TREIZIÈME COLLOQUE GRETSI - JUAN-LES-PINS DU 16 AU 20 SEPTEMBRE 1991

MODÈLE D'AUDITION ET RÉDUCTION DU DÉBIT NUMÉRIQUE POUR LE CODAGE DES SIGNAUX AUDIBLES: PAROLE ET MUSIQUE

T. Maier *, J. Soumagne **, B. Paillard ***

*: Rheinisch-Westfälische Technische Hochschule, Aachen, **: Ecole Supérieure d'Electricité, Metz,
***: Université de Sherbrooke (Québec)

RÉSUMÉ

Résumé : Deux modèles simples d'audition, liés au comportement de la membrane basilaire dans l'oreille interne, sont présentés. Leur utilisation permet de prendre en compte les phénomènes de masquage acoustique de l'ouie. Le débit de codage numérique peut alors être notablement réduit en se limitant à la transmission des seules composantes spectrales "non masquées".

ABSTRACT

Abstract : Two auditory models are presented. They are related to the behaviour of the basilar membrane which is a sensitive organ of the inner ear. These models permit to take into account the masking phenomena of the human hearing system. Therefore the digital coding rate can be seriously reduced if only "unmasked" frequencies are transmitted.

1. INTRODUCTION

Le développement des nouveaux services numériques concerne aussi le domaine des signaux de haute qualité. En particulier le signal de musique, dont la référence de qualité est actuellement le disque compact, ne se prête pas aisément à la transmission numérique. Le débit, par voie monophonique, est de l'ordre de 700 kbit/s ce qui est prohibitif pour les applications visées que ce soit la partie sonore de la télévision haute définition (*HDTV*) ou la future radiodiffusion numérique directe (*DAB: Digital Audio Broadcasting)* ou encore l'enregistrement (stockage) voire le visiophone. Des objectifs de réduction du débit, communs à de nombreux travaux, sont dès lors fixés pour ces applications dans une plage allant de 128 à 64 kbits/s. Les phénomènes de perception auditive (masquages fréquentiels et temporels) sont pris en compte afin de permettre une contribution très sensible à la réduction du débit numérique de transmission. La connaissance des phénomènes psychoacoustiques, issue de l'expérimentation, permet de développer des modèles de comportement de l'audition. Deux modèles de principe simple ont été conçus pour simuler les effets, qui conduisent à la perception d'un signal sonore. La "sensation basilaire" ainsi obtenue permet alors de prendre en compte les phénomènes de masquage fréquentiel. Les composantes du spectre sonore, considérées comme masquées sans altération notable de la "sensation basilaire" elle-même, peuvent alors être éliminées donc ni codées ni transmises.

Le traitement du signal sonore au niveau du modèle d'audition impose une analyse spectrale préalable. Le "spectre", ou du moins une représentation assez fidèle de celui-ci, nécessite d'importants calculs numériques. Des transformations Temps-Fréquence diverses peuvent être envisagées parmi les modèles classiques (FFT, DCT,..). Actuellement les transformées à "recouvrement" semblent préférables et notamment la MLT (*Modulated Lapped Transform*) [1] très récemment introduite qui s'avère particulièrement intéressante grâce à un algorithme de calcul rapide. De plus, le recouvrement inhérent à cette transformation limite considérablement l'effet négatif du traitement par bloc (*blocking effect*) que l'on connait avec les transformations classiques. Le modèle d'audition présenté utilise la répartition d'énergie obtenue en sortie du "banc de filtres" que constitue la MLT. Pour le modèle retenu un ensemble de 1024 filtres est utilisé avec chacun une réponse impulsionnelle formée de 2048 coefficients. Le vecteur de sortie fournit des échantillons dont la valeur, élevée au carré, représentera "l'énergie fréquentielle" du signal. On lui opposera "l'énergie basilaire" qui sera une distribution d'énergie évaluée tout au long de la membrane basilaire. A la fréquence d'échantillonnage des signaux (soit 44,1 kHz) la résolution temporelle est donc de l'ordre de 20 millisecondes, et correspond sensiblement au temps d'intégration des cellules sensorielles. La résolution spectrale, quant à elle, est de l'ordre de 20 Hertz. Ces valeurs ont aussi fourni les meilleurs résultats dans les travaux de simulation. D'un côté la représentation spectrale obtenue est uniforme sur une échelle graduée en fréquence et d'un autre côté la "localisation", sur la membrane basilaire, de chaque fréquence du spectre, est très largement non uniforme particulièrement vers les hautes fréquences.

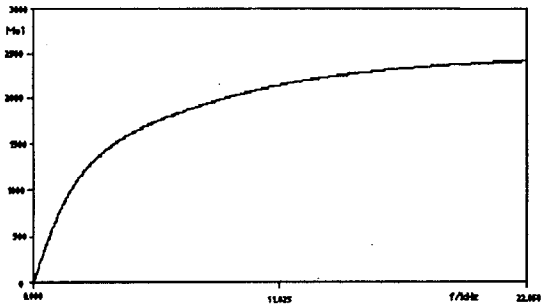

FIGURE 1 Relation entre fréquence et hauteur tonale perçue en mels.

La loi dite "loi de tonie" [2] définit la "localisation" de l'excitation des cellules distribuées tout au long de la membrane basilaire.

Une expression [3] en est donnée par la formule :

$$V = 13 \arctan(0{,}76\, f) + 3{,}5 \arctan((f/7{,}5)^2) \qquad (1)$$

où la fréquence est exprimée en kHz; la tonie obtenue ainsi est exprimée en Bark (l'échelle des Bark correspond à une subdivision uniforme de la membrane basilaire- en 24 sections- correspondant aux 24 bandes critiques). Une autre unité le "mel" est aussi utilisée (100 mels=1 Bark). La figure 1 montre la courbe calculée à partir de la relation (1) :

Les phénomènes de perception ne sont pas seulement liés aux caractéristiques du masquage fréquentiel mais aussi à certains paramètres de masquage temporel qui peuvent être introduits dans le développement d'un modèle d'audition. Deux fonctions exponentielles e^{aT} et e^{bT}, (extraites de [2]), l'une pour le masquage antérieur l'autre pour le masquage postérieur sont utilisées pour déterminer une courbe enveloppe du signal. Avec un échantillonnage à 44,1 kHz les arguments des fonctions sont alors $aT = -6{,}8027\ 10^{-3}$ et $bT = -2{,}2675\ 10^{-3}$.

La figure 2 montre la détermination d'enveloppe d'un signal avec les coefficients des masquages temporels :

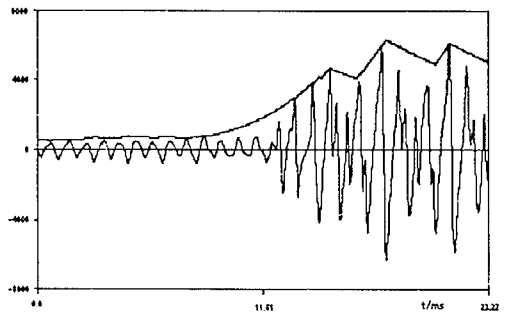

FIGURE 2 Enveloppe d'un signal

De plus et avant même d'influencer les cellules sensorielles de la membrane basilaire le signal acoustique est altéré par la fonction de transfert de l'oreille externe, pavillon et conduit auditif. Une expression analytique est proposée dans [4] et s'exprime par la relation:

$$H(f) = 10^{-3} f^4 - 6{,}5\, e^{-0{,}6\,(f-3{,}3)^2} \quad dB_N \qquad (2)$$

La figure 3 présente la forme de la réponse en fréquence de cette fonction de transfert.

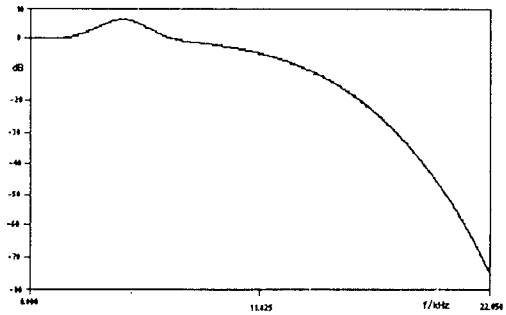

FIGURE 3 Réponse en fréquence de l'oreille externe pavillon et conduit auditif

Enfin les cellules sensorielles elles-mêmes sont affectées d'un certain niveau de bruit de fond limitant considérablement la perception des sons en basse fréquence.

Une formulation due à [4] est rappelée ici :

$$E_R = 3{,}65\, f^{-0{,}8} \quad dB_N \qquad (3)$$

avec la fréquence exprimée en kHz. On fixe à 0 dB_N la valeur 1 représentant le minimum de l'amplitude du signal. Dans la littérature [2], on définit un seuil de masquage inférieur au niveau du son masquant (de l'ordre de 4 dB_N). La formule (3) est basée sur des mesures de seuil auditif absolu; pour compenser cet effet il faut augmenter les valeurs estimées pour l'énergie du bruit de fond. Il est admis que ces 4 dB_N sont à peu près équivalents à une augmentation de 1 dB pour les résultats expérimentaux. La formule (3) est modifiée selon la relation (4) pour définir l'énergie de bruit de fond sur la membrane basilaire :

$$E_{RB} = 1{,}2589\, E_R \qquad (4)$$

2. MODÈLE D'AUDITION D'ORDRE 1

Initialement proposé par [5] ce modèle est directement lié aux résultats des expériences développées dans [2] et concernant les phénomènes de masquage de tonalités pures par des bruits à bande étroite et inversement. Ce modèle tient compte, dans l'ordre, de :

- l'effet de filtrage du pavillon et du conduit auditif
- l'effet de la loi de tonie ou "localisation"
- l'effet d'excitation des cellules de la membrane basilaire ou "dispersion" et celui du bruit de fond.

2.1 Atténuation

L'énergie du signal obtenue en sortie du banc de filtres de la MLT est modifiée par le filtre (figure 3). Lorsque les signaux proviennent d'un disque compact la "désaccentuation" des hautes fréquences est indispensable (loi EIAJ 15/50 microsecondes).

2.2 Localisation

La loi de tonie répartit de manière non uniforme, particulièrement en hautes fréquences, les points d'action de chaque fréquence du spectre. A chaque raie du spectre un point de la membrane basilaire est donc "localisé" sur une échelle évaluée en mels. Cette localisation est bien plus "dense" vers l'extrémité de l'échelle mel qui est traditionnellement graduée à partir de son origine, placée au niveau de l'hélicotrème (0 mel - lieu des basses fréquences), en direction de la fenêtre ovale (2549 mels - lieu des hautes fréquences).

2.3 Dispersion

Chaque fréquence pure reçue au niveau de l'oreille interne sollicite l'ensemble des cellules sensorielles. L'onde de propagation qui s'y est installée présente un maximum au point dit de localisation et un affaiblissement tout au long de la membrane. Inspirée des résultats expérimentaux [2] (pentes des courbes de masquage), la fonction de dispersion de l'énergie sur la membrane basilaire est fixée, et ce pour chaque composante spectrale, comme étant représentée par deux fonctions exponentielles décroissantes :

-de 0,1 dB par mel du côté des hautes fréquences
-de 0,27 dB par mel du côté des basses fréquences.

La figure 4 donne une représentation du principe.

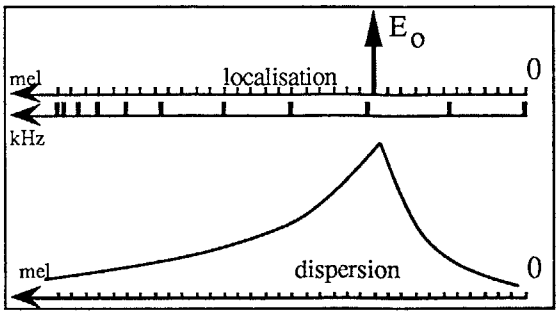

FIGURE 4 Dispersion d'énergie sur la membrane basilaire pour chaque composante spectrale

Sur une échelle x en mel, à partir d'une seule raie d'énergie spectrale E_0, chaque composante d'énergie basilaire peut être calculée de manière récursive par l'une des relations:

-à fréquence croissante $E^+(x) = E_0\,(x\text{-}1)\ 10^{-0,01}$ (5)
-à fréquence décroissante $E^-(x) = E_0\,(x+1)\ 10^{-0,027}$ (6)

Le vecteur d'énergie du signal est alors "transformé" en un vecteur d'énergie basilaire dispersée . La transformation peut s'écrire par une relation matricielle, où la matrice de transformation inclut à la fois les opérations de filtrage (oreille externe) et de localisation (loi de tonie). Les calculs représentent donc N^2 opérations. Or étant donné le choix spécifique des fonctions exponentielles pour le modèle de dispersion il est nettement préférable d'effectuer le calcul de dispersion en remarquant que chaque fonction de dispersion est une réponse impulsionnelle de filtre récursif d'ordre 1 (sur une échelle en mel et non en temps) . Ainsi la dispersion d'énergie peut être réalisée en deux étapes :

- un filtrage du vecteur d'énergie, par fréquence croissante par "pas" de 1 mel avec l'argument , a_1=0,027
- un filtrage du vecteur d'énergie, par fréquence décroissante par "pas" de 1 mel avec l'argument , a_2=0,01.

La somme des deux vecteurs donne, après 2N opérations, l'énergie totale "distribuée" sur la membrane basilaire.
Il est aussi possible de regrouper ces deux filtres, dont les sorties sont additionnées, en un ensemble de deux filtres placés en cascade. En effet, le filtrage retour (par fréquence décroissante) du vecteur déjà filtré par fréquence croissante, est permis en introduisant un facteur de normalisation , N_0= 1-a_1a_2.
La transformation inverse, pour retrouver "l'énergie fréquentielle" à partir de "l'énergie basilaire", correspond aussi à un filtrage de même nature.

2.4 Sensation basilaire

Le vecteur de sensation est défini en tenant compte de la propriété de perception logarithmique de l'ouie et de la présence de l'énergie de bruit de fond sur la membrane basilaire.

$$W(x) = \log_{10}(E_S(x) + E_R(x)) - \log_{10}(E_R(x)) \qquad (7)$$

où E_S et E_R représentent les énergies du signal et du bruit de fond sur la membrane basilaire.

2.5 Algorithme de masquage fréquentiel

On admet, en fonction de l'expérience, qu'une variation de l'énergie basilaire distribuée (signal et bruit de fond) ne dépassant pas 1 décibel est imperceptible. Ainsi on définit un algorithme de détection des fréquences masquées. Les fréquences du spectre d'énergie sont classées par ordre d'influence masquante. La variation d'énergie excitatrice juste détectable est définie par

$$\Delta\Phi = 0,1\,(E_S + E_R) \qquad (8)$$

Le niveau d'énergie du spectre réduit (aux seules fréquences masquantes) doit alors vérifier en tout point l'inégalité :

$$\Phi \geq 0,9\,E_S - 0,1\,E_R \qquad (9)$$

La recherche des fréquences masquantes, par ordre d'influence, est basée sur l'utilisation des seuils de masquage. Chaque composante d'énergie "localisée" sur la membrane basilaire crée de part et d'autre du point de localisation deux pentes de masquage (0,1 et 0,27 db/mel). Toute autre composante "localisée" en un autre point (voisin immédiat ou non) dont l'amplitude est inférieure à ces seuils de masquage est d'abord considérée comme non dominante. L'analyse par fréquence croissante puis décroissante permet de retenir en premier lieu les fréquences masquantes "dominantes". Le spectre complet d'énergie basilaire distribuée est alors recalculé uniquement à partir de ces fréquences. Si en tout point de la membrane basilaire, la relation (9) n'est pas vérifiée, une nouvelle analyse est refaite avec les fréquences "non dominantes" rejetées au tour précédent et ainsi de suite. Seules les fréquences masquantes sont conservées ; les autres sont éliminées (mise à zéro de leur amplitude). La MLT inverse ne sera donc effectuée que sur un ensemble réduit de raies, dont les signes ont été conservés, pour retrouver le signal temporel.

3. MODÈLE D'AUDITION D'ORDRE 2

Un second modèle est basé sur une modélisation des ondes de propagation sur la membrane basilaire. On fera l'hypothèse que l'excitation des cellules sensorielles est proportionnelle à l'énergie de l'onde en chacun des points de la membrane basilaire.

3.1 Dispersion

La forme (enveloppe) des ondes de propagation relevée dans [2] peut être représentée par une fonction du type:

$$U(x) = a\,.\,\sqrt{\sin(bx+c).e^{-dx+e}} \qquad (10)$$

Les constantes (a,b,c,d,e) sont déterminées à partir d'une forme d'onde, "localisée" à l'abscisse 0 mel, étendue sur toute la membrane basilaire (2551,35 mels auxquels on ajoute les 150 mels de dispersion précédents, et strictement positive sur cette distance).
On trouve a=15,5033, b=1,16297 10^{-3}, d=6,5989 10^{-3}, c=e=0.
Le facteur (a) normalise l'amplitude maximale à 1.
La figure 5 en montre l'allure générale. On notera que le point de "localisation" de la fréquence qui correspond au maximum de la courbe tient compte d'une dispersion d'énergie sur 150 mels vers les zones de fréquence basse.
On suppose que la même forme d'onde concerne individuellement chaque raie spectrale par simple translation d'une abscisse x sur la membrane basilaire correspondant à la variation de tonie T_i.

La loi de tonie est non linéaire et aux fréquences équidistantes du spectre correspondent des distances variables, exprimées en mels, traduisant les variations de tonie T_i.

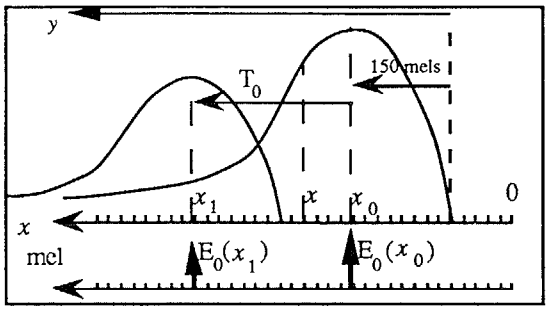

FIGURE 5 exemple d'enveloppe d'onde de propagation

A partir de la relation (10) on peut définir un nombre complexe (ou vecteur E_*) pour représenter la répartition d'énergie basilaire directement pour chaque fréquence et pour une seule onde sonore. Si x est l'abscisse courante et x_0 le point de localisation, en fixant $y = x + 150 - x_0$, on obtient :

$$E_*(x) = 15{,}5033\ E_0(x_0)\ e^{(-a + jb)y}\ u(y) \quad (11)$$

où u( ) est la fonction échelon. Alors:

$$E(x) = \text{Imag}\ [E_*(x)] \quad (12)$$

Il s'agit d'un vecteur (E_*) complexe tournant caractérisant une forme d'onde sonore. Ce vecteur peut être mis sous forme récursive en le calculant avec des déplacements de 1 mel sur la membrane basilaire (dans le sens croissant) :

$$E_*(x) = E_*(x\text{-}1)\ e^{(a + jb)} \quad (13)$$

A chaque addition d'une autre onde sonore, pour une autre fréquence, un autre vecteur tournant est ajouté à ce vecteur complexe en tenant compte du décalage de 150 mels soit:

$$E_*(x) = E_*(x\text{-}1)\ e^{(-a+jb)} + 15{,}5033\ E_0(x+150) \quad (14)$$

L'énergie totale distribuée sur la membrane basilaire est alors calculée de manière récursive (par un filtrage à coefficient complexe - ordre 2) et en une seule opération. Il n'y a pas lieu de distinguer la dispersion du côté des fréquences croissantes et décroissantes séparément comme dans le premier modèle.
En pratique on tiendra compte de la répartition non linéaire des points de localisation des fréquences en calculant directement la formule de récursion où l'on introduit des déplacements égaux à la différence de tonie entre deux fréquences. Soient x_{i-1} et x_i les deux abscisses de localisation et $T_{i-1} = x_i - x_{i-1}$ la différence de tonie en mels. Pour chaque nouvelle fréquence localisée à x_i la formule (14) prend la forme:

$$E_*(x_i\text{-}150) = E_*(x_{i-1})\ e^{(-a+jb)^{T_{i-1}}} + 15{,}5\ E_0(x_i) \quad (15)$$

Connaissant la loi de tonie, les angles de rotation du vecteur complexe sont calculés et tabulés pour une exploitation rapide. La normalisation de l'énergie de la forme d'onde (fig.5) introduit un facteur de correction (valant $b/(a^2+b^2)$)

3.2 Algorithme de masquage fréquentiel

Les variations de dynamique du signal ont pour effet (néfaste) d'étaler le spectre d'énergie calculé au moment des transitions. Le masquage appliqué trop brutalement provoque une reconstitution perturbée du signal qui est audible. Il s'agira donc d'introduire un facteur lié à la dynamique du signal. On le définit par le rapport entre le maximum et le minimum atteints par la courbe enveloppe présentée figure 2 et liée aux caractéristiques des masquages temporels. Soit D ce facteur.
La subdivision en bandes critiques, chacune d'une tonie évaluée à environ 100 mels, est aussi introduite. Ainsi pour chaque fréquence dont l'énergie "localisée" sur la membrane basilaire est E_i, en un point où la différence de tonie (entre deux fréquences consécutives est T_i, on définit une "énergie critique" associée à la bande critique au même point par :

$$E_{Ci} = 100\ .\ (E_i / T_i) \quad (16)$$

L'algorithme de masquage ne nécessite plus qu'un seul calcul, sans aucune itération comme le précédent. A chaque contribution nouvelle E_i d'une composante spectrale, dans le calcul de la dispersion (rotation du vecteur complexe, relation (15)) l'énergie "critique", calculée par (16), est comparée au seuil de détection. Ce seuil est fixé à 1 dB du niveau de l'énergie complète du signal $E_S(x_i)$ en ce point ; celle-ci correspond à la partie imaginaire du vecteur complexe tournant calculé en ce point, voir (12), à partir de l'origine de l'échelle mel.
Pour tenir compte du facteur de dynamique D on diminue le seuil de masquage par celui-ci. De la sorte seront éliminées les raies dont l'énergie "critique" correspond à:

$$E_{Ci} \leq \alpha\ E_S(x_i)/(10.D) \quad (17)$$

α est un facteur de sécurité ajouté pour tenir compte des écarts entre spectre réel et spectre calculé par MLT ($\alpha = 1/2$).

4. CONCLUSION

L'utilisation des modèles d'audition présentés permet de réduire le contenu spectral, en termes de nombre de composantes à coder et transmettre, de manière appréciable. De 60 à 80 % de ces composantes sont éliminées sans phénomènes audibles particulièrement présents. Un codage efficace et à allocation dynamique des bits permet une transmission à des débits moyens inférieurs à 64 kbit/s. Dans certains cas le débit peut être réduit à 35 ou 40 kbit/s à la limite de perception.

RÉFÉRENCES

[1] H.S. Malvar, "Lapped Transforms for Efficient Transform/ Subband Coding" IEEE Trans. on ASSP Vol.38, N°6, June 90.
[2] E. Zwicker, E. Feldtkeller, "Psychoacoustique; L'oreille récepteur d'information", Coll. CNET ENST , Masson Ed.
[3] E. Zwicker, E. Terhardt, "Analytical Expressions for Critical Band Rate", Journal of Acoust. Soc. of America, Vol 68, Nov 80
[4] E. Terhardt, G. Stoll, M. Sweeman, "Algorithm for Extraction of Pitch and Pitch Salience from Complex Tonal Signals", JASA, Vol.73, March 82.
[5] B. Paillard, J. Soumagne, P. Mabilleau , S. Morissette, "A Perceptual Distance Criterion for the Coding of Speech or Music Signals" EUSIPCO 88, Signal Processing IV, EURASIP 88.