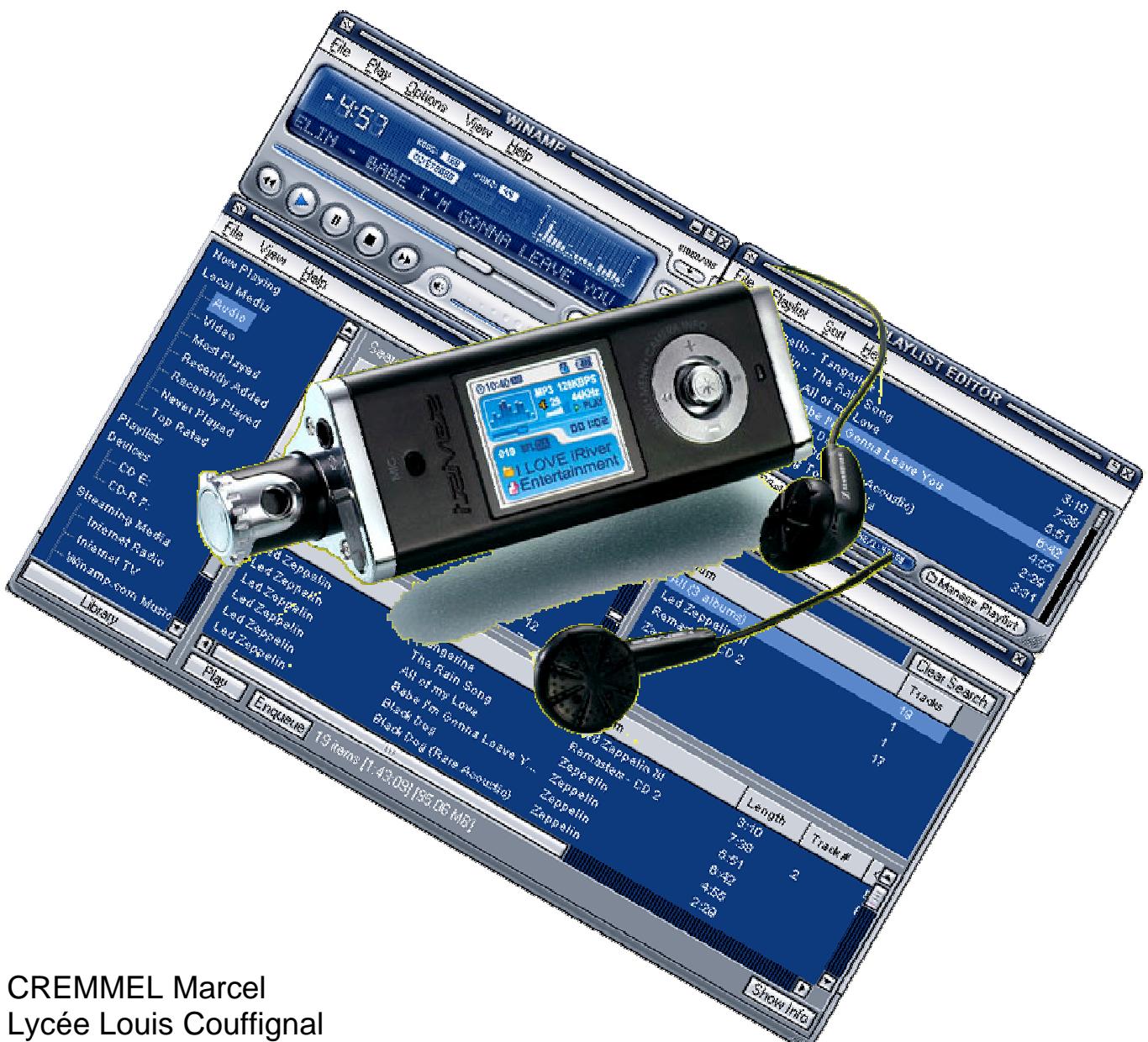


LA COMPRESSION AUDIO NUMÉRIQUE

PRINCIPE DU CODAGE

MPEG LAYER 3 / MP3



CREMMEL Marcel
Lycée Louis Couffignal
STRASBOURG

LA COMPRESSION AUDIO

MPEG-LAYER3 / MP3

Les enregistrements sonores ne se font plus aujourd'hui dans un format **analogique** malgré la résistance de quelques nostalgiques. Les derniers progrès réalisés avec les microsillons et surtout les enregistreurs à bande permettaient pourtant d'atteindre d'excellentes performances (bande passante et rapport signal sur bruit). Par contre le support (disque ou bande) était volumineux et fragile et l'information enregistrée se dégradait dans le temps. De plus, les équipements étaient plutôt encombrants (même si le "walkman" a eu beaucoup de succès) et la transmission (par câble ou radio) entraînait nécessairement une dégradation de l'information.

L'arrivée du "CD" dans les années 80 a déclenché le basculement au format numérique. Ses caractéristiques principales sont les suivantes (valables encore aujourd'hui) :

- stéréophonique
- fréquence d'échantillonnage : 44,1kHz
- résolution : 16 bits par échantillon

Si on veut transmettre un message musical codé avec ce format, le canal doit accepter un débit de :

$$44100\text{Hz} \times 16 \times 2 = 1\,411\,200 \text{ bits par s, soit } \mathbf{1411,2 \text{ kbps}}$$

Ainsi, **60 minutes** de musique codées dans ce format sont représentées par :

$$60\text{mn} \times 60\text{s} \times 44100\text{Hz} \times 16\text{bits} \times 2\text{voies} = 5\,080\,320\,000 \text{ bits} = \mathbf{605 \text{ Moctets}} \quad (1\text{M}=1\,048\,576)$$

La conversion analogique/numérique ne posait pas de problème technique à l'époque mais il a fallu inventer un support bon marché permettant de mémoriser autant de bits : le Compact Disc !

Les défauts du support (ne serait-ce que les rayures !) engendrent des erreurs de lecture qui ont obligé les inventeurs (Philips et Sony) d'ajouter aux échantillons "musicaux" des bits de contrôle permettant de détecter et même corriger les erreurs de lecture.

Mais aujourd'hui, avec le développement d'internet et des mémoires flash de forte capacité, la musique doit pouvoir circuler sur la toile et se mémoriser sur ces nouveaux supports. Or, si on utilisait le format numérique du CD tel quel, une chanson de 3mn :

- représente : $3\text{mn} \times 60\text{s} \times 44100\text{Hz} \times 16\text{bits} \times 2\text{voies} = 254\,016\,000 \text{ bits}$ soit **30,3 Moctets**
- et est transmise, avec une connexion à 512kbits/s en : 496s, soit **8mn et 16s**

Malgré la qualité sonore optimale obtenue, ces performances sont jugées médiocres !

On a donc développé dès la fin des années 80 des procédés permettant de réduire la quantité de bits à mémoriser ou le débit binaire : on parle de **compression**. Les formats les plus utilisés aujourd'hui (MP3, AAC et WMA) ont été développés par l'**institut FRANHOFER** en Allemagne.

La complexité des procédés de compression est liée à la technologie disponible de sorte que les constructeurs puissent proposer des appareils portatifs à faible consommation et faible coût.

Les algorithmes actuellement exploités permettent d'obtenir des

**taux de compression de 1/11 environ (soit 128 kbps en stéréophonie)
sans perte de qualité audible**

(aux dires des inventeurs ! et de nombreuses "oreilles" expertes).

1. Bref historique

En 1986, K. Brandenburg et son équipe, travaillant à l'institut Franhofer, sont chargés du projet Eurêka qui est la création d'une radio numérique (du nom de DAB : Digital Audio Broadcasting). Le problème était que le son ne pouvait être transmis intégralement. Il fallait créer un moyen de le compresser pour le transmettre.

En 1992, le "Motion Picture Experts Group" (MPEG) adopte ce format pour la compression du son de qualité "haute fidélité". Le standard propose 3 modes de compression : les "Layer 1, 2 et 3", du plus simple au plus efficace.

Le "Layer 1" est utilisé par Philips dans son système de cassette DCC, aujourd'hui abandonné. Le "Layer 2" est utilisé dans les CD multimédias (déjà disparus !) et les "Video Disc".

Le "Layer 3" est le plus complexe et le plus performant et peut être traité aujourd'hui en temps réel par les microcontrôleurs RISC des lecteurs et enregistreurs portables. Un PC actuel prend même de l'avance sur le temps réel ! Rapidement, le MPEG audio layer 3 est renommé MP3.

Les spécifications de ces standards sont libres de droits (le MPEG publie même les sources des programmes en "C").

2. Principe de la compression du débit binaire

Le grand principe de la compression audio:

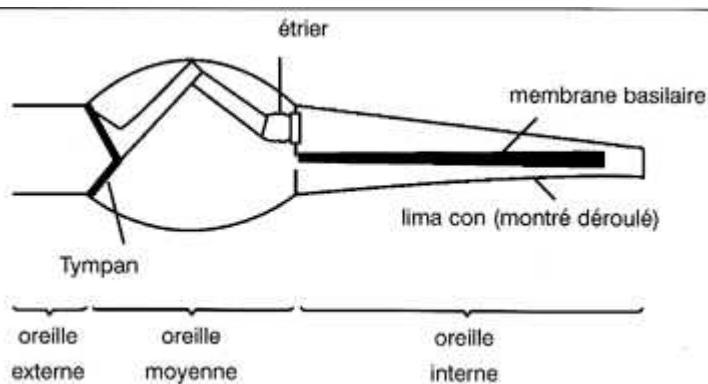
"Ne jamais transmettre ce que l'on ne peut pas entendre."

Le mp3 est fondé sur le principe du "Codage perceptuel". Cela consiste à réduire au maximum la quantité d'informations nécessaires à la perception intégrale du son par l'oreille humaine. D'un point de vue strictement technique, il s'agit d'un procédé destructif ; mais, et c'est là la nouveauté, cette perte est quasiment imperceptible car elle est fondée sur les limites connues du système auditif humain. Le décodeur est bien moins complexe car son seul travail est de reconstruire le signal audio à partir des composantes codées. C'est pourquoi on s'intéressera seulement au processus de codage.

2.1 Le mécanisme de l'audition

L'audition se compose d'un processus physique à l'intérieur de l'oreille et d'un processus nerveux et mental qui se combinent pour donner une impression sonore.

Le mécanisme physique de l'audition se répartit en trois parties: l'oreille externe, l'oreille moyenne et l'oreille interne. En plus du pavillon, l'oreille externe comprend le conduit auditif et le tympan. Le tympan transforme les sons incidents en une vibration comme le fait un diaphragme de microphone. L'oreille interne opère en utilisant ces vibrations transmises à travers un fluide.



On voit ci dessus que les vibrations sont transférées à l'oreille interne par l'étrier qui agit sur la fenêtre ovale. Les vibrations du fluide de l'oreille interne parviennent au limaçon, une cavité du crâne en forme de spirale (montrée déroulée sur la figure pour plus de clarté). La membrane basilaire est étirée sur toute la longueur du limaçon.

Le poids et la consistance de cette membrane varient d'un bout à l'autre. Près de la fenêtre ovale, la membrane est rigide et légère et sa fréquence de résonance est élevée : elle est sensible aux sons aigus. A

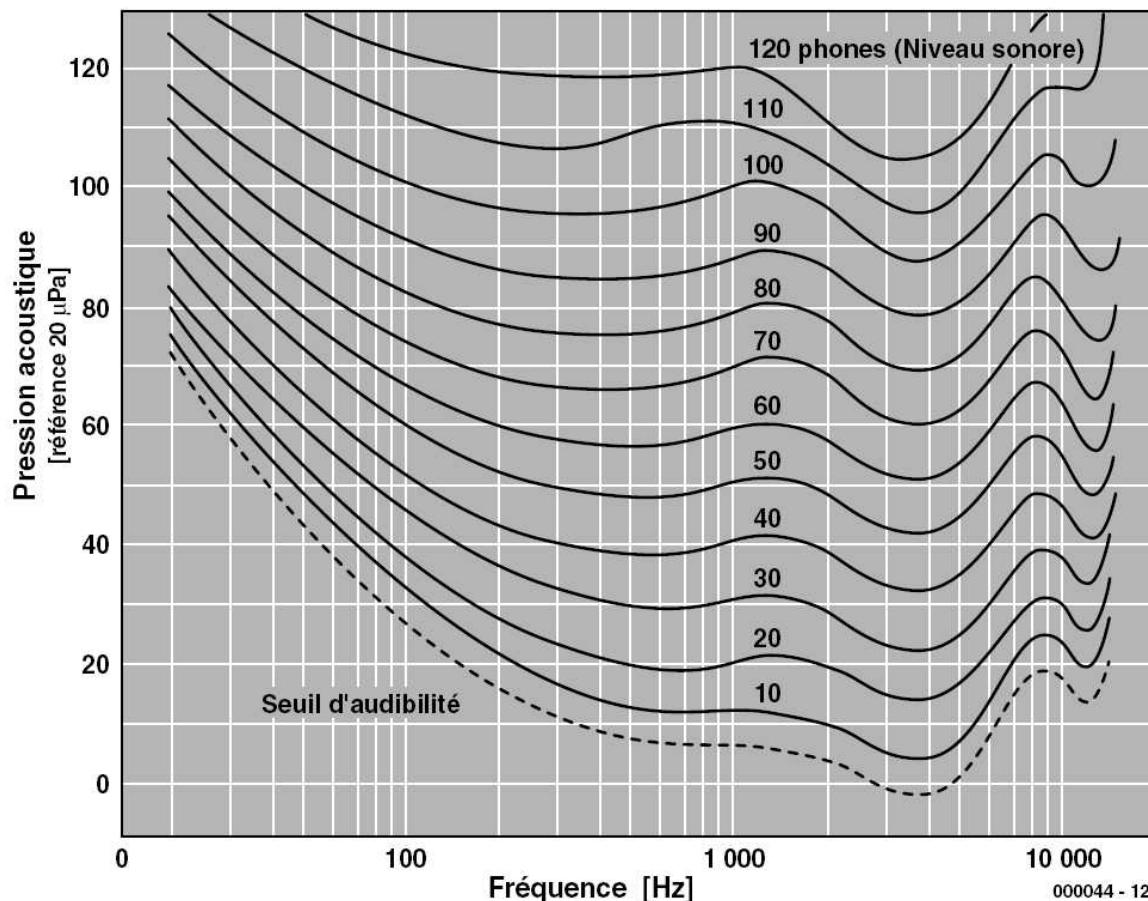
Le codage de compression MPEG-Layer3

l'autre extrémité, la membrane est lourde et souple, ce qui fait qu'elle résonne aux fréquences basses : elle est sensible aux sons graves.

La gamme de fréquences disponible détermine la plage de l'audition humaine qui, pour la plupart des gens, s'étend de 20 Hz à 15 KHz.

La figure ci-dessous montre le seuil d'audibilité et la sensibilité de l'ouïe. Le seuil d'audibilité indique le niveau d'une tonalité pure détectable dans un environnement silencieux.

On constate que le seuil d'audition est fonction de la fréquence et que la plus grande sensibilité se situe naturellement dans la gamme de fréquences de la parole (300-3000Hz).



→ Toute composante inférieure au seuil d'audibilité peut être éliminée.

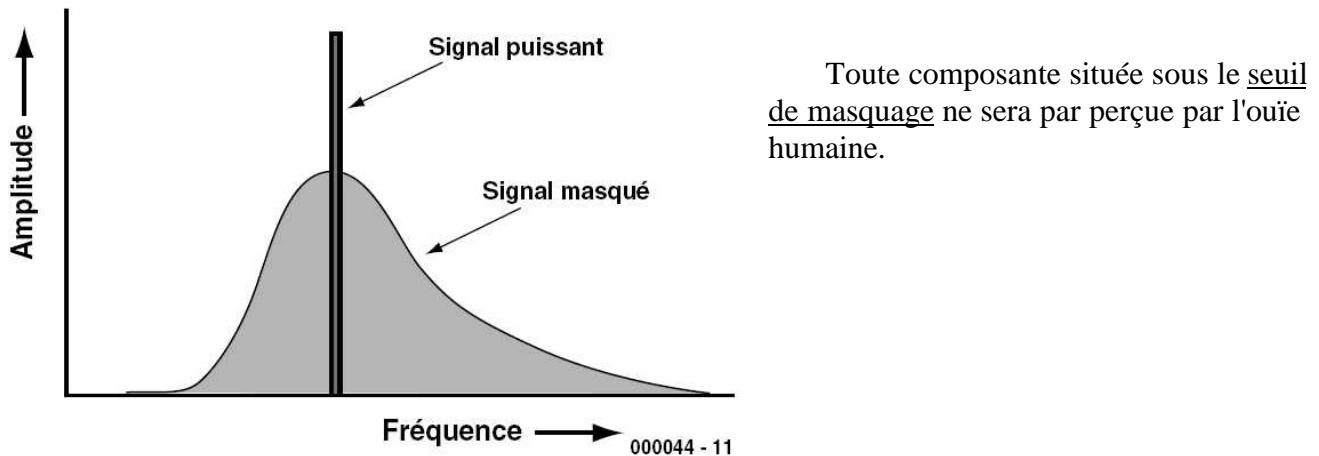
2.2 Psycho-acoustique

Les caractéristiques de l'ouïe humaine d'une part et le traitement par le cerveau des informations acoustiques (ce que l'on appelle la psycho-acoustique) de l'autre, jouent un rôle important dans le cas d'un processus de compression intelligent. En raison du comportement sélectif de la membrane basilaire, l'ouïe subdivise la plage des fréquences audibles en 25 sous-bandes définies par leurs fréquences centrales et leurs largeurs :

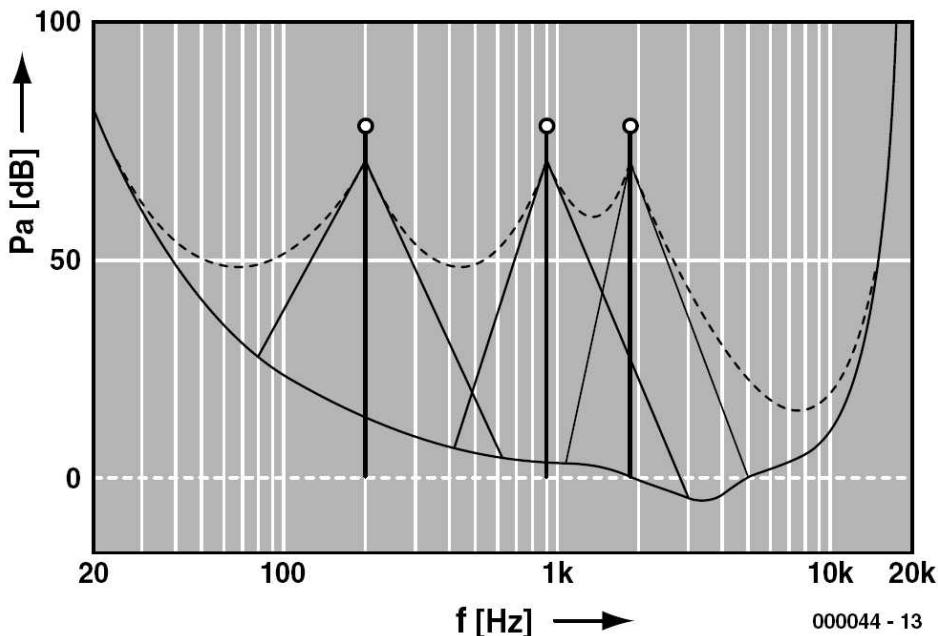
Band No.	Center Freq. (Hz)	Bandwidth (Hz)	Band No.	Center Freq. (Hz)	Bandwidth (Hz)	Band No.	Center Freq. (Hz)	Bandwidth (Hz)
1	50	-100	10	1175	1080-1270	19	4800	4400-5300
2	150	100-200	11	1370	1270-1480	20	5800	5300-6400
3	250	200-300	12	1600	1480-1720	21	7000	6400-7700
4	350	300-400	13	1850	1720-2000	22	8500	7700-9500
5	450	400-510	14	2150	2000-2320	23	10,500	9500-12000
6	570	510-630	15	2500	2320-2700	24	13,500	12000-15500
7	700	630-770	16	2900	2700-3150	25	19,500	15500-
8	840	770-920	17	3400	3150-3700			
9	1000	920-1080	18	4000	3700-4400			

Ce sont les **bandes critiques** (nommées **barks** dans la littérature spécialisée).

Sons masqueurs : en présence d'une tonalité puissante centrée sur une des bandes critiques, la zone concernée de la membrane basilaire de l'oreille est saturée. Si une autre tonalité, moins puissante mais située dans la même bande, se présente, **elle ne sera pas perçue : elle est masquée par la tonalité principale**.



Sur la totalité du spectre, le seuil de masquage est déterminé en fonction du seuil d'audibilité dans le silence (voir figure du §2.1) et du niveau et de la fréquence de chaque masqueur.



Exemple :

La figure ci-contre montre en pointillé **la courbe du seuil de masquage** déduite du seuil d'audibilité. Il est, en quelque sorte, "remonté" par la présence de 3 masqueurs.

Toute composante spectrale située sous cette courbe ne sera pas perçue par l'ouïe humaine. Elle pourra donc être éliminée sans dénaturer le message sonore.

→ *Tous les procédés de compression exploitent ces "anomalies" psycho-acoustiques de l'oreille humaine.*

La courbe du seuil de masquage étant une fonction de la fréquence, les codeurs (ou compresseurs) décomposent le spectre du signal audio en bandes de fréquence aussi proches que possible des bandes critiques. Ils calculent alors cette courbe de masquage et l'exploitent pour compresser le débit binaire en éliminant les composantes masquées.

Mais, pour atteindre une qualité "Haute Fidélité" cette méthode de compression se révèle insuffisante pour atteindre l'objectif de 64 kbps par voie. On met alors en œuvre des procédés complémentaires décrits dans le paragraphe suivant.

3. Principe du codeur MPEG-Layer3

3.1 Caractéristiques et performances du codeur

- Flux d'entrée : échantillons 16 bits ou 24 bits avec FE=32kHz, 44,1kHz ou 48kHz
Ce flux est généralement produit par la lecture de fichiers "wave"
- Flux de sortie : conforme à la norme avec des débits de 32kbps à 160kbps (en stéréo)
Ce flux peut être produit en temps réel (ex : radios sur internet) ou stocké dans un fichier ".mp3"
La compression est optimisée pour un débit qui correspond à 1,33 bit / échantillon environ
- Le format du flux de sortie permet une lecture aléatoire (début de lecture à une position quelconque), en avance et retour rapides.
- Prise en compte des modes "mono", "stéréo" et "double mono" (2 langues par exemple)
- Possibilité d'insérer des informations de contrôle permettant la détection d'erreurs pour les transmissions peu fiables (par porteuse radio par exemple).

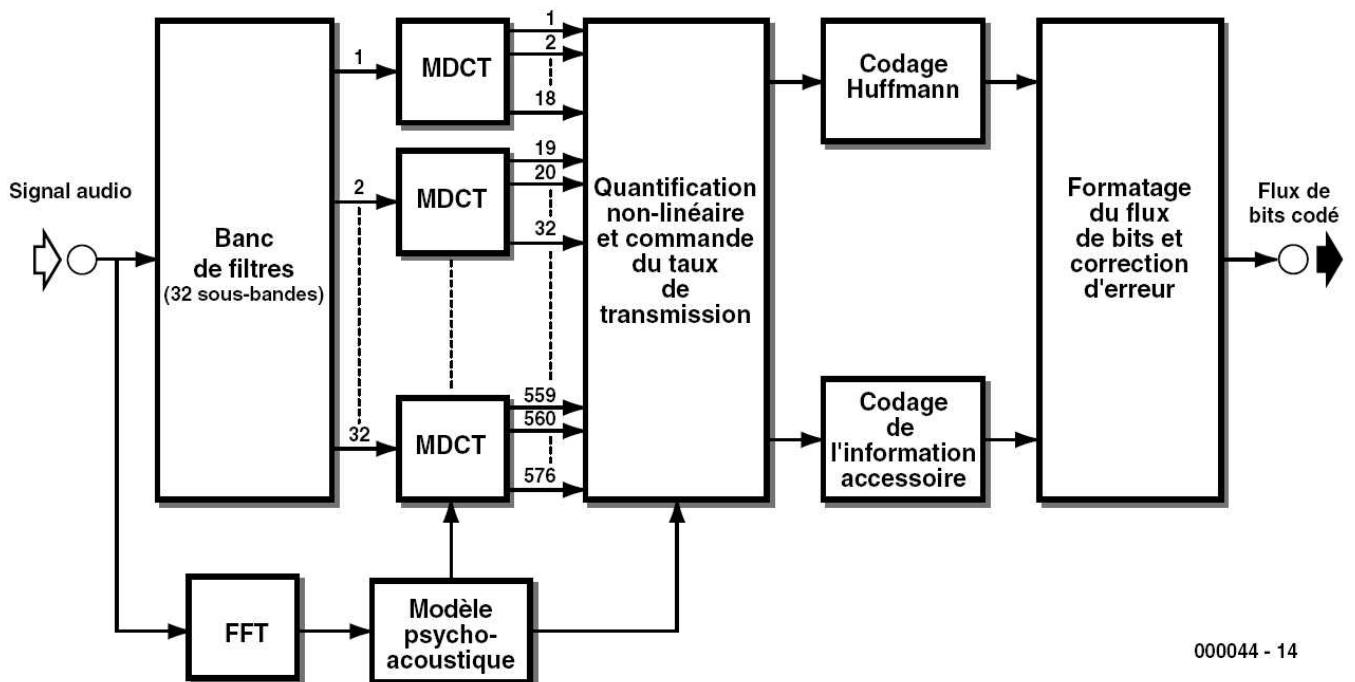
Note : les concepteurs du standard MPEG ont cherché à rendre le format MP3 aussi souple que possible, donc à limiter au maximum les éléments **normatifs**. Le flux comporte ainsi de nombreux éléments **informatifs** comme des formules de calcul et le format des données (virgule fixe ou flottante par ex.). Cela permet de faire progresser les performances audio avec l'évolution de la technologie sans remettre en cause les définitions de la norme.

3.2 Étude fonctionnelle

Le signal d'entrée du codeur est un flux numérique d'échantillons sonores, codés sur 16 bits et échantillonnés avec une des fréquences standards. Dans cette étude, on choisit FE = 44,1kHz, ce qui donne un débit binaire de 705,6 kbps par voie; le codeur va le réduire à 64 kbps.

Les fonctions décrites ci-dessous ne traitent pas les échantillons un par un mais nécessitent une connaissance plus globale du son. Les échantillons d'entrée sont donc regroupés par **trame** d'une taille de **1152 éléments** (soit une durée du son de 26mS environ avec FE=44,1kHz).

Schéma fonctionnel du codeur MPEG-Layer3



Toutes ces fonctions sont réalisées par des structures logicielles

Note : un traitement en stéréophonie nécessite 2 codeurs identiques.

- **Banc de filtres :**

Les 32 filtres passe-bande de cette fonction décomposent en fréquence le signal audio en autant de sous-bandes. Les 32 filtres sont similaires et ont tous une bande passante de 689Hz ($FE/(2*32)$) avec une fréquence d'échantillonnage de $FE=44,1\text{kHz}$). Cette décomposition est réversible et la fonction opposée est utilisée dans le décodeur pour reconstituer le signal audio.

La sélectivité des filtres n'est pas parfaite et les réponses en fréquence se chevauchent quelque peu. Les bandes de fréquence obtenues diffèrent sensiblement des bandes critiques mais les filtres sont simples et ne demandent pas beaucoup de calculs.

Pour chaque bande, le niveau d'entrée est amplifié par multiplication de sorte que la sortie soit la plus forte possible, sans débordement. Le coefficient d'amplification est déterminé pour chaque trame et il est intégré dans le flux de sortie pour que le décodeur puisse ramener le niveau à sa valeur d'origine. Ce procédé permet d'optimiser le rapport "signal/bruit de quantification" à la sortie des filtres. Autrement dit : on profite au mieux du nombre de bits utilisés pour représenter le signal numérique.

Une nouvelle valeur est calculée en sortie de chaque filtre tous les 32 échantillons

→ *Le rythme des signaux numériques à la sortie de ces filtres (repères 1 à 32) est de : 1378 nombres par seconde pour chaque filtre (ou 36 nombres par trame de 1152 bits)*

Si ces nombres sont codés sur 20 bits (cas le plus courant), on obtient un débit total à la sortie du banc de filtre de 882kbps ! A ce niveau, il n'y a pas de compression, au contraire !

- **FFT et modèle psycho-acoustique :**

La fonction FFT est réalisée sur 1024 points, ce qui donne une résolution de 43Hz avec $FE=44,1\text{kHz}$. La fonction "Modèle psycho-acoustique" utilise le contenu spectral actuel du signal audio fourni par la FFT et le seuil d'audibilité de l'ouïe pour :

- identifier les "masqueurs"
- et en déduire la courbe du seuil de masquage

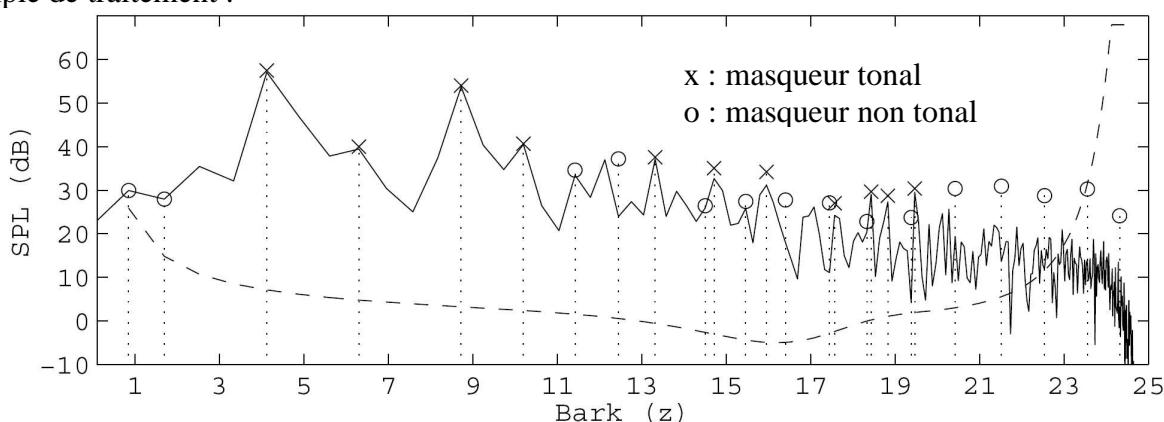
Ces informations sont calculée à chaque trame de 1152 échantillons d'entrée.

1. Identification des masqueurs

La fonction analyse **chaque bande critique** à partir de la composition spectrale fournie par la FFT. Elle commence par détecter le niveau maximum dans chaque bande critique pour ensuite identifier deux types de masqueurs :

- **masqueur tonal** (repéré par x dans l'exemple) : la raie spectrale du niveau maximum est un masqueur si les raies voisines (à une distance de 2 à 6 barks suivant la fréquence) ont un niveau plus faible de 7 dB. D'une certaine manière, on identifie les sons "purs".
- **masqueur non tonal** (repéré par o dans l'exemple) : on combine toutes les raies spectrales d'une bande critique qui n'ont pas été identifiées comme masqueur tonal. On crée alors une composante virtuelle, située au centre de la bande critique, de même énergie que l'ensemble des raies.

Exemple de traitement :



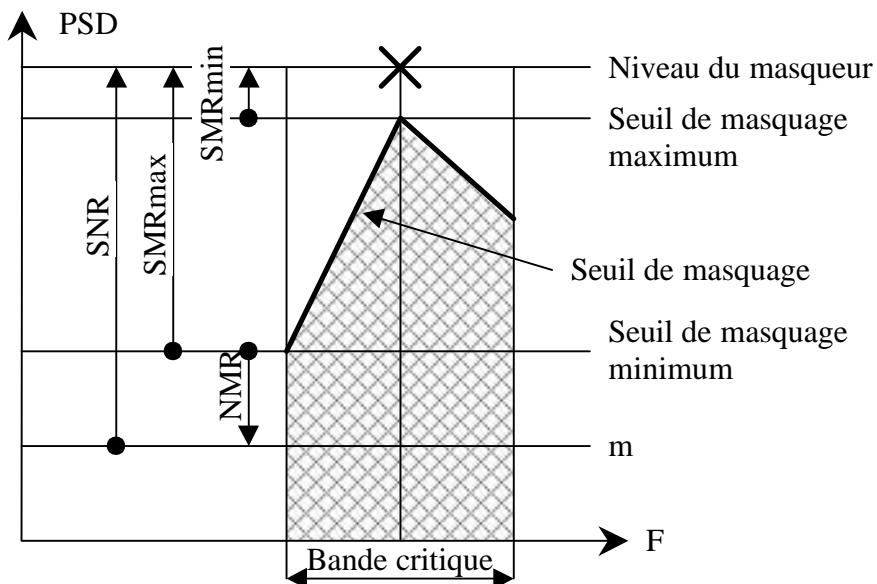
Le message analysé est de la "pop music". L'échelle de fréquence est donnée en "bark" : voir §2.2
La courbe en pointillé est le seuil absolu d'audibilité mis à l'échelle.

La résolution de la FFT est fixe (43Hz) : on obtient donc une analyse plus fine en haute fréquence qu'en basse fréquence. Ceci est constaté sur le graphe. Malgré ce défaut, on admet que l'identification des marqueurs en très basse fréquence est bonne.

2. Courbe du seuil de masquage

Comme cela a été vu au §2.2, l'effet de masquage s'étend de part et d'autre de chaque masqueur. Le MPEG a établi des lois permettant de calculer la courbe du seuil de masquage sur toute l'étendue du spectre à partir des coordonnées de chaque masqueur.

Cas d'un masqueur isolé



La courbe de masquage est définie par 2 droites de pentes différentes obtenues par des lois assez simples :

- Seuil de masquage maximum = $14,5\text{dB} + \text{"N° bande critique"}$ pour un masqueur tonal
3 à 5dB pour un masqueur non tonal
- Pente montante : 25dB/bark environ (en fait dépend du niveau du masqueur)
- Pente descendante : -10dB/bark environ (même remarque)

Le niveau repéré "m" correspond au bruit de quantification du masqueur (voir plus loin). Il est d'autant plus faible que le nombre de bits représentatifs est élevé. On en déduit les rapports suivants :

- SMR : Signal to Mask Ratio (en dB)
- NMR : Noise to Mask Ratio (dB). Ce rapport doit toujours être négatif.
- SNR : Signal to Noise Ratio (dB)

Synthèse de tous les masqueurs d'une analyse

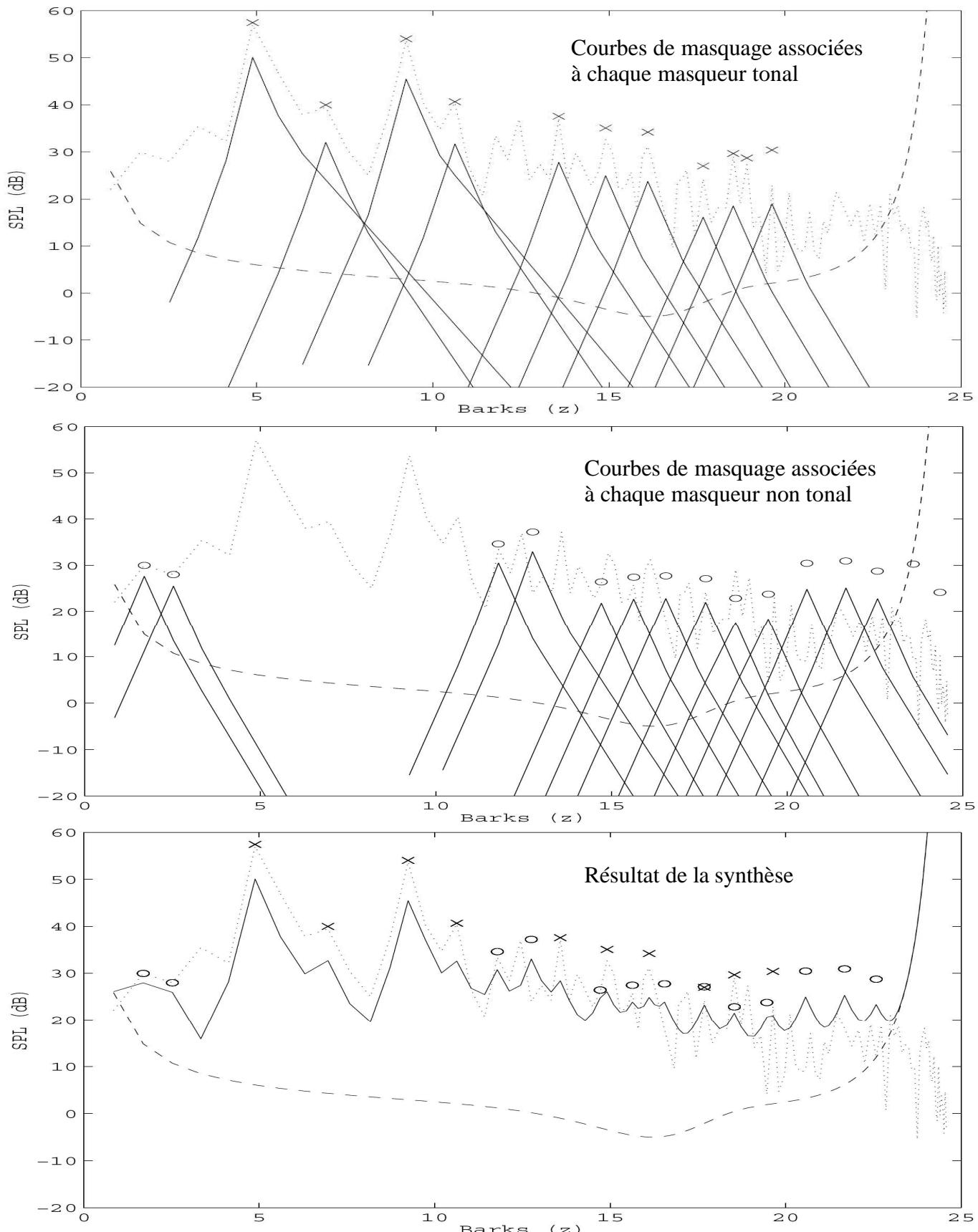
Les 3 figures suivantes montrent dans l'ordre :

- Les courbes de masquage associées à chaque masqueur tonal. On observe que les pointes sont à au moins 15dB environ des masqueurs et que l'écart augmente avec la fréquence : c'est conforme aux relations données ci-dessus.
- Les courbes de masquage associées à chaque masqueur non tonal. On constate que les pointes sont situées à 5dB environ sous les masqueurs.
- La synthèse qui en fait par le codeur MPEG. On constate qu'il a tenu compte du seuil absolu d'audibilité qui prend le dessus aux fréquences les plus élevées.

En observant attentivement, on remarque que quelques masqueurs de type tonal ont disparu. Ceci est normal car dans le cas où deux masqueurs sont très proches (moins qu'un 1/2 bark), le modèle perceptuel ne retient que le masqueur le plus puissant.

Sont aussi éliminés tous les masqueurs situés sous le seuil absolu d'audibilité.

Le codage de compression MPEG-Layer3



- **MDCT :**

L'acronyme MDCT signifie : "Modified Discrete Cosinus Transform". Il s'agit en fait de filtres passe-bande numériques réalisés à base d'algorithmes permettant de réduire la quantité de calculs. Ils sont en effet très nombreux (576 !) pour décomposer chaque bande du premier banc de filtres en 18 nouvelles bandes de fréquences. Les 576 filtres ont tous une bande passante de 38Hz environ ($FE/(2*32*18)$ avec $FE=44,1\text{kHz}$) et couvrent ainsi le spectre audio de façon très fine. Cette résolution permet ainsi de grouper ces filtres pour se conformer au mieux avec les bandes critiques de l'ouïe humaine (voir §2.2).

Pour chaque filtre, une nouvelle valeur du signal numérique est calculée tous les 18 échantillons à son entrée :

→ *Le rythme des signaux numériques à la sortie de ces filtres (repères 1 à 576) est de :*

$$1378/18 = 76 \text{ nombres par seconde environ pour chaque filtre}$$

ou 2 nombres par trame de 1152 bits

Si ces nombres sont codés sur 20 bits (cas le plus courant), le débit total à la sortie de ces 576 filtres est toujours de 882kbps ! Les fonctions suivantes vont compressés ce débit.

- **Quantification non linéaire :**

Cette fonction réalise en grande partie la compression du débit binaire. Le codage MPEG utilise simultanément 2 méthodes :

- La suppression complète des composantes totalement masquées. Il s'agit de celles situées sous le seuil de masquage calculé par la fonction "FFT et Modèle psycho-acoustique". Cela est rendu possible par le nombre de filtres MDCT qui permet une décomposition en fréquence plus fine que les bandes critiques. Les sorties des filtres MDCT concernés sont simplement ignorées.
Mais, pour atteindre une qualité "Haute Fidélité" cette méthode de compression se révèle insuffisante pour atteindre l'objectif de 64 kbps car trop peu de raies spectrales se situent sous le seuil de masquage.
- Les autres composantes ne peuvent évidemment pas être éliminées (il n'y aurait plus de son !) mais on constate qu'un bruit plus important peut être toléré dans la bande de fréquence critique concernée. Le codeur MPEG exploite ce constat.

On comprend aisément que le débit binaire global est réduit si on ignore les signaux numériques issus de quelques filtres MDCT. Mais comment un bruit plus important peut aboutir au même résultat ?

Mis à part le bruit présent dans le signal audio d'entrée, le "plancher" de bruit est déterminé par la taille des représentations des échantillons et des nombres dans le processus : c'est le **bruit de quantification**.

Par exemple, dans un codage en puissances de 2, le rapport "signal/bruit de quantification" augmente de 6dB environ par bit ajouté dans la représentation des nombres.

→ *Ainsi, si on peut augmenter le bruit de quantification tout en restant inaudible, on peut se permettre de réduire le nombre de bits représentatifs des nombres des signaux numériques.*

Ce traitement est effectué pour chaque bande critique en fonction du seuil de masquage fourni par la fonction "FFT et modèle psycho-acoustique" : le bruit de quantification doit toujours rester inférieur au seuil de masquage, tout en s'en rapprochant le plus possible : **NMR < 0dB** (voir description du modèle psychoacoustique).

Rappel : rapport S/B de quantification $\text{SNR}_{\text{dB}} = 20 \log(2^n \cdot \sqrt{1,5})$, n étant le nombre de bits.

Ex : n = 8 bits → SNR = 50dB

n = 16 bits → SNR = 98dB

Exemple

On se réfère à la courbe de masquage reproduite sur la page 9 (dernière figure).

L'écart entre la composition spectrale du signal audio (courbe en pointillé) et le seuil de masquage s'étend de quelque dB à une vingtaine de dB. On ignore évidemment les composantes situées en-dessous car elles sont totalement masquées.

Pour obtenir un rapport "Signal/bruit de quantification" (SNR) satisfaisant, il doit rester supérieur à cette quantité pour toutes les composantes quantifiées dans la bande critique correspondante.

Ainsi :

- pour les écarts les plus faibles, les composantes ne seront quantifiées que sur 1 ou 2 bits !
- pour les écarts les plus forts, on peut se contenter d'une quantification sur 4 bits

La compression est très efficace car l'origine les valeurs des signaux numériques sont codés sur 20 bits.

Le seuil de masquage est déterminé pour chaque trame d'entrée, il en résulte que le choix du nombre de bits représentifs n'est renouvelé que tous les 1152 échantillons d'entrée, donc toutes les 26mS environ. Ceci peut provoquer une petite dégradation du message dans certains cas (musique très rythmée par exemple).

L'information de résolution doit être insérée dans le flux de sortie MP3 pour que le décodeur puisse correctement reconstituer les signaux numériques.

→ *Malgré tous ces efforts, le débit global à ce niveau reste encore excessif par rapport à l'objectif de 64 kbps*

• **Codage de Huffmann**

Il s'agit d'un algorithme de compression sans perte.

La compression Huffman consiste à coder les données selon leur récurrence statistique. Plus la valeur à coder est courante, plus le code qui lui est associé est court. Au moment de la décompression, ces codes de longueurs variables sont confrontés à une table de correspondance qui restitue leur valeur initiale. Cette méthode de compression, qui n'est pas spécifique au MP3, assure à elle seule une compression de l'ordre de 20 à 25%.

→ *On abouti enfin au débit visé de 64 kbps*

• **Formatage du flux**

Le décodeur doit être capable de fonctionner même s'il ne reçoit pas le flux dès le début.

Le flux est donc organisé en trames de 1152 bits (soit 26mS environ à 44,1kHz) qui contiennent chacune toutes les informations nécessaires au décodage, c'est à dire :

- Une entête de 32 bits :
 - Un mot de synchronisation pour identifier le début de la trame
 - Le débit total (par exemple 128kbps)
 - La fréquence d'échantillonnage
 - Le n° du "Layer" (1, 2 ou 3)
 - Le mode de codage : mono, stéréo, dual mono, ...
- Eventuellement les informations de protection sur 16 bits
- Les coefficients d'amplification et les informations de quantification utilisées dans cette trame sur 136 à 256 bits
- Les données audio : il s'agit de l'ensemble des raies de la composition spectrale quantifiées suivant les règles décrites précédemment. On utilise tous les bits restants (> 848 bits).