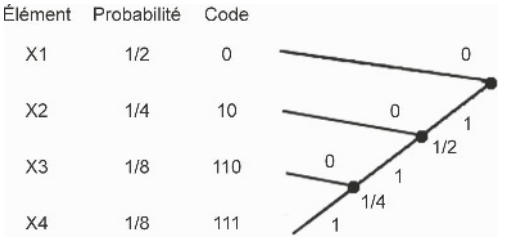


Figure 2.1.
Arborescence d'un codage de Huffman.

2.1.3 Les codages à dictionnaire

Le plus célèbre de ces codages est connu sous les initiales de ses auteurs LZW (Lempel, Ziv, Welch). Il est apparu en 1977 dans sa première version (LZ 77), spécialisée dans les données textuelles et qui est utilisée dans les programmes d'archivages de donnés comme PKZIP.

L'originalité du codage LZW réside dans la construction dynamique du dictionnaire au fur et à mesure de l'analyse du fichier.

Voici son fonctionnement pour une suite de caractères, mais il peut s'appliquer à tous types de données. On le trouve notamment dans les formats d'images GIF et TIFF.

Le code LZW, appliqué au texte, construit son dictionnaire à partir de la table des 256 signes et caractères du code ASCII définis par 8 bits ; il étend cette table à 4 096 « cases » définies grâce à l'utilisation de 12 bits.

Il est basé sur le fait que des successions de caractères se retrouvent plus souvent que d'autres, par exemple les digrammes ES ou NE en français ou le trigramme ENT.

Le dictionnaire est construit dynamiquement d'après les caractères rencontrés. Lorsque l'algorithme rencontre pour la première fois un motif, il le repère par un index qui est le numéro d'une case vide de la table. Lorsque ce motif est à nouveau rencontré l'algorithme le remplace automatiquement par le numéro de la case.

La structure de la table est donc implicitement contenue dans le fichier à traiter. Elle pourra être reconstruite d'une manière analogue lors de la relecture du fichier. Il n'est donc pas nécessaire de la transmettre séparément.

Prenons la suite de caractères E A M O E A E G :

- première entrée E qui se trouve dans la table des caractères ASCII sous le code 69 ;
- deuxième entrée la chaîne EA, ne se trouve pas dans la table et devient le motif 257 ;
- troisième entrée la chaîne AM, ne se trouve pas dans la table et devient le motif 258 ;
- quatrième entrée la chaîne MO, ne se trouve pas dans la table et devient le motif 259 ;

- cinquième entrée la chaîne OE, ne se trouve pas dans la table et devient le motif 260 ;
- sixième entrée la chaîne EA, se trouve dans la table case 257, elle est remplacée par son index 257 ;
- septième entrée, l'entrée précédente étant indexée, on s'intéresse à la chaîne EAE qui devient le motif 261.

C'est ainsi que pas à pas le signal établit lui même la table qui le décrit.

On remarque que de la case 256 à la case 512 les 256 motifs indexés, ce sont ceux qui apparaissent le plus souvent, sont définis à l'aide de 9 bits. Lorsque ces motifs sont des suites de 2 bits, le passage par la table fait gagner 7 bits (9 bits au lieu de 2×8). Lorsqu'il s'agit d'un motif de 3 lettres le gain est de 15 bits (9 bits au lieu de 3×8). S'il s'agit d'un motif de 4 lettres le gain est de 23 bits. Plus le motif est long plus le gain est important.

Il suffit de remplacer les séquences de deux lettres par celles de deux pixels pour comprendre l'intérêt du codage pour les images.

Cet algorithme astucieux et peu exigeant donne de bons résultats sur des images simples définies sur peu de bits (par exemple sur des images N & B ou en 256 couleurs). Mais il devient peu efficace dès que les images deviennent plus complexes. Il ne dépasse pas un taux de compression de 2:1 pour des images en 8 bits par couleur (24 bits).

Les algorithmes JPEG qui seront étudiés plus en détail sont de type codage à dictionnaire, mais les motifs retenus sont bien plus complexes.

2.1.4 Les codages de type psychophysiologique

Le mot d'ordre est ici « ne conserver que ce que le spectateur, ou l'auditeur, peut effectivement utiliser ». La connaissance des limites de nos possibilités de

perception peut permettre d'éliminer, sans dommages apparents, une partie des informations initialement contenue dans le signal.

Cette élimination vient en appui des algorithmes de réduction de débit sans pertes. Elle permet, pour la distribution des programmes d'obtenir des débits faibles, voire très faibles, au prix de dégradations peu perceptibles.

Ces méthodes de réduction de débit entraînant des pertes peu sensibles doivent être évitées (ou maniées avec la plus extrême réserve) dans le cadre de la production ou de la post-production où la perte de certaines données peut devenir très pénalisante lors des traitements ultérieurs.

2.2 LES TECHNIQUES NUMÉRIQUES

2.2.1 Les méthodes analogiques

Depuis un siècle et demi (1875, début du téléphone de Graham Bell) on sait transporter certaines données assez complexes grâce à l'électricité.

La variation de pression d'une colonne d'air qui constitue le signal sonore, est ainsi traduite en une variation analogue d'un signal électrique, grâce à un dispositif approprié : un transducteur (dans ce cas, un microphone, schématiquement une petite bobine qui se déplace face à un aimant).

À la réception un autre transducteur (ici un écouteur ou un haut-parleur, schématiquement une petite plaque métallique ou une petite bobine activée par un électro-aimant) recrée une variation de pression identique (plus ou moins) à celle d'origine.

Toutes les transmissions, dites analogiques, fonctionnent selon le même schéma : transducteur d'entrée puis de sortie permettent d'emprunter le domaine électrique pour la durée du voyage.

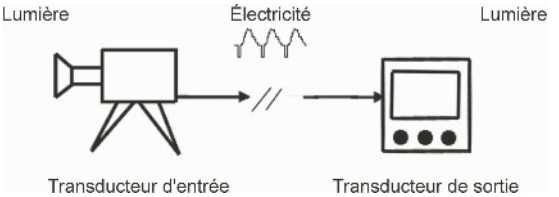


Figure 2.2.
Principe d'une chaîne de transmission analogique.

Ces méthodes analogiques, progressivement améliorées, ont donné pleine satisfaction. Elles souffrent cependant d'un handicap considérable qui est celui du « bruit de fond ». Il s'agit d'un terme qui rappelle les années TSF où l'on entendait Frehel chanter sur fond de crachotements, sifflements…

Le bruit de fond ou simplement le bruit (*noise* en anglais) représente l'ensemble des signaux parasites, acquis au cours d'une transmission, qui viennent perturber la réception et la compréhension d'un message.

Chaque étape de la transmission apporte sa contribution au bruit de fond. Une fois celui-ci acquis, il est très difficile, voire impossible, en mode analogique, de l'éliminer.

La qualité d'un élément de la chaîne peut être qualifiée par le rapport signal sur bruit (S/B, ou S/N en anglais).

Ce rapport est défini en acoustique comme le rapport W_u / W_b entre la puissance du signal utile W_u et celle du bruit W_b. On l'exprime en général en décibels :

$$S/B_{dB} = 10 \log W_u / W_b$$

3 dB correspondent à un facteur 2 car le logarithme décimal de 2 est égal à 0,3.

La puissance électrique étant proportionnelle au carré de la tension du signal on utilise généralement pour caractériser la qualité d'une chaîne de transmission des valeurs de tensions électriques et le rapport devient :

$$S/B = 20 \log U_u / U_b.$$

Le rapport S/B d'une caméra est de l'ordre de 60 dB, celui d'un mélangeur de l'ordre de 80 quand celui d'un

magnétoscope analogique est de l'ordre de 45 dB. Les magnétoscopes et les magnétophones étaient donc le maillon le plus faible de la chaîne analogique. Ceci explique qu'il était impossible de réaliser plus de quelques générations successives d'un programme car chaque génération (surtout l'enregistrement) apportait sa contribution au bruit de fond.

Le numérique était indispensable. En effet, dès que l'on a traversé le miroir pour pénétrer dans l'univers des nombres, et que l'on y demeure, il n'y a plus de modification du bruit (au moins théoriquement !).

2.2.2 Échantillonner et quantifier

J'ai l'habitude de faire remonter les règles de base de la numérisation aux environs de l'an mille lorsque le moine Guido d'Arezzo, perfectionnant des tentatives antérieures, institua la notation musicale moderne en proposant de représenter une composition musicale grâce à la portée. Les différentes notes codifient la hauteur (la fréquence) des sons. Les mesures, avec leurs sous multiples (rondes, noires, croches…), liées à une base de temps ou une horloge (le métronome), définissent l'instant où ces notes doivent être jouées. Une véritable réussite qui permit de conserver et transmettre le patrimoine musical.

La numérisation consiste à transformer un signal continu en une suite d'éléments discontinus ou « discrets » représentés en musique par des symboles ou, plus largement aujourd'hui, par des nombres.

Ces nombres sont eux mêmes physiquement codés sous forme d'impulsions électriques, optiques…

Ces opérations sont effectuées sur le signal analogique par un Convertisseur Analogique/Numérique ou CAN. Après traitement, enregistrement, transmission, un Convertisseur Numérique/Analogique (CNA) permet de revenir à un domaine analogique, seul accessible à nos sens.

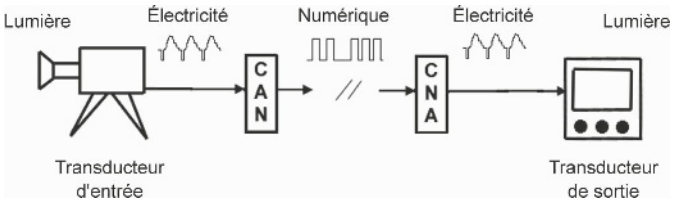


Figure 2.3. Principe d'une chaîne de transmission numérique.

Dans la première étape de la numérisation, on sélectionne à intervalles réguliers (c'est le pas d'échantillonnage) des échantillons (*samples* en anglais) dont la collection doit pouvoir se substituer au signal de manière à en donner une représentation satisfaisante.

L'échantillonnage (*sampling*) ne doit retenir ni trop d'échantillons, ce qui alourdirait inutilement le message à transmettre, ni trop peu, ce qui entraînerait une représentation non conforme du signal.

La deuxième étape consiste à affecter à ces échantillons des valeurs numériques, c'est la quantification.

Cette quantification utilise une numérisation binaire ne comportant que deux symboles 0 et 1 (on parle de bits, abréviation de *binary digits* : chiffres binaires). Cette numérisation binaire est plus simple à manipuler avec fiabilité par les systèmes informatiques que notre numérisation décimale (mieux adaptée à des intelligences supérieures), mais elle génère par contre des messages plus « encombrants ».

C'est ainsi qu'à la suite des nombres décimaux correspond la suite binaire :

Décimaux	0	1	2	3	4	5	6	7	8	9	10
Binaires	0	1	10	11	100	101	110	111	1000	1001	1010

Il faut utiliser 9 bits pour compter entre 256 et 512, dix pour compter entre 512 et 1024, etc.

On peut dire que la notation décimale propose un système de réduction de débit grâce à un codage à dictionnaire (voir chapitre 2.1).

2.2.3 Théorème de Nyquist et bruit de quantification

Le nombre d'échantillons à retenir à chaque seconde dépend évidemment de la vitesse de variation du phénomène étudié. En d'autres termes la fréquence d'échantillonnage F_e dépend de la composante à la plus haute fréquence contenue dans le signal.

On exprime d'une manière simple cette relation par le théorème de Nyquist (également connu sous le nom de théorème de Shannon ; il aurait été initialement exprimé par Nyquist en 1928 et reformulé par Shannon en 1945).

Une fréquence d'échantillonnage F_e ne permet pas de transmettre correctement un signal dont la fréquence est supérieure à $F_e/2$. Cette fréquence limite est appelée fréquence de Nyquist F_N. On peut dire, en d'autres termes, qu'il faut disposer d'au moins deux échantillons par période pour la composante à la plus haute fréquence contenue dans le signal à transmettre.

Si cette condition n'est pas remplie (on parle alors de sous-échantillonnage), on assiste au phénomène de repliement de spectre ou d'*aliasing*. Des échantillons en nombre insuffisant sont interprétés comme provenant d'un signal à fréquence plus basse.

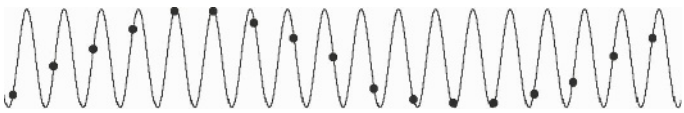


Figure 2.4.
Aliasing.

Si une partie du spectre du signal s'étend au delà de la fréquence de Nyquist, elle sera reproduite comme si elle avait été repliée autour de cette fréquence. Une composante telle que $F_N + f$ sera perçue comme $F_N - f$.

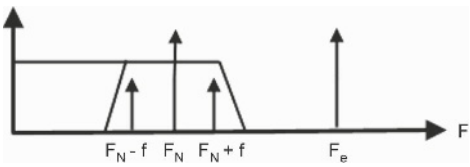


Figure 2.5.
Repliement de spectre.

Une manifestation bien connue de repliement de spectre, dans le domaine temporel, est le phénomène de stroboscopie, qui fait tourner à l'envers les roues des chariots de western. C'est la conséquence de l'échantillonnage à 24 images par seconde que réalise le cinématographe.

Afin d'éviter cette redoutable introduction d'informations erronées (des fausses notes s'il s'agit de musique par exemple) il convient de limiter, par filtrage passe bas avant échantillonnage, le spectre du signal à traiter.

Pour ce qui concerne les signaux audio, la bande passante utile est limitée à 20 kHz afin de s'adapter au mieux à nos possibilités physiologiques ; une fréquence d'échantillonnage supérieure à 40 kHz doit donc être choisie. Par contre on a, par souci d'économie de la bande passante de l'émetteur, limité à 30 kHz la fréquence d'échantillonnage retenue pour les émissions de radio numérique ; la bande passante audio du signal à transmettre a donc été limitée par filtrage à 15 kHz.

La quantification ne doit de son côté retenir que le nombre de bits nécessaires. Cette « profondeur de quantification » nécessaire peut être déterminée en prenant en compte le bruit de quantification. Il s'agit en fait d'une incertitude, d'une imprécision ou d'une erreur de quantification qui perturbe la valeur finale du signal à la manière du bruit de fond.

Le niveau de bruit correspond à la valeur en décibels de $6{,}02\,N + 1{,}76$; N étant le nombre de bits utilisés. Nous avons dit précédemment que le niveau de bruit des caméras se situait aux environs d'une soixantaine de décibels. Une quantification sur 10 bits apparaît donc comme suffisante pour les signaux vidéo puisqu'elle conduit à un niveau de bruit de quantification qui se situe en dessous du bruit initial du signal.

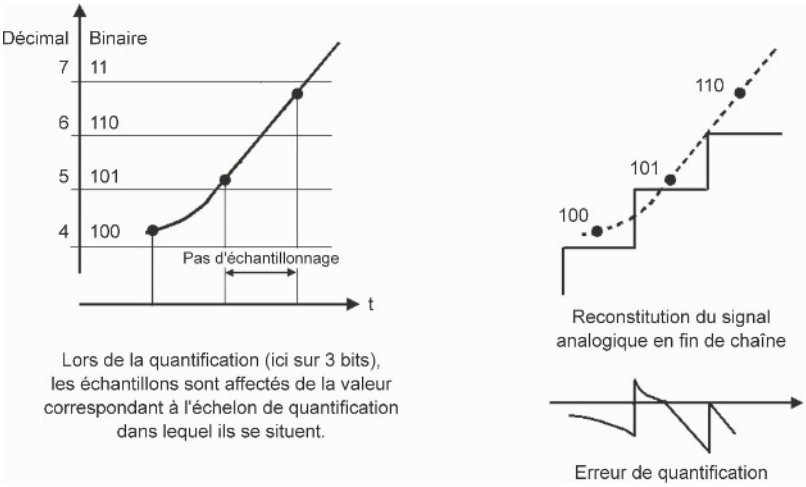


Figure 2.6.
Erreur de quantification.

2.2.4 Avantages et inconvénients du numérique

Le signal numérique est par essence très robuste. On a vu qu'une fois dans le domaine numérique, une fois que l'on travaille sur des nombres, il n'y a plus à craindre d'augmentation du bruit (tout au moins si l'on a pris la précaution de travailler dans un espace numérique susceptible d'utiliser un nombre de bits suffisant pour éviter les erreurs d'arrondi – en anglais *rouding* –, lors des opérations effectuées sur les données).

Les signaux numériques sont par ailleurs codés sous la forme de trains d'impulsions qui peuvent avoir des niveaux élevés ce qui est un gage de pérennité. Leur forme ainsi que leur position temporelle théorique sont connues ; il est donc possible d'effectuer une restauration en cas d'avarie.

On a su en outre mettre en œuvre de puissantes méthodes mathématiques de corrections d'erreurs. L'entrée dans le monde de l'informatique et des méthodes algorithmiques a d'une part donné naissance à de nouveaux modes de traitement (dont ceux des effets spéciaux et ceux de la réduction de débit) et a d'autre part fait largement baisser le coût du matériel grâce à l'utilisation de produits de masse.

Le seul inconvénient du numérique réside dans l'encombrement des messages. Les avancées des puces et des programmes (*hardware* et *software*), notamment les avancées des méthodes de réduction de débit, permettent aujourd'hui de s'affranchir dans une large mesure de cet inconvénient.

2.2.5 Les normes pour les images animées

Après quelques débuts en même temps enthousiasmants et anarchiques on s'est rendu compte qu'il importait de définir des normes assurant la cohérence et l'interopérabilité de l'ensemble de la chaîne de production et de diffusion des images télévisées.

Cette normalisation est intervenue sous l'égide du CCIR (Comité Consultatif International de Radiodiffusion, aujourd'hui ITU-R *International Telecommunications Union-Radiocommunications*). C'est la célèbre recommandation BT 601, complétée par la recommandation BT 656.

Cette norme concerne les systèmes de télévision standard à 576 ou 480 lignes actives (systèmes 50 ou 60 Hz). Elle prend en compte des systèmes « composantes » YUV dans lesquels les informations de chrominance sont divisées par deux.

On trouvera sur chaque ligne, deux fois moins d'échantillons pour ces signaux de chrominance U et V que pour celui de luminance Y.

Ceci est symbolisé dans la désignation 4:2:2. (quatre échantillons de luminance, deux pour U et deux pour V) devenue courante pour la norme.

Il s'agit d'une réduction de débit avec pertes (*lossy*), mais avec pertes peu pénalisantes pour le simple spectateur qui n'a pas à intervenir sur le contenu des images.

La fréquence d'échantillonnage est de 13,5 Mhz pour la luminance et de 6,75 Mhz pour U et V. La fréquence

de Nyquist est donc de 6,75 Mhz pour la luminance ce qui assure la possibilité d'une résolution légèrement supérieure à celle des systèmes analogiques précédents dont la bande passante pour la luminance plafonnait aux environs de 5,5 Mhz.

Ceci conduit à 720 échantillons de luminance pour la partie utile de la ligne (864 pour la durée totale de la ligne) et 360 pour chaque signal de chrominance.

La quantification a été définie (pour Y, U et V) sur 8 bits ou 10 bits ce qui définit 256 ou 1 024 niveaux et conduit à un débit global de 216 ou 270 Mb/s.

Il a été établi par la SMPTE (*Society of Motion Picture and Television Engineers*) une norme d'interface pour la transmission « parallèle » du signal 4:2:2/10 bits (SMPTE 125 M) ainsi qu'une norme pour la transmission « série » (SMPTE 259 M) connue sous le nom de SDI (*Serial Digital Interface*).

On trouve, à côté du 4:2:2 deux modes d'échantillonnage plus économiques en termes de débit.

L'un comme l'autre réalisent cette réduction de débit en réduisant un peu plus le nombre des échantillons de chrominance. Une fois encore il y a déperdition d'informations, mais déperdition maîtrisée.

Le mode 4:2:0 divise par deux la résolution verticale ; un seul des deux signaux U et V est transmis, alternativement, à chaque ligne. C'est un choix qui apparaît assez logique dans la mesure où l'œil n'est ni plus, ni moins, exigeant en vertical qu'en horizontal. Il est retenu pour les balayages à 576 lignes utiles.

Il a été considéré comme conduisant à une résolution verticale insuffisante pour les systèmes à 480 lignes utiles pour lesquels il a été retenu un échantillonnage 4:1:1.

La Haute Définition avec un format d'image élargi (16/9) a été prise en compte en 2000 par la norme ITU-R BT 709/SMPTE 240 M. Celle ci validait les deux sys-

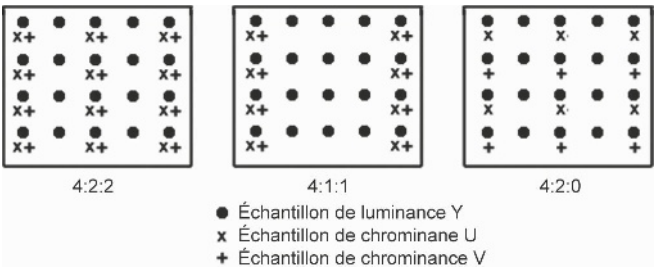


Figure 2.7. Modes d'échantillonnage 4:2:2 ; 4:1:1 et 4:2:0.

tèmes historiques concurrents, et aujourd'hui abandonnés, à 1035 et 1152 lignes actives en balayage entrelacé à 60 et 50 Hz.

La norme SMPTE 274 M a pris en compte, peu après, le standard, devenu commun (CIF, *Common Image Format*), à 1080 lignes actives et 1920 pixels par ligne en balayage entrelacé à 50 ou 60 Hz. (fréquences d'échantillonnage 74,25 et 37,125 MHz) ainsi que des modes en balayage progressif à 50 et 60 Hz (objectif encore un peu éloigné, mais de moins en moins) et surtout un balayage progressif à 24 images par seconde, ce qui est comme on le sait la fréquence des images cinématographiques.

Ce mode 24p (dit également SF pour *Segmented Frame* ou image segmentée) présente des avantages évidents pour la production cinématographique (passage plus facile de l'univers numérique à l'univers argentique sans risque d'artefacts temporels) ; il est par ailleurs totalement compatible avec l'ensemble du matériel HD : l'image est divisée en deux segments qui se comportent comme les deux trames de l'entrelacé mais qui auraient été saisies au même instant.

En 2001 la norme SMPTE 296 M a retenu un format $1\,280 \times 720$ en mode progressif.

Les deux formats $1\,920 \times 1\,080$ i et $1\,280 \times 720$ p, chacun avec ses avantages et ses supporters, cohabitent aux USA où la diffusion HD est depuis quelques années une vraie réalité ; les grands réseaux se sont convertis à peu près également à l'un ou l'autre de ces deux formats.

La SMPTE a produit la normalisation d'interfaces HD-SDI (SMPTE 292 M) au débit de 1,485 Gb/s, pour des utilisations classiques ; elle a publié ensuite la norme « Dual Link » (SMPTE 372 M) au débit de 2,97 Gb/s qui vise des applications haut de gamme en production, telles qu'un échantillonnage RVB 4:4:4, une quantification sur 12 bits ou un format 1 920 × 1 080 en mode progressif… Un futur de rêve !

2.3 LES CODAGES AUDIO

2.3.1 Du téléphone au Disque Compact, la modulation PCM

Les signaux audio ont été, grâce notamment aux progrès de la téléphonie, les premiers signaux (si l'on met à part les domaines scientifique et technique) à bénéficier de la numérisation. On les a alors qualifiés en français de signaux MIC (Modulation par Impulsions Codées) ou PCM en anglais (*Pulse Code Modulation*).

La première application sur le marché grand public fut lancée en 1982 par Philips et Sony avec le Disque Compact Audio ou CD-DA (*Compact Disk Digital Audio*).

Le signal est numérisé avec une fréquence d'échantillonnage de 44,1 kHz pour une quantification de 16 bits. Le disque de 12 cm de diamètre supporte un signal stéréophonique. Des codes correcteurs d'erreur puissants, qui permettent « d'oublier » dans une large mesure empreintes de doigts et rayures sur la surface du disque, sont ajoutés aux données audio. Le débit total atteint 1,5 Mb/s. La capacité du disque atteint de 680 à 700 Megaoctets pour une durée d'enregistrement de 74 à 80 minutes. L'histoire dit que cette durée a été choisie par l'épouse d'Akio Morita, fondateur de Sony, afin de pouvoir enregistrer la plus longue pièce de musique classique, la Neuvième Symphonie de Beethoven.

2.3.2 L'interface AES/EBU

L'AES (*Audio Engineering Society*) et l'EBU (*European Broadcasting Union*) ont établi une norme audio numérique de base connue sous le terme d'interface AES/EBU qui assure l'interopérabilité entre les différents appareils audio-numériques professionnels.

L'interface série AES/EBU définit deux canaux d'information audio, gauche et droit, transmis sur une paire de fils torsadés (il existe également des versions sur câble coaxial et sur fibre optique) Les signaux peuvent être numérisés jusqu'à 24 bits (quantification linéaire), à plusieurs fréquences d'échantillonnage : 32 kHz, 44,1 kHz ou 48 kHz.

Comme dans tout échantillonnage, la règle de Nyquist doit évidemment être respectée. Par conséquent, un échantillonnage à 32 kHz, retenu pour la radio numérique, limite la bande passante du signal audio à 15 kHz. La fréquence d'échantillonnage, bizarre, de 44,1 kHz a été choisie afin de permettre une possibilité d'enregistrement segmenté en trames sur des magnétoscopes institutionnels voire grand public (U-Matic et Betamax). Elle assure au plus juste la transmission de la bande passante de 20 kHz qui est celle de l'auditeur moyen, jeune et en bon état.

Les premiers enregistreurs numériques de studio utilisaient plutôt une fréquence d'échantillonnage de 48 kHz. Il existe bien entendu nombre d'appareils utilisant cette interface pour véhiculer des signaux moins performants.

Afin de s'adapter aux avancées techniques, la norme AES a été régulièrement amendée et peut véhiculer un signal comportant 24 bits à une fréquence d'échantillonnage de 192 kHz.

2.3.3 DPCM et ADPCM

On a ensuite cherché à raffiner la technique afin de transmettre moins de bits.

La modulation DPCM (*Differential Pulse Code Modulation*) code les différences entre les valeurs successives en PCM. Ces valeurs sont en général faibles et conduisent donc à un petit nombre de bits.

On estime que la réduction de débit audio peut atteindre 25 % par rapport à un signal PCM. Il est à noter que ce mode de réduction de débit n'entraîne aucune perte d'information.

L'ADPCM (*Adaptive* DPCM) est plus complexe. On utilise un algorithme de prédiction qui prend en compte « l'histoire du signal » en étudiant la variation de plusieurs échantillons antérieurs. Il est suivi d'un comparateur entre signal réel et signal prédit qui détermine une valeur d'erreur. C'est cette valeur d'erreur qui est transmise. Lors du décodage elle est ajoutée à la valeur prédite élaborée par l'algorithme de prédiction du décodeur. La valeur de l'erreur, généralement faible, détermine le pas de quantification qui lui est appliquée (souvent 3 ou 4 bits seulement).

L'ADPCM réduit le débit dans une forte proportion avec un taux supérieur à 4:1 dans le cas d'échantillons codés sur 16 bits et d'une différence transmises sur 4 bits. Cependant, elle peut générer des pertes d'information.

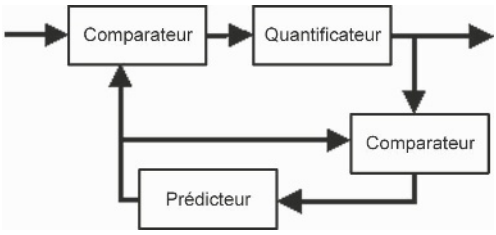


Figure 2.8.
Synoptique
d'un codeur ADPCM.

2.3.4 Les codages MPEG-audio

Compte tenu du relativement faible débit des signaux audio en mode PCM, par rapport à celui des signaux vidéo, compte tenu également des exigences de la perception acoustique, les méthodes de compression des signaux audio sont déconseillées en production.

Échantillonnage (kHz)	Quantification (bits)	1 canal (kb/s)	2 canaux (kb/s)	5 canaux (kb/s)
44,1	16	706	1 411	3 528
48	16	768	1 536	3 840
48	24	1 152	2 304	5 760

Tableau 2.1. Débits audio sans compression.

L'avantage du gain en débit est largement contrebalancé par les risques de dégradation du signal ainsi que par la limitation des possibilités de manipulation ultérieure de ce signal.

Les méthodes de codage MPEG sont donc spécifiquement destinées à la diffusion sur les ondes ou sur support numérique. Elles exploitent, pour l'essentiel, les limites psycho-acoustiques humaines ; elles éliminent ce qui ne peut être entendu.

On évitera de transmettre tout ce qui est en dessous du seuil de perception en tenant compte notamment des phénomènes de masquage.

Afin de bénéficier au mieux de ce phénomène, on découpe le spectre audio en plusieurs bandes de fréquences. C'est ce qu'on nomme codage en sous-bandes.

La norme MPEG-1 (1993) concerne un ou deux canaux audio à des débits allant de 32 kb/s à 912 kb/s avec une valeur moyenne de 384 kb/s.

Il existe quatre modes possibles : mono, stéréo, dual mono (paramètres gauche et droite indépendants) et « joint stéréo » (codage d'une voie et de la différence entre les deux voies).

Ce codage est basé sur la technologie Musicam (*Masking Pattern Adapted Universal Subband Integrated Coding and Multiplexing*), développée au CCETT (Centre Commun d'Études de Télévision et de Télécommunications), à Rennes et par l'IRT (*Institut für Rundfunk-Technik*) à Munich.

Le signal audio est divisé en 32 sous-bandes d'égale largeur en fréquence (environ 1 kHz pour les débits moyens). Les sous-bandes sont élaborées grâce à une batterie de filtres de bande.

Pour chaque bande, l'encodeur analyse par Transformée de Fourier Rapide (FFT, *Fast Fourier Transform*) le spectre du signal et évalue le niveau de bruit de ce signal par l'intermédiaire d'un modèle psycho-acoustique. L'encodeur détermine le nombre minimal de bits nécessaire dans chaque bande afin que les erreurs de quantification (ou bruit de quantification) demeurent en dessous de la courbe de masquage, donc qu'elles n'apportent pas de perturbation perceptible. L'information concernant la quantification mise en œuvre dans chaque sous-bande est transmise avec les échantillons de la sous-bande codée.

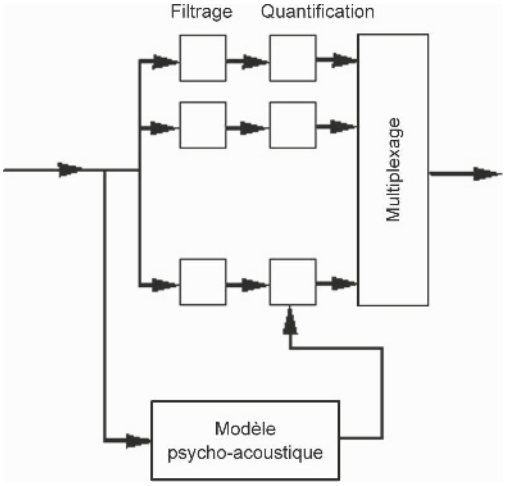


Figure 2.9.
Codage en sous-bandes.

MPEG-audio, finalisé dès l'introduction de MPEG-1 en 1992, définit trois modes optionnels ou couches (*layers*) connus sous les dénominations Layer I, Layer II, et Layer III. Les taux de compression vont en croissant de couche en couche, mais il en va de même pour la complexité de l'encodage.

La couche I, connue sous le nom de « MUSICAM simplifié », est le mode le plus simple qui convient à une utilisation domestique. Le débit peut varier de 32 kb/s à 448 kb/s pour des fréquences d'échantillonnage classiques : 32 ; 44,1 et 48 kHz. Une haute qualité audio (analogue à celle d'un CD) implique un débit compris entre 256 et 384 kb/s par programme stéréo ; la valeur idéale se situant aux alentours de 192 kb/s par canal.

La couche II, connue sous le nom de MUSICAM, permet un taux de compression plus élevé que le mode précédent ; cette couche est utilisée dans différentes applications aussi bien domestiques que professionnelles (émission de radio, télécommunications, multimédia…) Le débit varie de 32 à 192 kb/s pour un signal mono, et 64 à 384 kb/s pour la stéréo. Une haute qualité audio implique un débit entre 192 et 256 kb/s par programme stéréo.

La couche III, connue sous le nom de MP3, ajoute une quantification non-uniforme et un codage de Huffman. Elle étend encore les possibilités de compression avec des débits pouvant aller de 320 kb/s jusqu'à des valeurs aussi basses que 8 kb/s et des fréquences d'échantillonnage pouvant descendre jusqu'à 24 et 16 kHz. Chacune des 32 bandes de fréquence est elle même découpée en 18 « tranches », soit au total 576 bandes d'une trentaine de Hertz.

Elle est utilisée pour les applications de télécommunications à bande étroite certaines applications professionnelles, mais surtout, pour des applications grand public. Dans ce cas on estime qu'une bonne qualité peut être obtenue aux environs de 128 kb/s en stéréo.

On peut atteindre, tout en conservant une bonne qualité d'écoute (qualité CD), une réduction du débit des données de :

- 1:4 pour la couche 1,
- 1:6 à 1:8 pour la couche 2,
- 1:10 à 1:12 pour la couche 3,

En 1996, est apparue une extension des normes précédentes, sous le nom de MPEG-2 BC (*Backward Compatible*) qui incluait des fréquences d'échantillonnage plus basses pour des débits plus faibles et qui surtout permettait de traiter des signaux audio 5+1 ; ce codage restait compatible avec les précédents grâce à la possibilité de restituer sur 2 canaux stéréo l'ensemble des données du flux multicanal MPEG-2. Ce codage avait été prévu pour la restitution du son multicanal sur les DVD.

Finalisé en 1997 grâce, en particulier, aux travaux du Fraunhofer Institut d'Erlangen, le standard international ISO/IEC 13818-7 ou MPEG-2 *Advanced Audio Coding* (AAC) a été publié en 1998. C'est une des retombées du chantier de travail sur MPEG-4 dont il constitue un élément important.

Ce standard étend la gamme des fréquences d'échantillonnage utilisables de 8 kHz à 96 kHz ainsi que le nombres des bandes de fréquences d'analyse spectrale qui passe à 1024. Le nombre de canaux disponibles passe quant à lui à 48 (le traitement du son multicanal a été un élément déterminant du projet). On trouve trois niveaux : *Main*, LC (*Low Complexity*) et SSR (*Scaleable Sampling Rate*). Le débit à qualité subjective équivalente, est la moitié de celui du codage Musicam. L'AAC *Main* à 96 kb/s serait supérieur au MP3 à 128 kbps.

Les tests de L'ITU (*International Telecommunications Union*) ont montré que la qualité d'un signal stéréo est à 128 kb/s indiscernable de celle de l'original.

Ce codage n'est pas compatible avec les codages MPEG précédents, c'est pourquoi il est parfois nommé MPEG-2NBC (*Non Backward Compatible*).

Le format MP3 a récemment été étendu (2004) par le Fraunhofer Institut, sous l'appellation MP3SX (*Stereo eXtended*) ou MP3 *Surround*, au traitement du son multicanal 5+1, tout en conservant la compatibilité stéréophonique avec le MP3 classique (les 6 canaux sont transcrits sur deux canaux stéréo) dont il conserve les débits. Ce mode de codage « *Spatial Audio Coding* » est, bien entendu, le résultat des travaux sur les codages MPEG-4.

2.3.5 Le codage AC-3

L'AC-3 (*Audio Coding* 3e génération, nom commercial Dolby Digital) est un codage muticanal (6 canaux distincts, mode connu sous l'appellation 5.1) qui met en œuvre des méthodes analogues à celles utilisées par MPEG, mais des méthodes « propriétaire ». Il a été mis au point (après Dolby Surround à 4 canaux) par les laboratoires Dolby aux USA où il a été choisi en 1995 pour les systèmes de télévision numérique ATSC (*Advanced Television Systems Commitee*). Il a réussi à s'implanter également dans le reste du monde, notamment en Europe, par le biais du DVD dont il constitue, depuis 1977, un mode obligatoire d'enregistrement audio en parallèle avec MPEG Layer 2.

Une version améliorée dite E-AC3 (*Extended* AC-3, nom commercial Dolby Digital Plus) étend le nombre de canaux à 7+1.

2.3.6 Le système DTS (*Digital Theater System*)

Très largement utilisé dans les appareils grand-public (consoles de jeu, DVD…), mais également en diffusion cinématographique (d'où son nom), le système

DTS permet un enregistrement 5.1, comme le Dolby Digital mais avec un taux de compression plus faible.

Il est très fortement inspiré (pour ne pas dire qu'il en a détourné les brevets) du format LC Concept, conçu à la fin des années 80 par des Français (L pour Élisabeth Lochen et C pour Pascal Chedeville).

Le fonctionnement du LC Concept repose sur la synchronisation entre un disque de type CD et la pellicule d'un film en projection. Cette synchronisation est assurée par un code temporel imprimé sur le bord de la pellicule à côté de la piste sonore optique. Ce procédé qui permettait une amélioration très sensible de la restitution sonore maintenait la compatibilité avec les modes de diffusion analogiques classiques puisque les pistes sonores n'étaient pas modifiées.

2.4 LES CODAGES JPEG

2.4.1 Indispensables

La quantité d'information nécessaire pour décrire en mode basique, pixel après pixel, une image est submergeante, aussi bien pour les systèmes d'enregistrement que pour les canaux de transmission. Prenons le cas d'une image de télévision en 4:2:2, assez médiocre pourtant : on compte 576 lignes de 720 pixels, soit 414 720 pixels ; une quantification satisfaisante compte 10 bits par pixel et par couleur soit un total de près de 12,5 Mbits en RVB, ou près de 8,3 Mbits en YUV.

Des images bien plus complexes, photographies de presse, images médicales ou scientifiques… sont évidemment encore plus exigeantes.

Il est vrai que la capacité des systèmes d'enregistrement numérique ne cesse de croître, mais dans le même temps l'appétit d'images stockées ou transmises est également en pleine expansion.

Réduire la quantité d'information sans (trop) détériorer la qualité est, par conséquent, indispensable.

On a donc cherché à réaliser une description plus intelligente de l'image tenant compte de ses redondances spatiales. On parle alors de codage intra-image par opposition au codage inter-images adapté à des séquences animées, qui sera abordé un peu plus loin. On désigne en abrégé ces deux types de codage par les termes « intra » et « inter ».

Un groupe de travail s'est réuni dès 1986 à l'initiative conjointe de deux organismes de normalisation : ISO (*International Standards Organization*) et IEC (*International Electrotechnical Commission*).

Ce groupe a proposé en 1991 un standard de réduction de débit pour « les images fixes en tons continus », ISO/IEC 10918 plus connu sous le nom de JPEG (*Joint Photographic Experts Group*). Ce standard a été validé en septembre 1992 en tant que standard international (*Digital Compression and Coding of Continuous-tone Still Images*) par le CCITT (Comité Consultatif International pour le Télégraphe et le Téléphone) de l'ITU (*International Telecommunications Union*) selon la recommandation T 81.

2.4.2 L'analyse harmonique ou spectrale et la DCT

La première étape a consisté à mettre au point une méthode permettant de décrire l'image suivant un catalogue ou un dictionnaire de formes ou de motifs prédéfinis (voir 2.1 : Principes de base de la réduction de débit) dont on peut repérer la présence dans l'image grâce à des algorithmes mathématiques relativement simples et qui permettent de regrouper et classer les informations selon les fréquences spatiales horizontale (plus ou moins de barreaux dans une grille) et verticale (plus ou moins de marches dans un escalier).

La DCT utilise une opération mathématique appelée Transformation de Fourier, en français, et *Discrete Cosine Transform*, en anglo-saxon.

Il s'agit d'analyser le signal $f(t)$ à l'aide d'une famille de fonctions mathématiques, dites fonctions de base, assez simples et telles qu'une série d'entre elles puisse représenter le signal considéré :

$$f(t) = \sum c_i \, \Psi_i(t)$$

Les fonctions de base Ψ_i étant fixées, l'information relative au signal est portée par les coefficients c_i représentant le poids de chacun des termes de la série.

Les bonnes fonctions mathématiques sinusoïdales, sinus et cosinus (c'est la même chose à un décalage près dans le temps), assez simples d'emploi et bien adaptées à la description des phénomènes physiques périodiques, représentaient un bon candidat pour ce type de calcul.

Ceci d'autant que le physicien Joseph Fourier a élaboré en 1822 une méthode permettant de représenter un signal périodique quelconque par un ensemble de fonctions sinusoïdales.

C'est l'analyse harmonique, déjà utilisée expérimentalement par Helmholtz pour l'analyse des sons musicaux grâce à une batterie de résonateurs mécaniques.

La « décomposition en série de Fourier » conduit à représenter la fonction $f(t)$ de période :

$$T = 2\,\pi/\omega$$

par l'expression :

$$f(t) = a_0 + \sum (a_k \sin k\,\omega\, t + b_k \cos k\,\omega\, t)$$

Elle permet d'obtenir la représentation d'un phénomène périodique par une superposition de composantes sinusoïdales dont les fréquences, ainsi que les amplitudes et les phases sont bien spécifiées.

Pour $k = 1$, on obtient la fondamentale de pulsation ω dont la période $T = 2\,\pi/\omega$ est identique à celle de la fonction $f(t)$; pour $k = 2, 3\ldots$, on obtient les harmoniques dont les fréquences sont des multiples de celle

de la fondamentale à laquelle ils viennent ajouter les détails qui affinent la coïncidence avec la fonction à représenter. Plus on utilisera d'harmoniques meilleure sera la représentation.

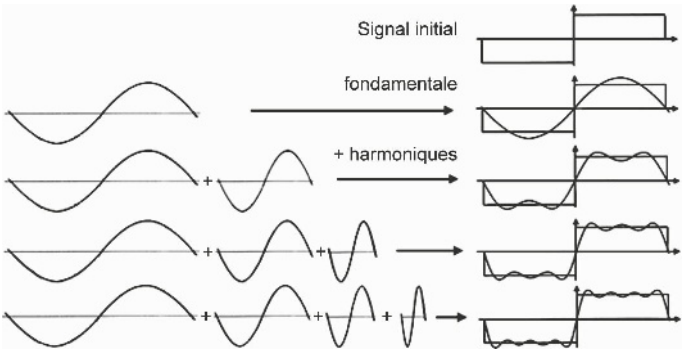


Figure 2.10. Représentation d'une fonction par une série de Fourier.

On a su étendre, sous le nom de Transformation de Fourier, cette méthode à des signaux non périodiques (ou non stationnaires) en supposant que leur période est infinie, ce qui conduit à un spectre de composantes pouvant comporter… une infinité de composantes. Il s'agit d'une représentation dont l'existence n'est pas localisée dans le temps (sons) ou dans l'espace (image). C'est le jeu des interférences (additives ou destructrices) entre les différentes composantes qui fait apparaître ou disparaître à un instant (ou un lieu) donné une composante (une note de musique ou un motif graphique).

Les informaticiens ont su mettre au point des algorithmes relativement simples (FFT : Fast Fourier Transform) pour réaliser ces opérations sur ordinateur. C'est ce type de transformée qui est notamment utilisé dans la compression JPEG.

Dans le cas des images, la compression JPEG met en œuvre un algorithme dénommé en français Transformation Cosinus Discrète (TCD) ou en anglais *Discrete Cosine Transform* (DCT).

Le premier terme indique que l'on va modifier la manière de caractériser le signal ; le deuxième terme indique que l'on va utiliser des fonctions mathématiques périodiques sinusoïdales (ici, uniquement des cosinus) qui sont bien connues et pleines de qualités ; le troisième terme réfère au caractère discontinu ou discret de l'information à traiter ; il indique que le processus portera sur des signaux numérisés. Quant aux fréquences mises en jeu il s'agit bien sûr de fréquences spatiales dans le cas es images.

Les fonctions cosinus utilisées pour « décrire » l'image n'ont, si j'ose dire, « ni queue ni tête », elles représentent des ondes « angéliques » dont l'intervalle ou l'espace d'existence s'étend d'un infini à l'autre.

Le traitement de l'ensemble d'une image exigerait la manipulation d'énormes quantités de données. On a donc été amené à découper en morceaux ces opérations.

On analyse successivement des petits morceaux d'image ou blocs de 8×8 pixels, c'est-à-dire qu'on fait subir au signal à transformer un fenêtrage préalable (STFT, *Short Time Fourier Transform*, Transformation de Fourier à Court Terme ou transformation de Gabor). L'analyse en fréquences spatiales s'effectue successivement suivant l'axe vertical puis suivant l'axe horizontal.

Il faut bien entendu effectuer ces opérations pour les différentes composantes définissant l'image, généralement en couleur : soit les trois composantes R,V,B, soit les composantes Y,U,V.

2.4.3 Le catalogue des tartans

Les différentes fréquences correspondent à des motifs comportant des bandes (verticales ou horizontales) régulièrement disposées, ce qui donne naissance à un catalogue des 64 motifs géométriques simples que l'on

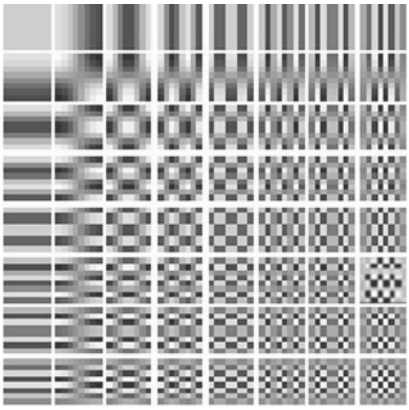


Figure 2.11.
Catalogue
de tartans écossais.

peut réaliser avec les 64 pixels d'un bloc 8×8. Il me fit penser à un catalogue de tartans écossais, d'où le nom que je me plais à lui attribuer.

La DCT permet de savoir si chacun de ces motifs peut être décelé dans le bloc analysé. On substitue alors au tableau constitué par les 64 nombres affectés aux points du bloc (domaine spatial) un autre tableau comportant les coefficients d'amplitude des composantes fréquentielles permettant de caractériser la contribution dans l'image de chacun des motifs dont la présence a été détectée (domaine fréquentiel).

On peut dire d'une manière simpliste que l'image initiale du bloc qui aurait été réalisée par un peintre, point par point, peut être reconstruite par un spécialiste du dessin animé qui empilerait quelques celluloïds portant, de manière plus ou moins affirmée, les différents motifs détectés.

Le tableau est tel que les fréquences horizontales croissent de la gauche vers la droite et les fréquences verticales de haut en bas.

Lorsqu'on s'éloignera en diagonale en partant du coin haut à gauche pour aller vers le coin bas à droite on rencontrera des motifs de plus en plus fins correspondant à des détails de plus en plus ténus ; le motif haut gauche représente la valeur moyenne du bloc.

Tableau 2.2. Valeurs des échantillons de luminance (domaine spatial) (d'après document Tektronix).

123	138	145	156	137	121	139	120
134	145	158	136	129	140	151	154
156	167	153	132	110	121	130	134
141	112	154	148	141	137	128	128
141	140	139	137	135	132	134	138
138	137	136	133	131	133	138	140
141	144	141	136	133	138	140	145
144	138	133	137	138	141	144	147

Tableau 2.3. Coefficients DCT (domaine fréquentiel) (d'après document Tektronix).

86	-23	13	7	5	6	-1	0
121	15	-14	10	3	7	1	0
89	-4	-20	6	5	-1	1	-1
65	-15	8	1	-1	2	0	0
18	-10	-1	-1	-1	2	0	1
9	-4	-3	-2	1	-1	0	0
-4	2	-4	1	-3	2	1	0
-7	1	0	0	-1	1	1	0

On constate, pour la plupart des images naturelles, que les coefficients de la partie du tableau située en dessous et à droite de la diagonale ont presque toujours des valeurs soit très faibles soit nulles. On con-

çoit donc déjà qu'il est plus économique, en termes de flux numérique, de transmettre la fiche d'identité des quelques fréquences présentes au dessus de la diagonale (les fréquences basses qui correspondent à des blocs peu riches en détails fins) plutôt que celles des 64 pixels individuels du bloc. On obtient ainsi une transformation théoriquement transparente, c'est-à-dire que toute l'information présente dans le bloc est conservée ; tout au moins si la fenêtre a été bien choisie, c'est-à-dire avec une fonction de pondération convenable corrigeant les effets de troncature qui se manifestent par des phénomènes de Gibbs (oscillations au niveau des contours) et si l'on opère sur un nombre de bits suffisants permettant d'obtenir lors des calculs des erreurs d'arrondi (*rounding*) négligeables. Les applications JPEG traitent des fichiers 8 ou 10 bits par couleur, quantification considérée comme un peu juste pour assurer une compression véritablement sans pertes.

2.4.4 Avec ou sans pertes (*lossy* ou *lossless*)

On peut, au prix de quelques sacrifices, augmenter le taux de compression en réalisant un compression virtuellement transparente (c'est-à-dire que le spectateur moyen n'y verra que du feu).

Ceci est obtenu en négligeant tout d'abord les coefficients inférieurs à un certain seuil puis en affectant à chaque composante fréquentielle conservée un poids tenant compte de l'importance que représente pour la psychophysiologie de la perception la fréquence correspondante ; les coefficients les moins importants (les hautes fréquences c'est-à-dire les détails peu perceptibles) sont codés avec une faible précision sur peu de bits, leur valeur initiale étant arrondie à la valeur la plus proche d'un des larges échelons de quantification ainsi créés.

Un taux de compression important et peu pénalisant pour le spectateur peut être obtenu par un choix adapté

Tableau 2.4.
Coefficients après requantification (d'après document Tektronix).

87	-2	3	1	0	0	0	0
8	3	-1	0	0	0	0	0
4	-1	-1	1	0	0	0	0
2	-1	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

des valeurs des seuils et des coefficients de pondération permettant de n'éliminer que des informations peu pertinentes (*irrelevant* en anglais) pour l'œil.

2.4.5 Les couches finales (lecture en zig-zag, RLC, codage entropique)

Le tableau est ensuite lu en zig-zag en commençant par le haut à gauche (les basses fréquences).

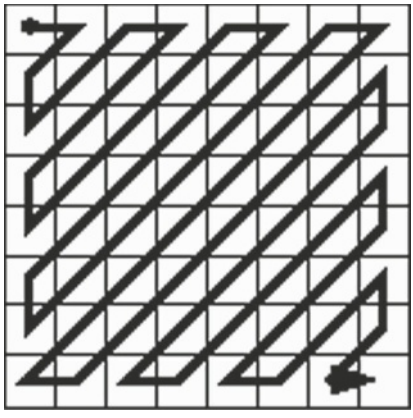


Figure 2.12.
Lecture en zig-zag.

Cette analyse donne la suite des valeurs ci-dessous :

87 -2 8 4 3 3 1 -1 -1 2 0 -1 -1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Ce mode d'exploration conduit à rencontrer de longues séquences de zéros qui seront bien évidemment codées économiquement sous une forme condensée (longueur de la séquence) par un « codage à longueur variable » (RLC : *Run Length Coding*).

$$87 ; -2 ; 8 ; 4 ; (2 \times 3) ; 1 ; (2 \times -1) ; 2 ; 0 ; (2 \times -1) ; (4 \times 0) ; 1 ; (53 \times 0)$$

On peut voir sur cet exemple que l'ensemble du système peut être extrêmement efficace.

On effectue ensuite une analyse statistique de l'occurrence des différents coefficients pour effectuer un « codage entropique ».

Les coefficients intervenant le plus souvent sont codés, par l'intermédiaire d'une table d'indexation ou table de correspondances (LUT, *Look Up Table*), en utilisant peu de bits (mots courts) tandis que les mots longs sont affectés à ceux dont la probabilité d'occurrence est faible.

Ces trois dernières étapes, purement informatiques, n'entraînent aucune perte d'information. C'est exclusivement au niveau du seuillage et de la pondération qu'on peut avoir une « déperdition » d'information. On peut donc, si l'on se contente d'un faible taux, assurer compression et décompression pratiquement sans pertes.

Voici le synoptique d'un codeur. Le décodeur est symétrique.

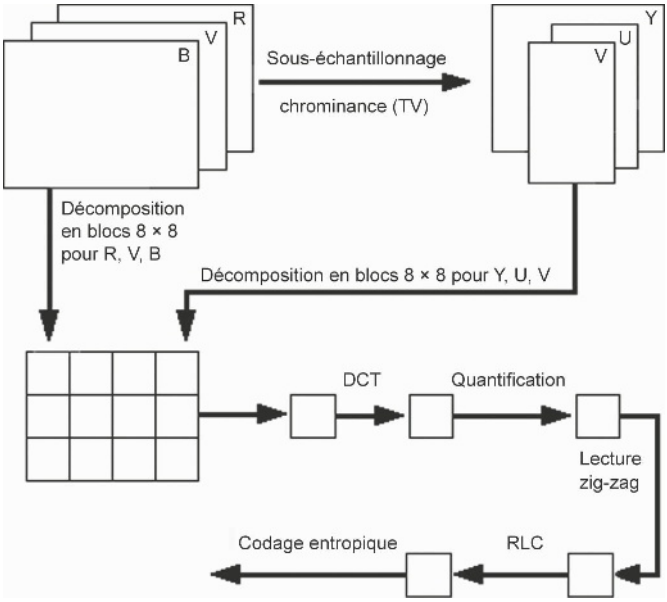


Figure 2.13.
Synoptique d'un codage JPEG.

2.4.6 Performances, le prix à payer

Le traitement par DCT fonctionne bien. L'expérience montre qu'un facteur de compression de 10 produit généralement une image indiscernable de l'original.

Les résultats dépendent évidemment de la complexité de l'image ; des images fortement bruitées (le bruit est un phénomène complètement incorrellé qui ne présente pas de redondances) seront notamment difficiles à traiter.

On peut par ailleurs noter que le système ne donne *a priori* totale satisfaction que pour les structures dont les dimensions sont de l'ordre de celle de la fenêtre, de largeur constante (ici, la fenêtre de 8 pixels semble avoir été judicieusement choisie), et qui peuvent être représentées correctement par des fréquences donnant des longueurs d'onde du même ordre que cette largeur.

Toute erreur sur le premier motif « en haut à gauche » qui représente la valeur moyenne du bloc entraîne un effet de mosaïque ou effet de bloc (*blockiness*) ; et c'est bien ce qui se produit dès lors que le taux de compres-

sion devient trop élevé. Cette structure régulière des artefacts peut évidemment devenir gênante voire désastreuse.

Les algorithmes JPEG ont prouvé leur efficacité ainsi que leur souplesse (notamment dans le choix du taux de réduction) aussi bien dans les applications professionnelles (bibliothèques numérisées, stockage d'images médicales, traitement d'images…) que pour des utilisations « grand public » : photographie numérique, transmission de documents par Internet…

JPEG est devenu un standard incontournable dans ces domaines de l'archivage et de la transmission, mais ce n'est pas un standard de production..

Il faut noter, par ailleurs, que l'intitulé du standard précise « tons continus » ; il existe des méthodes plus efficaces pour les documents purement graphiques ne comportant pas (ou peu) de nuances.

2.4.7 JPEG-LS et JBIG

Le standard JPEG-LS (LS pour *LosslesS*), ISO 14495-1 /ITU-T.87, a été proposé en 1999 comme permettant une compression sans pertes ou presque sans pertes (*lossless / near-lossless*). Il est basé sur des algorithmes LOCO-I (*LOw COmplexity LOssless COmpression for Images*) développés par Hewlett-Packard.

Dans ce type d'algorithmes, il s'agit de définir si le pixel à coder fait partie d'une zone de l'image uniforme où s'il est proche d'une transition ; cette évaluation se fait en comparant la valeur x du pixel à celles de 4 pixels précédents.

Si les différences d-b, b-c, c-a sont nulles (sans pertes) ou inférieures à un seuil prédéfini (presque sans pertes) le pixel est réputé faire partie d'une zone homogène ; dans le cas contraire il est proche d'une transition.

On élabore ensuite une valeur prédite pour le pixel qui est la moyenne de a, b, c, d et on la compare à la valeur

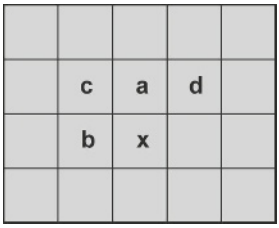


Figure 2.14.
JPEG-LS.

vraie pour élaborer l'erreur de prédiction. Si la zone est homogène, la valeur du pixel est transmise en mode RLC ; dans le cas contraire on transmet la valeur de l'erreur de prédiction.

Un codage entropique est enfin appliqué aux données ainsi élaborées avant qu'elles ne soient transmises.

Le codage JPEG-LS, qui avait été proposé pour répondre notamment à certaines exigences dans les domaines de l'imagerie médicale et scientifique, n'a pas connu un développement notable.

Le standard JBIG (*Joint Bi-level Image experts Group*), ISO 11544 /ITU-T82, a été créé, comme le fax, pour le codage sans pertes d'images à deux niveaux qui peuvent être définies par un seul bit par pixel. Il peut cependant être étendu à des images définies sur un nombre limité de bits en utilisant autant de « plans » que de bits codant cette image.

Chaque pixel est codé, comme pour JPEG-LS, par rapport à un motif (*template*) constitué par les 8 pixels adjacents. Bien que fort efficace ce standard n'a pas connu un succès foudroyant. Une des raisons pourrait être qu'il repose sur des brevets d'IBM, Mitsubishi et Lucent.

L'arrivée de JPEG 2000, qui sera étudié dans le chapitre 6.3, n'a sans doute pas été non plus sans influence sur ce manque de réussite.

3.1 VERS L'IMAGE ANIMÉE : M-JPEG, H.261

3.1.1 Les dialectes M-JPEG

Dès les années 1970 le montage non linéaire (NLE, *Non Linear Editing*), celui qui supprime le temps de latence consacré à la recherche des séquences linéairement sur une bande, était à l'ordre du jour. Les noms de CMX-600, Ediflex ou Montage évoqueront pour certains des systèmes audacieux et, pour tout dire, inutilisables.

En 1985, Quantel proposait Harry, une superbe machine, mais onéreuse et encombrante et qui ne supportait que des séquences de 80 secondes !

Les méthodes de réduction de débit étaient encore dans les langes.

Quelques années plus tard, fin 1988, et quasi simultanément, deux constructeurs Editing Machines Corp. (EMC) et AVID Technology introduisirent les premiers systèmes NLE réellement opérationnels. Il s'agissait de EMC2, sur plate-forme IBM PC et de Media Composer sur plate-forme Apple Macintosh II.

L'idée de base était de considérer une séquence d'images animées comme une suite d'images fixes auxquelles on pourrait appliquer des modes de réduction de débit de type JPEG (alors en voie de normalisation).

Cette méthode se développa sous l'appellation M-JPEG (*Motion* JPEG).

Elle souffrait de plusieurs inconvénients : la compression JPEG effectuée par un ensemble de logiciels et de cartes de codage et décodage était lente (pas de temps réel) et peu efficace, la capacité des disques était trop faible ce qui ne permettait qu'une résolution de type VHS (une référence à cette époque) en ne prenant en compte qu'une trame sur deux.

Il s'agissait de montages « off-line » qui devaient être suivis (en utilisant une liste informatisée des points de montage ou EDL, *Edit Decision List*), d'une conformation sur des machines de qualité « broadcast » mais qui allaient révolutionner les habitudes de post-production dans les univers du film et de la télévision.

En 1990, C-Cube Microsystems proposa un circuit intégré VLSI (référence CL550) destiné à la compression d'images en mode JPEG. Ce processeur spécialisé visait des application comme les photocopieuses et les scanners mais également l'enregistrement et le traitement d'images animées en mode M-JPEG avec une qualité toujours de type VHS mais cette fois en temps réel.

Le problème essentiel n'était cependant pas résolu. M-JPEG n'est pas un standard établi. C'est un ensemble d'algorithmes que chaque constructeur assemble à sa manière. Les différents systèmes mis en œuvre étaient très proches mais rarement interopérables. On a pu parler de l'existence de plusieurs dialectes, ce qui résume bien la situation.

Il faut ajouter que ces systèmes ne bénéficient évidemment pas de l'importante possibilité de réduction de débit qu'apporte la redondance temporelle entre les différentes images des séquences animées.

3.1.2 La visioconférence ouvre la voie

Le standard H.261, développé à partir de 1988 et publié en 1990, est la partie consacrée à l'image animée du protocole ITU-R H.320 relatif à la visioconférence sur réseau téléphonique numérique RNIS (Réseau Numérique à Intégration de Services). Il exploite des débits compris entre 64 kbit/s et 2 Mbit/s, par tranches de 64 kb/s (débit d'un canal RNIS).

Il s'agit du premier standard de réduction de débit prenant en compte les spécificités des images animées. Tout était à tester ou à inventer, depuis les méthodes

efficaces de travail coopératif entre des développeurs provenant d'horizons divers (et parfois concurrents) jusqu'aux algorithmes de compression.

Les travaux du comité de développement ont irrigué ceux menés en parallèle, notamment pour les standards MPEG.

Le codage H.261 supporte les formats CIF et QCIF (352×288 et 176×144) avec échantillonnage 4:2:0 et met en œuvre des techniques intra-image et inter-images.

Le codage « intra » est similaire à celui utilisé pour le codage JPEG qui était alors en cours de développement. L'image est divisée en blocs de 8×8 pixels auxquels est appliquée une transformation DCT. Cette étape est suivie d'un codage à longueur variable (RLC) et d'un codage entropique.

Le codage « inter » fait appel à deux types d'images. Les images I sont les images de référence codées uniquement en intra ; les images P (Prédites) sont seulement définies par la différence par rapport à l'image I ou P qui les précède. Des techniques d'analyse des déplacements de macroblocs 16×16 sont utilisées avec transmission de vecteurs déplacements.

Cette technique est particulièrement efficace dans le type d'application visé, la visioconférence où l'image est plutôt statique.

Le codage H.261 est capable d'adapter le débit de source à celui du canal en faisant varier les seuils de quantification.

Il n'est pas sûr que le projet de visioconférence sur RNIS ait été un succès économique et commercial, mais ce qui est certain c'est que les connaissances qu'il a contribué à développer ont eu une influence considérable.

3.2 LA BOÎTE À OUTILS MPEG-2 POUR LES SIGNAUX VIDÉO

Le groupe de travail MPEG (*Moving Pictures Experts Group*) démarre en 1988 sous l'égide des organismes internationaux ISO (*International Standards Organization*) et IEC (*International Electrotechnical Commission*).

Le nom de baptême initial du projet « Coding of Moving Pictures and audio » définit bien ses objectifs : l'encodage des images animées et des sons qui les accompagnent.

Il est important de rappeler qu'il s'agissait alors d'une démarche prospective osée, dans la mesure où la télévision numérique n'était alors qu'une hypothèse visionnaire. À cette époque, elle exigeait, pour s'imposer, une remise à plat des modes de pensée, des méthodes de travail – de la production à la diffusion –, du matériel ainsi que de l'économie du secteur.

Le premier standard international fut publié en 1993 sous le nom de MPEG-1 (ISO/IEC 13818-2). Il définissait un débit de l'ordre de 1,5 Mb/s (dont 1,2 Mb/s pour l'image animée, 256 kb/s pour un couple audio stéréo et quelques kb/s pour les données de service) pour le stockage ou la diffusion d'un programme de type télévision.

Ce groupe de travail a opté pour ce débit car il correspondait à celui des disques compacts audio, seuls supports numériques réellement opérationnels à cette époque.

Ce standard montra rapidement ses limites : résolution médiocre (format CIF, *Common Image Format*, 360 × 288 pour les systèmes à 50 trames, ou pire, QCIF, *Quarter* CIF, 180 × 144), codage d'une trame sur deux seulement – la deuxième étant simplement répétée – et trop faible capacité du support (680 Mo, soit 74 minutes) qui ne permettait pas de stocker un long métrage sur un seul disque.

Il ne permettait en aucun cas de faire face à la versatilité et à la qualité qui devaient être celles des futurs programmes numériques.

En 1994, fut publié sous la référence MPEG-2 un ensemble de standards répartis en profils et niveaux (qui intègrent l'ancien MPEG-1). On en parle souvent sous le nom de « boîte à outils MPEG ». Il s'agit en effet d'un ensemble d'outils permettant, en principe, de faire face à tous les besoins de production, stockage et diffusion de programmes audiovisuels pendant encore une bonne dizaine d'années.

3.2.1 Redondance spatiale et redondance temporelle

Les algorithmes MPEG sont les héritiers des algorithmes JPEG. Ceux-ci exploitent les redondances existant à l'intérieur de l'image.

La redondance spatiale permet aux algorithmes JPEG d'atteindre un taux de réduction de débit de l'ordre de 10:1 sans trop de problèmes. Mais il faut atteindre des taux bien plus élevés pour réussir à faire passer dans les réseaux les 216 Mb/s (8 bits) ou 270 Mb/s (10 bits) de l'image TV standard (sans parler des 1,5 Gb/s de la HD).

Fort heureusement, les différentes images d'une séquence animée présentent toujours (sauf s'il y a eu un point de montage) une grande parenté : plan fixe traversé par une automobile ou un ballon, déplacement régulier des éléments de l'image dans un panoramique…

C'est l'exploitation de cet important gisement de redondance temporelle qui permet aux algorithmes MPEG d'obtenir les taux élevés de réduction de débit qu'exigent les images animées.

La réduction de débit JPEG est dite « intra » pour intra-image, la réduction de débit MPEG est dite « inter » pour inter-images.

On notera qu'il convient de distinguer en fait des possibilités de codages inter-trames et inter-images, tous deux retenus par MPEG-2 pour les systèmes à balayage entrelacé. Le mode inter-trame donne de meilleurs résultats lorsqu'il se produit un déplacement important des images, en particulier un panoramique vertical.

Il faut noter que tout événement brutal survenant au cours d'une séquence (le flash d'un appareil photographique au cours d'une conférence de presse par exemple) affecte la redondance temporelle, réduit donc l'efficacité de la réduction de débit et peut devenir source de graves perturbations dans l'encodage.

3.2.2 Blocs, macroblocs et tranches (*slices*)

Chaque image commence par être traitée comme en JPEG après découpage en blocs de 8×8 pixels.

Ces blocs sont ensuite groupés par 4 pour former un macrobloc. Les macroblocs forment des tranches (*slices*).

L'algorithme de compression cherche à repérer des déplacements éventuels des différents macroblocs en comparant deux images successives. Une fois ces macroblocs identifiés, on leur associe des « vecteurs de déplacement » (*motion vectors*) qui, décrivent en direction, sens et amplitude le déplacement de ces éléments d'image.

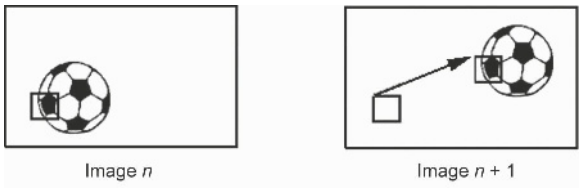


Figure 3.1.
Vecteur déplacement.

Il est à noter que cette technique ne s'applique pas à des mouvements de rotation, à vrai dire bien moins fréquents que les panoramiques… Mais il serait bon de prévenir les opérateurs de prise de vues de cet inconvénient.. La méthode sera également prise en défaut s'il existe dans l'image de nombreux mouvements incorellés en direction : un plan fixe d'un lac dont la surface est animée d'un doux clapotis peut devenir un piège redoutable… Il serait bon de prévenir les réalisateurs.

3.2.3 Des GOPs (*Group Of Pictures*) avec trois types d'images

Les séquences sont segmentées en groupes d'images, en abrégé GOPs, plus ou moins longs à l'intérieur desquels on va rechercher les parentés entre images (redondance temporelle).

Dans un GOP : trois types d'images peuvent être mis en œuvre : Images I pour Intra, P pour Prédites, B pour interpolées Bidirectionnelles.

Chaque GOP commence par une image I que j'appelle « image pivot », car elle va servir à définir les autres images du GOP.

Les images I subissent uniquement une réduction de débit de type « Intra » comme en JPEG. On notera l'existence dans le codeur d'une boucle de contrôle du dispositif de quantification qui réduit la profondeur de celle-ci lorsque le débit maximal autorisé est atteint ; il y a dans ce cas, bien entendu perte de qualité. Perte d'autant plus importante que l'image I sert de base au calcul de toutes les autres images du GOP.

Chaque image P d'un GOP est prédite soit à partir à partir de l'image I qui la précède, soit à partir de l'image P qui la précède. On peut dire de manière schématique qu'elle est définie ainsi : « c'est la même image à telles différences près » ; il suffit donc de transmettre ces dif-

férences pour savoir reconstituer la nouvelle image ; il est inutile de retransmettre ce qui est identique. Les vecteurs déplacement jouent évidemment un rôle fondamental dans l'expression de ces différences.

Les images B sont élaborées en référence avec le passé et le futur. L'encodeur réalise une interpolation bidirectionnelle, utilisant des vecteurs déplacement avant et arrière, entre les images I ou P qui encadrent l'image B, afin de définir des macroblocs intermédiaires. Les valeurs de ces macroblocs interpolés sont soustraites de celles des macroblocs réels et seules les différences, généralement faibles, sont transmises. C'est un élément essentiel quant à l' efficacité du codage MPEG dans la réduction de débit.

Voici un exemple de GOP de 12 images I B B P B B P B B P B B I B B P… C'est un GOP long, classique en diffusion, mais on peut également rencontrer des GOP de 16 images, voire plus. Le GOP le plus court est de 1 image I I I I… On parle alors de codage « I only ». Il est évidemment moins efficace quant à la réduction de débit mais présente d'autres avantages pour la production.

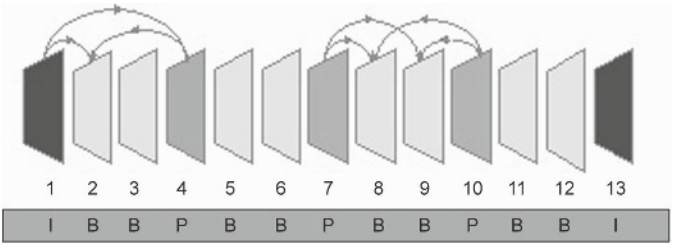


Figure 3.2.
Un GOP de longueur 12 images ; avec images I, P, B.

Il faut noter que les images du GOP ne sont pas transmises dans l'ordre de lecture et d'affichage mais dans celui qui permet la reconstruction la plus rapide :

I(1) P(4) B(2) B(3) P(7) B(5) B(6), etc.

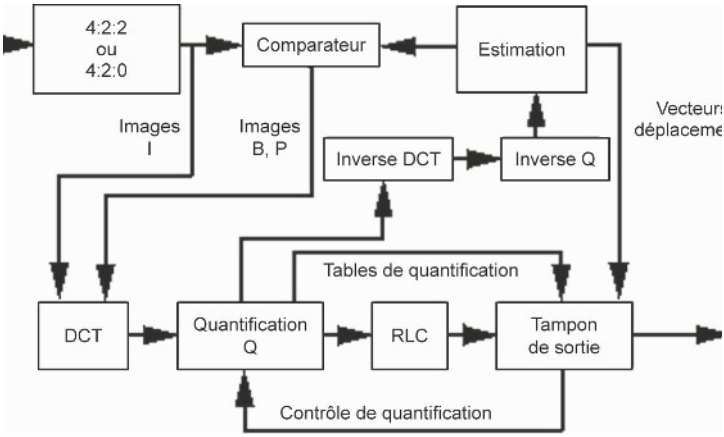


Figure 3.3.
Synoptique
du codage MPEG.

3.2.4 Des puzzles bien entremêlés

Le codage des séquences s'effectue selon une syntaxe hiérarchique bien précise dans laquelle chaque élément est identifié par un en-tête ou *header* qui permet la reconstruction des séquences lors du décodage. L'ensemble des données audio ou vidéo d'une séquence de programme constitue un « flux élémentaire » ES (*Elementary Stream*) dont la longueur n'est pas spécifiée. Il s'agit d'un puzzle complexe mais (en principe) facilement reconstructible par le décodeur grâce aux données de service contenues dans les en-tête.

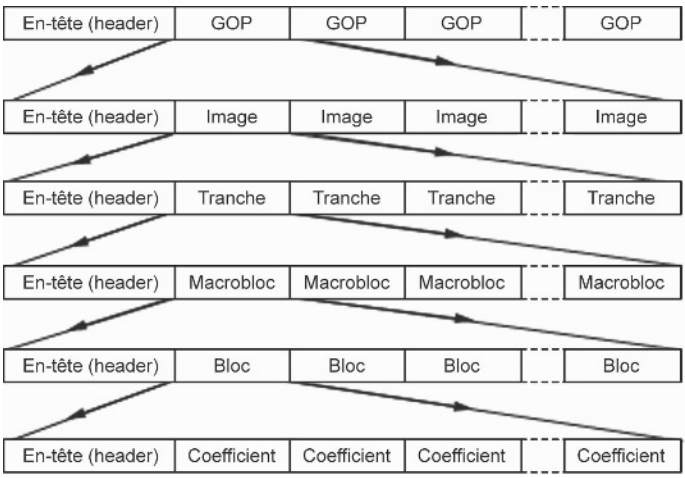


Figure 3.4.
Structure
du codage MPEG.

Les flux élémentaires sont ensuite « découpés » en paquets de données, de longueur fixe pour l'audio et variable pour la vidéo, qui constituent les PES (*Packetized Elementary Stream*) et qui disposent, bien entendu, d'en-têtes permettant de les identifier.

Les PES audio et vidéo d'un programme sont multiplexés pour constituer le flux de programme PS (*Program Stream*). Plusieurs programmes peuvent être multiplexés pour constituer un flux de transport TS (*Transport Stream*).

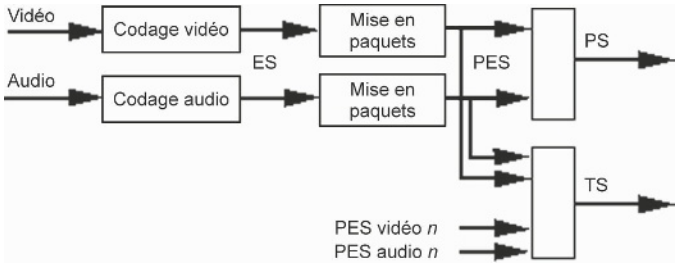


Figure 3.5. Les différents flux MPEG.

C'est donc un puzzle encore bien plus complexe que le décodeur du récepteur de télévision numérique doit reconstituer… Encore faut-il prendre en compte les données de service du type PSI (*Program Specific Information*), PAT (*Program Association Table*), EPS (*Electronic Program Guide*), qui doivent être ajoutées aux programmes afin que le spectateur puisse utiliser ceux-ci confortablement.

Eh bien, en général, ça fonctionne parfaitement !

3.2.5 Choix du meilleur GOP

Christoper D. Bennet de la Compagnie Hewlett Packard a présenté, à Las Vegas en 1996 des résultats présentant le poids moyen des images pour plusieurs GOPs, en prenant pour référence le poids des images I. Ce poids est divisé environ par 3 pour un GOP de

12 images I B B P B B P… ; divisé environ par deux pour des GOPs de type I B P B I… ou I B I B I…

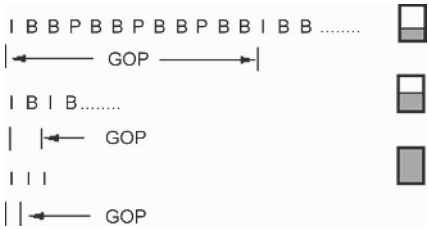


Figure 3.6.
Valeur moyenne du poids des images selon le type de GOP (d'après C. D. Bennett).

On a donc *a priori* tout intérêt pour une efficacité maximale à utiliser des GOPs étendus. Mais, ce n'est pas aussi simple.

La restauration des images s'effectue de proche en proche ; la restauration d'une image B exige que celles des images P adjacentes aient été réalisées etc. L'ordre de lecture et d'affichage des images et l'ordre de transmission sont différents ; on transmet une série réorganisée I, P, B, B, P… afin de permettre la restauration du GOP la plus rapide. Ceci entraîne un retard de l'ordre de la durée du GOP qui est certes tolérable en diffusion ou l'efficacité de la compression est primordiale (à condition qu'il soit tenu compte de ce retard pour maintenir synchrones les composantes audio du programme) mais qui est difficilement supportable en post-production. Dans ce domaine ce sont les codage *I-only* qui font la loi.

Il faut également tenir compte de la distance de propagation des erreurs ; une simple erreur sur la première image P d'un GOP long peut affecter plus d'une dizaine d'images… Gênant !

3.2.6 Une méthode asymétrique

Les codeurs MPEG sont relativement compliqués et lents, ils doivent en effet déterminer les vecteurs déplacement, élaborer l'image prédite, la valider par

comparaison avec l'image « vraie » après que les données des différents blocs aient été codés par DCT puis décodés, enfin multiplexer ces images selon l'ordre de transmission.

Fort heureusement les manipulations exigées des décodeurs, placés dans les foyers, sont plus simples, même s'il faut une capacité importante de stockage ainsi qu'une puissance de calcul non négligeable afin de reconstruire le puzzle complexe que nous avons évoqué.

3.2.7 Profils et niveaux

Les différents standards MPEG-2 constituent un ensemble d'outils qui ont été conçus afin de pouvoir répondre à tous les besoins prévisibles. Ils sont répartis en profils et niveaux.

Les niveaux (*levels*) spécifient la résolution des images.

- Niveau Bas (*Low Level*) : résolution CIF ; 360 pixels par ligne sur 288 lignes ;
- Niveau Principal (*Main Level*) : résolution CCIR 601, 720 pixels sur 576 lignes ;
- Niveau Haut 1440 (*High 1440 level*) : haute définition à 1 440 pixels par ligne sur 1 080 lignes utiles ;
- Niveau Haut (*High Level*) : haute définition pleine résolution, 1 920 pixels sur 1 080 lignes utiles.

Les profils (*profiles*) offrent une collection d'outils propres à des utilisations spécifiques : types d'images utilisables parmi la panoplie I,P,B et type d'échantillonnage des images.

- Profil Simple (*Simple Profile*) : images I,P et échantillonnage 4:2:0 ;
- Profil Principal (*Main Profile*) : images I,P,B et échantillonnage 4:2:0 ;

- Profil Haut (*High Profile*) : images I,P,B et échantillonnage 4:2:2 ;
- Profil 4:2:2 : il ne comporte qu'un seul niveau correspondant à la norme CCIR 601 de numérisation du signal de télévision standard. On note qu'il code 608 lignes, c'est-à-dire que les lignes de l'intervalle de suppression trame (*vertical blanking*) qui peuvent notamment supporter le code temporel VITC sont prises en compte ; il s'agit d'un véritable outil de production.

On trouve également deux profils qui possèdent des outils supplémentaires permettant un codage hiérarchique : codage d'une image de base, robuste, à laquelle on peut venir ajouter des couches supplémentaires permettant d'améliorer soit la résolution de l'image (*Spatially scalable*), soit son rapport Signal sur Bruit (SNR, *Scalability, Signal to Noise Ratio*). Ces deux profils assez complexes n'ont pas, à ce jour, fait l'objet de mises en œuvre ; ces fonctions sont par contre au cœur des codages MPEG 4. Quant au profil HIGH, il commence seulement à être utilisé.

Tableau 3.1.
Profils et niveaux.

Niveaux	Profils					
	Simple	Principal	4 :2 :2	SNR scal. choix S/B	Spatial scal. choix résolut.	High
Types d'images	4 :2 :0 / I,P	4 :2 :0 / I,P,B	4 :2 :2 / I,P,B	4 :2 :0 / I,P,B	4 :2 :0 / I,P,B	4 :2 :2 / I,P,B
Haut		1920x1080 80 Mb/s				1920x1080 100 Mb/s
1440		1440x1080 60 Mb/s			1440x1080 60 Mb/s	1440x1080 80 Mb/s
Principal	720x576 15 Mb/s	720x576 15 Mb/s	720x608 50 Mb/s	720x576 15 Mb/s		720x576 20 Mb/s
Bas		352x288 4 Mb/s		352x288 4 Mb/s		

Cet ensemble de profils et niveaux constitue la « boîte à outils MPEG », chaque combinaison conduit à un débit maximum toléré, allant de 4 à 100 Mb/s.

Seules 11 combinaisons ont été retenues et seules deux ont aujourd'hui été réellement exploitées. On les connaît sous l'étiquette MP@ML (*Main Profile at Main Level*) pour la diffusion, avec des débits de l'ordre de 5 Mb/s ainsi que pour l'édition de DVD avec des débits pouvant approcher 10 Mb/s et sous l'étiquette 4:2:2P@ML pour la production en résolution standard (SDTV).

On trouvera également des liaisons entre studios de production, dites « liaisons de contribution » à des débits de l'ordre de 30 Mb/s.

3.2.8 Performances en fonction du débit

On aura compris que l'on dispose de nombreux paramètres : le débit bien sûr mais également le type d'images I,P,B ainsi que le type d'échantillonnage 4:2:2 ou 4:2:0.

Le débit est un paramètre directement lié à l'économie ; plus il sera faible plus on pourra diffuser de programmes sur un canal donné. Il faut rester toutefois extrêmement prudent, ne pas chercher à obtenir à tout prix un débit faible ; celui-ci pourrait devenir synonyme de perturbations importantes des programmes, surtout s'ils sont très animés (programmes sportifs) où s'il existe trop de différences entre les images successives (bruit de fond, mouvements peu corrélés, donc peu prédictibles : le clapotis sur un lac, le vent dans les branches, la fin d'un travelling…) pouvant aller jusqu'au terrible écran noir pendant quelques instants.

Il n'est cependant pas indispensable d'aller au delà d'une certaine limite. On constate en effet que la qualité augmente linéairement en fonction du débit puis tend vers une asymptote.

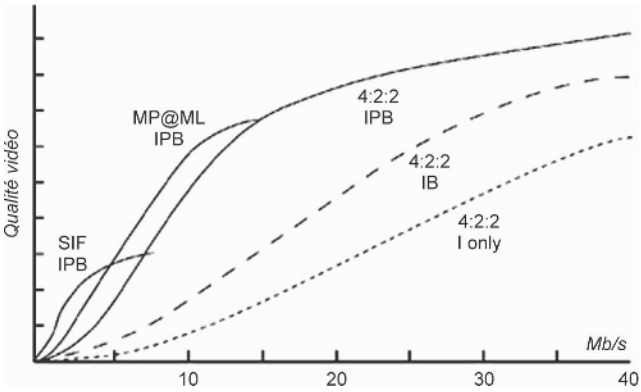


Figure 3.7.
Qualité en fonction du débit (d'après C.D. Bennett).

N.K. Lodge (*Independent Televisison Commission*, UK) et D. Wood (UER/EBU) ont pu montrer lors d'une conférence (IBC 1994) que 84 % des séquences réellement diffusées sur les antennes de la BBC demeurent satisfaisantes avec un débit total correspondant à 1,5 bit par pixel ; seul 1 % d'entre elles, les plus complexes, nécessitent 3 bits par pixel. Une image 4:2:2 : comptant 414 720 pixels on aura 10 368 000 pixels/s ; un débit de l'ordre de 30 Mb/s apparaît donc comme suffisant pour une transmission virtuellement transparente.

C'est encore une fois le couple « débit/type d'utilisation » qu'il convient de prendre en compte.

3.2.9 Débit Constant CBR (*Constant Bit Rate*) ou débit Variable VBR (*Variable Bit Rate*)

Les spécifications MPEG proposent deux types d'encodage à débit constant (CBR) ou à débit variable (VBR) ; le deuxième semble plus intelligent, pourquoi bloquer de la capacité de transmission ou d'enregistrement si l'image est fixe ou quasi fixe et véhicule très peu d'information nouvelle ?

Si l'encodage CBR est moins efficace il est également plus simple et plus rapide : une seule passe suffit une fois le débit fixé (un peu à l'estime ou compte tenu de

la durée du programme et de la capacité de stockage ou des disponibilités du lien de transmission).

La méthode VBR stipule trois taux fixés par l'utilisateur : débit maximal qu'il ne sera jamais possible de dépasser quels que soient les besoins réels en termes de qualité d'image ; débit moyen autour duquel le débit instantané oscillera ; taux minimum au dessous duquel le débit ne descendra pas.

Pour afficher ces données, il faut qu'elles aient été déterminées lors d'une première passe du programme dans l'encodeur. Le mode VBR est donc plus complexe à mettre en œuvre pour des transmissions « temps réel » ; il demande souvent des « mémoires tampon » importantes afin d'éliminer des irrégularités temporelles lors de la transmission.

3.2.10 Le problème du bruit

Pour que le système fonctionne il faut qu'il y ait une bonne corrélation entre les images successives. Or les signaux parasites connus sous le terme de bruit de fond (faisant référence aux années TSF), sont totalement incorrélés d'une image à l'autre. Il peut en être de même si la numérisation du signal a introduit des artefacts dus à une sous-quantification ou à un sous-échantillonnage. Ces signaux vont donc perturber le processus de réduction de débit.

On constate par exemple qu'un fichier numérique issu du transfert bien réalisé d'une superbe copie de film 35 mm peut conduire, bien que la résolution initiale soit plus élevée, à des débits moins importants que ceux résultant de l'encodage d'une bande VHS très bruitée.

Il est donc indispensable de commencer le traitement par un « débruitage » (ou NR, *Noise Reduction*) aussi efficace que possible.

Encore faut-il que cette réduction de bruit ne s'effectue pas grâce à un simple filtrage passe-bas entraînant

une dégradation de l'image. Il conviendra de mettre en œuvre des filtres numériques successivement dans le domaine spatial (de type « median », par exemple : la valeur de chaque pixel est comparée à celle des pixels adjacents dont on prend la valeur moyenne) et dans le domaine temporel de type MCTF (*Motion Compensated Temporal Filtering*).

3.2.11 Les problèmes de concaténation

Plier et déplier un papier laisse des traces indélébiles. Il en va de même pour la concaténation (du latin *catena*, chaîne), ou l'enchaînement, ou encore la réalisation en cascade, de séries de codages/décodages.

C'est pourtant ce qui risque de se produire lorsque l'on réalise la post-production de programmes. Chaque traitement graphique porte sur les pixels individuels ; ceux-ci ne sont plus accessibles en mode JPEG ou MPEG ; on doit donc procéder à un décodage pour revenir à une description spatiale des images. La concaténation entraîne l'apparition d'artefacts visibles dès lors que la réduction de débit est forte.

La « Task Force for Harmonized Standards for the Exchange of Program Material as Bitstreams » (harmonisation des standards pour l'échange de programmes sous forme de flux numériques) de la SMPTE (*Society of Motion Picture and Television Engineers*) et de l'UER /EBU (Union Européenne de Radiodiffusion / *European Broadcasting Union*) a réalisé des tests subjectifs de comparaison de séquences normalisées Ce groupe de travail est arrivé à la conclusion qu'il convenait de disposer d'un débit d'environ 50 Mb/s afin de réaliser une post-production exigeante, mais qu'un enregistrement à environ 20 Mb/s permettait une post-production simple sans que n'apparaissent des artefacts perturbants.

3.2.12 Interopérabilité et « compliance »

Il est important de préciser que les standards MPEG ne précisent que la syntaxe des flux numériques codés sans intervenir, de quelque manière que ce soit, dans le processus de codage ; on dit qu'il s'agit de systèmes ouverts.

Chaque constructeur est libre de mettre en œuvre les méthodes qu'il estime les plus performantes. Il en résulte que les problèmes de compatibilité des équipements avec le standard (*compliance*) et d'interopérabilité entre eux sont essentiels. Tous les équipements doivent utiliser le même protocole (notamment en termes de structure et d'indexation des différents flux numériques audio et vidéo ainsi que des données de services annexées) et être capables de fonctionner aux débits limites fixées par ce protocole.

3.2.13 La nécessité de mesures

Un des problèmes essentiels du numérique est que « ça passe ou ça casse ». En analogique on pouvait assister à une dégradation progressive de l'image ; en numérique, elle s'effondre brutalement c'est ce qu'on appelle « l'effet de falaise » (*cliff effect*). On a donc tout intérêt à vérifier en temps réel les performances de la chaîne afin de prévenir la catastrophe.

Dans les systèmes MPEG la syntaxe hiérarchique est extrêmement complexe. Il est essentiel de pouvoir localiser où se situe un éventuel problème tant au niveau de la syntaxe qu'au niveau des différents éléments de la chaîne. Il existe des appareils permettant une surveillance des flux en temps réel suivie d'une analyse détaillée en différé après enregistrement d'une séquence.

Il est également impératif de pouvoir analyser les artefacts introduits dans l'image par un codage inadapté (débit trop faible par exemple).

Une première méthode consiste à faire visionner les séquences après codage par un collège de spectateurs qui déterminent à partir de quel moment ils notent des dégradations perceptibles sur 5 niveaux : il s'agit d'une méthode subjective et en temps différé.

Une deuxième méthode consiste à calculer la différence entre l'image initiale et l'image après codage et à établir une cartographie des différences ou un indice de qualité un peu à la manière dont on définit un rapport signal/bruit. Il s'agit d'une méthode objective en temps réel mais qui ne donne pas forcément de renseignements sur la dégradation visible des images.

Une méthode plus fine consiste à « injecter » dans cette mesure objective un modèle du système visuel humain. Le modèle élaboré par le laboratoire Sarnoff, aux USA, a conduit à une méthode dite JND (*Just Noticeable Différence*) mise en œuvre par Tektronix qui donne une valeur de qualité dite PQR (*Picture Quality Rating*) sur une échelle de 0 à 25 (0 : copie codée identique à l'original ; 10 : dégradation très perceptible ; 25 : image complètement dégradée) et qui permet de définir très rapidement la qualité de l'image encodée.

Jusqu'au niveau 3, la dégradation est imperceptible. Une image issue d'un Betacam Numérique se situe environ au niveau 0,8 ; celle d'un très bon DVD au niveau 4 ; les premiers codeurs MPEG se situaient au niveau 8 à 10 avec un débit de 4 Mb/s, ils sont maintenant en dessous du niveau 3 avec un débit de 2,5 Mb/s. Bravo !

3.3 LA FAMILLE DV

En juillet 1993 Matsushita, Philips, Sony et Thomson annoncèrent la naissance du premier format d'enregistrement numérique (allant jusqu'à la haute définition) destiné au grand public (*Consumer Use Digital VCR*, Video Cassette Recorder). Ce format, avec compression, prit le nom de DV (*Digital Video*, tout simplement).

Très rapidement la majorité des constructeurs d'électronique domestique adoptèrent le format et les premières machines apparurent sur le marché en 1996. Un projet très audacieux, à peine crédible, mais dont l'histoire a prouvé amplement qu'il était réaliste et qui mérite un chapitre spécial.

3.3.1 Les spécifications

Le projet adoptait délibérément la compacité avec l'utilisation d'une cassette de très faible dimensions contenant une bande magnétique minuscule de largeur 6,35 mm (1/4 de pouce). Cette bande possédait des propriétés magnétiques remarquables. Il ne s'agissait plus de l'utilisation d'une enduit composé d'un liant enrobant des particules de métaux ferromagnétiques (fer, nickel ou cobalt) ou de leurs oxydes mais d'une couche continue et d'épaisseur moléculaire de métal ferromagnétique, obtenue par vaporisation sous vide : d'où le nom de bandes « métal évaporé » ou ME.

La vitesse de défilement de la bande est très faible (18,81 mm/s) ce qui conduit à une largeur de piste magnétique inscrite de seulement 10 µm (environ un dixième du diamètre d'un cheveu)

Le standard DV met en œuvre une réduction de débit « intra » basée sur une DCT, de type JPEG (mais ce n'est pas du JPEG) et de taux de compression 5:1.

Dans le système PAL, à 50 trames par secondes et 575 lignes utiles, le signal vidéo est le 4:2:0 : les échantillons de chrominance ne sont présents qu'une ligne sur deux ; dans le système NTSC à 60 trames par seconde et 480 lignes utiles, le signal vidéo est de type 4:1:1 : les échantillons de chrominance sont présents sur toutes les lignes, mais seulement pour 1 pixel sur 4. Dans les deux cas, la quantification est effectuée sur 8 bits.

Le choix d'une compression « intra » a été retenu pour ce format grand public afin de permettre des solutions de montage simples.

Le débit vidéo est finalement réduit de 125 Mbits/s (4:2:0 ou 4:1:1 / 8 bits, sans les suppressions) à 25 Mbits/s, mais le débit total enregistré s'élève à 41,85 Mbits/s (avec l'audio, le code de correction d'erreurs et les données auxiliaires, notamment le code correcteur d'erreurs).

En ce qui concerne l'audio, le DV peut traiter soit 4 pistes échantillonnées à 32 kHz codées sur 12 bits, soit 2 pistes échantillonnées à 48 kHz codées sur 16 bits.

Deux types d cassettes existent, la cassette normale de durée d'enregistrement de 4,5 h et la petite cassette dite « mini DV » destinée aux caméscopes de paume, d'une durée d'enregistrement de 1 h.

Il avait été initialement prévu par les pères du projet un développement futur vers la Haute Définition rendu possible par une augmentation de la vitesse de défilement de la bande.

3.3.2 Pour quelques micromètres de plus

En 1996, Panasonic proposait, sous le nom de DVCPRO, une déclinaison vers le haut du DV, dont lé dénomination affichait clairement qu'elle était destinée à une utilisation professionnelle.

Les modifications étaient mineures, pour l'essentiel une multiplication par 1,8 de la vitesse de défilement ce qui sécurise l'enregistrement en portant la largeur de piste à 18 µm contre 10 en DV. Le signal reste identique à celui du DV (mêmes algorithmes de compression, mêmes débits).

Sony suivit peu après un chemin analogue avec le format DVCAM, dont la vitesse défilement était multipliée par 1,5 (largeur de piste 15 µm).

On notera que les machines de Sony utilisent les cassettes définies pour le DV quant les machines DVCPRO de Panasonic utilisent, pour la bande de 6,35 mm, un boîtier spécifique. C'est une petite guerre industrielle dont seuls les utilisateurs ont fait les frais.

Panasonic a ensuite proposé, suite aux recommandations de la « Task Force » SMPTE/UER (voir chapitre 5) une version DVCPRO-50. Tout était prêt… Il suffisait d'utiliser deux codecs DV fonctionnant en parallèle pour porter le débit vidéo de 25 à 50 Mb/s (débit total 99 Mb/s). La vitesse de défilement de la bande est double de celle du DVCPRO-25.

Le signal vidéo est de type 4:2:2. Le taux de compression passe à 3,3 :1. Le nombre de pistes audio PCM passe de 2 à 4 (toujours en 48 kHz/16 bits).

3.3.3 L'interface numérique

Le codage « intra » a été choisi afin de simplifier les problèmes de montage. Les développements rapides de l'informatique ont fait qu'il est devenu très vite possible d'effectuer la post-production de séquences DV sur des ordinateurs domestiques équipés de codecs DV. Il a été développé, initialement sous l'impulsion d'Apple, une interface numérique, spécifique au signal DV, permettant la liaison entre un magnétoscope et un ordinateur. Cette interface est connue sous une appellation générique IEEE 1394 (*Institute of Electrical and Electronics Engineers*) et deux appellations d'origine contrôlées : FireWire (origine Apple), I-Link (origine Sony). Il s'agit d'un bus de type « plug and play » analogue au bus USB avec des débits pouvant aller jusqu'à 400 Mb/s pour la version *a* et 800 Mb/s pour la version *b*.

3.3.4 Les formats Haute Définition

Panasonic a réalisé les machines DVCPRO-100 en utilisant 4 codecs DV en parallèle (donc débit vidéo de 100 Mb/s), une méthode plutôt économique sinon optimale.

Enfin est apparu en 2003 le format HDV (consortium HDV : Canon, JVC, Sharp et Sony) en principe destiné au grand public. Les premières machines sont apparues en 2004.

Ce format n'a plus grand chose à voir avec le DV d'origine, si ce n'est la cassette et la cinématique.

Toute l'électronique numérique a été modifiée. C'est une véritable gageure qui permet de conserver le débit de 25 Mb/s du signal DV pour véhiculer une image 1 440 × 1 080i au format 16/9 avec des pixels rectangulaires au rapport 1,33 (Sony et Canon). Il s'agit d'un signal 4:2:0 quantifié sur 8 bits. Il ne s'agit donc pas d'un véritable enregistrement « full HD », mais la résolution demeure cependant près de 5 fois plus élevée qu'avec l'enregistrement DV.

Le standard HDV peut également supporter le format 720p : 1 280 × 720 pixels carrés (format retenu par JVC) avec un débit de 19.7 Mbit/s. Quelques machines supportent également un format progressif SF 1080p.

La réduction de débit vidéo est effectuée en MPEG-2 avec un GOP de 12 images pour les systèmes à 50 Hz et 15 pour les systèmes à 60 Hz. L'enregistrement audio utilise MPEG-1 Layer 2.

Les caméscopes HDV actuellement sur le marché, à l'allure (et au prix) grand public haut de gamme, semblent avoir conquis, à juste raison, un large public professionnel (news, reportages, clips, court métrages, etc.).

Cependant, si l'acquisition des images est facile, la post-production a été longtemps problématique en raison

du GOP long : il était courant de transcoder aux formats HD-CAM.

Aujourd'hui, plusieurs systèmes de montage non linéaire sont capables de traiter le signal HDV natif.

Tous les caméscopes HDV utilisent des cassettes DV de type « master quality », (60 minutes d'enregistrement sur une cassette MiniDV) mais de nouveaux modèles sont apparus pouvant utiliser en option une enregistrement sur disque dur ou sur carte Compact-Flash.

4.1 LE CODAGE MULTIMÉDIA POUR LE TROISIÈME MILLÉNAIRE

Les nouveaux modes d'utilisation des programmes multimédia (Télévision Numérique Standard ou à Haute Définition, ADSL, VOD, *Video on Demand*, DVD Haute Définition – Blu-ray –, Télévision pour mobiles, etc.) qui sont apparus au début du millénaire, exploitent une gamme de débits numériques très étendue de quelques dizaines de kilobits par seconde (kb/s) pour la transmission vers les mobiles à quelques dizaines de mégabits par seconde (Mb/s) pour les applications HD.

Dans cet univers mouvant et foisonnant les notions de flexibilité et d'évolutivité en même temps que celle de compromis débit/qualité ont pris une importance fondamentale.

L'arrivée des méthodes de compression des données d'images animées a été, il y a une vingtaine d'années, quelque chose de véritablement magique. Comment imaginer que l'on pouvait éliminer la moitié ou les trois quarts des données tout en conservant la possibilité de visionner confortablement une séquence ? Depuis lors, nous avons appris le pourquoi et le comment de cette affaire !

La réduction de débit a par ailleurs été une réussite intellectuelle (avec le rassemblement d'un immense éventail de compétences) technique et économique exemplaire.

MPEG-1 et surtout MPEG-2 ont complètement modifié le panorama de la production et de la distribution des programmes audiovisuels.

Et pourtant…

4.1.1 Tous les pixels ne sont pas égaux

Les méthodes précédentes souffrent d'un handicap ; elles traitent de manière identique tous les pixels des images. Or ceux-ci n'ont pas tous la même fonction, la même origine, la même existence, ni la même importance.

Il est évident que le décor d'un studio de journal télévisé, apporte moins d'information que le visage et la voix du journaliste. De plus, il ne change pas d'une minute à l'autre ni du jour au lendemain..

Il est donc venu aux esprits des spécialistes, afin d'améliorer les performances des codages, de représenter un programme audiovisuel comme une collection d'éléments identifiables, séparables et regroupables, mais de natures ou de fonctions différentes.

C'est la notion d'objets audiovisuels (image naturelles, textes, graphismes, environnements « 3D », sons naturels ou synthétiques) qui apparaît.

Chacun de ces objets qui constituent une scène, sera soumis au type de codage le plus efficace compte tenu de sa nature.

Chaque lettre d'un titre ou d'un sous-titre, qui peut correspondre à plusieurs milliers de pixels (c'est-à-dire plusieurs milliers de fois 3×8 bits, au minimum, en RVB), peut être parfaitement défini grâce aux 8 bits du code ASCII, plus quelques bits d'attributs (couleur, police, taille…).

On sait également tout l'intérêt des traitements en mode vectoriel pour les graphismes. Ces éléments peuvent être rendus interdépendants temporellement et spatialement par l'intermédiaire d'un « descripteur de scène » (*scene descriptor*).

4.1.2 Les objectifs

Les travaux sur le codage MPEG-4 (ISO/IEC 14496), « le codage multimédia pour le troisième millénaire » comme on le nomme parfois, ont commencé en 1995, une première parution a eu lieu en 1998. Il est devenu Standard International en 2000 ; il continue à évoluer.

Il est défini comme « orienté objets ». On pourra, par exemple, construire une scène comme la collection hiérarchique des objets suivants :

- un présentateur, enregistré sur fond bleu, et sa parole (deux objets) ;
- un décor en images de synthèse 3D constitué de deux objets, un fauteuil et un lampadaire ;
- une image de synthèse 2D ;
- un logo de chaîne codé ASCII ;
- une horloge provenant d'un générateur électronique codé ;
- une musique MIDI.

Il s'agit en quelque sorte des éléments d'un programme avant passage dans le mélangeur.

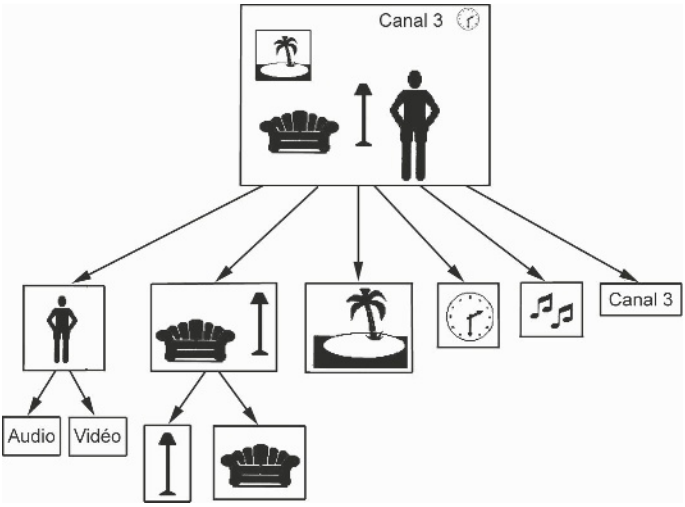


Figure 4.1.
Décomposition d'une scène en objets audiovisuels.

Le but est d'assurer, pour les programmes multimédia (depuis la télésurveillance jusqu'à la haute définition), une standardisation technologique à tous les niveaux : production, distribution et diffusion.

Si l'objectif d'un faible débit (inférieur à 64 kb/s) a été essentiel au début des travaux, on s'est assez vite rendu compte que la notion d'objets indépendants assurait bien d'autres avantages, notamment des possibilités de décodage des programmes à plusieurs niveaux de qualité (effet d'échelle, adaptabilité, ou *scalability*) ainsi que des possibilités d'interaction du spectateur avec le contenu de la scène.

Le développement de nouveaux types d'outils de création et de manipulation d'objets de synthèse ouvre également des applications dans les domaines des jeux, de la réalité virtuelle, de la téléprésence, de l'imagerie scientifique, etc.

La norme MPEG-4 est extrêmement complète puisqu'elle comporte 21 parties qui définissent depuis les généralités du système (*part 1*) jusqu'aux applications graphiques interactives en langage Java (*Java Graphic Framework eXtension*, *part 21*), la spécification de la transmission des polices de caractères (*part 18*), les méthodes de transport du flux des données sur Internet (*part 8*) en passant par les codages vidéo (*part 2*) et audio (*part 3*). La partie 10, sans doute la plus importante aujourd'hui, concerne la transmission ou l'enregistrement de programmes multimédia.

MPEG-4 a prouvé son efficacité dans le domaine de la télévision numérique ainsi que dans celui des applications graphiques et multimédia interactives avec des débits pouvant aller de 64 kb/s jusqu'à 4 Mb/s pour les applications les plus exigeantes (limite supérieure qui a été récemment étendue à 1,2 Gb/s pour les besoins du cinéma numérique).

4.2 PRINCIPES DE BASE

Les différents objets constituant la scène audiovisuelle peuvent être stockés, puis transmis, indépendamment vers le récepteur où ils seront assemblés grâce aux informations de composition du descripteur de scène. Celui-ci, qui utilise le format BIFS (*BInary Format for Scenes*) est transmis parallèlement afin d'informer le décodeur des relations spatio-temporelles qui régissent la reconstitution de la scène.

Le format BIFS est proche du langage VRML (*Virtual Reality Modeling Language*) utilisé depuis une bonne dizaine d'année pour la présentation et l'animation d'espaces virtuels 3D, interactifs ou non.

On parlera notamment d'*Audio Visual Objects* (AVO), *Video Objects* (VO) et de *Video Objects Planes* (VOP), ces derniers représentant un ensemble d'objets visuels (VO) qui doivent être manipulés simultanément dans une scène. Il peut s'agir d'objets naturels (d'un aspect apparenté à la photographie) ou d'objets réalisés en images de synthèse.

On peut dire, d'une manière imagée, que le récepteur recevra un kit d'éléments qu'il assemblera suivant le plan de montage joint. Le spectateur pourra éventuellement décider de ne pas prendre en considération tel ou tel objet (sous-titre, médaillon avec traduction en langage des signes…) parce qu'il(s) ne l'intéresse(nt) pas ou parce qu'il est (ils sont) incompatible(s) avec les performances de son installation.

Un certain nombre d'éléments pourront être stockés localement afin de réduire le débit nécessaire à la transmission.

Quel est l'intérêt de retransmettre chaque soir ou chaque semaine le décor du studio d'un Journal Télévisé ou le générique de « Thalassa » ? Ce stockage d'informations chez l'utilisateur pourrait être développé et d'autres éléments réutilisables, comme des expressions

faciales liées à des phonèmes ou des pictogrammes d'une carte météo, pourraient eux aussi être « délocalisés ».

Il sera également possible de modifier en temps réel la résolution ou la taille de tel ou tel objet suivant le contexte.

L'effet de changement d'échelle (*scalability*), est fondamental dans MPEG-4.

Il peut être spatial (un décodeur peut n'utiliser qu'une fraction du flux de données afin d'afficher des images dans une résolution réduite), temporel (une séquence vidéo peut être reproduite avec moins d'images par seconde qu'elle n'en comportait à l'origine) ou encore qualitatif (les données sont organisées en couches : une couche de base accompagnée de couches d'extension, *Enhancement Layer*, permettant un affichage progressif de la qualité).

Cette technique d'adaptabilité aux besoins de l'utilisateur ou aux performances du matériel peut porter sur l'image entière ou sur un seul objet.

4.2.1 Profils et niveaux

Comme pour MPEG-2, des profils et niveaux ont été définis pour MPEG-4. Les possibilités étant plus grandes, le nombre des profils est plus élevé.

Les profils sont d'ordre qualitatif. Chaque profil est adapté à un type d'application pour laquelle il regroupe les outils appropriés.

Les niveaux sont d'ordre quantitatif.

La première version du standard a défini neuf profils visuels ; cinq pour les images naturelles et quatre pour les images de synthèse.

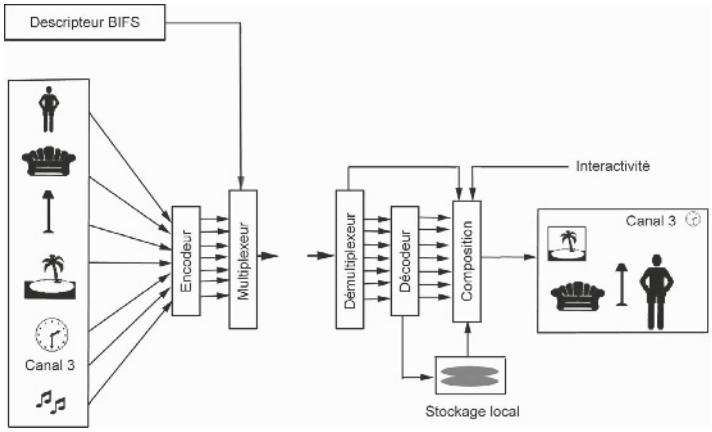


Figure 4.2.
Synoptique d'une transmission MPEG-4.

Les images naturelles disposaient des profils « Simple Visual Profile » (destiné aux objets vidéo rectangulaires pour applications mobiles) ; « Scalable Visual Profile » (adapté à plusieurs niveaux de transmission) ; « Core Visual » (analogue au *Simple* mais pouvant gérer des objets de taille variable avec un niveau limité d'interactivité) ; « Main », (analogue au *Simple* mais pouvant gérer des images entrelacées ainsi que des *sprites*) ; « N-Bit » (codage d'objets sur peu de bits pour des applications de télésurveillance et certaines applications médicales).

Les images de synthèses disposaient des profils « Simple Facial Animation » (création de « têtes parlantes » simples) ; «Scalable Texture » (textures d'images fixes à niveaux de qualité variables, utilisé pour manipuler des objets dans les jeux) ; « Basic Animated 2D Texture » (animation faciale simple et manipulation d'objets de taille variable) ; « Hybrid » (associant les possibilités du profil *Core* à la gestion d'objets de synthèse du profil précédent).

La version 2 a défini six nouveaux profils : « Core Scalable » (ajoute au profil *Core* l'adaptabilité spatiale et temporelle afin de rendre les programmes utilisables pour des applications diverses : mobiles, Internet, télévision) ; « Advanced Core » (combine des images animées et des images fixes pour des applications

multimédia) ; « Advanced Coding Efficiency » (réception mobile de télévision, enregistrement sur caméscopes) ; « Advanced Real Time Simple » (applications « temps réel », visioconférence) ; « Advanced Scalable Texture » (accès aléatoire aux images fixes à plusieurs niveau de qualité : consultation de banques d'images) ; et « Simple Face and Body Animation » (amélioration du SFA).

Au fil des avancées techniques et de l'apparition de nouveaux besoins, d'autres profils ont été ajoutés plus récemment. Le profil « Fine Granularity Scalable » vise à adapter la qualité à la transmission sur des réseaux à débits fortement variables dans le temps. Le profil « Advanced Simple Profile » rassemble les meilleurs outils permettant une diffusion de qualité à faible débit. Deux autres profils visent le Cinéma Numérique : « Simple Studio » (acquisition d'images à haute résolution jusqu'au format 4K – 4 000 pixels/ligne – et un débit de 1,2 Gb/s, pour la production) et « Core Studio » (pour la distribution).

En fait, parmi cette panoplie de possibilités couvrant tous les champs multimédia, seuls les profils « Simple Profile » et « Advanced Simple Profile » sont réellement utilisés aujourd'hui.

Chaque profil comporte plusieurs niveaux (*levels*). Ils imposent des limites aux paramètres du flux numérique tels que : débit maximal autorisé, résolution des images, nombre d'objets utilisables.

Pour le Profil Simple on a par exemple 4 niveaux (du niveau 1 – résolution QCIF ; 1 objet ; 1,4850 Mb/s –, au niveau 4 – résolution CIF ; 4 objets ; 12 Mb/s) ; pour le *Main*, 3 niveaux (résolution CIF, 720 × 576, 1920 × 1080 ; de 16 à 32 objets ; débits de 24 à 49 Mb/s).

La partie 10 des normes MPEG-4 a introduit le profil AVC, également connu sous la référence H.264, qui sera étudié spécifiquement compte tenu de son importance.

4.2.2 LASeR et SAF

La partie 20 a introduit les spécifications LASeR et SAF (*Lightweight Application Scene Representation* ; *Simple Aggregation Format*) « visant la représentation et la distribution de services de type *Rich-Media* à destination d'appareils à faibles ressources tels que les téléphones mobiles ».

Les services « Rich-Media » concernent la diffusion sur réseaux (*streaming*) de contenus audiovisuels synchronisés avec des textes (commentaires, annotations…), des tableaux, des diapositives (*slides*), des graphismes en mode vectoriel. Ce sont des services dynamiques et interactifs.

Des outils existent pour gérer des documents *Rich Media* sur Internet, mais ils se sont révélés trop lourds et complexes pour une utilisation sur téléphones mobiles. LASeR et SAF ont les dimensions et les fonctionnalités optimales pour ce type d'applications.

Il s'agit de deux formats binaires :

- Le premier (LASeR) permet l'encodage de scènes 2D en mode SVG (*Scalable Vector Graphics*) ainsi que des indications de modifications temporelles de la scène. Il s'agit en quelque sorte d'un descripteur BIFS « allégé » (*lightweight*).
- Le deuxième (SAF) permet de rassembler (agréger) dans un flux unique des contenus audiovisuels et des contenus de type LASeR.

4.3 LE CODAGE DES OBJETS VISUELS

4.3.1 Le codage des objets visuels naturels

Les objets visuels naturels (ceux issus d'une caméra) sont codés sous la forme de deux éléments : la forme (*shape*) et la texture, c'est-à-dire la partie de l'image qui se trouve à l'intérieur de la forme.

Les formes les plus simples sont rectangulaires (correspondant au cadre d'une image), mais à la différence de MPEG-2, elles ne sont pas de dimensions fixes ; elles sont définies à l'aide des macroblocs 16 × 16 dont elles doivent être des multiples.

Mais il peut exister des formes plus complexes, telles que des silhouettes de personnage. Elles sont définies comme un assemblage irrégulier de macroblocs.

On peut penser à ce qui se passe pour des incrustations qui utilisent un signal de découpe (*key* ou *alpha signal*) qui représente la silhouette d'un personnage et un signal de remplissage de la découpe.

La partie intérieure de la forme peut être codée sur un seul bit (transparence ou opacité totale de la texture de remplissage) ou sur plusieurs bits (plusieurs niveaux de transparence de la texture, notamment en bordure – comme pour les incrustations à bords flous).

Le codage des textures s'effectue en 4:2:0 (sauf pour les profils « studio »).

Les outils, dont chaque profil retiendra tout ou partie, utilisent pour l'essentiel, mais en les améliorant, les techniques déjà mises en œuvre dans MPEG-1 et MPEG-2 : division de l'image en blocs, DCT, codages entropiques et à longueur variable.

Même si la DCT et le codage à longueur variable ont été améliorés, c'est surtout en termes d'estimation de mouvement que les progrès ont été décisifs pour les versions les plus récentes.

Les images seront, selon les profils, de type I, IP ou IPB.

L'estimation des images non codées en Intra peut prendre en compte jusqu'à 32 images, passées ou futures, (au lieu de 1 ou 2 en MPEG-2). Elle est de ce fait plus efficace.

La partie 10 de la norme autorise l'utilisation de plusieurs types de blocs : des blocs 4 × 4, 4 × 8, 8 × 4, 8 × 8, 8 × 16, 16 × 8 et 16 × 16.

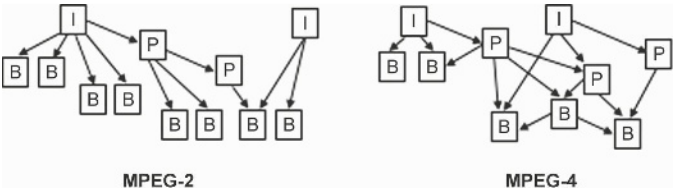


Figure 4.3.
Les relations entre les différents types d'images (d'après document Apple).

L'estimation de mouvement est plus fine, elle peut atteindre une précision du quart de pixel. Dans le cas, assez fréquent, où l'on peut détecter un déplacement global de l'image (panoramique, *travelling*), on ne transmet qu'un seul vecteur mouvement pour l'ensemble de l'image, c'est le système « *Global Motion Compensation* » qui permet d'économiser sur le crédit de bits nécessaire à la transmission de l'information de déplacement.

Le codage entropique et le codage à longueur variable peuvent être de type probabiliste. Le premier est dit « *Context-based Adaptative Binary Arithmetic Coding* » (CABAC), le second « *Context-based Adaptative Variable Length Coding* » (CAVLC) Ils choisissent, parmi une sélection disponible, un modèle qui dépend statistiquement des derniers symboles codés.

Enfin une boucle de contrôle de l'effet de bloc (talon d'Achille des systèmes découpant l'image en petits rectangles) a été ajoutée.

On estime que l'amélioration apportée par l'ensemble de ces outils est importante puisque le débit peut être

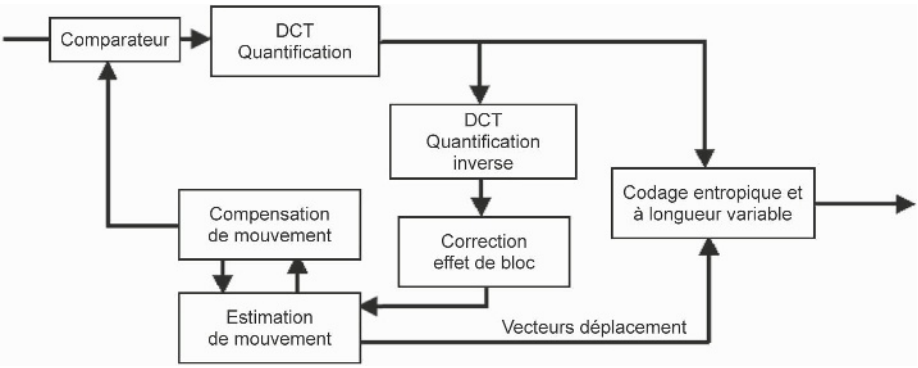


Figure 4.4.
Synoptique du codeur : adjonction d'une boucle de correction d'effet de bloc.

divisé par un coefficient de l'ordre de 2 par rapport à celui de MPEG-2.

Il est à noter que des images fixes texturées sont codées par ondelettes ce qui rend faciles les opérations de mise à l'échelle (décrites dans le chapitre 6.2).

4.3.2 Les *Sprites*

Les *sprites* (fantômes) constituent une classe particulière d'objets visuels. Ce sont des objets statiques de grandes dimensions qui apparaissent dans une scène de manière parcellaire. Il s'agit souvent d'un arrière plan : décor d'un studio, paysage, tribunes d'un stade, etc. Les *sprites*, dans la mesure où ils sont statiques, peuvent être transmis une seule fois en début de séquence.

Il suffira par la suite d'en sélectionner des zones afin de reconstituer l'arrière plan de la scène sur lequel on viendra placer les objets de premier plan. La sélection de la zone à prendre en compte est effectuée grâce à des informations de déplacement, prenant en compte les paramètres géométriques et optiques de la caméra de prise de vues (position, axe de prise de vue, focale de l'objectif), qui seront transmises par le descripteur de scène,

Cette fonction est particulièrement sollicitée dans les jeux vidéo, mais elle peut également être utilisée pour des débats en studio, un journal télévisé…

La possibilité de segmenter une prise de vues réelle en plusieurs objets est aujourd'hui à l'ordre du jour, soit grâce à des informations de profondeur Z relevées à la prise de vues, soit grâce à des algorithmes de suivi de forme.

On pourra donc « séparer » l'image d'un journaliste de celle de la côte bretonne victime d'une marée noire qu'il commente afin d'échanger, selon le contexte, les bits du flux (donc la résolution), entre les deux objets

« journaliste » et « côte souillée », sans dépasser le débit toléré.

On pourra de même séparer à l'intérieur des images les zones sombres, qui tolèrent une quantification réduite, des zones éclairées, qui exigent une quantification plus fine.

C'est ce qu'on nomme quantification adaptative.

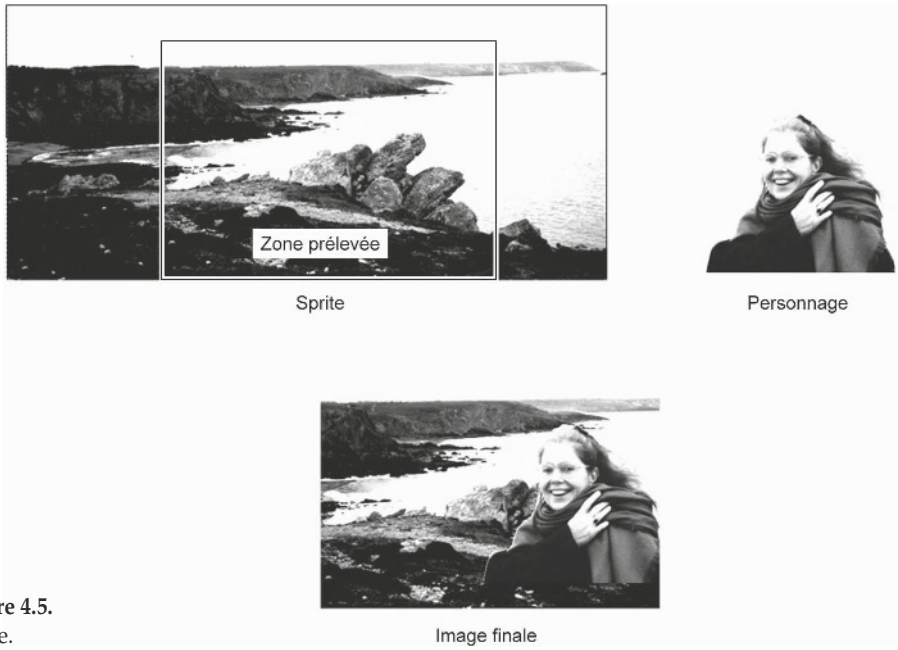


Figure 4.5.
Sprite.

4.3.3 *Advanced Simple Profile* (ASP)

C'est un des profils les plus utilisés. Il se décline en 6 niveaux avec des débits pouvant aller de 128 kb/s (niveau 1) à 8 Mb/s (niveau 6), des résolutions allant de QCIF à 720 × 576.

Ce profil est destiné à l'enregistrement de programmes, en particulier sur DVD, avec une haute qualité et avec des débits plus faibles que ceux autorisés classiquement par MPEG-2.

Il ne s'embarrasse pas de tout ce qui peut toucher à l'interactivité mais utilise les outils permettant simultanément qualité et faible débit.

Il peut utiliser des images de type P et B et supporte les formats entrelacés ; il bénéficie de plusieurs outils avancés que sont la compensation globale de mouvement, l'estimation de mouvement au quart de pixel ainsi que de la quantification adaptative.

Par contre il ne dispose pas de possibilités de mise à l'échelle ni de gestion de *sprites*.

C'est véritablement un format efficace d'enregistrement de programmes. On le trouve au cœur des encodeurs les plus répandus comme DivX ou son concurrent XviD.

4.3.4 Le codage des objets de synthèse

Le codage des objets de synthèse a constitué dès le début une cible primordiale pour les experts dont beaucoup venaient de l'univers de l'infographie.

Ces objets synthétiques, importants aujourd'hui, le deviendront certainement encore plus demain, que ce soit dans le domaine de jeux ou dans celui des présentateurs virtuels.

Les codages MPEG-4 les décrivent à l'aide de quatre types d'outils :

- la description synthétique du visage et du corps humain ;
- l'animation du visage et du corps ;
- le codage par maillage 2D ou 3D (*mesh*) ;
- le codage des textures.

L'animation faciale est réalisée d'une manière très efficace. On utilise un modèle de visage standard neutre, totalement inexpressif (muscles détendus, regard de face, bouche fermée…), ainsi que quelques

paramètres (*Facial Animation Parameters*) qui permettront de le modifier.

Ces paramètres d'animation prennent en compte la position de points de référence sur un modèle maillé du visage ainsi que des configurations labiales liées aux différents phonèmes. On transmettra ensuite, en fonction du type de personnage recherché, des textures de visage pour « habiller » le modèle.

L'animation corporelle prévue dans le profil « Simple Face and Body Animation » dispose d'une panoplie de méthodes du même genre, très largement rodées dans les jeux interactifs.

Les maillages 2D ou 3D décomposent les surfaces planes ou volumiques en triangles, ou mailles (*meshes*) qui sont simplement décrits par les coordonnées de leurs sommets qui constituent les nœuds du maillage. La modification de quelques dizaines de paramètres de ces nœuds permet de manipuler économiquement des modèles « fil de fer » d'objets de forme assez complexe. Elle permet notamment de déterminer, suivant le point de vue, les parties de l'objet qui ne seront pas perçues par le spectateur.

4.4 LES CODAGES MPEG-4 AUDIO

4.4.1 Buts et principes de base

Le standard MPEG-4 (ISO/IEC 14496-3) a été proposé en 1999. Il est, « orienté objets ».

Un objet sonore peut être constitué de plusieurs canaux audio habituels au sens où l'on parle du canal d'un haut-parleur. Plusieurs objets peuvent être combinés mais un objet ne peut pas (ou très approximativement) être décomposé en plusieurs objets.

La partie audio de la norme vise à gérer, pour une très large gamme d'applications, du téléphone mobile à la diffusion haute définition, tout système audio, quel

qu'en soit le débit ; elle concerne la représentation, la manipulation et la combinaison des sons :

- sons naturels, définis comme ceux provenant de microphones ;
- sons synthétiques pour lesquels seront mis en œuvre des outils permettant de coder musique ou parole grâce à une symbolique définie conduisant à des débits très faibles.

Pour la parole il s'agit de « Voix de synthèse », éventuellement liée à un texte, TTS (*Text To Speech*) ou une animation faciale (FA, aujourd'hui très performante).

Pour la musique on utilise la description SA (*Structured Audio*), avec des outils comme SAOL (SA *Orchestra Language* : description des méthodes de synthèse et des effets sonores) ; SASBF (SA *Sample Bank Format* : constitution de dictionnaires de motifs sonores instrumentaux ou *Sample Banks*) ; SASL (SA *Score Language* : organisation temporelle des éléments qui inclut l'interface MIDI – *Musical Instrument Digital Interface*).

Il est à noter que MPEG-4 ne standardise pas une méthode de synthèse sonore (parole ou musique) mais la possibilité de décrire une de ces méthodes (existante ou future) dont on est certain qu'elle pourra ainsi être utilisée par le décodeur.

On peut également comprendre qu'aucun codeur ne pourra prendre en charge l'ensemble de ces exigences et qu'il sera nécessaire de faire appel à plusieurs types de codeurs spécialisés. Une partie localisation 3D permet par ailleurs de créer des environnements dans lesquels sont localisées des sources sonores.

MPEG-4 Audio propose également des fonctions avancées, comme le contrôle de la vitesse de lecture, le changement de tonalité, la résistance aux erreurs, ainsi que des fonctions de « choix d'échelon de qualité » ou de « facteur d'échelle » ou encore d'adaptabilité (*scalability*) ; il s'agit d'organiser les données en sous-ensembles tels

que l'on puisse toujours obtenir un son utilisable (mais de qualité moindre) quelles que soient les performances de la chaîne de transmission :

- possibilité d'élaborer, à l'intérieur du flux global, un flux faible débit, mais représentatif du signal et décodable ;
- possibilité de sélectionner une bande passante audio analogique réduite dans le flux global ;
- possibilité pour un décodeur de bas niveau de pouvoir utiliser, pour partie, un flux de niveau supérieur.

La norme MPEG-4 permet des codages, de 2 kbits/s à plus de 64 kbits/s par canal, qui assurent la transmission de sons dans toutes les qualités, depuis une voix synthétique jusqu'au son multicanal de haute qualité.

4.4.2 Le codage des objets sonores naturels : AAC, CELP, HVXC, HILN

Le codage AAC (*Advanced Audio Coding*) dont les premières applications ont concerné des améliorations des codages MPEG-1 (voir le chapitre 2.3) est au cœur des codages MPEG-4 audio. Il s'agit d'un codage de base qui ne présume pas du type de sons à transmettre. C'est le codage qui est utilisé, entre autres, pour les systèmes de radiodiffusion numérique DRM (Digital Radio Mondiale) et l'Ipod d'Apple.

La norme a été étendue à une version HE (*High Efficiency*) qui intègre l'AAC, avec des débits pouvant aller jusqu'à 256 kb/s.

Des outils plus spécifiques ont été élaborés. Ce sont plutôt les bas, voire les très bas débits, qui ont fait l'objet de recherches avancées. Celles-ci sont basées sur la décomposition du signal en composants qui sont définis par des modèles appropriés qui peuvent être facilement indexés dans une table grâce à quelques paramètres seulement ; on parle de codages

paramétriques. C'est le savoir faire résultant des travaux sur la synthèse de la parole qui a notamment permis l'élaboration de ces méthodes.

Les échantillons numérisés sont groupés en paquets appelés trames.

Pour la parole (*Speech Coding*), à des débits moyens de 6 à 16 kb/s, c'est un codage CELP (*Code Excited Linear Predictive*) qui est mis en œuvre. Il est mal adapté à la musique. Il s'agit en effet d'un codage des différents phonèmes grâce à ce qu'on nomme des « formants » modélisant l'action des cordes vocales (excitation) ainsi que celle du larynx et de la bouche (filtrage et mise en forme spectrale de l'excitation).

On met en œuvre une méthode dite VQ (*Vector Quantization*). À l'intérieur de chaque trame du signal audio, on constitue des groupes d'échantillons dont la moyenne constitue un vecteur dont on trouvera une représentation symbolique compatible dans un dictionnaire ou *codebook*.

Ce formant « excitation » est appliqué au filtrage de mise en forme afin de fournir un signal audio synthétique prédit. Celui-ci est comparé au signal réel pour affiner la prédiction.

Le codeur et le décodeur possèdent évidemment le même dictionnaire ; il suffit donc, pour reconstruire le signal audio initial, de transmettre la référence de l'entrée choisie pour l'excitation ainsi que celle du filtre sélectionné.

On a pu réussir à coder de manière assez satisfaisante la parole à de très bas débits, de 2 kbit/s à 4 kbit/s, grâce au codage HVXC (*Harmonic Vector eXcitation Coding*). Il s'agit d'une version du codage précédent dans lequel c'est à l'enveloppe de l'erreur de prédiction, préalablement transférée dans le domaine fréquentiel, qu'est appliquée la quantification vectorielle VQ.

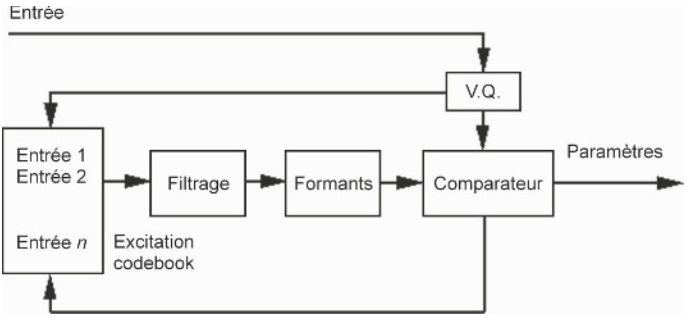


Figure 4.6.
Synoptique d'un codeur CELP.

Ces deux types de codage peuvent supporter, grâce au système BSAC (*Bit Sliced Arithmetic Coding*) des flux annexes assurant l'effet d'échelle (*scalabilty*).

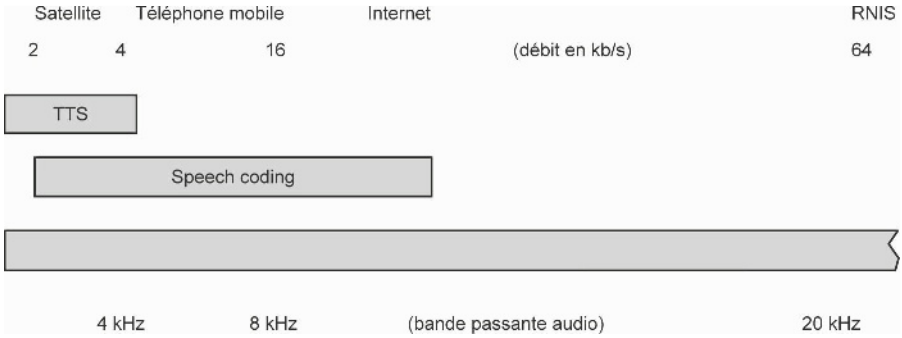


Figure 4.7.
Débit des codages MPEG-4.

Pour les sons musicaux, on utilise un codage HILN (*Harmonics Individual Lines and Noise*) dans lequel les composants sont la fondamentale et ses harmoniques. Chacun de ces composants peut être indexé par trois paramètres (fréquence, amplitude et phase) dans les banques de motifs (SBF : *Sample Bank Format*). Il est également pris en compte une composante « bruit ».

On peut ajouter un composant dit OCS (*One Channel Stereo*) afin de représenter efficacement les informations stéréophoniques.

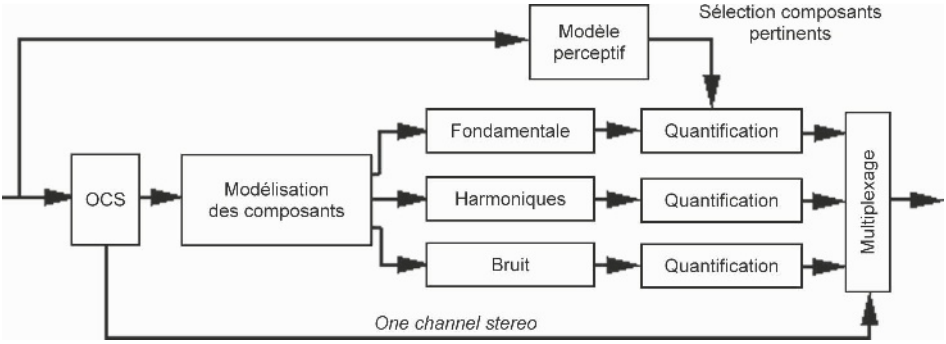


Figure 4.8.
Synoptique d'un codeur harmonique.

4.4.3 Le codage des objets sonores de synthèse

Ces codages font appel à un ensemble d'outils regroupés sous le titre SA (*Structured Audio*). Ils mettent en œuvre des algorithmes donnant accès à des débits extrêmement faibles, de moins de 1 b/s à 10 kb/s. Ils permettent également de transmettre de manière concise des paramètres de post-production et d'effets sonores.

Les principaux outils sont :

- SASBF (*Structured Audio Sample Bank Format*) : fournit une bibliothèque d'échantillons de sons PCM (*wavetable*) permettant de réaliser économiquement (en termes de débit) la synthèse d'instruments de musique traditionnels ;
- SAOL (*Structured Audio Orchestra Language*) : permet de constituer un orchestre à partir de ces synthèses ;
- TTS (*Text-To-Speech*) : permet de faire correspondre à un texte, quelle que soit la langue (grâce à l'utilisation du dictionnaire phonétique international), une suite de phonèmes liés à une base de données d'éléments sonores (intégrant des notions d'age, de sexe, de débit…) qui permettent la synthèse de la voix en tenant compte de la prosodie. Ces éléments peuvent également être liés à des outils d'animation faciale.

Voici quelques applications possibles de cette technique : lecture automatique de textes pour les mal voyants, messages synchronisés avec une animation faciale pour des publications multimédia, voix d'avatars en « réalité virtuelle », doublage de films à partir d'un fichier texte, etc.

4.4.4 La composition des scènes

Comme les objets image, les objets sonores individualisés peuvent être combinés spatialement et temporellement grâce à un descripteur de scène (*scene descriptor*) transmis au début du programme au format BIFS (*Binary Format For Scenes*). Ce descripteur est envoyé au récepteur en même temps que les différents objets qu'il permet de ré-assembler.

L'auditeur aura, quant à lui, la possibilité de choisir les objets qu'il veut écouter et d'en éliminer d'autres. On peut ainsi envisager le choix de la langue pour un commentaire ou un dialogue, la suppression d'un commentaire jugé trop intrusif, la possibilité pour un violoncelliste de jouer sa propre partition dans un quatuor à cordes, ou encore la sélection exclusive des dialogues d'une séquence cinématographique pour faciliter l'apprentissage d'une langue…

4.4.5 Les dernières avancées : le codage HE-AAC

Ce codage, aussi connu sous le nom de AAC Plus, résulte de l'addition de deux nouvelles techniques au noyau AAC.

Il s'agit, dans la version 2 actuelle (ISO/IEC 14496-3), d'une part, d'une « reconstruction du spectre par duplication » ou SBR (*Spectral Band Replication*), d'autre part, d'un « codage stéréo paramétrique » ou PS (*Parametric Stereo*).

La technologie SBR repose sur la constatation qu'il existe une forte corrélation entre la partie « hautes

fréquences » du spectre sonore et sa partie « basses fréquences ». Ceci n'est certainement une surprise lorsque l'on pense à la structure harmonique sonore ; mais cela fonctionne aussi pour les signaux de type bruit de fond.

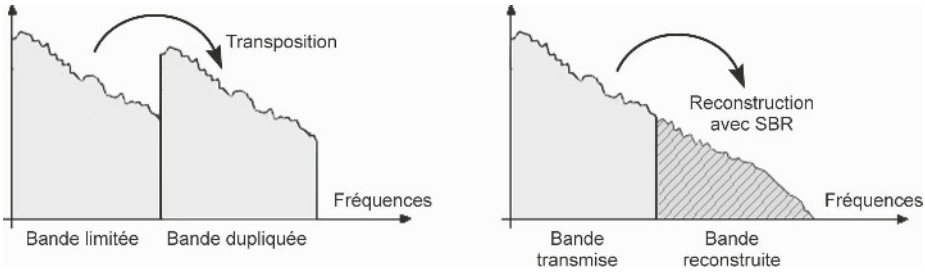


Figure 4.9. Schéma de principe de la SBR (d'après document UER/EBU).

On ne transmet donc qu'une bande inférieure limitée du spectre du signal et le décodeur reconstruit la partie supérieure grâce à une simple duplication/transposition assistée de quelques éléments d'information, tels que l'enveloppe spectrale du signal original, transmis à très bas débit qui permettent une mise en forme plus précise pour la reconstruction des hautes fréquences.

Cette solution permet de transmettre un signal à large spectre avec des débits très faibles.

Débit stéréo (b/s)	Bande passante AAC (Hz)	Bande passante SBR (Hz)
20 000	0 - 4 500	4 500 - 15 400
32 000	0 - 6 800	6 800 - 16 900
48 000	0 - 8 300	8 300 - 24 900

Tableau 4.1. Débits et bandes passantes (d'après document UER/EBU).

La technologie PS est basée sur le mode de perception stéréophonique de l'être humain.

Le codeur transmet un signal mono élaboré à partir du signal initial ; le codeur extrait également des informations, transmises à bas débit qui permettent au

décodeur de reconstituer un signal stéréophonique de grande qualité.

Les informations paramétriques correspondent à celles que notre cerveau retient pour élaborer une représentation spatiale des sources à partir des signaux sonores reçus par les deux oreilles :

- *Inter-Channel Intensity Difference* (IID) qui spécifie la différence de niveaux entre les deux signaux ;
- *Inter-Channel Cross Correlation* (ICC) qui spécifie la cohérence entre les deux signaux ;
- *Inter-Channel Phase Difference* (IPD) qui spécifie le délai d'arrivée des deux signaux.

On estime que le codage HE-AAC permet la transmission d'un signal 5+1 de bonne qualité à 128 kb/s, d'un signal stéréo de qualité proche du CD à 32 kb/s et que 16 kb/s suffisent pour une transmission stéréo satisfaisante.

Quel chemin parcouru depuis les 150 à 200 kb/s nécessaires pour une transmission de qualité CD avec les codages MP3 ou Musicam du début des années 90 !

4.5 LE CODAGE DES PROGRAMMES VIDÉO : MPEG-4 *PART 10*, AVC, H.264

Le codage AVC (ISO/IEC 14496-10), est le fruit du travail de deux familles d'experts réunis au sein du JVT (*Joint Video Team*). D'une part, le *Video Coding Experts Groups* (VCEG) de l'ITU (*International Telecommunications Union*) qui s'intéresse aux applications de télécommunications comme la visioconférence. D'autre part, le comité MPEG de l'ISO/IEC, qui s'intéresse aux applications de type Télévision. Ils ont finalisé en 2003 ce standard connu sous trois appellations : ITU-T H.264, ISO MPEG-4 part 10 ou encore AVC (*Advanced Video Coding*).

Il concerne l'encodage de programmes multimédia, pour la transmission ou l'enregistrement, sur une large

gamme de débits, depuis la transmission sur réseaux jusqu'à la diffusion HD.

En raison de cet objectif bien précis, très actuel, mais limité, il ne s'encombre pas des fonctionnalités les plus évoluées et les plus prometteuses, à plus long terme (ou les plus irréalistes), retenues par MPEG-4 comme, simplement, la gestion d'objets audiovisuels.

Le codage AVC n'est pas en rupture avec les codages MPEG précédents desquels il a hérité la décomposition des images en blocs, macroblocs et tranches, les méthodes de prédiction temporelles (images de type I, P et B) avec compensation de mouvements ainsi qu'une transformation de type DCT.

Son avantage résulte d'un ensemble d'améliorations ainsi que d'une plus grande flexibilité dans la mise en œuvre des différents « outils » qui ont permis de gagner un coefficient d'efficacité estimé aujourd'hui entre 2 et 3 (selon les versions de codeurs et selon le type de séquences à traiter) par rapport à ses prédécesseurs.

4.5.1 Profils et niveaux

Plusieurs profils ont été définis :

- *Baseline* (BP) : applications temps réel ; visioconférence ; réception sur mobiles ;
- *Extended* (XP) : diffusion sur réseaux (*streaming*) ;
- *Main* (MP) : diffusion TV standard (entrelacé) ;
- *High* (HP) : diffusion HDTV.

Ces profils ont bien entendu été subdivisés en niveaux.

Le profil de base comporte 4 niveaux. Le niveau 1 concerne le format QCIF à 15 images par seconde avec un débit de 64 kb/s. Le niveau 4 passe à un débit de 768 kb/s pour un format CIF à 25 ou 30 images par seconde.

Le Profil Haut (Télévision à Haute Définition) comporte également quatre niveaux :

- HP (*High Profile*) retient un échantillonnage 4 :2 :0/8 bits ;
- H10 retient un échantillonnage 4 :2 :0/10 bits ;
- H4:2:2 retient un échantillonnage 4:2:2/10 bits ;
- H4:4:4 retient un échantillonnage 4:4:4/12 bits.

Tableau 4.2. Caractéristiques principales des niveaux et profils.

Niveaux	Résolution	Fréquence images par seconde	Débit maximum *Baseline Profile* *Extended Profile*	Débit maximum *High Profile*	Débit maximum *High Profile* 4:2:2 / 4:4:4
1	QCIF	15	64 kb/s	80 kb/s	256 kb/s
1b	QCIF	15	128 kb/s	160 kb/s	512 kb/s
1.1	CIF ou QCIF	7,5 (CIF) 30 (QCIF)	192 kb/s	240 kb/s	768 kb/s
1.2	CIF	15	384 kb/s	480 kb/s	1536 kb/s
1.3	CIF	30	768 kb/s	960 kb/s	307 kb/s
2	CIF	30	2 Mb/s	2,5 Mb/s	8 Mb/s
2.1	HHR (Half Horizontal Resolution) 352 × 480/352 × 576	30/25	4 Mb/s	5 Mb/s	16 Mb/s
2.2	SDTV	15	4 Mb/s	5 Mb/s	16 Mb/s
3	SDTV	30/25	10 Mb/s	12,5 Mb/s	40 Mb/s
3.1	1280 × 720p	30	14 Mb/s	17,5 Mb/s	56 Mb/s
3.2	1280 × 720p	60	20 Mb/s	25 Mb/s	80 Mb/s
4	HD 1080i / 720p	60p/30i	20 Mb/s	25 Mb/s	80 Mb/s
4.1	HD 1080i / 720p	60p/30i	50 Mb/s	62,5 Mb/s	200 Mb/s
4.2	HD 1920 × 1080p	60p	50 Mb/s	62,5 Mb/s	200 Mb/s
5	2k	72	135 Mb/s	168 Mb/s	540 Mb/s
5.1	4k	120/30	240 Mb/s	300 Mb/s	960 Mb/s

4.5.2 Les avancées

Le codage AVC permet de mettre en œuvre, au moins pour le Profil Principal (*Main*), la plupart des outils susceptibles d'améliorer son efficacité et de lui offrir un champ d'application élargi :

- Il traite simultanément des séquences d'images en modes entrelacé ou progressif.
- Il peut utiliser des images I, P et B avec gestion dynamique des GOPs (*Dynamic* GOP) ; cette technique permet à l'encodeur d'adapter la structure ainsi que la longueur des GOPs en fonction de la nature du contenu de la séquence : on pourra ainsi avoir des GOPS à trois, sept ou même 15 images de type B.
- Il peut également gérer les macrobloc de 16×16 pixels de manière dynamique puisque ceux-ci peuvent segmentés (macrobloc sub-portion) en sous-blocs 16×8, 8×16, 8×8, 8×4, 4×8, et 4×4 (*Tree Structured Motion Compensation*), ce qui augmente les possibilités de retrouver des éléments identiques, avec des possibilités de compensation de mouvement (locale ou globale) au quart de pixel.

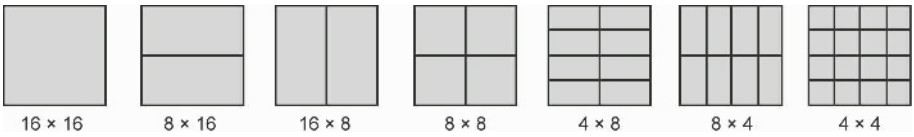


Figure 4.10. Segmentation d'un macrobloc 16×16 en sous-blocs.

- Il sait mettre en œuvre une prédiction intra-image dans le domaine spatial (nouvel outil) qui est assez subtile. Les différents macroblocs d'une tranche peuvent être prédits à partir des macroblocs précédemment codés. Les 64 sous-blocs 4×4 du macrobloc courant (macrobloc « cible ») sont définis par référence à ceux du macrobloc précédent (macrobloc « source ») qui peuvent être individuellement déplacés et réorganisés en utilisant des vecteurs déplacement sur 8 directions afin d'obtenir la meilleure coïncidence possible entre les deux macroblocs.

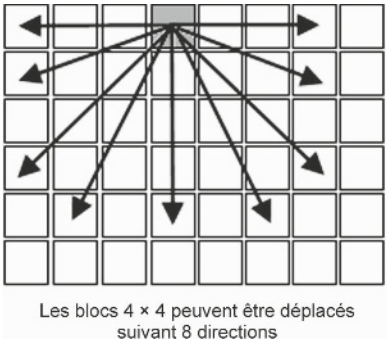


Figure 4.11.
Déplacement
des blocs 4x4.

- Il est également possible, dans le cas de zones homogènes de simplifier le processus en réalisant une prédiction sur un macrobloc 16 × 16 avec 3 vecteurs déplacement (horizontal, vertical, diagonal). On peut penser à une méthode de compression fractale (voir chapitre 6.4) simplifiée à laquelle je reprends les termes cible et source.
- Il bénéficie des codecs CAVLC et CABAC.
- Enfin, le décodeur met en œuvre un filtre de réduction des effets de bloc (*deblocking filter*) qui « adoucit » les frontières entre blocs dans le domaine spatial (après la transformation inverse).

Ce sont ces avancées qui permettent de multiplier par deux ou trois l'efficacité de la réduction de débit par rapport à celle de MPEG-2, au prix d'une puissance de calcul multipliée par 8 pour le codeur et par 4 pour le décodeur. Les progrès des microprocesseurs ont permis de faire face à ces exigences et les décodeurs HD apparaissent à des prix raisonnables. Les performances des codeurs évoluent sans cesse, qu'il s'agisse d'évolutions concernant le matériel (notamment les circuits intégrés dédiés ou ASICs, *Application Specific Integrated Circuit*) ou les logiciels.

Le codage AVC, performant, souple, et constitué en standard international est en train d'être très largement adopté aux dépends de son concurrent « propriétaire » VC-1 (*Video Coding 1*), proposé par Microsoft, qui était

pourtant assez proche aussi bien quant aux techniques mises en œuvre que quant aux performances.

C'est sans doute dans le domaine de la diffusion de la Haute Définition, aujourd'hui réalité, que les atouts du codage AVC sont les plus évidents. Alors qu'un débit d'environ 19 Mb/s (valeur retenue pour les canaux HD aux USA) était nécessaire en diffusion avec un codage MPEG-2, on peut espérer descendre aux environ de 6 Mb/s avec un codage AVC High Profile. Ceci signifie que le nombre de programmes HD sur un transpondeur satellite passera de 2 à 6 (avec l'espérance de passer à 8 ou 10 au fur et à mesure des améliorations des codeurs). Intéressant bien sûr !

La puissance des encodeurs intéresse évidemment aussi les distributeurs de programmes de télévision sur ADSL (IPTV).

Ce codage AVC High Profile a été retenu pour les DVD à Haute Définition, qu'il s'agisse du Blu-ray DVD ou de son concurrent HD-DVD, aujourd'hui disparu du paysage.

Il a par ailleurs été retenu par le Projet DVB (*Digital Video Broadcasting*) pour la diffusion HD par satellite ; la BBC et SAT-1 (Grande Bretagne), DirectTV (USA), Euro 1080, Premiere (Allemagne), Sky Italia (Italie) devraient être les premières chaînes satellitaires concernées.

La France a choisi AVC pour la diffusion HD sur la TNT. D'autres pays devraient suivre ; citons en vrac le Brésil, l'Estonie, la Lituanie, la Corée, le Japon…

En juillet 2006, Sony et Panasonic ont conjointement annoncé la mise sur le marché de caméscopes AVC-HD. Il s'agit d'un format « full HD » (1920 × 1080i ou 1280 × 720p) qui utilise le profil AVC HP (*High Profile*) avec un échantillonnage 4:2:0/8 bits et un débit vidéo pouvant atteindre 18 Mb/s. L'enregistrement ne s'effectue plus sur une bande magnétique mais sur un disque optique (DVD Blu-ray de 8 cm de diamètre) ou un

disque magnétique de 30 Go, chez Sony, (modèles HDR-UX1E et HDR-SR1E), ou encore sur des cartes mémoire flash SD de 4 Go, chez Panasonic.

Enfin en septembre de la même année Panasonic a présenté un enregistreur sur cartes P2 (la version professionnelle des cartes SD) fonctionnant en mode « AVC-I only » (AVC-I) ; cet enregistreur est disponible depuis la fin 2007. Ce format supporte des séquences 4:2:2 1920 × 1080i et 1280 × 720p au débit de 100 Mb/s. Un débit limité à 50 Mb/s permet d'enregistrer avec une résolution limitée à 1440 pixels par ligne… ce qui est déjà fantastique pour des applications de reportage.

5 LES ENREGISTREURS NUMÉRIQUES ET LA RÉDUCTION DE DÉBIT

Je n'aborderai pas ici la description des serveurs dont les possibilités techniques font qu'ils n'ont pas (ou peu) de problèmes pour s'adapter à l'enregistrement de séquences d'images numérisées, de quelque type qu'elles soient. Je n'aborderai que le cas de ce qu'on appelait jusqu'à une date récente « magnétoscopes » et pour lesquels on devra forger un nouveau terme (Discoscope ? Ramscope ?) depuis que la bande magnétique est tombée en désuétude pour être remplacée par des disques magnétiques ou optiques, ou encore par des cartes à mémoire flash.

En 1976, on a commencé à se poser la question de la faisabilité d'un magnétoscope numérique ; l'objectif fondamental était celui de la création de multiples générations pour un programme vidéo, sans perte de qualité. C'était la condition d'une post-production évoluée.

En 1979, Ampex, Sony et Bosch présentèrent indépendamment le produit de leurs recherches. Ils comprirent rapidement la nécessité d'un standard universel. Ce standard, qui utilise la numérisation préconisée par la recommandation CCIR 601, 4:2:2 sur 8 bits (voir chapitre 2.2), conduisit en 1986 (bientôt un quart de siècle !), à la présentation par Sony, rapidement suivi par Bosch, du magnétoscope DVR-1000 au format D1 (Digital 1) qui utilisait des cassettes « trois quart de pouce ».

Tout était à découvrir. Tout fut analysé ou inventé ; notamment les codes numériques (NRZI, Manchester, Miller et Miller 2, EFM…), la segmentation des données sur la bande en secteurs, ainsi que les systèmes de masquage d'erreurs, sans oublier les codes correcteurs d'erreurs (code CRC, *Cyclic Redundancy Check*, code Reed-Solomon) qui sécurisent, au prix d'un accroissement du débit par des bits de redondance, les données face aux risques inhérents à l'enregistrement magnétique.

En 1988, Ampex proposa, pour répondre à la demande du marché américain, le format D2 qui fonctionnait en composite. Un assez grand nombre d'exemplaires furent effectivement vendus aux USA, pour venir en relève des magnétoscopes analogiques NTSC, 1 pouce, à bout de potentiel. Sony développa alors une machine à ce standard. Mais les systèmes « composantes » avaient fait la preuve de leurs avantages et étaient à cette époque en train de reléguer les systèmes « composites » au rayon musée. Le D2 ne pouvait avoir qu'une espérance de vie limitée.

En 1990 Panasonic réussit le tour de force de commercialiser un caméscope numérique (format D3) sur cassette un demi pouce ; il était également en composite et n'eut de ce fait que peu de succès.

Il était d'évidence devenu impératif de réaliser des magnétoscopes en composantes et de faible encombrement.

Les nouvelles générations de magnétoscopes devaient forcément utiliser la réduction de débit.

5.1 L'ENREGISTREMENT NUMÉRIQUE AVEC COMPRESSION, PREMIER ACTE : DIGITAL BETACAM

En 1993, Sony a introduit le premier format numérique avec compression. Il s'agissait du « Digital Betacam » (souvent nommé DigiBeta). J'ai été complètement impressionné ! Ce fut un succès !

Le Digital Betacam enregistre un signal 4:2:2 sur 10 bits, considéré comme la référence ultime en définition standard (SDTV), avec une compression Intra de type DCT (ça ressemble au JPEG, mais c'est un format « propriétaire » développé en parallèle aux algorithmes JPEG finalisés en 1991) avec un très faible taux de compression 2,3:1 (on en était aux balbutiements des techniques de réduction de débit, souvent totalement incomprises et fort décriées par les concurrents !) pour

un débit vidéo de 90 Mb/s. Il supporte également 4 canaux audio en PCM à 48 kHz. Le DigiBeta utilise la bande demi pouce ainsi que la cinématique éprouvées de la gamme Betacam antérieure.

Le faible taux de compression présentait des avantages de robustesse et transparence. Sony a présenté, à cette époque, lors du Symposium de Télévision de Montreux une séquence (Alice et 99 clones défilant au pays des merveilles électroniques) présentant 100 niveaux successifs de cycles de compression/décompression en « Betacam Numérique » sans que le moindre artefact ne soit perceptible à l'œil.

Pour être tout à fait honnête il faut signaler qu'à cette même date Ampex avait proposé un format qui mettait également en œuvre une compression de type DCT 2:1 et qui était dénommé – hasard ou humour – DCT pour *Digital Component Tape*… Mais ces machines n'ont jamais été commercialisées pour la vidéo ne serait-ce qu'en raison du fait qu'elles utilisaient des cassettes trois-quart de pouce trop encombrantes. Machines superbes, elles sont devenues des mémoires de masse pour systèmes informatiques. Le développement du Digital Betacam ouvrait une ère nouvelle…

5.2 L'ENREGISTREMENT NUMÉRIQUE AVEC COMPRESSION, DEUXIÈME ACTE : DV, DVCPRO, DVCAM ET BETACAM SX

En 1993, également (mais les premières machines apparurent sur le marché en 1996) tous les plus grands constructeurs d'électronique domestique définirent un format grand public avec compression, dénommé DV (*Digital Video*, tout simplement), basé sur une bande magnétique de largeur 6,35 mm (1/4 de pouce). Un projet très audacieux mais dont l'histoire a prouvé amplement qu'il était réaliste.

Le format DV utilise une réduction de débit « intra » de type JPEG de facteur 5:1 (mais ce n'est pas du JPEG).

Dans les systèmes, à 50 trames par seconde et 625 lignes, le signal vidéo est le 4:2:0 : les échantillons de chrominance ne sont présents qu'une ligne sur deux ; dans les systèmes à 60 trames par seconde et 525 lignes, le signal vidéo est de type 4:1:1 : les échantillons de chrominance sont présents sur toutes les lignes, mais seulement pour 1 pixel sur 4. Dans les deux cas, la quantification est effectuée sur 8 bits.

Le choix d'une compression « intra » a été retenu pour ce format grand public afin de permettre des solutions de montage simples.

Le débit vidéo est finalement réduit de 125 Mbits/s (4:2:0 ou 4:1:1/8 bits, sans les suppressions ligne et trames) à 25 Mbits/s, mais le débit total enregistré s'élève à 41,85 Mbits/s (avec l'audio, le code de correction d'erreurs et les données auxiliaires). En ce qui concerne l'audio, le DV peut traiter soit 4 pistes échantillonnées à 32 kHz codées sur 12 bits, soit 2 pistes échantillonnées à 48 kHz codées sur 16 bits.

Le format « Betacam SX » fut introduit par Sony en 1996 comme une alternative meilleure marché au Digital Betacam.

Il met en œuvre une compression « inter » MPEG-2 de type 4:2:2P@ML. La puissance de la réduction de débit MPEG permet de passer pour le signal vidéo d'un débit de 156 Mb/s (4:2:2/8 bits sans les suppressions) à 18 Mb/s. Le GOP est court, 2 images : I, B, I, B…

Il enregistre également 4 canaux audio PCM 48 kHz /16 bits pour un débit total de 3,6 Mb/s.

Le code correcteur d'erreurs vidéo a été particulièrement soigné avec 42 % de bits de redondance (contre 19 % en DV et 22 % en Digital Betacam) afin de sécuriser au maximum les images : un argument commercial de poids quand on sait que Sony positionna ce format sur le reportage d'actualité.

Le débit total enregistré atteint 40 Mbits/s.

Le choix d'une compression « inter » entraîne, même si le GOP est très court, une complication du montage.

Chaque image « bidirectionnelle » B dépend en effet des images I précédente et suivante. Tout point de coupe laissera donc une image « orpheline » qui ne pourra plus être reconstruite.

La parade à cet inconvénient majeur a été trouvée au prix d'une relative complexification des machines. Celles-ci possèdent une tête de prélecture (*pre-read*) qui décode l'image B en image Bu (bidirectionnelle-unidirectionnelle : un oxymore !), cette nouvelle image qui ne dépend plus que de l'image I qui la précède ou de celle qui la suit (suivant l'emplacement du point de montage choisi) remplace alors l'image B initiale. C'est un processus qui fonctionne très bien mais qui complexifie les machines et qui était bien entendu inimaginable pour les machines grand public de type DV qui devaient demeurer simples.

En cette même année 1996, Panasonic proposait, sous le nom de DVCPRO (également connu sous l'appellation D7), une déclinaison du DV vers le haut, destinée à une utilisation professionnelle. Les modifications étaient mineures, pour l'essentiel une multiplication par 1,8 de la vitesse de défilement ce qui sécurise l'enregistrement en portant la largeur de piste à 18 µm contre 10 en DV. Le signal reste identique à celui du DV (mêmes algorithmes de compression, mêmes débits). Par contre, les cassettes sont différentes ce qui entraîne une incompatibilité (certainement volontaire) avec les autres machines DV.

Sony suivit peu après un chemin analogue avec le format DVCAM, dont la vitesse de défilement était multipliée par 1,5. Les dimensions des cassettes sont identiques à celles du DV, ce qui permet une assez large compatibilité avec les machines à ce format, mais pas avec les machines DVCPRO !

5.3 LE RAPPORT DE LA « TASK FORCE » SMPTE/UER

La « Task Force for Harmonized Standards for the Exchange of Program Material as Bitstreams » (harmonisation des standards pour l'échange de programmes sous forme de flux numériques) est un groupe de travail composé de nombreux experts, réunis sous l'égide de la SMPTE (*Society of Motion Picture and Television Engineers*) et de l'UER/EBU (Union Européenne de Radiodiffusion, *European Broadcasting Union*) afin de mettre un peu d'ordre dans le foisonnement attendu des formats numériques.

Ce groupe a publié un rapport très complet en septembre 1998 qui vise à assurer l'interopérabilité entre les systèmes de circulation des programmes.

Seule la partie concernant les débits vidéo nous intéresse ici.

Afin de limiter les dérives, la *Task Force* n'a retenu que deux types de compression le DV et le MPEG-2.

Des tests subjectifs de comparaison par rapport à un traitement en Betacam SP ou Digital Betacam, pris en référence, ont été réalisés sur des séquences normalisées ayant subi des compressions à deux débits différents : 18 et 50 Mb/s pour le MPEG-2 (4:2:2 P), et d'autre part 25 et 50 Mb/s pour les systèmes DV.

Trois scénarios ont été étudiés :

- une génération correspondant à une acquisition et une relecture simple ;
- quatre générations dont deux avec un décalage temporel et un décalage spatial ;
- sept générations avec de nombreux décalages spatiaux et temporels.

Les résultats (en DV ou 4:2:2 P) montrent que des débits de l'ordre de 18 à 25 sont suffisants lorsqu'on envisage une simple acquisition suivie d'une relecture ; qu'ils le

demeurent pour une post-production légère (disons jusqu'à trois ou quatre générations) ; mais qu'ils sont insuffisants si l'on envisage une post-production plus évoluée qui ne peut être valablement réalisée avec des débits inférieurs à 50 Mb/s en raison des artefacts provoqués par les compressions/décompressions successives.

Sony, Thomson (au moins théoriquement) ainsi que Snell & Wilcox (avec le système *Mole*, la taupe) avaient déclaré pouvoir contourner cet inconvénient en transmettant aux différents codecs (codeurs-décodeurs) de la chaîne de post-production les décisions prises par le codeur initial. Cela semblait en effet être efficace mais n'a pas donné lieu à de véritables applications.

5.4 L'ENREGISTREMENT NUMÉRIQUE AVEC COMPRESSION, TROISIÈME ACTE : DVCPRO-50 ET BETACAM IMX

Panasonic proposa très rapidement une réponse au rapport de la *Task Force* avec le DVCPRO-50. C'est une solution simple mais élégante ; il s'agit en fait de coupler deux codecs DV fonctionnant en tandem pour passer à un débit de 50 Mb/s. Le signal est alors du 4:2:2/8 bits avec un taux de compression faible de 3,3:1.

Il fallut attendre 2001 pour que Sony propose le format Betacam IMX fonctionnant en MPEG-2 4:2:2 Profile@ML, mais en n'utilisant que des images I avec un débit atteignant les fatidiques 50 Mb/s (taux de compression 3,3:1). Il n' y a bien entendu plus aucun problème pour le montage.

En raison de ce « retard à l'allumage », ainsi que de l'arrivée des serveurs, Betacam SX et IMX n'ont pas eu le succès de leurs glorieux et impériaux ancêtres : Betacam et Digital Beta.

5.5 L'ENREGISTREMENT NUMÉRIQUE AVEC COMPRESSION, QUATRIÈME ACTE : LA HAUTE DÉFINITION

En 1997, Sony, qui visait dès cette époque le cinéma électronique, introduisit sous la dénomination « HD-Cam », une version HD du Digital Betacam. Le débit vidéo atteignait 144 Mb/s, ce qui est faible quand on sait que la norme SMPTE 274 M (image HD à 1 080 lignes, 1 920 échantillons par ligne pour la luminance, 960 échantillons par ligne pour les signaux de chrominance, 10 bits) conduit à un débit de près de 1,5 Gb/s. Il a donc fallu faire des concessions…

Les images du HD-Cam subissent d'abord une réduction de la définition de 4:2:2 en 3:1:1 (1440 et 480 échantillons par ligne), elles sont quantifiées sur 8 bits, avant de subir une DCT de taux 5:1. La réduction totale de débit par rapport au signal HD (1920 × 1080, 10 bits) est de l'ordre de 10:1. Ces machines peuvent enregistrer 4 canaux audio 20 bits/48 kHz.

Les caméscopes baptisés « CineAlta » qui permettent un enregistrement en mode 24p ont effectivement été largement utilisés, avec succès, par les professionnels du cinématographe.

Panasonic avait proposé en 1994 un standard d'enregistrement numérique sur cassettes ∫ pouce noté D5. Il fonctionnait en 4:2:2/10 bits et se positionnait donc en concurrent du Digital Betacam de Sony. Mais, pari difficile à tenir, il n'utilisait pas de compression, ce qui entraînait un débit énorme de 270 Mb/s (certaines machines pouvaient même aller jusqu'à 323 Mb/s). Ce standard, mécaniquement complexe, difficile à finaliser, qui ne proposait pas de caméscope (Panasonic se contentait de vanter son caméscope D3), et dont le prix de revient du « consommable » (les cassettes) était très élevé n'a pas réussi à s'imposer face au Digital Betacam.

Quelques années plus tard, les méthodes de compression ayant largement fait leur preuve, le D5 est devenu le D5-HD. Il peut enregistrer des images 1920 × 1080 sur 8 bits ou 10 bits avec des taux de compression en mode Intra très modérés de 4:1 et de 5:1. Il est très utilisé par les laboratoires et les sociétés de post-production, notamment pour la masterisation en sortie des télécinémas HD.

En 2003, Sony proposa une nouvelle gamme dite HDCAM-SR (*Superior Resolution*) capable de répondre aux nouveaux défis de la cinématographie électronique.

Les magnétoscopes de cette gamme peuvent en effet enregistrer en mode SQ (*Standard Quality*) un signal RVB en 4:4:4 sur 10 bits avec un débit vidéo de 440 Mb/s et un taux de compression de 4:1 ou un signal HD 4:2:2 avec un taux de 2,7:1 Le débit total avoisine les 600 Mb/s

Ce format utilise une réduction de débit de type MPEG-4. Studio Profile.

Certaines machines peuvent fonctionner en mode HG (*High Quality*), avec un débit de 880 Mb/s et une vitesse de défilement de la bande doublée, ce qui rend possible l'enregistrement d'un signal RVB HD avec un taux de compression de 2:1 seulement, ou encore deux signaux 4:2:2 HD (on pense à des enregistrements stéréoscopiques) avec un taux de réduction de 2,7:1.

Le mode de compression mis en œuvre, MPEG-4. Studio Profile, est original puisqu'il comporte deux voies parallèles.

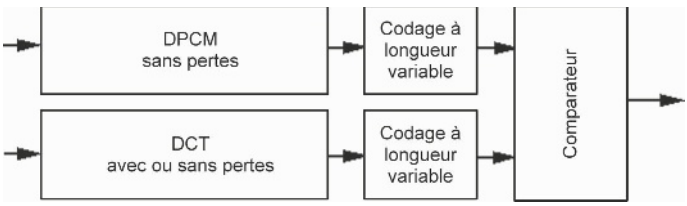


Figure 5.1. Compression HDCAM SR.

L'une d'elles utilise une compression assez classique par DCT, qui peut être avec ou sans pertes selon la complexité de l'image. La deuxième utilise une méthode DPCM (*Differential Pulse Code Modulation*) qui est par essence sans pertes. Un comparateur de sortie analyse les débits des deux voies et choisit la plus efficace. La DPCM est très efficace pour les grandes zones uniformes des images, elle permet dans la majorité des cas d'atteindre, lorsqu'elle peut entrer en fonction, un débit faible et, en conséquence, cela permet d'allouer plus de bits pour la quantification lorsque la DCT doit être choisie, ce qui en améliore la qualité.

Il ne faut pas oublier dans ce panorama les magnétoscopes : DVCPRO-100 de Panasonic, que l'on peut considérer comme l'assemblage de 4 magnétoscopes DV de base, et qui ont fait (et continuent à faire) une belle carrière pour les images HD en 720p.

Notons pour être complet que Toshiba et BTS avaient développé en 1993 un magnifique magnétoscope numérique HD sans compression (débit 1,2 Gb/s) qui a été amélioré par Philips sous le nom de Voodoo et commercialisé, lui aussi, pour des application « télécinéma » mais dont la fabrication a été arrêtée en 2003.

5.6 DES ENREGISTREURS « EXOTIQUES »

Sony et JVC ont adapté le codage DV à des supports déjà existants (ce qui réduit à zéro les frais de conception cinématique). Ces formats « exotiques » n'ont pas réussi à s'implanter.

Chez Sony il s'agissait d'enregistrer un signal DV sur une cassette précédemment prévue pour des signaux analogiques « Hi-8 ». Certaines machines étant susceptibles de relire les anciennes cassettes analogiques au format « Hi-8 ». C'était bien astucieux mais ça n'a pas marché ! Ce format est parfois appelé D 8.

Chez JVC, il s'agissait du « Digital-S » directement dérivé du « S-VHS » analogique. C'était en fait du DVCPRO-50 sur bande demi-pouce VHS. Ce format intéressant est également connu sous la dénomination D9.

5.7 LES FORMATS « PROSUMER » HDV ET AVC-HD

En 2003 apparut le format HDV, (Canon, JVC, Sharp, Sony) à vocation « grand public », mais dont les professionnels se sont rapidement emparés. Le prix et l'encombrement de ces petites merveilles les situent dans l'espace que les anglo-saxons nomment « Prosumer », à l'intersection des domaines *Professional et Consumer*.

Ce format n'a plus d'autre rapport avec le format DV que la cassette et le débit. La réduction de débit est en effet de type MPEG-2 avec des GOPs longs : 12 images pour les systèmes à 50 Hz et 15 pour les systèmes à 60 Hz ce qui a rendu plus délicat le montage et plus généralement la post-production en format natif, même sur des machines puissantes ! Aujourd'hui, plusieurs systèmes de montage supportent le signal HDV natif.

La numérisation est de type 4:2:0 sur 8 bits.

Deux modes sont disponibles : entrelacé 1440 × 1280 et progressif 1280 × 720, avec des débits respectivement de 25 et 19,7 Mb/s.

L'enregistrement audio stéréo est en MPEG Layer II (48 kHz, 16 bits) avec un débit de 384 kb/s (mode optionnel : 4 canaux à 96 kb/s par canal).

En juillet 2006, Sony et Panasonic ont introduit sous le nom d'AVC-HD un format « full HD » qui utilise le format AVC HP (*High Profile*) avec un échantillonnage 4:2:0/8 bits et un débit vidéo pouvant atteindre 18 Mb/s. L'enregistrement ne s'effectue plus sur une

bande magnétique mais sur un disque optique (DVD Blu-ray de 8 cm de diamètre) ou un disque magnétique de 30 Go, chez Sony, ou encore sur des cartes mémoire flash SD de 4 Go, chez Panasonic.

On notera qu'il s'agit, comme pour le HDV, d'un système de compression avec GOP long.

L'enregistrement audio s'effectue soit en Dolby Digital jusqu'à 5.1, soit en PCM jusqu'à 7.1.

Les séquences HD sont soit en 1920 × 1080i, soit en 1280 × 720p. Plusieurs modes, correspondant à plusieurs débits (donc qualités) ont été retenus.

Pour les modèles avec disque optique, on trouvera le mode standard, « par défaut », avec un débit de 7 Mb/s encadré du mode LP (*Long Play*) à 5 Mb/s et du mode HQ (*High Quality*) à 12 Mb/s. Les durées d'enregistrement sont de 60 minutes en mode LP, 45 minutes en mode standard, et 27 minutes en mode HQ.

Les modèles à disque dur disposent quant à eux d'un mode XP-Quality à 15 Mb/s. Les durées sont pour ces enregistreurs de 11 heures en mode LP, 8 heures 30 en mode standard, 7 heures en mode 9 Mb (spécifique au modèles HDD) et 4 heures en mode XP-Q.

Les modèles à carte flash de 4 Go sont annoncés avec des durées de 85 minutes au débit de 6 Mb/s ou 55 minutes à 9 Mb/s. Il est à noter que la capacité de ce type de cartes est en augmentation constante et que les durées d'enregistrement vont rapidement pouvoir devenir très suffisantes, voire excédentaires.

Il convient de rappeler que Sony avait déjà, en 2004, introduit des caméscopes au format DVCAM IMX ou HDCAM enregistrant sur disque optique Blu-ray.

5.8 DERNIERS FORMATS À VOIR LE JOUR (AVANT LE PROCHAIN) : PANASONIC AVC-I ET THOMSON INFINITY

Panasonic a annoncé en septembre 2006 et mis sur le marché fin 2007 un enregistrement sur carte P2 (la version professionnelle des cartes SD) bénéficiant d'un nouveau mode de compression AVC-Intra (AVC-I).

Le caméscope AG-HVX2100 supporte tous les formats DVCPRO classiques ainsi que, grâce à une carte optionnelle, ce nouveau standard à deux débits (100 et 50 Mb/s) en 1080i ou 720p ou 1080 SF. Il peut embarquer jusqu'à 5 cartes P2.

Il s'agit, comme pour les machines IMX d'utiliser un GOP très court, puisqu'il est de une image I (les images de type P ou B ne sont pas utilisées). Ce type de GOP est évidemment optimal pour toutes les opérations de montage et post-production.

Il s'agit d'un format HD qui vient se positionner au dessus du DVCPRO-100 déjà bien ancien.

Le mode de compression est donné pour être deux fois plus efficace que celui du DVCPRO. Au débit de 100 Mb/s, le codeur AVC-I donnera respectivement 1920 pixels par ligne pour la luminance en mode 1080i, et 1280 en mode 720p. Il est également prévu pour des applications de type reportage un fonctionnement avec un débit de 50 Mb/s correspondant à « seulement » 1440 pixels par ligne.

Les cartes P2 ont actuellement une capacité de 32 Go (capacité confortable qui permet un enregistrement de 32 minutes par carte au débit de 100 Mb/s), mais celle-ci double tous les ans et on table sur 128 Go en 2010.

La conjonction des savoir-faire des équipes de Thomson, Philips (avec l'héritage de Bosch Fernseh) et Grass Valley, regroupées dans une entité connue, suivant les latitudes et les habitudes, sous les noms de Thomson ou de Grass Valley, vient d'aboutir à la réalisation d'un

caméscope HD dénommé, en toute humilité, « Infinity ». Ce caméscope se singularise en proposant un mode d'enregistrement JPEG 2000 : on voit qu'il vise le cinéma numérique (mais on n'est pas encore aux qualités 2k et 4k).

Les formats d'enregistrement sont :

- SD et HD : 525-60i, 525-50i, 720p50 et 60, 1080i 50 et 60 ;
- en DV 25, en 4/2/0/8 bits en 50 Hz et en 4/1/1/8 bits en 60 Hz ;
- en JPEG 2000 SD et HD, en 4:2:2/10 bits.

L'enregistrement MPEG-2 est possible via une carte optionnelle.

Les débits, sont modulables suivant les besoins : 50, 75 et 100 Mb/s.

Le caméscope Infinity enregistre sur des supports informatiques comme les cartouches REV Ioméga de 35 Go, des cartes mémoire Compact Flash, ou des disques durs externes en liaison USB ou FireWire.

Les enregistreurs REV sont les successeurs des enregistreurs Jazz, bien connus depuis plus d'une décennie, mais dont la capacité ne dépassait pas 2 Go. Grass Valley a revisité les machines REV pour réaliser des enregistreurs REV PRO dédiés au secteur audiovisuel.

Les cartouches REV sont constituées de mini-disques durs de 2 pouces 1/2 et sont disponibles en versions 35, 75 et 120 Go (une cartouche de 35 Go autorise un enregistrement de 45 minutes au débit de 75 Mb/s), avec un débit supérieur à 100 Mb/s.

La solution JPEG 2000 est séduisante même si l'on est en droit de s'interroger aujourd'hui sur le mode de post-production qui sera en mesure de traiter les séquences produites… et dans quels délais.

6 LES NOUVELLES GÉNÉRATIONS DE CODAGE NUMÉRIQUE

6.1 LA FAMILLE MPEG ÉLARGIE

L'intelligence collective constituée par le groupe MPEG ne s'est pas arrêtée de fonctionner. On a vu par exemple que les standards MPEG-4 sont toujours en évolution. D'autres propositions méritent également d'être citées, même si elles ne font pas intrinsèquement partie des systèmes de réduction de débit. Il s'agit des standards MPEG-7, MPEG-21 et MPEG-A.

Le groupe MPEG est loin d'avoir dit son dernier mot. De nouveaux chantiers sont déjà ouverts : MPEG-B, MPEG-C, MPEG-D et MPEG-E. Ils portent notamment su l'amélioration de la Transformée Cosinus Discrète Inverse (IDCT), sur l'amélioration du codega XML pour les métadonnées, sur la spatialisation sonore, sur la transmission et l'enregistrement d'images stéréoscopiques (SSV, *StereoScopic Video*)… Et comme l'alphabet doit avoir 26 lettres, nous n'avons pas fini de nous amuser !

6.1.1 MPEG-7

Après MPEG-1 et MPEG-2, il était prévu initialement un standard MPEG-3 relatif à la Haute Définition dont les recommandations se sont en fait trouvées naturellement intégrées à la « boîte à outils » MPEG-2. La référence MPEG-3 est aujourd'hui souvent utilisée sous les initiales MP3. Il s'agit en fait de la « compression » des termes MPEG *audio coding Layer 3*.

Il a donc été logiquement choisi la référence 4 pour la nouvelle génération de codage multimédia qui a été étudié au préalable. Les membres du groupe MPEG ont alors constaté que les références 1, 2, 4, formaient une progression géométrique de raison 2, ce qui aurait pu impliquer que le standard suivant aurait porté la référence 8. Comme un clin d'œil pour faire mentir cette « obligation numérique », le 7 aurait été choisi… c'est du moins ce qu'on nous a dit un jour.

Le standard MPEG-7, ISO/IEC 15398 a été proposé en 2001. Il est également connu sous le titre « Multimedia Content Description Interface » (Interface de description des contenus multimédia). « Il vise à décrire les données des contenus multimédia qui présentent des niveaux d'interprétation susceptibles d'être traités sous forme de code informatique. »

Cette description qui devrait permettre de « trouver ce que l'on cherche en évitant de trouver ce que l'on ne cherche pas », est nécessaire parce que le volume des données multimédia est en spectaculaire progression, parce qu'elles elles sont éparpillées dans le monde entier et parce qu'elles intéressent, essentiellement pour des raisons économiques, un nombre croissant de personnes.

Cette description est rendue possible par l'utilisation de données annexes structurées, associées aux documents de tous types et notamment multimédia. On parle de métadonnées ou de *metadata*. Ces termes expriment le fait qu'il s'agit de données décrivant des données (« bits about the bits » disent les anglo-saxons).

Ces informations documentaires permettent d'indexer, classer, archiver des documents et d'en optimiser la recherche.

MPEG-7 s'est préoccupé de définir une représentation de l'information sur le contenu audiovisuel dénommé « Essence ». Il a standardisé une grammaire pour exprimer les descriptions ainsi qu'une syntaxe pour les coder.

Les métadonnées peuvent être d'ordre technique et d'ordre sémantique.

Dans le cas de documents multimédia les informations techniques pourront être par exemple le standard, le support, le type de compression… auxquelles on pourra ajouter le lieu de tournage (il existe des caméscopes avec GPS) la date et l'heure des séquences (code temporel)…

Les données sémantiques sont plus complexes à déterminer ; elles exigent souvent l'intervention d'un documentaliste et se matérialisent par des mots clefs. Il existe cependant des systèmes automatiques permettant d'identifier des mots dans un texte ou un enregistrement sonore ainsi que de détecter dans une séquence la présence de personnages « célèbres » : Bill Clinton, ou Joey Star, mais sans doute pas Claudia Cardinale ou Pierre Bourdieu. Ces méthodes de reconnaissance faciale sont des retombées de systèmes de sécurité.

MPEG -7 utilise un langage dit DDL (*Description Definition Language*) basé sur XML avec des adjonctions spécifiques aux contenus audiovisuels.

L'élément de base est le Descripteur (D) qui représente différentes propriétés d'un contenu audiovisuel, avec des niveaux simples (LLD, *Low Level Descriptor*) ou plus élaborés (HLD, *High Level Descriptor*). Les Descripteurs sont regroupés en éléments plus complexes qui sont des *Description Schemes* (DS).

Voici quelques descripteurs :

- pour les images : distribution spatiale de luminance ; histogrammes de luminance ; espace colorimétrique ; histogramme colorimétrique ; homogénéité de texture ; analyse de formes ; mouvement de caméra ; mouvement d'objets dans la scène…
- pour les sons : l'analyse du langage parlé (mots, phonèmes) ; la hauteur ; le timbre (fondamentale et harmoniques) ; l'attaque ; le rythme…

MPEG-7 vise à réaliser des interfaces intuitives constituant une sorte d'extension des modes hypertexte tels qu'ils ont été popularisés par le Web. Grâce à ses descripteurs spécifiques MPEG-7 doit permettre de faire des requêtes indépendantes du langage, en dessinant un croquis rapide afin de retrouver un tableau de Gauguin ou l'escalier d'Odessa du Cuirassé Potemkine ou bien encore d'obtenir dans l'instant « Y a d'la

joie » de Trénet ou le Requiem de Mozart en sifflotant quelques notes.

6.1.2 MPEG-21

L'origine du nom est claire : le MPEG du XXIe siècle. Leonardo Chariglione le charismatique animateur du Groupe MPEG l'a ainsi défini : « MPEG-21 (finalisé dans sa version 2 en 2005) a pour objectif de décrire une architecture multimédia et d'asseoir notre vision d'un futur environnement qui est capable de gérer la livraison et l'utilisation de tous types de contenu par différentes catégories d'utilisateurs dans des domaines variés d'applications »… vaste programme !

De nombreux éléments existent aujourd'hui qui peuvent permettre de construire une infrastructure pour l'utilisation d'éléments multimédia : séquences vidéo, éléments sonores, images fixes… accompagnés de *metadata*. Il n'existait aucune vue d'ensemble permettant de décrire comment les spécifications de ces éléments (on parle d'« Items Numériques ») pourraient entrer en relation.

Le but de MPEG-21 est de comprendre comment ces divers Items Numériques (DI, *Digital Items*) peuvent interagir.

MPEG-21 a pour ambition de définir une architecture multimédia qui permettrait une utilisation (création, distribution, commercialisation, consommation) « transparente et plus importante des ressources multimédia au travers d'appareils et de réseaux utilisés par différentes communautés ».

Les solutions envisagées prennent en compte des transactions interopérables et hautement automatisées qui sont indispensables à ce nouveau type de circulation et de commercialisation d'objets mutimédia.

MPEG-21 définit des Utilisateurs, des Items Numériques, sur lesquels les Utilisateurs exécutent des

Actions qui génèrent d'autres Items qui peuvent devenir l'objet de Transactions.

Un Item Numérique (DI) comprend un contenu multimédia appelé « Resource », des métadonnées indissociablement attachées à ce contenu, appelées « descriptor/statement ». Ces deux éléments sont associés grâce à un « Composant » (*component*). Les descripteurs comportent impérativement des informations structurelles (débit, codage…) ; ils peuvent également comporter des informations sémantiques telles que mots-clef, imagettes…

MPEG-21 prend en compte la déclaration (DID, *Digital Item Declaration*) qui signale la localisation géographique des Items ainsi que l'identification et la description (DIID, *Digital Item Identification and Description*) de ces Items Numériques ; ces deux ensembles de métadonnées permettront l'utilisation des Items et leur gestion commerciale (*Content Handling and Usage*) par toute application conforme au standard.

MPEG-21 permet de mettre en place la Gestion et la Protection de la Propriété Intellectuelle (IPMP, *Intellectual Property Management and Protection*). À cette fin le Dictionnaire des Droits (RDD, *Rights Data Dictionary*) contiendra les termes nécessaires pour décrire les droits de tous les utilisateurs grâce à un Langage d'Expression des Droits (REL, *Rights Expression Language*), compatible machine permettant de gérer automatiquement droits et autorisations et de contrôler l'utilisation des Items.

C'est grâce à MPEG-21 que vous pourrez retrouver sur le Web la trace d'un morceau de musique particulièrement intéressant et que vous pourrez simplement en négocier les droits d'utilisation si vous voulez l'utiliser dans un film.

6.1.3 MPEG-A

le standard MPEG-A (ISO/IEC 23000) a été développé par le groupe MPEG en sélectionnant différentes techniques issues des travaux du groupe et en les regroupant sous le terme de « Multimedia Applications Formats » (MAFs).

Il s'agit de faciliter (avec éventuellement l'aide de technologies proches, existantes ou à développer) l'utilisation des si nombreux outils offerts, de manière un peu dispersée, par les différents standards MPEG afin de créer des applications efficaces, innovantes et interopérables car standardisées. Ces applications sont très spécifiques mais également très largement utilisées.

Les spécification MAF retiennent un ou plusieurs codage pour le document multimédia envisagé ainsi qu'un outil documentaire basé sur MPEG-7 pour associer des métadonnées. Les outils de codage seront issus de MPEG-2 et MPEG-4 avec la possibilité de recourir à des outils tels que JPEG si utilisé pour les images fixes. Les applications MAF pourront également faire appel aux outils MPEG-21.

Deux chantiers principaux sont aujourd'hui en cours : « Music Player MAF » et « Photo Player MAF ». On voit immédiatement qu'il s'agit plutôt d'applications à vocation grand public, tout comme l'a été MP3 et ce n'est pas un hasard.

Music Player consiste à rassembler des données audio codées MP3 avec des *metadata* de type MPEG-7 afin de fournir des audiothèques facilement accessibles et gérables.

Photo Player de son côté s'intéresse à la gestion et la manipulation des quantités énormes de photographies numériques stockées ici et là. Il associera aux outils MPEG-4 les algorithmes JPEG associés à des métadonnées de type MPEG-7 afin de permettre, notamment, la création de photothèques facilement utilisables par le grand public.

6.2 LES ONDELETTES

On a vu lors de l'étude des algorithmes JPEG qu'il était pratique et efficace de définir, grâce à l'utilisation de fonctions mathématiques, un catalogue de motifs simples puis de déterminer si ces motifs apparaissent dans l'image.

On a vu également que la Transformation de Fourier, sous la forme de la DCT, conduit à mettre en œuvre des fonctions cosinus, qui ne sont pas limitées dans le temps. La transformée de Fourier est incapable de localiser les portions du signal dans lesquelles les variations sont rapides, ni celles où elles sont lentes. Il est nécessaire, pour maîtriser cette extension, de réaliser un fenêtrage, qui peut être pénalisant.

Il est bien évidemment venu à l'esprit des spécialistes du traitement du signal l'idée de trouver des motifs dont l'extension et la fréquence spatiale seraient liées à l'échelle des structures de l'image. Les structures de faible étendue, donc à variation rapide, correspondent à des fréquences spatiales élevées, les structures larges, à variation lente, correspondent à des fréquences basses.

On se heurte à une limitation liée à ce que l'on nomme « le principe d'incertitude de Heisenberg » :

« Toute tentative pour connaître la valeur d'un paramètre a pour conséquence de perturber d'une façon imprévisible les autres paramètres du système ». Ce principe a été initialement énoncé en mécanique quantique à propos du couple « énergie-position » d'une particule. Plus généralement, si l'on considère un signal variable dans le temps, comme un signal audio ou vidéo, il est impossible de connaître avec précision simultanément les deux variables conjuguées que sont la fréquence et le temps.

6.2.1 Les arbres et la forêt

De nouvelles techniques sont apparues à la fin des années 70 grâce (notamment) à l'ingénieur français Jean Morlet puis à Stéphane Mallat de l'École Polytechnique.

Elles ont été appliquées dans le domaine de la réduction de débit pour les images, sous le vocable « compression par ondelettes » (*wavelets based compression*).

Il s'agit d'utiliser des fonctions plus éphémères que les sinusoïdes de Fourier, ayant un domaine d'existence limité dans le temps ou l'espace (selon le type de signal traité), ce qui correspond mieux à notre univers périssable.

Les ondelettes sont des fonctions qui sont caractérisées par une fréquence et qui sont localisées dans le temps (ou dans l'espace). C'est Dennis Gabor (prix Nobel en 1971 pour ses travaux sur l'holographie) qui pressentit l'intérêt d'une analyse d'un signal sur les deux axes « temps » et « fréquence » grâce à des fonctions sinusoïdales modulées par des fonctions de Gauss.

Plusieurs types de ces fonctions ont été utilisés. On constitue pour chacun des types une famille autour d'une « ondelette mère » (*mother wavelet*) donnant naissance à des filles (*daughter wavelets*) obtenues par contraction temporelle.

Figure 6.1.
Ondelettes de Morlet, mère et fille.

Grâce à cette opération de contraction/dilatation, les ondelettes s'adaptent d'elles mêmes à la taille des motifs qu'elles recherchent (effet d'échelle). Elles sont étendues pour étudier les basses fréquences (grande échelle), et resserrées pour étudier les hautes fréquences (petite échelle).

L'ondelette mère permettra de déterminer les éléments de l'image à variation relativement lente, tandis que les ondelettes filles permettront, chacune, de rendre compte des détails de plus en plus fins.

On pourrait dire, d'une manière approximative, que l'on regarde à l'œil nu pour voir la forêt et avec des jumelles pour admirer les arbres.

Il s'agit d'un système d'analyse multi-résolution qui fonctionne comme un jeu de filtres conduisant à l'obtention de plusieurs fichiers correspondant à un découpage du spectre du signal en un ensemble de bandes de fréquences. C'est ce qu'on appelle un codage en sous-bandes (*subband coding*) ; ce type de codage est classiquement utilisé dans les systèmes de compression des signaux audio.

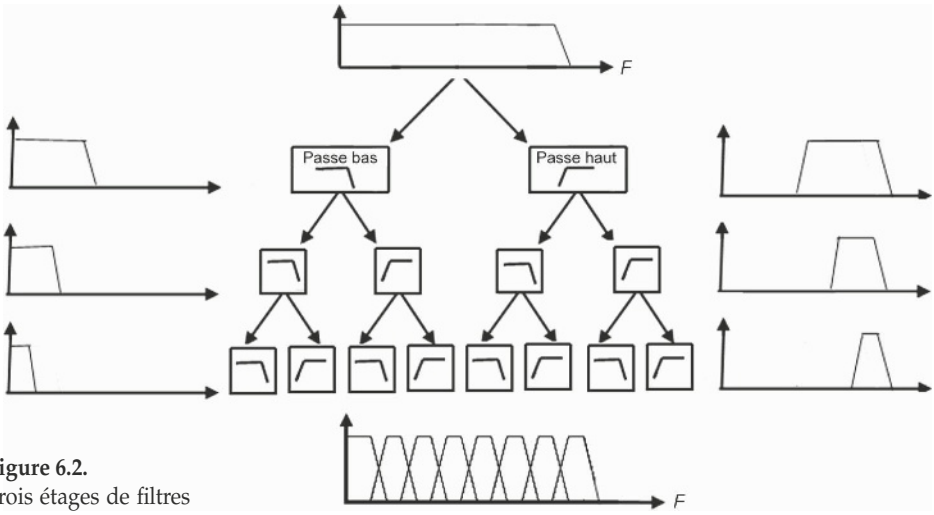


Figure 6.2.
Trois étages de filtres passe-bas et passe-haut permettent de définir 8 sous-bandes.

6.2.2 L'arbre de la connaissance

La réalisation d'un système de compression par ondelettes met en œuvre, selon une structure arborescente (arbre de Mallat), une batterie de filtres numériques

dont les réponses impulsionnelles correspondent à une famille d'ondelettes.

La structure pyramidale comprend, à chaque niveau, une paire de filtres : l'un passe-bas donne une approximation lissée du signal, l'autre passe-haut retient au contraire uniquement les détails. On parle de segmentation dyadique. L'opération s'effectue successivement selon les deux dimensions X et Y.

On obtient ainsi une série d'images, dérivées de l'image originale, chacune de celles-ci offrant une gamme d'information spécifique sur l'image origine.

Cette progression s'accompagne, à chaque étape, d'une division par deux du nombre des pixels retenus. Une simple interpolation, qui peut être précise après l'analyse en fréquence, permet de reconstituer l'intégralité de l'information.

De cette manière le nombre de pixels retenus est identique à celui de l'image avant traitement.

Une fois les informations « triées » en sous-fichiers selon leur fréquence, il est possible, comme pour la compression JPEG, de tenir compte des propriétés psychophysiologiques de la vision humaine pour négliger ou

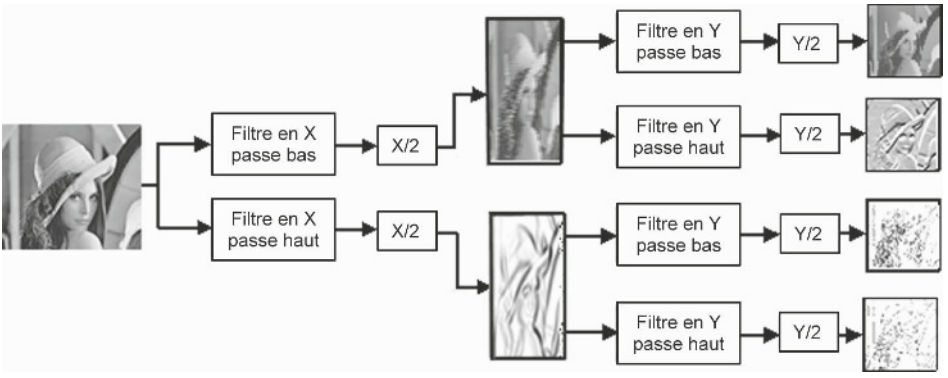


Figure 6.3.
Le visage de Lena, « playmate » du numéro de Playboy de novembre 1972 est une des images de référence en traitement de l'image.

tout au moins sous-quantifier les données les moins pertinentes pour le spectateur.

Comme en JPEG un codage RLC puis un codage entropique (codage de Huffman) sont ensuite appliqués aux données. On sait que ces étapes n'apportent pas de dégradation.

Le codage entropique est évidemment d'une grande efficacité lorsqu'il s'applique à des collections de données qui ont déjà été triées par les filtrages successifs.

6.2.3 Une méthode très efficace

Un taux de compression élevé peut être obtenu en analysant par des méthodes statistiques (valeur max et min du poids des pixels, valeur quadratique moyenne…). L'importance que chacun des sous-fichiers de données présente relativement à la vision humaine. Ces informations sont transmises à un algorithme de quantification. Celui-ci attribue aux différents sous-fichiers une sorte de « crédit de bits » qui dépend d'une part du débit binaire global tolérable par le système, d'autre part de l'importance « visuelle » du sous-fichier.

Dans le cas d'une forte compression (taux supérieur à 100:1, ce qui signifie quand même qu'il est nécessaire d'éliminer 99 % des informations initiales), seuls les sous-fichiers qui correspondent aux fréquences basses, auront un crédit confortable. Le reste du budget binaire sera distribué aux autres sous fichiers en fonction de l'importance qui a été déterminée quant à la contribution qu'ils apportent pour la lecture de l'image. Les détails les moins significatifs seront donc sous-quantifiés, c'est-à-dire qu'ils seront codés avec une dynamique réduite, sur très peu de niveaux n'utilisant qu'un faible nombre de bits.

Il est donc possible de contrôler l'impact visuel de l'incontournable dégradation de l'image imposée par les taux de compression élevé. Des taux supérieurs à

300:1 sont obtenus pour des applications de surveillance.

Les artefacts liés à la décomposition en blocs 8×8 du JPEG sont inconnus avec les ondelettes.

M. Hilton, B.D. Jawerth, A. Sengupta (*Multimedia Systems*) annonçaient dès 1994 que les algorithmes JPEG sont plus efficaces pour des taux allant jusqu'à 25:1 mais « qu'au dessus de 30:1 les performances des algorithmes JPEG se détériorent rapidement tandis que celles des codages par ondelettes se dégradent doucement et peuvent encore fonctionner bien au delà de rapports 100:1 ». Ils pensaient que des taux de compression de l'ordre de 1000:1 seront possibles lorsque les performances requises en termes de mémoire et de performances des processeurs auront été atteintes.

Le grand avantage du codage par ondelettes réside dans son caractère hiérarchique : l'image peut être constituée par couches successives (une couche de base puis des couches de « raffinement ») apportant chacune sa contribution à la résolution.

L'image, à sa résolution la plus grande, est égale à la somme d'une version floue, et des détails apparaissant à des échelles différentes, c'est-à-dire à des résolutions différentes.

Un seul fichier pourra donc être utilisé pour des utilisations sur plusieurs afficheurs de qualité différente.

La structure arborescente de l'information permet également d'obtenir facilement des modifications d'échelle (*scalability*).

Le codage par ondelettes a été retenu pour le système JPEG 2000 ainsi que pour le codage des images fixes en MPEG-4.

6.2.4 Le futur proche ?

Le département R&D de la BBC travaille actuellement sur le développement d'un codec dénommé DIRAC (en hommage au physicien britannique Paul Dirac, Prix Nobel en 1933 avec le physicien autrichien Erwin Schrödinger, tous deux spécialistes de physique quantique).

DIRAC est un codage vidéo « ouvert », accessible par tous les utilisateurs potentiels sans versement de droits. Il vise des applications de distribution allant du *streaming* (QCIF 180 × 144) à la Haute Définition (1920 × 1080) en modes progressif ou entrelacé, ainsi que des applications de post-production et d'archivage jusqu'à une résolution 4k sur 16 bits.

Les premières implémentations industrielles sont en cours de réalisation tandis que le codec est en cours d'homologation par la SMPTE sous la dénomination VC 2.

Plusieurs autres codecs sont également en développement mais aucun n'est arrivé à un stade véritablement opérationnel.

6.2.5 Les bandelettes

En 2005, Erwan Le Pennec et Stéphane Mallat ont introduit la notion de « Bandlets Compression ». Il s'agit de prendre en compte des structures géométriques récurrentes contenues dans des images techniques et scientifiques comme celles produites par un écoulement tourbillonnaire ou dans des images naturelles comme celles contenues dans une chevelure (parce que vous le valez bien… !), les nervures d'un bois ou les plis d'un tissu.

Cette méthode réconcilie l'analyse harmonique et le pure description géométrique. En voici très succinctement (car les outils mathématiques sont plutôt coriaces) le fonctionnement.

La segmentation dyadique des méthodes par ondelettes impose une structure en carrés. Mallat et Le Pennec introduisent la notion géométrique de direction privilégiée d'une structure.

À la suite de la segmentation dyadique classique, un champ vectoriel est déterminé qui caractérise la direction générale de la structure dans chaque domaine où celle-ci apparaît. Ces vecteurs sont utilisés pour transformer les coefficients de la décomposition par ondelettes en de nouveaux coefficients (*warped wavelet coefficients*) qui tiennent compte de la « régularité anisotropique » locale et qui sont stockés dans un dictionnaire.

Cette méthode qui semble particulièrement performante n'est pas encore mise en œuvre dans des systèmes de compression pour la diffusion et le stockage d'images (sauf peut-être dans des domaines spécifiques scientifiques et médicaux) mais a donné naissance, grâce à le société « Let It Wave », créée par Stéphane Mallat, à des circuits intégrés programmables FGPA (*Field Programmable Gate Array*) utilisés dans des systèmes professionnels ou grand public de traitement et d'amélioration des images tels que les convertisseurs de formats, la conversion SD/HD (*upconverters*), le débruitage (notamment l'élimination des artefacts introduits par les compressions MPEG).

6.3 JPEG 2000 ET LE CINÉMA NUMÉRIQUE

JPEG 2000 (ISO/IEC 15444-1 et 15444-2) est une norme de l'ISO et de l'ITU-T partiellement finalisée, comme son nom l'indique, en 2000.

Il s'agit d'un standard de réduction de débit, avec ou sans pertes, pour les images fixes. Bien qu'il ait été défini par le *Joint Photographic Experts Group*, il ne constitue ni une extension ni une amélioration du standard JPEG publié une dizaine d'années auparavant

(le format JPEG a été normalisé en 1991 et les travaux sur JPEG 2000 ont commencé en 1997).

Il a été conçu, en mettant en œuvre une méthode entièrement nouvelle, afin de pouvoir traiter tous les types d'images : photographies, images scientifiques et médicales, images de télédétection… Il concerne aussi bien l'impression que la transmission d'images sur réseaux à faible débit en passant, bien sûr, par toutes les applications à la photographie numérique.

Les principaux objectifs initiaux étaient de fournir une qualité d'image élevée, notamment aux très faibles débits, sans utiliser d'algorithmes trop gourmands en calcul, et de fournir de nouvelles possibilités telles que la définition de « régions d'intérêt ».

Les performances de compression de JPEG 2000 dépassent celles de JPEG, en particulier pour les forts taux de compression, mais au prix d'algorithmes environ cinq fois plus complexes. Les avancées des possibilités informatiques font que cela ne devrait pas constituer un frein à son adoption, sauf pour les systèmes mobiles.

Cette norme est composée de plusieurs parties dont certaines ne sont pas encore totalement finalisées. Comme toutes les normes JPEG ou MPEG, elle est écrite du point de vue du décodeur ; elle ne définit pas les détails du codeur mais les spécifications essentielles exigées de lui afin que le fichier compressé puisse être décodé.

- La première partie définit l'algorithme de décodage de base ainsi que le format de ce qu'on appelle le *codestream* (flux codé) ainsi que le format JP2 qui permet d'ajouter au *codestream* des informations techniques relatives à l'image codée (type de fichier, détection d'erreurs, taille de l'image, résolution d'acquisition et d'affichage par défaut, espace colorimétrique, etc.).

- La deuxième partie et la troisième partie définissent des extensions de la partie 1, telles que l'utilisation des métadonnées ainsi que le codage de séquences audiovisuelles, considérées comme une succession d'images, compressées par un système respectant la partie 1 de la norme ; on nomme cette extension « Motion JPEG 2000 ».
- La quatrième et la cinquième parties sont consacrées aux règles de « conformance » ainsi qu'à des logiciels de référence.
- La sixième partie définit un format de fichier pour les images dites « compound » composées de textes, d'images et de graphiques.
- La huitième partie (JPSEC) est consacrée aux problèmes essentiels de la sécurité de transmission.
- La neuvième partie (JPIP) est consacrée aux outils d'interactivité.
- La dixième partie (JP3D) est consacrée au codage d'objets en trois dimensions, essentiellement pour des applications médicales et techniques.
- La onzième partie (JPWL) est consacrée à la diffusion sur réseaux hertziens.

6.3.1 Les algorithmes

Les algorithmes de codage JPEG 2000 mettent successivement en œuvre une transformation par ondelettes et un codage entropique.

Si l'image est de grandes dimensions il est parfois nécessaire, afin de soulager la mémoire du codeur et du décodeur, de la découper préalablement en plusieurs morceaux appelés pavés ou tuiles (*tiles*).

JPEG 2000 peut traiter des images RVB sur trois plans, mais on applique fréquemment, avant codage, une transformée couleur afin de faire passer une image RVB

en YUV. Ceci est simple à réaliser et conduit à des données peu corrélées ce qui augmente l'efficacité du codage.

La transformation par ondelettes présente l'intérêt de conduire à une codage hiérarchique ou codage en sous-bandes qui permet d'afficher instantanément (et si nécessaire simultanément) l'image à différentes résolutions.

Il est habituel d'en donner la représentation symbolique ci-dessous dans laquelle B, H_2 représente le résultat du filtrage passe-bas horizontal et passe-haut vertical de niveau 2 ; H, H_3 le résultat du filtrage passe-haut horizontal et passe-haut vertical de niveau 3… etc. Le fichier final compressé se compose d'une couche de base et de couches d'amélioration (*enhancement layers*).

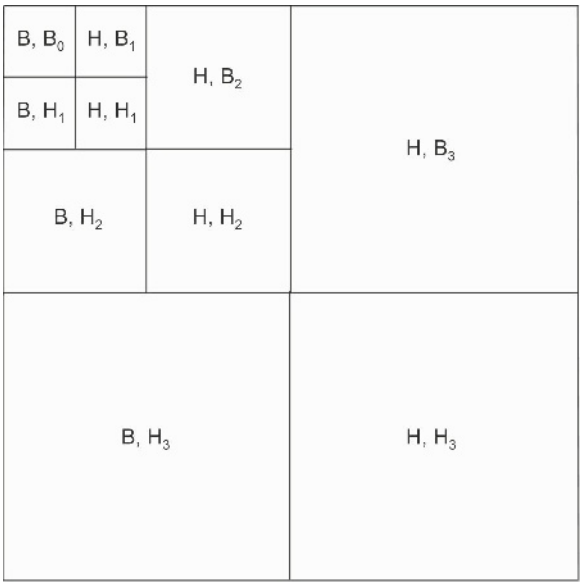


Figure 6.4

La norme prévoit jusqu'à 32 niveaux soit 97 sous-bandes.

Comme pour les codages audio, les coefficients dans chaque sous-bande peuvent être quantifiés avec un nombre de bits plus ou moins important. La quantifi-

cation doit tenir compte du modèle visuel humain. C'est à ce niveau que peuvent intervenir des pertes d'information.

Un codage entropique est ensuite appliqué. Il s'agit d'un codage adaptatif contextuel utilisant l'algorithme EBCOT (*Embedded Block Coding with Optimized Truncation*).

Chacune des sous-bandes est décomposée en blocs (généralement 64 × 64 coefficients). Chaque bloc est compressé indépendamment en trois passes successives.

Une passe signifiante ou « signifiance pass » ; une passe d'affinage ou « refinement pass » ; une passe de nettoyage ou « cleaning pass ». Chacune de ces passes récupère des informations contextuelles, c'est-à-dire une façon de tenir compte des blocs voisins. Si les valeurs des blocs voisins sont grandes, le coefficient sera fortement quantifié.

Chaque passe prend en compte des contextes différents qui permettent de classer les coefficients selon leur importance. La dernière passe permet d'identifier les coefficients de très peu d'importance qui peuvent être, si besoin, éliminés sans grand risque.

Les données compressées forment des paquets qui sont constitués en fonction de l'importance des divers coefficients correspondant à un niveau de qualité donné. Il disposent chacun d'une en-tête identifiant leur contenu.

On peut ainsi obtenir un fichier final flexible avec des possibilités de troncature selon la résolution.

Ceci explique la mise en œuvre d'une fonction « allocation de débit » qui adapte automatiquement, en supprimant les paquets notés comme les moins significatifs, la nature du flux en fonction des spécifications de la chaîne de transmission.

6.3.2 Les fonctionnalités nouvelles

Décodage progressif

Les flux (*codestream*) JPEG 2000 sont, par essence, multi-résolution. On peut donc afficher l'image, en fonction des possibilités de l'écran ou en fonction de celle du canal avec une amélioration progressive de la qualité.

Régions d'intérêt (ROI : *Region Of Interest*)

Il peut être utile de pouvoir coder et décoder une image avec des degrés de précision variables. JPEG 2000 permet de décider qu'une zone de l'image de forme rectangulaire ou elliptique doit être privilégiée. Il peut s'agir d'un visage, d'une région pathologique dans une image médicale… qui constituent la partie informative de l'image. Le reste de l'image, le fond (*background*) est codé avec pertes. De cette manière, on obtient un taux de compression élevé sans perdre d'informations essentielles.

Résistance aux erreurs

JPEG 2000 possède des outils de résistance aux erreurs de transmission. Ils s'appliquent au codage EBCOT. Le découpage en blocs, dont les en-têtes sont bien protégées, permet en lui même de limiter les erreurs à un seul bloc sans risque de propagation aux blocs voisins.

Accès aléatoire rapide

Il est possible de ne décoder qu'une partie spécifiée d'une image lorsque celle-ci est de très grandes dimensions.

Sécurité, JPSEC

Confidentialité, vérification de l'intégrité des données transmises, authentification des sources, accès conditionnel, vérification que des opérations comme l'échelonnabilité et le transcodage peuvent être effectuées sans perte de protection, telles sont les fonctions recherchées. Elles mettent en œuvre des systèmes de cryp-

tage, de signature numérique, de tatouage des images (*watermarking*)…

Interactivité et protocoles (JPIP)

Un ensemble d'outils, exploitant les possibilités d'accès aléatoire et de décodage hiérarchique du standard, pour assurer la diffusion sur réseaux.

Extension 3D, JP3D

Vise essentiellement des applications scientifiques, techniques, médicales.

Transmission radio, JPWL (Wireless)

JPEG 2000 en raison de ses propriétés d'efficacité du codage, de résistance aux erreurs et d'accès aléatoire est parfaitement adapté à la transmission sur des réseaux hertziens susceptibles d'être « bruités » par des signaux parasites.

Cette fonctionnalité met en œuvre des améliorations spécifiques de résistance aux erreurs : protection renforcée des en-têtes, protection hiérarchique des données. La couche de base, essentielle, est fortement protégée ; les couches d'amélioration (*enhancement layers*) le sont de moins en moins au fur et à mesure que leur rang s'élève.

Il a été défini un système « Motion JPEG 2000 » (comme JPEG avait donné naissance à des algorithmes « Motion JPEG »). Les données audio sont au format MP4 et les modes de synchronisation avec les images des sons et des données annexes comme les sous-titres sont ici bien normalisés, ce qui n'était pas le cas pour M-JPEG.

6.3.3 Le cinéma numérique

De nombreuses avancées techniques, aussi bien dans le domaine de la projection numérique que dans celui des caméras électroniques ou celui des transferts à

haute résolution d'images de films 35 mm vers l'espace numérique, la facilité d'intégrer dans un document des éléments en images de synthèse, les nombreuses possibilités d'effets spéciaux remarquables, font que la production cinématographique « tout numérique », remplace rapidement la production « argentique » classique et que l'ouverture dans le monde entier (notamment aux USA) de milliers de salles de projection de cinéma numérique (*Digital Cinema*) est aujourd'hui une réalité.

« Digital Cinema Initiatives », (DCI) est une structure créée par les sept plus importants studios de production cinématographique des USA : Disney, Fox, Metro-Goldwyn-Mayer, Paramount Pictures, Sony Pictures Entertainment, Universal Studios, et Warner Bros. Studios.

Cette structure a eu une importance déterminante même si les européens ont tenté, avec l'EDCF (*European Digital Cinema Forum*), de faire entendre leur voix.

Le but principal de DCI était de publier des spécifications « universelles » indispensables pour l'établissement d'une industrie mondiale cohérente du cinéma numérique.

Celles-ci ont été publiées en juillet 2005. La traduction française de ces 170 pages est disponible sur le site de la CST. Cette publication note que « le Cinéma Numérique aura la possibilité de fournir une projection en salle de qualité supérieure à ce que l'on pourrait réaliser avec une copie 35 mm traditionnelle ».

Les spécifications portent sur plusieurs chapitres : Master de distribution de cinéma numérique (DCDM, *Digital Cinema Distribution Master*) ; Compression de l'image ; Conditionnement (*Packaging*) ; Transport ; Équipement des Salles (*Theater Systems*) ; Projection ; Securité (*Security*). Nous ne nous intéresserons ici qu'aux deux premiers, même si le troisième est considéré comme fondamental par les détenteurs de droits.

Le master de distribution, DCDM, est le produit numérique final de la post-production. Il comprend les éléments : images, sons et sous-titres : ces éléments sont placés dans des fichiers de données qui constituent le DCDM. Celui-ci est ensuite compressé, crypté et conditionné sous la forme d'une copie de distribution numérique ou DCP (*Digital Cinema Package*) qui peut être déconditionnée, décryptée et décompressée pour créer le fichier DCDM*, qui sera diffusé dans les salles et qui est identique au DCDM original.

Le DCDM utilise une structure d'image hiérarchique qui supporte aussi bien des fichiers de résolution 2K (2048 × 1080 ; 24 ou 48 images/seconde) que des fichiers de résolution 4K (4096 × 2160 ; 24 images/seconde) afin de pouvoir être utilisés par des types de projecteurs 2 ou 4K. Les images sont de type RVB quantifiées avec 12 bits par couleur. La distribution des programmes sera possible aussi bien en résolution 2K qu'en résolution 4K.

- Pour le format 1,85, l'image ne retiendra que 3996 × 2160 pixels en 4K et 1998 × 1080 en 2K.
- Pour le format 2,39, elle ne retiendra que 4096 × 1714 pixels en 4K et 2048 × 858 en 2K.

Une compression (indispensable pour la transmission et le stockage : une image 4K « pèse » 32 Mo, soit environ 9 To pour un programme de 3 heures) de type JPEG 2000 a été choisie après de nombreux tests.

Plusieurs spécifications ont contribué à ce choix :

- il s'agit d'un codage « intra » qui, nous l'avons vu, simplifie la post-production ;
- on peut mettre en œuvre un décodage hiérarchique qui permet de véhiculer simultanément dans un seul fichier plusieurs résolutions, avec ou sans pertes) ce qui est apparu comme essentiel afin de simplifier la distribution 2K/4K et de permettre éventuellement des visionnages de contrôle rapides ;

- la résistance aux erreurs est élevée ;
- les éléments de sécurité et de confidentialité des programmes sont développés.

Il existera deux types de décodeurs, 4K et 2K.

Les décodeurs 2K devront extraire du fichier les données 2K dans un mode de distribution 4K en éliminant les données correspondant au niveau supérieur de la résolution ; les décodeurs 4K fourniront en sortie toutes les données.

C'est au système de projection qu'il conviendra éventuellement, s'il n'est pas adapté à la résolution du décodeur, de sur-échantillonner ou sous-échantillonner les données fournies par celui-ci.

Chaque image est composée de 3 « pavés » correspondant aux trois primaires (espace colorimétrique XYZ) codées sur 12 bits, la compression s'effectue sur 5 niveaux d'ondelettes en 2K et 6 en 4K.

Les données audio (16 canaux possibles) ne sont pas compressées, elles sont codées (AES 3) sur 24 bits aux fréquences d'échantillonnage de 48 ou 96 kHz.

6.4 LES MÉTHODES DE COMPRESSION FRACTALES

C'est le mathématicien français Benoit Mandelbrot qui inventa en 1975 l'adjectif « fractal » (du latin *fractus*, interrompu, irrégulier) pour nommer des êtres mathématiques un peu bizarres mais qui pouvaient être définis d'une manière itérative grâce à des règles génératives simples.

Voici un exemples élémentaire et classique d'objet fractal : le flocon de Von Koch. Cette courbe s'obtient en partant d'un triangle équilatéral dont on remplace sur chaque côté le tiers central par deux segments de longueur identique à celle du morceau enlevé.

Figure 6.5.
Flocon de Von Koch.

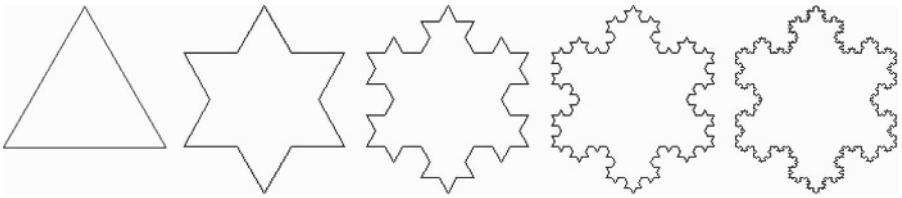


Figure 6.6.
Construction itérative du flocon de Von Koch.

Cette figure présente un caractère d'autosimilarité. Si on réalise un zoom, fut-il infini, on retrouve des détails identiques à ceux observés à une échelle plus grande ou plus petite.

Si on mesure la longueur de la courbe on s'aperçoit qu'elle tend vers l'infini, lorsque le nombre des éléments augmente, bien qu'elle reste contenue à l'intérieur d'un cercle de rayon fixe et fini et qu'elle enferme une surface finie.

Mandelbrot pose la question « combien mesure la côte de la Bretagne ? » La réponse est : cela dépend de la longueur de l'instrument de mesure choisi. Si l'on utilise un mètre on pourra évaluer des détails beaucoup plus fins que si l'on utilise un décamètre et la longueur de la côte sera plus importante.

6.4.1 La notion de dimension fractale

Prenons un cube et multiplions sa taille par 3 ; c'est en termes mathématiques une homothétie de rapport 3. La longueur de ses arrêtes sera multipliée par 3, la surface de ses faces sera multipliée par 9 (3^2), son volume sera multiplié par 27 (3^3). Il en va de même pour la circonférence d'un cercle de rayon R, ou sa surface, ou le volume de la sphère de rayon R. On dit que les longueurs sont de dimension 1, les surfaces de dimension 2 et les volumes de dimension 3. La dimension D est l'exposant qui affecte le coefficient de l'homothétie.

On peut écrire :

• pour les longueurs D = log 3 / log 3 = 1 ;

• pour les surfaces D = log 9 / log 3 = 2 ;

• pour les volumes D = log 27 / log 3 = 3.

Les objets fractals dérogent à cette règle que l'on croyait universelle. Si on fait la même opération sur le flocon de Von Koch on constate que la longueur de la courbe est multipliée par 4. On dira que sa dimension D est égale à log 4/ log 3 = 1,26.

6.4.2 Générer des formes naturelles

On peut obtenir des formes moins schématiques que celle du flocon en se donnant des règles génératrices un peu plus complexes ou en introduisant une pincée d'aléatoire dans le processus.

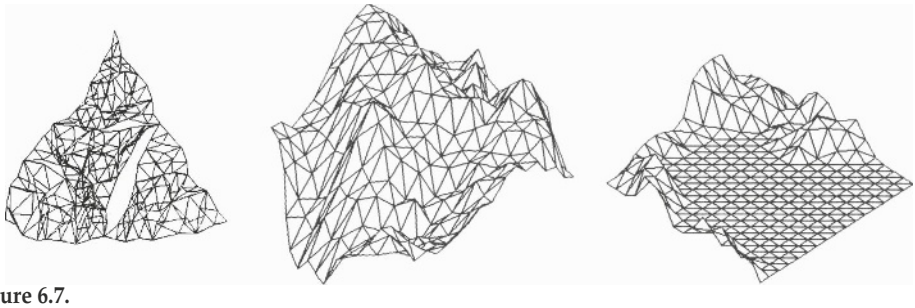


Figure 6.7.
Montagnes fractales créées par Jean-Baptiste Touchard.

On s'est vite aperçu que certaines des images obtenues par ces procédés mathématiques itératifs pouvaient ressembler fortement à des images naturelles.

Ainsi Jean Baptiste Touchard, dans les années 80, s'amusait-il à créer des montagnes en quelques triangles itératifs sur un simple Macintosh.

Benoit Mandelbrot, dans son livre «The fracatal geometry of nature » (1982), montra que des objets naturels aussi divers que des galaxies, des montagnes, des arbres ou la côte de Bretagne, possédaient des propriétés

d'autosimilarité et pouvaient être décrits économiquement en termes de structures fractales.

Le motif fractal parfaitement itératif et régulier ci-dessous est une photographie d'un choux romanesco.

Figure 6.8.
Choux romanesco.

6.4.3 Analyser et « compresser » des images naturelles

Le mathématicien américain Michael F. Barnsley déposa en 1988 un brevet décrivant un mode de fabrication d'images en motifs fractals.

Il décrit avec humour le principe de la méthode sous le nom « d'algorithme de la photocopieuse » : une photocopieuse bénéficiant de la fonction de changement d'échelle.

Son exemple porte sur la construction d'une feuille de fougère, élément naturel simple et fortement autosimilaire.

Plus généralement, la méthode de compression fractale consiste à trouver des motifs récurrents à l'intérieur de l'image, à en dresser un catalogue puis à spécifier les zones de l'image où ils ont été rencontrés.

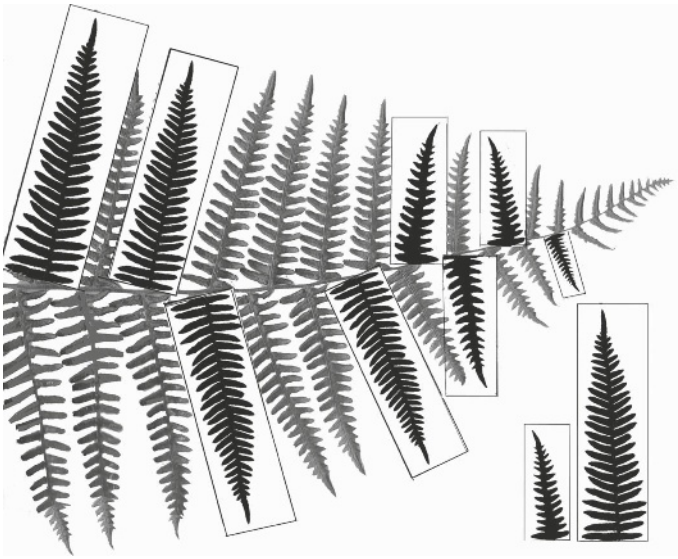


Figure 6.9.
L'algorithme de la photocopieuse. Quelques motifs suffisent, après rotation et mise à l'échelle, pour dessiner une feuille de fougère.

À la différence de la méthode JPEG, les motifs du catalogue ne relèvent pas de fonctions mathématiques mais sont des éléments graphiques provenant de l'image elle-même.

On prélève un petit élément de l'image (un carré ou un triangle), c'est l'élément « source ». Puis on le déplace sur l'ensemble de l'image en le comparant et en tentant de le faire coïncider avec les autres parties de l'image (éléments « destination » ou « cible »).

La procédure de recherche de coïncidence mise en œuvre dans les logiciels de compression utilise l'ensemble des transformations géométriques dites affines en ajoutant aux effets de translation, mentionnés plus haut, des effets de rotation, de symétrie et d'homothétie (ou changement d'échelle). On peut également jouer sur des modifications de niveau de luminance.

Il est ainsi possible de commencer par une recherche rapide en comparant un bloc de 64×64 pixels à des blocs 128×128 puis à des blocs 64×64, puis à des blocs 32×32, etc. jusqu'à des blocs 4×4 ou même 2×2.

Figure 6.10.
Les éléments A et A′ peuvent être appairés après symétrie, les éléments B et B′, après mise à l'échelle.

Il serait exceptionnel que l'on obtienne une coïncidence parfaite entre éléments « source » et « destination ». On estimera que deux régions sont identiques si la somme des différences de valeurs des pixels des deux éléments est inférieure à un seuil prédéfini. Plus le seuil sera élevé moins la ressemblance entre les deux éléments sera précise, mais plus on aura de chance de trouver rapidement des zones susceptibles de s'apparier. Le taux de compression obtenu sera plus élevé mais l'image reconstruite à partir du catalogue des motifs sera dégradée par rapport à l'image originale. C'est toujours le rapport qualité/prix qui en fin de compte décide !

Il est bien évident que ces tentatives multiples pour apparier ces zones sont assez lourdes et longues. Par contre, lorsque le catalogue des motifs qui ont été trouvés dupliqués a été établi et qu'on a simultanément connaissance de l'endroit où ils ont été reconnus ainsi que des indications de transformation qui doivent leur être appliqués, il devient facile de reconstruire l'image initiale. La méthode est totalement asymétrique et c'est à l'utilisateur final qu'il revient d'effectuer les tâches les plus simples.

Le méthodes de compression fractales sont extrêmement efficaces, elles permettent d'obtenir des taux de compression virtuellement transparentes de 100:1 ce qui est particulièrement intéressant pour le stockage de gros volumes d'images, par exemple, dans le domaine médical.

Ces méthodes n'ont pas réussi à s'imposer, bien que des logiciels aient été proposés (notamment par la compagnie de M. Barnsley, *Iterated Systems Inc.*), face aux méthodes de type JPEG, moins performantes certes mais plus banalisées ou, plus récemment, face aux méthodes de compression par ondelettes mises en œuvre dans JPEG 2000.

INDEX