RÉALISATION ET ÉVALUATION DE CODAGES NUMÉRIQUES DU SON DE HAUTE QUALITE POUR LA RADIODIFFUSION

PHASE III

Rapport final

CENTRE DE RECHERCHE SUR LES COMMUNICATIONS

DÉPARTEMENT DE GÉNIE ÉLECTRIQUE

FACULTÉ DES SCIENCES APPLIQUÉES

UNIVERSITÉ DE SHERBROOKE

TÉL.: 819-821-7141

TÉLEX 05-836149

FAX: 821-7903

SHERBROOKE, QUÉBEC, CANADA, J1K 2R1

CENTRE DE RECHERCHE SUR LES COMMUNICATIONS

Faculté des sciences appliquées

Université de Sherbrooke

RÉALISATION ET ÉVALUATION DE CODAGES NUMÉRIQUES DU SON DE HAUTE QUALITE POUR LA RADIODIFFUSION

PHASE III

Rapport final

Ministère des Communications du Canada
Ottawa - Canada

dans le cadre du

programme de Centres d'excellence

Contrat no. 36100-9-0208

Rédigé par Bruno Paillard

Responsables du projet:

Philippe Mabilleau
Sarto Morissette, Dir. CRCS

Sherbrooke, Québec

Le 26 mars 1990

SOMMAIRE

Les sections 1, 2 et 3 de ce rapport établissent l'état de l'art dans les domaines du codage en sous-bandes et par transformées des sons audio.

Les sections 4, 5 et 6 font état des travaux effectués par le Centre de Recherche sur les Communications de l'Université de Sherbrooke.

La description technique détaillée de ces travaux se retrouvent surtout dans les annexes 1 et 2.

TABLE DES MATIERES

1 Introduction

Depuis quelques années, plusieurs laboratoires, tant en Europe qu'en Amérique du Nord, se sont lancés dans l'utilisation de propriétés de perception de l'oreille pour le codage des signaux musicaux de haute qualité.

Dans le principe, tous les codeurs perceptuels fonctionnent de la même façon. Ils procèdent tout d'abord à une analyse temps-fréquence du signal (transformée de Fourier, décomposition en sous-bandes, etc...), puis en fonction de la répartition spectrale d'énergie du signal, et en faisant intervenir des règles de perception auditive (fonctions de masquages fréquentiels, sensibilité logarithmique, etc...), la précision de la quantification est ajustée pour chaque composante fréquentielle. Après quantification des composantes fréquentielles, le signal temporel est reconstruit.

Dans le principe donc, tous les codeurs perceptuels sont semblables à des codeurs par transformée, à la différence que l'attribution du nombre de bits de quantification suivant les composantes fréquentielles fait intervenir des lois de perception auditive, et n'est pas basée uniquement sur la répartition fréquentielle d'énergie.

Toutefois, il faut se garder de ne considérer ces codeurs que comme des "*codeurs par transformée améliorés*". En effet, dans le cas des codeurs par transformée, la décomposition temps-fréquence est utilisée pour déséquilibrer la répartition moyenne du signal suivant les différentes composantes, et ainsi tirer parti de la redondance (donc des propriétés statistiques) du signal. En ce qui concerne les codeurs perceptuels, la décomposition temps-fréquence est utilisée pour se placer dans un espace d'analyse proche de l'espace d'analyse de l'oreille, et donc pour pouvoir appliquer facilement et efficacement des règles de perception et des critères de distance perceptuels. On cherche dans ce cas à tirer parti des différences de pertinence, pour l'oreille, entre les différentes composantes du signal.

L'objectif n'est donc pas le même du tout. De même les propriétés souhaitables d'une décomposition temps-fréquence ne sont pas identiques. Dans le premier cas on cherchait une décomposition qui fournisse une répartition fréquentielle la plus déséquilibrée possible, alors que dans le cas des codeurs perceptuels, on désire une décomposition pour laquelle les composantes fréquentielles du signal sont bien séparées spectralement (les filtres d'analyse ont des transitions abruptes et une bonne atténuation en bande coupée).

Dans la pratique, les auteurs utilisent une vaste gamme de décompositions temps-fréquence. C'est pourquoi dans un premier temps nous essaierons de faire le point sur ces décompositions. Dans un deuxième temps, nous aborderons le problème du codage proprement dit en essayant de mettre en lumière les différentes qualités des solutions et des approches existantes.

2 Techniques d'analyse temps-fréquence

DTF (Discrete Fourier Transform), DCT (Discrete Cosine Transform), décompositions en sous-bandes, filtres polyphases, TDAC (Time Domain Aliasing Cancelation), etc... toutes ces techniques d'analyse temps-fréquence peuvent être décrites par le formalisme général de la décomposition en sous-bandes, elles constituent donc toutes, des cas plus ou moins particuliers de décomposition en sous-bandes. Cette similitude formelle entre les différentes techniques d'analyse temps-fréquence est très bien décrite dans [25]; une autre référence pourrrait être [20]. En pratique, des différences notables les distinguent, au point de vue de la qualité de la reconstruction (exacte ou approximative), la rapidité des algorithmes de calcul, l'orthogonalité de la transformation, le rapport longueur d'analyse/nombre de composantes fréquentielles, le taux de décimation, les caractéristiques spectrales des filtres d'analyse, etc... Dans le cadre du codage des signaux de haute qualité, il n'y a apparemment pas de consensus quant à l'utilisation de l'une ou l'autre de ces techniques, chacune ayant, bien entendu, ses avantages et ses inconvénients.

2.1 Décomposition en sous-bandes: cas général

Dans le cas le plus général de décomposition en sous-bandes, le nombre de sous-bandes peut être quelconque. Il n'y a pas non plus de contraintes sur les largeurs spectrales des bandes (elles peuvent être différentes les unes des autres). Récemment une méthode a été proposée [26] pour trouver le banc de reconstruction parfaite correspondant à un banc de décomposition arbitraire satisfaisant à un critère d'inversibilité. Il est à noter que ce critère d'inversibilité est satisfait en général si les taux de décimation dans chaque sous-bande sont en rapport inverse des largeurs de bande (critère de Nyquist), et si il existe un certain recouvrement entre les spectres de bandes adjacentes. Cette méthode permet, en principe, de construire les filtres d'analyse individuellement par des méthodes classiques de conception de filtre, et par une méthode itérative, calcule le banc de reconstruction optimal correspondant. La reconstruction n'est pas exacte, mais si on autorise des réponses impulsionnelles suffisamment longues pour les filtres de reconstruction, la qualité de la reconstruction

obtenue peut être arbitrairement bonne. Le plus gros avantage de cette solution est la grande liberté que l'on a quant à la conception du banc de filtre d'analyse. On peut imaginer, par exemple, une solution où les filtres d'analyse reproduiraient fidèlement le comportement mécanique de chacune des quelques milliers de cellules détectrices de la membrane basilaire. On aurait alors la possibilité de travailler directement dans l'espace d'analyse de l'oreille (et non pas dans un espace d'analyse proche de celui de l'oreille) et de mettre en oeuvre des modèles d'audition beaucoup plus performants.

La décimation est critique pour cette solution (il y a autant d'échantillons par unité de temps pour l'ensemble des sous-bandes que dans le signal temporel original). Malheureusement il n'existe pas d'algorithmes rapides pour la mise en oeuvre de ces bancs de filtres dans le cas général, ce qui rend cette solution difficile à utiliser dans l'état actuel de la technologie.

2.2 Transformations unitaires

Mise à part la solution générale que nous venons de décrire, toutes les décompositions existantes constituent des transformations unitaires du signal. Les termes utilisés pour décrire ces systèmes sont: "sans perte" ("lossless" en anglais), "paraunitaires", "unitaires" ou "orthogonaux". Les transformations unitaires ont certaines propriétés intéressantes, deux des plus notables sont:

- Si on introduit du bruit sur chaque composante fréquentielle, (dû à une quantification par exemple), l'énergie de bruit sur le signal temporel reconstruit est simplement égale à la somme des énergies de bruit sur chacune des composantes fréquentielles.

- Les réponses impulsionnelles des filtres de reconstruction sont obtenues par retournement temporel de celles des filtres de décomposition.

2.2.1 Galand/Johnston - Smith et Barnwell - Vaidyanathan et al

A part les transformations segmentaires (DFT, DCT ...) les premières réalisations de décompositions unitaires ont été les structures de Galand [5], avec une optimisation des coefficients faite par Johnston [4]. Un peu plus tard, Smith et Barnwell [10] ont proposé une solution similaire plus générale.

Toutes ces structures sont basées sur une décomposition en deux sous-bandes. Des décompositions plus importantes peuvent être réalisées par applications successives, en arbre binaire, de cette décomposition de base.

La qualité de la reconstruction est parfaite pour la solution de Smith et Barnwell, et (bien qu'approximative) excellente pour la solution de Galand/Johnston.

La décomposition en arbre binaire peut donner une décomposition en N sous-bandes de mêmes largeurs (N étant une puissance de 2), ou une décomposition en progression géométrique (figure 1).

Cette décomposition en progression géométrique est intéressante pour les codeurs perceptuels car elle suit approximativement la loi de conversion fréquence → lieu basilaire (en fait les auteurs font en général référence à la largeur des bandes critiques, mais cette dernière suit aussi la loi de conversion fréquence → lieu basilaire).

Plus récemment, Vaidyanathan et al [16], [24] ont proposé des méthodes plus générales de synthèse de bancs de filtres unitaires. Les principaux avantages de ces méthodes par rapport aux deux précédentes sont les suivants:

- En principe toutes les décompositions unitaires sont synthétisables par ces méthodes, les solutions obtenues ne sont donc pas des cas particuliers.

- Le nombre de sous-bandes peut être quelconque, toutefois les sous-bandes doivent avoir la même largeur.

  Ces solutions autorisent toujours une décomposition en arbre, et en particulier une décomposition en progression géométrique, mais ne sont pas limitées au cas des arbres binaires (des arbres d'ordre supérieur sont possibles).

- La reconstruction est parfaite.

Ces solutions sont donc toujours avantageuses par rapport aux solutions plus anciennes de Galand ou Smith et Barnwell.

Pour ces trois groupes de solutions (Galand/Johnston, Smith et Barnwell, Vaidyanathan et al), la décimation est critique, la séparation spectrale entre les sous-bandes (largeur de la transition, atténuation en bande coupée) peut être arbitrairement

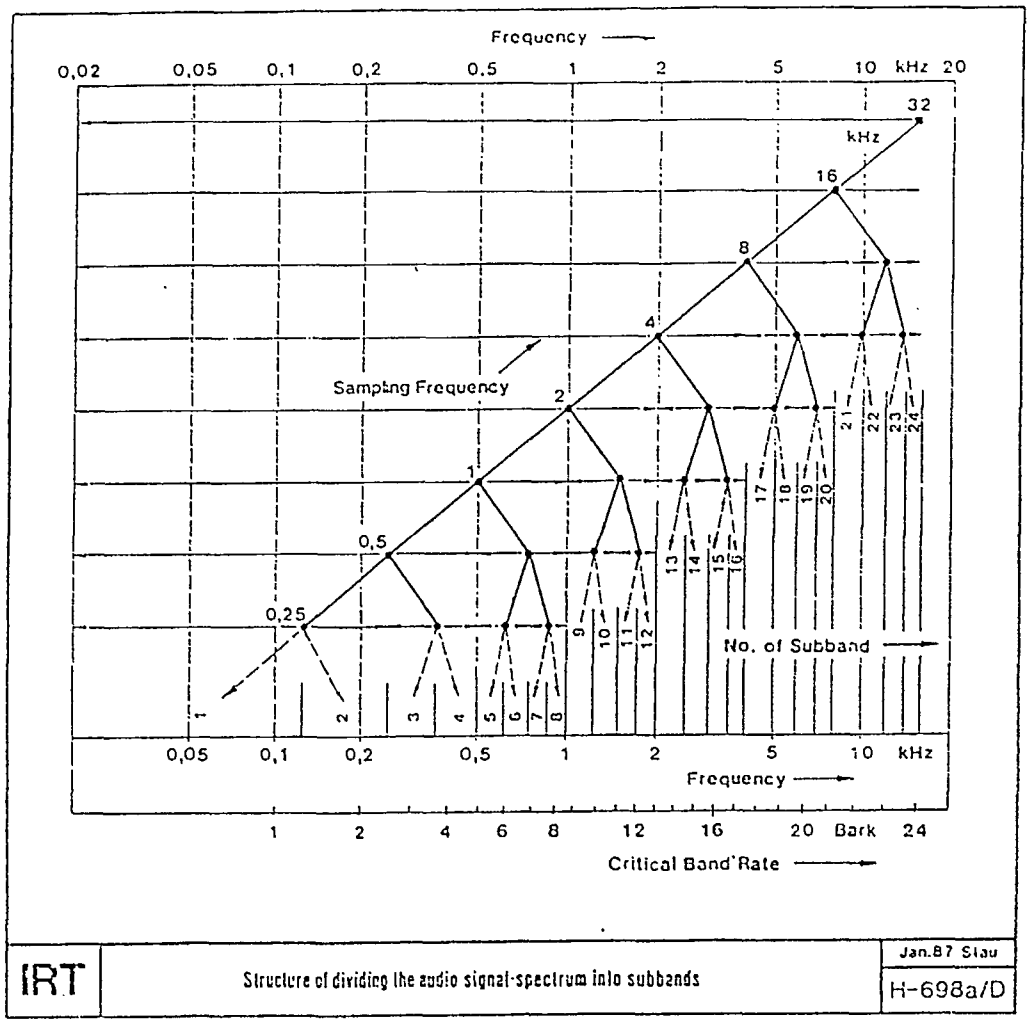


Figure 1 Décomposition en arbre binaire en progression géométrique

bonne si on tolère des filtres ayant des réponses impulsionnelles suffisamment longues.

Dans certains cas, la symétrie particulière des filtres permet d'accélérer un peu les calculs, mais il n'existe pas d'algorithme de calcul rapide à proprement parler. L'absence d'algorithmes de calcul rapide rend ces solutions mal adaptées au cas où on veut une décomposition en un grand nombre de sous-bandes (512, 1024, ...).

2.2.2 Les filtres polyphase

Proposés par Rothveiler [8], ils sont une évolution de solutions plus anciennes dues à Darlington [1], puis Bélanger [3]. Le banc de filtres est ici synthétisé en modulant un filtre passe-bas prototype par des fonctions sinusoïdales. La décimation est critique. La décomposition peut être effectuée en un nombre quelconque de sous-bandes d'égales largeurs. Ici aussi, si on accepte des filtres ayant une réponse impulsionnelle suffisamment longue, on peut obtenir une séparation spectrale arbitrairement bonne. La reconstruction est approximative pour 2 raisons:

- La méthode ne permet d'annuler que la distorsion de recouvrement entre deux bandes adjacentes. Entre deux bandes plus éloignées, on suppose que l'énergie de recouvrement est suffisamment faible pour être négligée.

- Le filtre passe-bas prototype est calculé par un algorithme d'optimisation (du même type que l'algorithme utilisé par Johnston pour optimiser les filtres de Galand).

  La solution peut donc être qualifiée de quasi-unitaire.

Bien qu'il n'existe pas à proprement parler d'algorithme de calcul rapide, les symétries particulières de la solution permettent un calcul plus rapide que dans le cas des filtres de Galand, Smith et Barnwell ou Vaidynathan et al. Il est tout de même difficile de mettre en oeuvre ces solutions pour un grand nombre de sous-bandes (512, 1024 ...).

2.2.3 Transformations segmentaires

Généralement seules la DFT (Discrete Fourier Transform) ou la DCT (Discrete Cosine Transform) sont utilisées.

Pour ces solutions bien connues, et en utilisant la terminologie des décompositions en sous-bandes, la décimation est critique, la reconstruction est exacte et il existe des

algorithmes de calcul rapide permettant une décomposition en un nombre important de sous-bandes.

Le nombre de sous-bandes (en général on les désigne par "composantes fréquentielles") doit être une puissance de 2 et toutes les sous-bandes ont la même largeur spectrale. Le plus gros inconvénient de ces transformations est leur mauvaise résolution spectrale. En effet, pour ces solutions, les longueurs des réponses impulsionnelles des filtres d'analyse (qui sont ici des tronçons de sinusoïdes ou d'exponentielles complexes) ne peuvent être plus grandes que le nombre de composantes spectrales. En toute exactitude, cette longueur d'analyse est strictement égale au nombre de composantes spectrales puisque ces transformations sont carrées (sans recouvrement). Cette relative petite taille de la réponse impulsionnelle des filtres d'analyse limite énormément leur qualité d'analyse en termes de largeur de la bande de transition, et atténuation en bande coupée.

Pour pallier à cet inconvénient, il est possible d'appliquer à ces transformations un fenêtrage adoucissant les bords du segment temporel analysé. Moyennant un léger recouvrement entre 2 transformations successives, la qualité de l'analyse est améliorée et la reconstruction est toujours exacte. Cependant, dans ce cas, la décimation n'est plus critique et il y a plus de coefficients spectraux par unité de temps pour l'ensemble des sous-bandes, qu'il y en a dans le signal temporel. Ce qui, dès le départ, est un handicap pour le codage.

2.2.4 TDAC, LOT et MLT

En 1987, Princen [15] proposait les TDAC (Time Domain Aliasing Cancelation). Ces solutions peuvent être vues comme des cas particuliers de filtres polyphase. Pour ces cas particuliers, la longueur de la réponse impulsionnelle des filtres d'analyse est égale au double du nombre de sous-bandes, ces solutions ont donc une bien meilleure résolution spectrale que les transformations segmentaires. La décimation est critique, la reconstruction est exacte et ce sont des transformations unitaires. Les sous-bandes sont de largeurs égales et il peut y en avoir un nombre quelconque.

Dans son article, Princen ne dit rien d'un possible algorithme de calcul rapide, cependant Brandenburg a apparemment adapté la méthode pour sa mise en oeuvre à partir d'une transformée en cosinus rapide. Dans ce cas, cette méthode se prête bien à une décomposition en un nombre important de sous-bandes, pourvu que ce nombre soit une puissance de 2.

Tout comme les TDAC, les LOT (Lapped Orthogonal Transforms) ont une longueur d'analyse égale à deux fois le nombre de sous-bandes, leur résolution spectrale est donc bonne aussi. Les LOT ont été introduites par Malvar et Staelin en 1988 pour les besoins du codage d'image [21].

Ces solutions permettent une décomposition en N sous-bandes (N étant une puissance de 2) de largeurs égales. La décimation est critique, la reconstruction est exacte et ce sont des transformations unitaires. Leur mise en oeuvre est basée sur des algorithmes de calcul rapide et donc elles se prêtent bien à une décomposition en un nombre important de sous-bandes.

Tout récemment, Malvar a introduit les MLT (Modulated Lapped Transforms) qui constituent un type particulier de LOT (article à paraître dans IEEE Transactions on ASSP) ces MLT sont basées sur les TDAC et ont des algorithmes de calcul encore plus rapides que les LOT. A noter que les algorithmes de calcul pour les LOT ou les MLT sont à base de transformée en cosinus rapide, et donc les temps de calcul sont du même ordre de grandeur que pour les FFT.

Les MLT ont des propriétés similaires aux LOT:

> **Note**: il est possible que l'algorithme rapide pour les TDAC auquel Brandenburg fait allusion dans [35] soit du même type que l'algorithme des MLT de Malvar. Cet algorithme aurait dû faire l'objet d'une publication à EUSIPCO 88 mais nous n'en avons pas trouvé trace dans les proceedings. Dans [35], Brandenburg indique toutefois qu'un fenêtrage explicite du segment temporel analysé doit être effectué, avant transformation par DCT modifiée. Pour les MLT, ce fenêtrage est implicitement contenu dans une transformation rapide.

En conclusion

De toutes ces solutions, les MLT (ou éventuellement l'algorithme rapide pour TDAC de Brandenburg) semblent le mieux adaptées aux besoins du codage perceptuel. En effet, elles ont l'avantage d'avoir des algorithmes de calcul rapide, rendant possible une décomposition en un nombre important de sous-bandes. Grâce à leur longueur d'analyse double du nombre de composantes, elles ont une résolution spectrale bien meilleure que les transformations segmentaires (DFT, DCT). La décimation est critique, et donc on n'a pas, dès le départ, le handicap d'avoir à coder plus d'échantillons dans le domaine spectral qu'on en aurait eu si on était resté dans le domaine temporel.

3. Perception et quantification

Comme nous le disions dans l'introduction, plusieurs laboratoires se sont lancés dans l'utilisation de propriétés de perception pour le codage des signaux musicaux de haute qualité. Ces laboratoires sont:

- L'IRT en RFA, avec les travaux de Theile, Link et Stoll, et le développement du "MASCAM', (Masking Pattern Adapted Subband Coding and Multiplex).
- L'Université de Erlangen en RFA, avec les travaux de Brandenburg, Seitzer, Kapust, Eberlein, Gerhauser, Popp et Schott, et le développement du "NMR" (Noise to Mask Ratio) et de l'"OCF" (Optimal Coding in the Frequency Domain).
- Les Bell Laboratories aux Etats-Unis avec les travaux de Johnston, et le développement du concept "d"entropie perceptuelle" et du PXFM (Perceptual Transform Coder).
- Le CNET en France avec les travaux de Mahieux, Petit et Charbonnier.
- Et, bien sûr, le CRCS de l'Université de Sherbrooke.

3.1 Travaux effectués à l'IRT

Le système de codage développé à l'IRT porte le nom de MASCAM pour "Masking Pattern Adapted Subband Coding and Multiplex".

La technique semble encore en pleine évolution, si bien que le nom 'MASCAM' ne fait pas référence à un système figé, mais à un ensemble de systèmes fonctionnant suivant le même principe général. Nous allons nous pencher sur 2 systèmes en particulier, le second faisant partie d'une proposition pour le codage son de la télévision avancée, faite en collaboration avec le CCETT (France) et Phillips.

3.1.1 Premier système

Pour ce premier système, la décomposition spectrale est effectuée par un banc de filtres QMF (apparamment la solution de Galand). Cette décomposition est effectuée en 24 sous-bandes, en progression géométrique. La progression géométrique est sans doute le moyen le plus élégant de tirer le meilleur parti (au point de vue perceptuel) du petit nombre de sous-bandes (il n'existe pas d'algorithme rapide pour ce genre de décomposition).

Toutes les 4 ms (donc à une cadence assez rapide en comparaison des autres systèmes de codage perceptuel), les maximums du signal dans chaque sous-bande sont détectés. Ces maximums vont servir à adapter la dynamique du quantificateur à celle du signal dans chaque sous-bande, et aussi à déterminer l'attribution des bits (donc la résolution de quantification dans chaque sous-bande).

La quantification est assez simplement mise en oeuvre par 24 quantificateurs scalaires, uniformes et indépendants les uns des autres.

L'attribution des bits est déterminée en fonction de la répartition fréquentielle et temporelle des maximums du signal. Elle fait donc intervenir non seulement des règles de masquages fréquentiels, mais aussi des règles de masquages temporels. Ces règles de masquages sont mises en oeuvre à partir de tables préenregistrées, mais les auteurs ne sont pas très loquaces à ce sujet.

Le codeur ne transmet au récepteur que la valeur des maximums (facteurs d'échelle) et les valeurs quantifiées des échantillons pour les 24 sous-bandes. Le récepteur doit donc reconstituer à partir des maximums (et en utilisant les mêmes règles de masquage que le codeur) la répartition optimale des bits dans les sous-bandes. Il peut ensuite reconstituer les valeurs quantifiées dans chaque sous-bande, puis reconstruire le signal temporel.

A noter dans les versions plus récentes, la présence d'un algorithme d'interpolation des facteurs d'échelle. Uniquement les facteurs d'échelle les plus critiques sont transmis au récepteur, qui effectue une interpolation pour retrouver les valeurs qui lui manquent.

Pour une qualité apparemment "transparente", le débit est fixe et de 160 Kbits/s.

> **Commentaire**: comme le souligne Theile dans [28], une opération d'égalisation (ou plus généralement de filtrage, voire de pré-emphase) peut nuire beaucoup à la qualité du signal codé par cet algorithme. En effet, ce filtrage va modifier la répartition spectrale de l'énergie du signal, et peut faire apparaître des dégradations que le codeur "pensait" inaudibles. Ce "défaut" du codeur existe pour tous les systèmes qui font intervenir efficacement des propriétés de perception dans le codage. On peut même mettre en doute l'efficacité à tenir compte des phénomènes de masquage, d'un codeur qui n'aurait pas ce problème. De même certains

codeurs sont sensibles et donc doivent connaître à l'avance le niveau absolu de reproduction de la pièce à coder. Fondamentalement, ces comportements sont donc rassurants, ils montrent (s'ils sont présents) que les propriétés de perception de l'oreille interviennent finement et efficacement dans le processus de codage.

Pour ce premier système, il faut noter:

- Que l'espace fréquentiel n'est découpé qu'en 24 zones spectrales, et bien que ce découpage soit fait de manière à épouser au mieux la loi de conversion fréquence → lieu basilaire, il est trop large pour faire apparaître éventuellement une structure harmonique fine du signal.

  Faire apparaître cette structure harmonique du signal devrait permettre de tirer parti avec plus de finesse et d'efficacité des phénomènes de masquages fréquentiels.

- Que la quantification est réalisée très simplement dans chaque sous-bande par des quantificateurs scalaires uniformes. Bien qu'il existe des techniques de quantification plus efficaces (quantification vectorielle par exemple), ce type simple de quantification permet si nécessaire une haute résolution (un nombre de bits important par échantillon) que ne permet pas la quantification vectorielle, où on est limité par une taille raisonnable de dictionnaires; et de plus, est extrêmement rapide et donc se prête bien aux grandes fréquences d'échantillonnages nécessaires à la représentation des signaux musicaux de haute qualité.

- Que la technique de codage fait intervenir les masquages temporels. Ce choix nous paraît discutable, les objections sont présentées à la section 4.3 Notons tout de même qu'il est possible que tenir compte des masquages temporels améliore la qualité subjective du codeur. Nous discutons ce choix surtout au plan formel, et compte tenu des hypothèses que nous avons faites quant au modèle d'audition.

3.1.2 Deuxième système

Récemment, l'IRT, dans le cadre du codage son de la télévision avancée, a proposé une technique de codage qui, bien qu'assez proche de celle que nous venons d'étudier, a tout de même ses particularités. Ce développement entre dans le cadre d'une coopération avec Phillips et le CCETT, et utilise le système de codage de canal COFDM (Coded Orthogonal Frequency Division Multiplex) réalisé par le CCETT.

Parmi les particularités de cette technique de codage, on trouve:

- Le codeur fonctionne sur un signal échantillonné à 48 kHz.
- Les débits proposés sont 64, 96 ou 128 Kb/s la qualité du codeur n'est donc certainement pas parfaitement "transparente".
- Le signal est décomposé en 32 sous-bandes d'égales largeurs par un banc de filtres polyphases (alors qu'il était décomposé en arbre binaire en progression géométrique par un banc de filtres de Galand pour le 1^{er} système).
- Parallèlement à cette décomposition en sous-bandes, une analyse de Fourier est effectuée, qui ne sert qu'à étudier les masquages fréquentiels et à décider de l'attribution des bits dans les sous-bandes. Bien que l'analyse fréquentielle soit bien plus fine que pour le système précédent, ce qui donc laisse supposer que les seuils de masquage peuvent être déterminés avec une plus grande justesse, l'attribution des bits est effectuée par blocs fréquentiels de largeur constante 750 Hz (les sous-bandes). En basse fréquence (et particulièrement pour la $1^{ière}$ sous-bande $0 \rightarrow 750$ Hz), cela semble un découpage très large de l'espace fréquentiel.

Les signaux transmis par le codeur au récepteur sont maintenant de 3 types:

- Les signaux quantifiés dans chaque sous-bande.
- Les facteurs d'échelle pour chaque sous-bande.
- L'attribution des bits pour chaque sous-bande.

Dans le système précédent, l'attribution des bits était calculée à partir des facteurs d'échelles, et donc seuls ces derniers étaient transmis au récepteur. La transmission explicite de l'attribution des bits simplifie énormément la tâche du récepteur qui n'a plus besoin de la calculer. La complexité du décodeur est donc dans ce cas, quasiment réduite à la reconstruction polyphase du signal, et donc est bien moindre que la complexité du codeur.

En conclusion, ce $2^{ième}$ système développé par l'IRT semble moins optimal que le précédent (découpage large et uniforme de l'espace fréquentiel, transmission explicite de l'attribution des bits alors qu'elle était déduite des facteurs d'échelle dans la solution

précédente), mais surtout moins compact (utilisation d'une analyse spectrale séparée pour le calcul de l'attribution des bits).

Par contre, il semble bien mieux adapté à une utilisation en situation de type "broadcasting" puisque la complexité du récepteur est très faible. L'utilisation d'une MLT à la place des filtres polyphases pour effectuer la décomposition et la reconstruction du signal, rendrait l'algorithme de décodage ridiculement peu complexe.

3.2 Travail effectué à l'université de Erlangen

3.2.1 Le NMR (Noise to Mask Ratio): un outil d'évaluation de codeur

A l'université de Erlangen, Brandenburg, Kapust, Eberlein, Gerhauser, Krageloh et Schott [34] ont développé un logiciel d'évaluation objective de codeur basé sur des règles de perception. Dans les grandes lignes, le NMR fonctionne comme suit:

Une analyse spectrale fine du signal est effectuée par une transformée de Fourier rapide sur 1024 points avec fenêtre de Hanning.

L'énergie du spectre est ensuite calculée et les énergies individuelles des composantes spectrales sont groupées en 27 bandes critiques qui couvrent le spectre de 0 à 20 kHz (ce découpage de l'espace basilaire en bandes critiques suit la loi de conversion fréquence → lieu basilaire).

Ensuite les seuils de masquage sont calculés pour chaque bande critique en faisant intervenir le niveau d'énergie dans la bande considérée, les masquages fréquentiels dûs aux autres bandes critiques du même bloc, ainsi que les masquages temporels dûs aux bandes critiques du bloc précédent (post-masking). Il semble que ces seuils de masquage soient calculés en faisant intervenir des "gabarits" de masquage tabulés, et apparemment dans une situation de "pire cas": on calcule pour une bande donnée les seuils de masquage dûs à chaque autre bande prise séparément, et le seuil le plus grand l'emporte. Finalement on mesure l'énergie de bruit dans chacune des bandes critiques, de la même façon qu'on a mesuré l'énergie du signal dans ces bandes (on prend la transformée de Fourier du bruit, puis on groupe les énergies des composantes spectrales du bruit dans les bandes critiques). Enfin, on compare pour chacune des 27 bandes critiques, le niveau de bruit au seuil de masquage.

Deux types d'information sont possibles:

- Le "masking-flag" indique à quels instants le niveau de bruit est supérieur au seuil de masquage dans au moins une des 27 bandes critiques.

- La valeur moyenne du rapport bruit/seuil de masquage est donnée (moyenne faite sur les 27 bandes critiques) en fonction du temps. Cette valeur indique donc l'évolution temporelle de la dégradation en fonction du temps.

En conclusion, le NMR est un outil simple, intéressant, et tout à fait novateur pour l'évaluation des codeurs. Certains points sont à noter toutefois:

- Bien que l'analyse spectrale du signal soit fine, le groupage des composantes spectrales en 27 bandes critiques effectue une discrétisation assez large de l'espace fréquentiel. Le seuil de masquage qui résulte de cette analyse est donc une fonction en escaliers (27 valeurs) qui couvre l'espace fréquentiel de 0 à 20 kHz.

- Les seuils de masquage sont obtenus par application de gabarits de masquage et en situation de "pire cas" (le seuil de masquage le plus haut obtenu est retenu pour chaque sous-bande). Mais surtout il semble (celà n'est pas très clair) que le seuil de masquage dans une bande x, dû à une bande y, soit calculé indépendamment de l'état d'énergie des autres bandes critiques, ce qui est une approximation.

Comme nous le verrons, l'utilisation d'un modèle explicite d'audition permet de remédier simplement à ces 2 problèmes, et donc de donner un seuil de masquage continu en fréquence et qui est calculé globalement pour la répartition spectrale d'énergie donnée.

3.2.2 OCF (Optimal Coding in the Frequency Domain)

Pour l'OCF, la décomposition spectrale est à base d'algorithmes de calcul rapides tels la DCT, et donc se fait en un nombre important de composantes (typiquement 512 coefficients pour une bande passante de 20 kHz). Pour les modèles anciens d'OCF ([32]), la décomposition était faite à partir d'une DCT pure et simple. Plus récemment, Brandenburg [33] utilise un algorithme rapide basé sur les TDAC. Cet algorithme est sans doute très proche de l'algorithme des MLT que nous utilisons au CRCS.

La quantification est très simple et est réalisée par un quantificateur scalaire non adaptif et indépendant pour chaque composante fréquentielle. Ici la complexité est reportée au niveau du codeur qui effectue un codage entropique (en fonction des statistiques d'apparition des valeurs quantifiées) de type codage de Huffman.

L'ajustement de la résolution de quantification pour chaque composante fréquentielle ne se fait pas par une attribution explicite de bits pour chaque composante comme c'est généralement le cas. Ici, tous les quantificateurs sont identiques, et ont un pas de quantification ainsi qu'une dynamique (donc un nombre de bits équivalent) fixés. Par contre le codeur entropique va représenter les valeurs quantifiées avec un nombre de bits variable d'autant plus faible que ces valeurs sont plus probables. Ici les faibles valeurs d'énergie sont les plus probables et donc nécessitent un nombre de bits plus faible. Ainsi, pour modifier artificiellement la résolution de la quantification d'une composante, il suffit de la multiplier par un coefficient avant la quantification (pré-emphase). Si ce coefficient est plus grand que 1, cela va augmenter la résolution pour la composante, et en même temps augmenter implicitement le nombre de bits (puisque la valeur quantifiée à coder devient moins probable). Si ce coefficient est plus petit que 1, la valeur quantifiée à coder s'approche de zéro. Elle devient donc plus probable (moins de bits) en même temps que la quantification devient plus grossière. Bien entendu, le récepteur doit connaître les valeurs de ces coefficients pré-emphase, pour être capable, après la reconstitution (déquantification) des composantes spectrales, d'effectuer la dé-emphase.

Cette façon d'aborder le problème d'ajustement spectral de la résolution de quantification est tout à fait originale et intéressante, et ressemble dans l'esprit aux techniques de réduction de bruit analogiques de type Dolby.

Le codage entropique délivre en sortie un débit variable dépendant de la probabilité d'apparition des symboles codés. Ici, le codeur ne dispose que d'une quantité fixe de bits pour coder chaque bloc. Les auteurs ont donc recours à une boucle de réaction qui va modifier uniformément la résolution de la quantification sur tout le spectre jusqu'à ce que le débit demandé par le codeur entropique soit dans la limite du nombre de bits disponibles pour le bloc. Il y a donc un nombre variable d'essais-erreurs pour décider de la précision de codage pour chaque bloc.

Critères de perception

Pour l'OCF, contrairement à d'autres algorithmes de codage perceptuel, il n'y a pas à proprement parler de mise en forme spectrale du bruit.

Les composantes spectrales sont groupées en bandes critiques, et on mesure l'énergie totale du signal dans chaque bande. A partir de cette répartition d'énergie, et sans doute par une procédure semblable au NMR, l'algorithme calcule le seuil de masquage du bruit pour chaque bande critique. Lorsque l'ensemble quantificateur-codeur a décidé d'une

résolution de quantification qui entre dans la limite de bits disponibles, les composantes spectrales quantifiées sont reconstituées (déquantifiées) puis l'énergie du bruit de quantification est estimée pour chaque bande critique, et comparée au seuil de masquage. Si pour une ou plusieurs bandes le bruit dépasse le seuil de masquage (et seulement dans ce cas), une résolution plus grande est accordée aux composantes qui sont dans ces bandes (relativement aux autres composantes), et une nouvelle recherche s'effectue pour obtenir une résolution qui entre dans la limite du nombre de bits disponibles. On voit donc que l'algorithme tend uniquement à baisser le niveau de bruit pour les bandes critiques "à risque", ce qui constitue une stratégie relativement différente de l'ensemble des autres systèmes de codage perceptuel.

L'information transmise au décodeur comprend les valeurs quantifiées et codées des composantes spectrales, ainsi que les valeurs des coefficients de pré-emphase pour chaque bande critique.

Le décodeur est donc très peu complexe, puisqu'il lui suffit de décoder les valeurs quantifiées des composantes spectrales (à l'aide d'une table de décodage de Huffman), puis de reconstituer ces composantes, de les diviser par le coefficient de pré-emphase, puis de reconstruire le signal temporel à partir de ces composantes.

Le débit final est de 2.5 bits/échantillon (soit environ 128 kb/s) pour une qualité proche de celle d'un compact-disk.

En conclusion, on voit que la fonction de masquage, et donc les attributions des bits sont calculées par blocs fréquentiels correspondant aux bandes critiques. Il n'y a pas de mise en forme spectrale du bruit à proprement parler, mais plutôt une limitation du bruit dans les zones spectrales "à risque".

La quantification est simple, mais le codage (codage entropique) est assez sophistiqué, et nécessite une procédure fastidieuse d'essais-erreurs pour décider de l'attribution finale des bits.

La complexité du codeur est donc assez importante. Par contre la complexité du décodeur est très faible, et donc l'OCF se prête bien à une utilisation en situation de type "broadcasting".

3.3 Travaux effectués aux Bell Laboratories

Les travaux effectués par Johnston [36, [37], [38] aux Bell Laboratories sont sur 2 plans: il s'agit tout d'abord de travaux fondamentaux sur le masquage du bruit par le signal, ainsi que le développement du concept d'entropie perceptuelle. Ensuite, ces travaux débouchent très naturellement sur des applications en codage perceptuel.

3.3.1 Entropie perceptuelle

L'entropie perceptuelle est définie dans [37] en relation avec le niveau maximum de bruit que l'on peut injecter dans le signal, ce bruit restant inaudible. Cette entropie perceptuelle est donc calculée pour un bruit injecté de manière à être tout juste masqué par le signal. Il serait possible de déduire du rapport signal/bruit obtenu dans ces conditions, une entropie différentielle qui représenterait l'entropie perceptuelle du signal de manière très générale. Toutefois Johnston définit formellement l'entropie perceptuelle d'une façon un peu plus restrictive, mais plus pratique, et qui mènera tout naturellement à son application au codage.

Tout d'abord le signal est analysé spectralement par une transformée de Fourier sur 2048 points avec fenêtre de Hanning.

L'énergie du signal apparaissant dans chacune des bandes critiques est ensuite calculée.

A partir de cette répartition d'énergie en bandes critiques, et en faisant intervenir les masquages entre bandes critiques, les seuils de masquage sont calculés pour chacune des bandes. On obtient donc pour chaque bande critique, le niveau maximum de bruit qui peut être injecté de manière à rester inaudible.

Jusqu'à présent (mis à part l'utilisation des masquages temporels pour le NMR), la procédure suit le même principe que le NMR.

A partir de ce niveau de bruit maximum pour chaque bande critique, il est facile d'évaluer la résolution (donc le nombre de bits) d'un quantificateur scalaire uniforme adaptatif (et parfaitement adapté) qui devrait quantifier chacune des composantes spectrales à la limite tolérable de bruit. Le nombre moyen de bits obtenu par composante spectrale donne l'entropie perceptuelle du signal.

On voit donc que cette définition est assez restrictive mais suggère quasiment une technique de codage.

Il est à remarquer que:

- cette étude ne fait pas intervenir de masquages temporels;
- l'obtention des seuils de masquage en fonction des bandes critiques ne se fait pas à partir de gabarits de masquage tabulés et de règles, mais plutôt à partir d'un modèle d'audition, similaire à plusieurs points de vue au modèle que nous avons développé au CRCS.

Les différences principales sont:

- l'analyse se fait en 24 bandes critiques alors qu'elle est effectuée en espace continu au CRCS;
- 2 seuils de masquage différents sont utilisés, pour les zones spectrales où le signal est harmonique, et celles où il est semblable à du bruit, alors que nous n'utilisons qu'un seuil de masquage.

Note: *Bien qu'il soit un peu tôt pour se prononcer, des essais informels semblent montrer que le modèle développé au CRCS tient compte "de lui-même" d'une différence de sensibilité entre des zones harmoniques et non harmoniques du spectre.*

- que le modèle de Johnston ne tient apparemment pas compte de la fonction d'atténuation de l'oreille moyenne alors que nous en tenons compte au CRCS;
- que la décomposition du signal effectuée n'est pas parfaitement unitaire (à cause de la fenêtre de Hanning et du recouvrement temporel) et donc, en toute exactitude, la somme des énergies mesurées dans l'espace fréquentiel n'est pas égale à l'énergie du signal. De même, la somme des énergies de bruit mesurées dans l'espace fréquentiel n'est pas égale à l'énergie de bruit du signal. En pratique toutefois, les différences doivent se compenser en moyenne, et donner quasiment le même résultat (à un coefficient multiplicatif près peut être) que si la décomposition était unitaire.

En conclusion: l'entropie perceptuelle est un concept novateur, et qui se révèle d'une importance essentielle en ce qui concerne l'utilisation des propriétés de perception pour le codage des signaux de haute qualité. Notons toutefois que la définition formelle restrictive donnée par Johnston de l'entropie perceptuelle devrait évoluer avec le temps au fur et à mesure que notre compréhension du fonctionnement auditif s'affinera. De plus, même s'il est pratique de l'associer à une méthode de codage pour les besoins de la

démonstration (et afin de la mesurer!), elle devrait peu à peu se placer à un niveau plus fondamental, et par là même, se détacher d'un principe de codage particulier.

3.3.2 Le PXFM: (Perceptual Transform Coder)

Les résultats des travaux sur l'entropie perceptuelle ont été utilisés simplement et directement pour la mise en oeuvre d'un codeur perceptuel: le PXFM. Le PXFM est simple et sans raffinements. Étant donné son manque de sophistication, ce codeur a sans doute été développé en premier lieu pour valider le concept d'entropie perceptuelle, et estimer les gains en débit et en qualité qu'il permettrait.

Le signal, sur une bande passante de 15 kHz est tout d'abord décomposé spectralement par une transformée de Fourier sur 2048 points, avec fenêtre d'analyse et recouvrement de 1/16 (soit 128 échantillons) entre 2 analyses successives. La fenêtre d'analyse est la racine carrée d'une fenêtre d'analyse de Hanning de longueur double de la longueur d'analyse (soit 4096 points). Cette fenêtre d'analyse particulière et le recouvrement de 1/16 doivent permettre une reconstruction parfaite en l'absence de quantification.

Le spectre est donc représenté sur 1024 points complexes, couvrant la gamme $0 \rightarrow 15$ kHz. Comme dans le cas de la mesure d'entropie perceptuelle, l'énergie du signal est mesurée dans chacune des 25 bandes critiques qui couvrent le spectre. Les seuils de masquage pour chaque bande critique sont ensuite déduits de cette répartition d'énergie. Notons qu'ici aussi la distinction est faite, pour le calcul du seuil de masquage, entre les bandes critiques à contenu spectral harmonique et celles à contenu spectral inharmonique. La distinction est faite à partir de la mesure d'uniformité spectrale qui est d'autant plus petite que le spectre est plus découpé. Si la bande critique est à contenu harmonique, le seuil de masquage du bruit sera environ 9 db plus bas que si elle ne l'est pas (les bandes à contenu spectral harmonique sont 9 db plus sensibles que les autres).

Le niveau spectral maximum du signal est ensuite mesuré pour un ensemble de 128 bandes spectrales couvrant la gamme de 0 à 15 kHz (donc 128 groupes de 8 composantes spectrales). Ces niveaux maximum vont servir (avec les seuils de masquage) à effectuer l'attribution des bits pour les composantes spectrales, ainsi qu'à adapter la dynamique des quantificateurs à celle des valeurs à quantifier.

Quatre types d'information sont transmis au récepteur:

- les valeurs quantifiées des 1024 composantes spectrales du signal;

- les valeurs quantifiées des 25 seuils de masquage;
- les valeurs quantifiées des 128 niveaux maximums.

A partir des valeurs des 25 seuils de masquage et des 128 niveaux maximums, le récepteur reconstitue l'attribution des bits de la même façon que l'émetteur l'a faite, et reconstitue (déquantifie) le spectre du signal.

Le signal temporel est ensuite reconstruit par transformation de Fourier inverse et recouvrement-addition .

Pour ce codeur, le débit total est 128 Kbits/s et la qualité est "transparente".

A noter, très récemment ([38]) Johnston a proposé un algorithme pour le codage d'un signal stéréophonique. Cet algorithme fonctionne suivant le même principe de codage perceptuel que le PXFM. Toutefois, en tirant parti de la redondance entre les 2 canaux, le débit peut être réduit d'un tiers par rapport au codage séparé de chacun des canaux.

L'algorithme est un peu plus sophistiqué que le PXFM, on peut noter l'utilisation de quantification vectorielle en dimension 2 pour les composantes spectrales de faible amplitude (partie réelle, partie imaginaire), ainsi que l'utilisation de codage entropique (codage de Huffman).

3.4 Les travaux effectués au CNET

Le codeur développé au CNET par Mahieux, Petit et Charbonnier a beaucoup en commun avec le PXFM de Johnston. Toutefois il est un peu plus sophistiqué et fait intervenir une prédiction linéaire des composantes spectrales.

Tout comme le PXFM, le codeur du CNET analyse le signal à l'aide d'une transformée de Fourier sur 512 points avec recouvrement de 1/16 des échantillons entre 2 analyses successives. Ici la fenêtre d'analyse est plate, à bords adoucis par une fonction sinusoïdale.

Tout ceci autorise une reconstruction parfaite. Comme dans le cas du PXFM, le recouvrement de 1/16 impose dès le départ un handicap au codeur avec 1/16 d'échantillons de plus à coder dans le domaine spectral qu'il y en a dans le domaine temporel.

A ce stade, l'algorithme du CNET procède à une prédiction linéaire des composantes spectrales. Cette prédiction linéaire est effectuée de manière très intéressante en coordonnées polaires. L'amplitude est simplement estimée par rapport à l'amplitude précédente, et la phase est estimée par rapport à sa vitesse de rotation moyenne, en remarquant que pour un signal harmonique stable cette phase tourne à vitesse quasi-constante, donc varie d'un angle constant entre 2 analyses successives. Cette prédiction linéaire permet de tirer parti de la redondance du signal, mais surtout, permet certainement un gain de qualité très important lors de la quantification des signaux harmoniques stables.

L'attribution des bits suivant les composantes fréquentielles est effectuée de manière très similaire au PXFM, à la différence près peut être que la distinction n'est pas faite, pour l'ajustement du seuil de masquage, entre les zones harmoniques et les zones de type "bruit" du spectre. Cela n'est pas très clair car bien que Mahieux n'en parle pas dans la procédure d'attribution des bits ([40]), il indique plus loin que le gain de prédiction (qui dépend de la nature harmonique du signal) intervient dans la procédure d'attribution des bits. Autre originalité de l'algorithme, le récepteur n'a besoin que des valeurs quantifiées des composantes spectrales. Toute l'information de contrôle (attribution des bits, répartition moyenne de l'énergie spectrale, etc...) est estimée localement au récepteur à partir du passé proche du codeur; à l'exception cependant des maximas des spectres sur 32 bandes fréquentielles uniformément réparties, qui sont transmises explicitement. Cette transmission explicite sert à réinitialiser les estimateurs locaux de l'information de contrôle en cas de transition abrupte du signal, améliorant ainsi le comportement dynamique du codeur.

Le débit de ce codeur est 96 Kbits/s. La qualité est excellente, sans toutefois être qualifiée de "transparente" par Mahieux.

Le récepteur est tout de même assez complexe comparé à un récepteur de type OCF, puisque toute l'information de contrôle (telle que l'attribution des bits) doit être calculée localement au récepteur.

3.5 Comparer les débits

Lorsque les auteurs donnent les débits des algorithmes qu'ils présentent, ils les donnent soit en bits/échantillon, soit en Kbits/seconde. Entre les deux, bien entendu, il y a une conversion directe faisant intervenir la fréquence d'échantillonnage.

Si on compare le débit donné par Brandenburg pour l'OCF (2.5 bit/échantillons) avec le débit donné par Johnston pour le PXFM (4 bit/échantillons), on note une nette différence entre les deux, alors que la qualité semble du même ordre. On serait tenté de conclure que le PXFM est bien moins efficace que l'OCF.

ATTENTION ...!

Pour les algorithmes de codage perceptuel, si les critères de perception sont bien utilisés, le codage de la partie haute du spectre (15 kHz → 20 kHz) ne devrait demander qu'un débit très faible comparé à la partie basse (0 → 15 kHz). En conséquence, le codage de cette partie du spectre étant quasiment gratuite au point de vue du débit, il est bien plus juste de comparer les débits en nombre total de bits/seconde (comme si les différents codeurs travaillaient à la même fréquence d'échantillonnage).

Si l'on fait la comparaison à ce niveau, l'OCF a un débit de 110 Kbits/seconde alors que le débit du PXFM est 128 Kbits/seconde. La différence est bien moins prononcée.

4. Travaux effectués au CRCS et orientations

Au CRCS, les travaux se déroulent sur 3 fronts:

- psychoacoustique;
- évaluation objective de codeur;
- codage.

4.1 Psychoacoustique

Ce premier volet consiste à étudier les phénomènes de perception acoustique (et en particulier les masquages fréquentiels), afin d'en avoir une meilleure compréhension. Plus précisément le CRCS a travaillé au développement d'un modèle d'audition: "OREILLE" dont les grandes lignes sont décrites dans l'annexe 1. La confrontation des résultats prédits par OREILLE avec les résultats d'expériences réelles décrites dans la littérature, nous permet de valider le modèle, mais aussi d'ajuster ses paramètres, et le cas échéant de cerner ses lacunes.

Ce modèle d'audition est, ou sera, au coeur des logiciels développés dans le cadre de l'évaluation objective de codeur, et celui du codage. Il est donc particulièrement important de mettre au point ce modèle de manière précise et méthodique, et de bien comprendre son comportement, de même que le comportement d'une oreille réelle. Par exemple, une question à laquelle il faudra répondre est: "*faut-il tenir compte*

explicitement du fait que les zones harmoniques du spectre sont plus sensibles au bruit que les zone inharmoniques? ou bien le modèle en tient-il compte de lui-même?" Notons que si le modèle prédit de lui-même une sensibilité plus grande des zones harmoniques du spectre, cela nous fournit en même temps une explication du phénomène.

4.2 Évaluation objective de codeur

Ces travaux sur la modélisation de l'audition servent à la mise en place d'un logiciel d'évaluation perceptuelle de codeur: "Perceval". En fait il y a deux versions de Perceval:

4.2.1 PERCEVAL-1

La première version de Perceval donne des résultats suivant la même méthodologie que le NMR de Brandenburg.

- le signal original et le signal codé sont décomposés spectralement par une MLT, de manière à avoir des composantes fréquentielles de largeur 15 Hz environ (1024 composantes pour un signal échantillonné à 32 kHz);
- la densité spectrale d'énergie de l'original est convertie (à l'aide de "OREILLE') en densité basilaire d'énergie;
- en reportant cette densité basilaire d'énergie dans l'espace fréquentiel, et en compensant pour l'atténuation fréquentielle de l'oreille moyenne et du conduit auditif, on obtient ce que nous appelons l'inverse de l'importance des composantes spectrales, mais qui est équivalent au seuil de masquage défini par Brandenburg pour le NMR. L'utilisation d'un modèle d'audition explicite permet d'obtenir un seuil de masquage continu en fréquence, qui ne fait pas intervenir de bandes critiques, mais surtout qui est calculé globalement pour la densité spectrale d'énergie originale, et non pas pour chaque composante prise indépendamment et dans une situation de "pire cas";
- parallèlement, l'énergie de bruit est mesurée pour chaque composante fréquentielle, et est comparée au seuil de masquage obtenu. Pour chaque bloc analysé, le maximum sur les composantes fréquentielles du rapport bruit/seuil de masquage est déterminé, et représente la valeur de la dégradation pour ce bloc.

On obtient plusieurs types d'informations:

- pour chaque bloc analysé on peut visualiser:
  - le spectre d'énergie du signal;
  - le seuil de masquage du bruit dû à ce spectre d'énergie;
  - le spectre d'énergie du bruit (figures 2 et 3);
- pour l'ensemble du fichier, on peut:
  - visualiser l'évolution de la dégradation en fonction du temps (figure 4);
  - obtenir la valeur moyenne de la dégradation sur l'ensemble du fichier.

4.2.2 PERCEVAL-2

Cette deuxième méthode a un principe de fonctionnement beaucoup plus proche d'un modèle d'audition.

- le signal original et le codé sont tous deux décomposés spectralement par une MLT, de la même façon que pour la 1ière méthode;
- la densité spectrale d'énergie du signal original est calculée et transformée par 'OREILLE' en une densité basilaire d'énergie originale;
- la densité spectrale d'énergie du bruit est calculée et ajoutée à la densité spectrale d'énergie du signal original. Cette densité spectrale d'énergie de "signal + bruit" est à son tour transformée par 'OREILLE' en densité basilaire d'énergie bruitée;
- pour tous les points de la membrane basilaire où la différence entre les deux densités basilaires dépasse un certain seuil, on ajoute cette différence à la valeur de la dégradation pour le bloc. Les différences sont calculées en échelles de sensation (échelles logarithmiques en tenant compte d'une légère énergie de bruit des détecteurs de la membrane basilaire);
- cette dégradation peut être affichée en fonction du temps (figure 5) ou être moyennée sur l'ensemble du fichier.

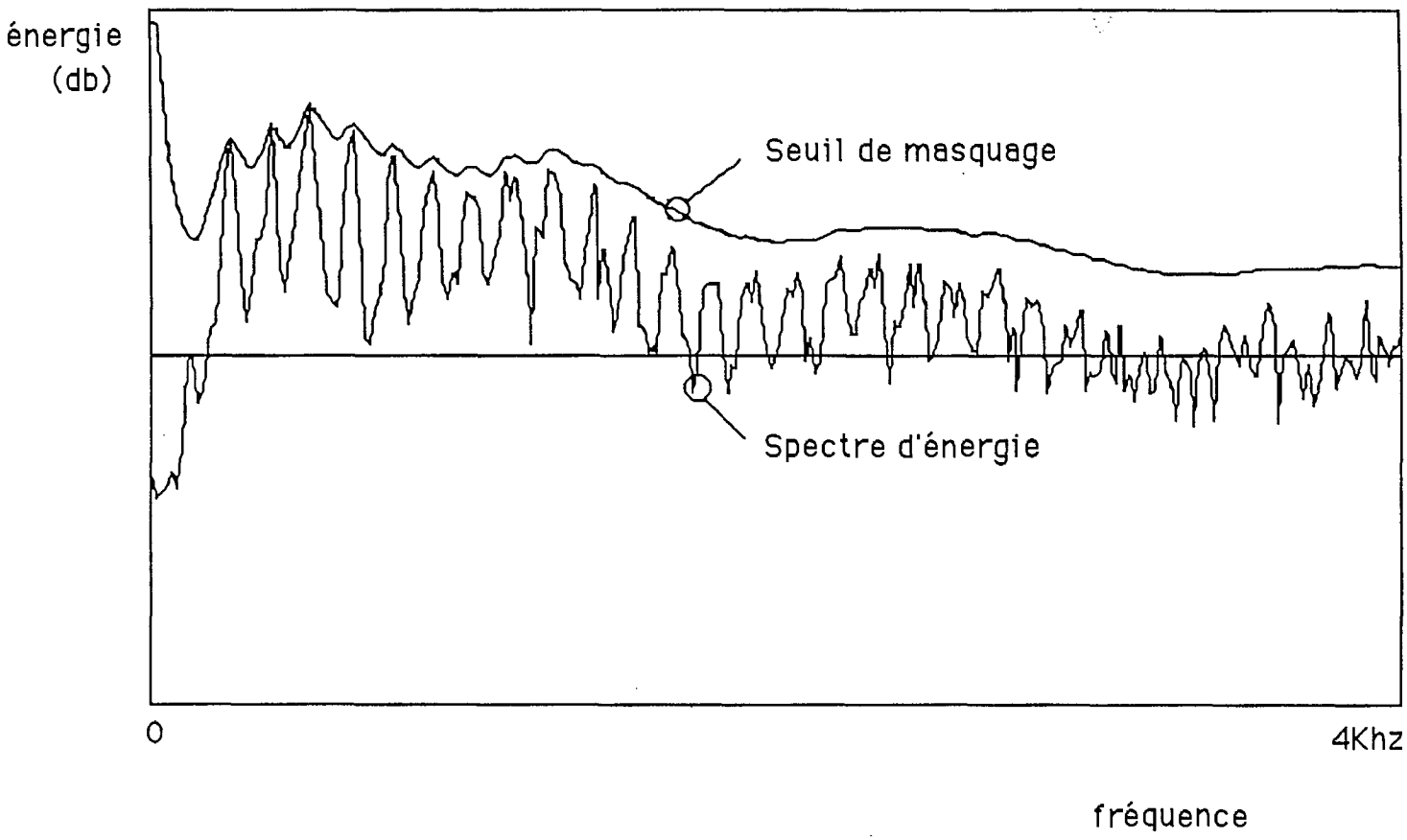


Figure 2 Spectre d'énergie du signal et seuil de masquage correspondant (calculé sur environ 100 ms de signal, soit 3 fenêtres)

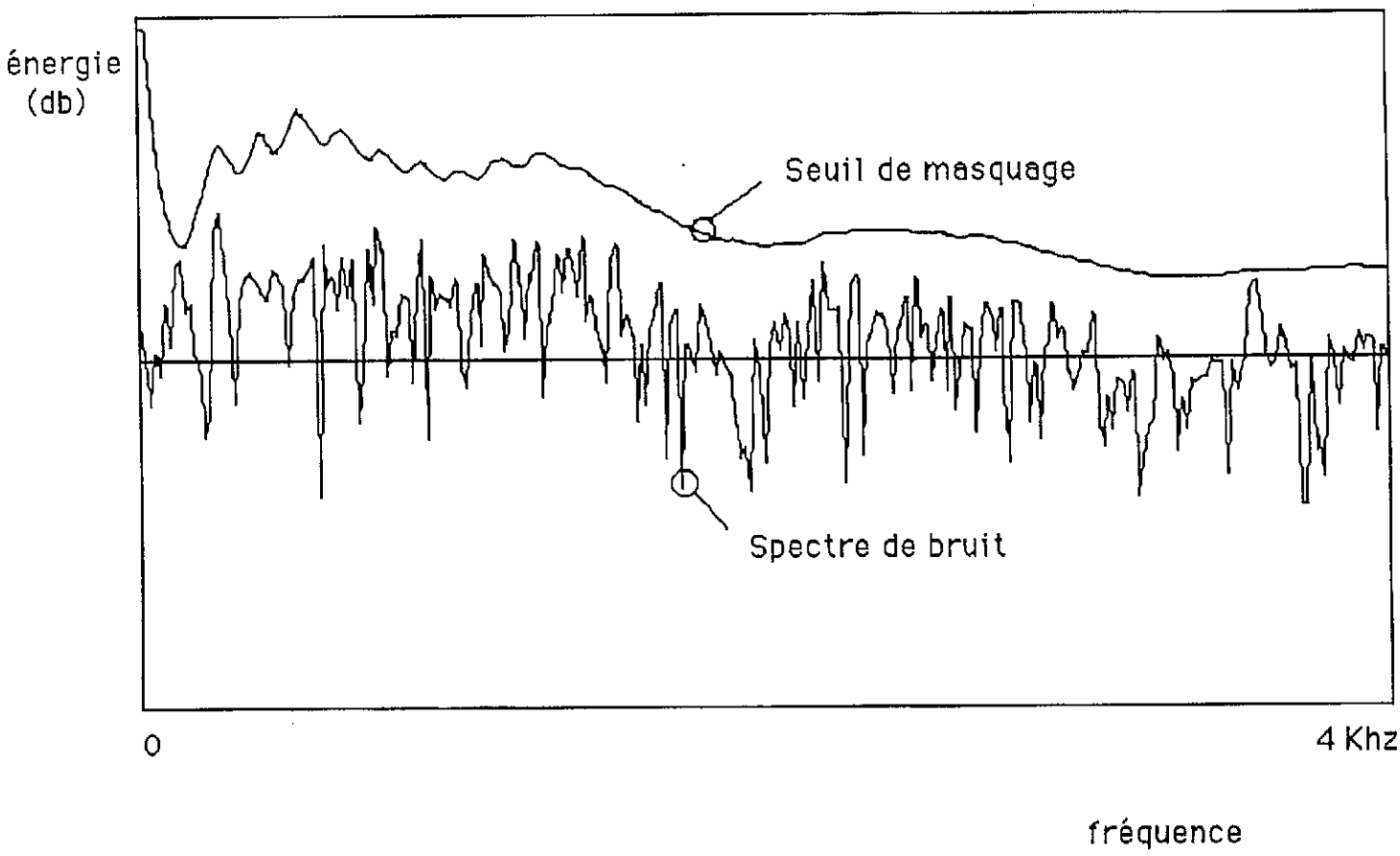


Figure 3 Seuil de masquage et niveau du bruit

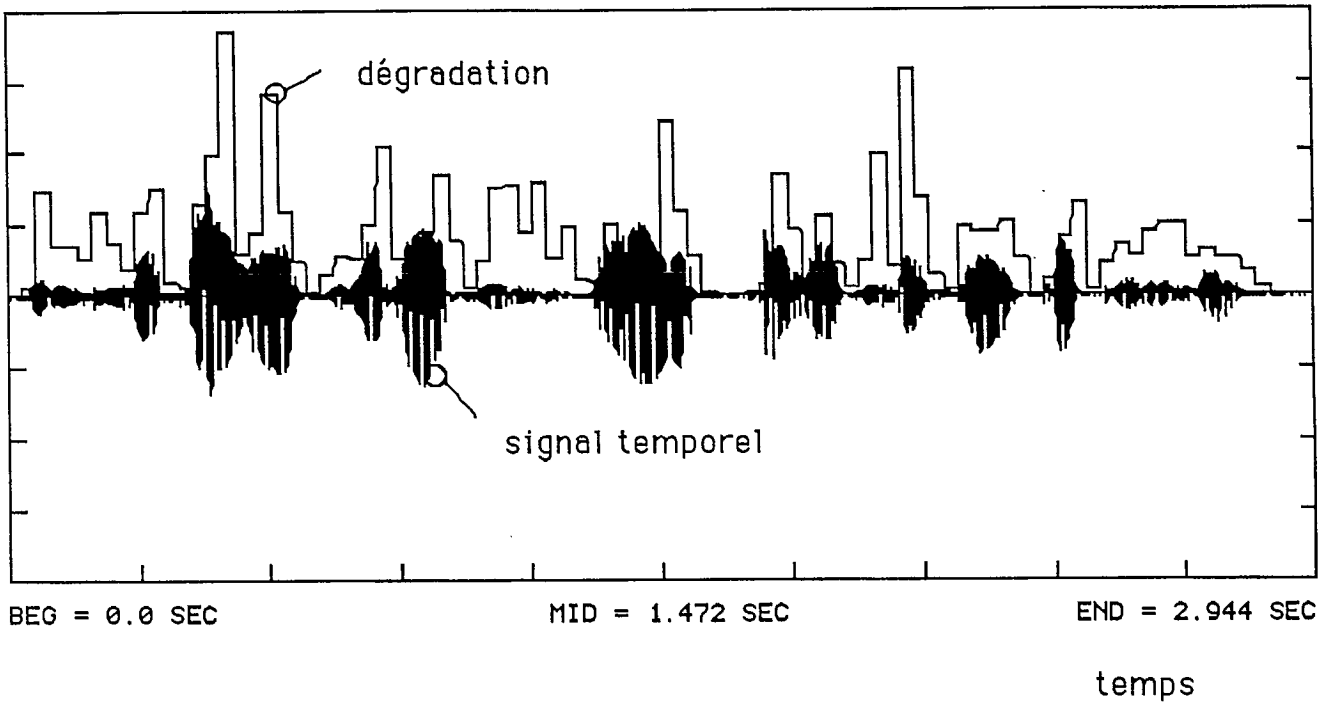


Figure 4 évolution de la dégradation perçue en fonction du temps (PERCEVAL_1)

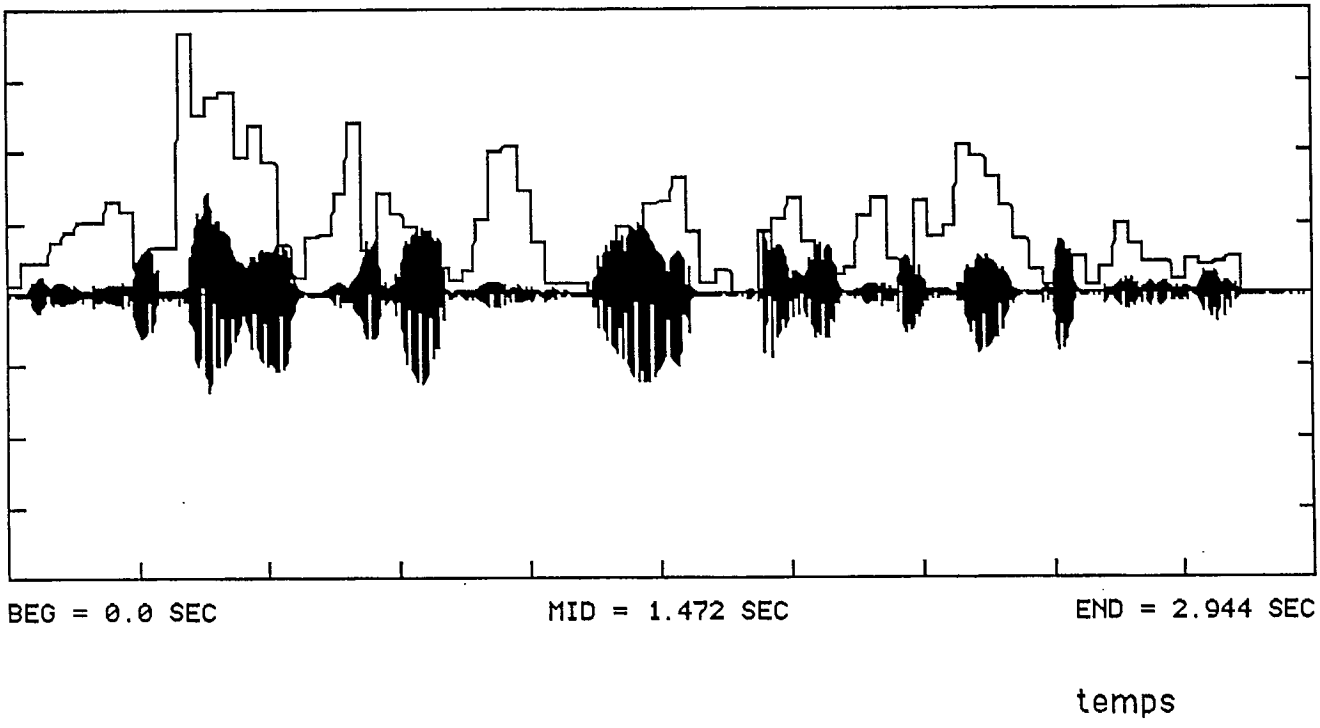


Figure 5 évolution de la dégradation perçue en fonction du temps (PERCEVAL_2)

4.3 Codage

A moyen terme, ces travaux vont déboucher sur le développement d'un codeur de haute qualité.

A la suite des résultats obtenus dans [42], certaines orientations ont été prises quant à la réalisation de ce codeur.

Résolution spectrale

Suite aux travaux décrits dans [42], il est apparu qu'une décomposition du spectre en 16 sous-bandes est trop grossière pour faire apparaître la structure harmonique du signal, et tirer parti des phénomènes de masquages fréquentiels de manière optimale. Le problème est critique pour la 1ière bande qui inclut à la fois des zones très sensibles au bruit (avant la 1ière harmonique), et des zones relativement bien masquées (après cette 1ière harmonique). Ce problème a d'ailleurs été noté aussi par Petit et Charbonnier du CNET [39]. Le découpage extrêmement large de l'espace fréquentiel (16 bandes de 500 Hz pour couvrir le spectre de 0 à 8 kHz) ne permet pas de distinguer les variations importantes de sensibilité au bruit à l'intérieur des sous-bandes (en particulier pour la 1ière) et ne permet pas non plus d'attribuer les bits de quantification plus finement que "en bloc" pour chaque sous-bande.

Pour cette raison, la décomposition spectrale se fait à présent en un nombre beaucoup plus important de composantes (1024 composantes de largeur environ 15 Hz pour un signal échantillonné à 32 kHz). Cette décomposition se fait à l'aide de l'algorithme des MLT qui combine les avantages d'une bonne résolution spectrale, une reconstruction exacte et un algorithme de calcul rapide.

Quantification

De même, il est apparu dans [42] que la quantification vectorielle, bien qu'étant plus efficace que la quantification scalaire, était moins bien adaptée aux besoins du codage des signaux de haute qualité. En raison de la taille, nécessairement limitée, des dictionnaires de quantification, la quantification vectorielle ne permet pas, si nécessaire, d'attribuer une précision (un nombre moyen de bits par composante) importante à une composante spectrale.

La complexité de la quantification vectorielle est aussi bien plus grande que celle de la quantification scalaire, et donc elle se prête moins bien à une utilisation aux grandes

fréquences d'échantillonnages caractéristiques des signaux de haute qualité. Pour ces deux raisons (précision et vitesse), la quantification doit se faire assez simplement, et en particulier être scalaire et uniforme. Par contre, il est indispensable que les quantificateurs soient adaptés dynamiquement au niveau des signaux à quantifier, ce qui demande qu'au moins un spectre d'énergie approximatif soit transmis au récepteur.

Modèle d'audition

Il semble clair que la mise en oeuvre des critères de décision perceptuels se fera à partir d'un modèle d'audition explicite. Ce choix permet, entre autres avantages, une utilisation plus cohérente des critères de perception ainsi que le calcul d'un seuil de masquage continu en fréquence, et donc une plus grande finesse pour l'attribution dynamique de la résolution de quantification.

Masquages temporels

Les masquages temporels sont sans doute dûs à un phénomène d'intégration temporelle (temps de réponse assez long des détecteurs de la membrane basilaire) de l'énergie par les détecteurs de la membrane basilaire. De même que les masquages fréquentiels sont dûs à une intégration spatiale de l'énergie (dans ce cas on fait plutôt référence à une "dispersion" de l'énergie sur la membrane basilaire). Il est expliqué dans [42], chapitre 2, comment ces phénomènes d'intégration associés à une loi de sensibilité des détecteurs de type "compression" (voire une loi logarithmique), expliquaient qualitativement (et quantativement pour les masquages fréquentiels) les phénomènes de masquage.

Le processus d'intégration temporelle n'est pas ignoré quant à la mise en oeuvre du modèle d'audition. Bien au contraire, c'est une des hypothèses de base qui nous permettent d'utiliser un modèle purement fréquentiel tel que décrit dans [42]. En effet, on suppose que le temps de réponse des détecteurs basilaires est suffisamment long pour que la localisation temporelle précise de l'énergie dans l'espace temps-membrane basilaire soit peu importante. A cette condition, il est possible de substituer à la décomposition spatio-temporelle exacte qu'effectue la membrane basilaire sur le signal (chaque point de la membrane basilaire agit sur le signal comme un filtre, à la sortie de ce filtre la cellule détectrice correspondante mesure l'énergie du signal), une décomposition approximée constituée d'une décomposition temps → fréquence, (décomposition en sous-bandes, DFT, DCT, MLT, etc...) suivie d'une mise au carré des composantes de sortie de l'analyse (extraction de l'énergie) suivie ensuite, pour chacune

des répartitions fréquentielles d'énergie de l'espace temps → fréquence, d'une transformation linéaire énergie fréquentielle → énergie basilaire. On obtient bien donc une répartition approximée de l'énergie dans l'espace temps → membrane basilaire. Comme il est souligné dans [42], l'approximation se situe au niveau de la localisation temporelle de l'énergie dans l'espace temps → membrane basilaire. Autrement dit, au fur et à mesure que la résolution de l'analyse spectrale utilisée augmente, la localisation basilaire de l'énergie dans l'espace temps → membrane basilaire devient de plus en plus exacte comparée à la répartition réelle, tandis que la localisation temporelle devient de plus en plus incertaine. On s'appuie donc sur l'hypothèse d'une intégration temporelle importante (un temps de réponse assez grand) des détecteurs de la membrane basilaire pour justifier l'approximation. En toute rigueur, un modèle fréquentiel n'est utilisable que si le temps de réponse des détecteurs est infiniment long.

On voit donc que ces phénomènes d'intégration temporelles qui sont sans doute à l'origine des masquages temporels, sont en fait une des hypothèses fondamentales qui permettent l'utilisation du modèle fréquentiel. On voit aussi que si la localisation temporelle approximée de l'énergie dans l'espace temps → membrane basilaire est incertaine, comparée à la localisation temporelle réelle, il est difficile de justifier l'utilisation des maquages temporels dans l'implémentation des critères de distance. C'est la raison pour laquelle, au CRCS, nous avons décidé de ne pas faire intervenir explicitement les masquages temporels dans la modélisation de la perception. Il faut toutefois noter que, bien qu'incertaine, et bien que dépendant de l'analyse fréquentielle utilisée (et en particulier du nombre de composantes fréquentielles), il est possible que la localisation temporelle approximée de l'énergie dans l'espace temps → membrane basilaire soit suffisamment proche de la localisation temporelle réelle (par rapport au temps de réponse des cellules détectrices) pour rendre l'utilisation des masquages temporels avantageux. A ce niveau, seule l'expérience peut justifier une telle utilisation.

5. Développement récents

5.1 Etat de l'art

Dans les sections précédentes, nous avons fait le point sur l'état de l'art dans le domaine des décompositions temps-fréquences, et dans celui du codage perceptuel. Certaines constantes sont apparues dans les solutions adoptées par les différents groupes actifs en codage perceptuel.

Par exemple, le signal à coder subit toujours une décomposition temps-fréquence, préalable à tous les autres traitements (analyse de l'énergie dans l'espace temps-fréquence, mise en oeuvre d'un modèle d'audition, calcul de la précision de codage en fonction des composantes fréquentielles ...etc.). Il est généralement admis que cette décomposition temps-fréquence, que ce soit une décomposition en sous-bandes, une transformation segmentaire, ou une autre forme de décomposition, est le moyen le plus direct et le plus efficace pour travailler dans un espace d'analyse proche de celui de l'oreille.

Un autre exemple est la forme spectrale que l'on cherche à donner au bruit de quantification, et la technique générale qui est utilisée pour y arriver. Cette technique peut se résumer comme suit :

- A partir du signal dans l'espace temps-fréquence, on évalue, pour une durée d'analyse donnée, la répartition fréquentielle de l'énergie du signal. A partir de cette répartition fréquentielle d'énergie, on détermine un *seuil de masquage fréquentiel.* Ce seuil de masquage peut être calculé par diverses techniques, plus ou moins sophistiquées, et plus ou moins approximatives. Il peut différer un peu, d'une technique à l'autre, dépendant des approximations faites et du nombre de composantes de l'espace fréquentiel, mais pour tous les groupes de recherche travaillant dans ce domaine, ce seuil de masquage représente bien le même concept. Il rend compte des *masquages fréquentiels* ou *masquages simultanés*, et en particulier des *pentes de masquages* qui apparaissent de part et d'autre des composantes harmoniques du signal, et qui sont plus douces vers les hautes fréquences (masquage plus efficace) que vers les basses fréquences.

- La stratégie de codage consiste ensuite à "moduler" la précision de la quantification en fonction des composantes fréquentielles de telle manière que le spectre d'énergie du bruit de quantification suive le *seuil de masquage fréquentiel* calculé précédement. Généralement, on obtient ce résultat en attribuant à chaque composante fréquentielle, un nombre de bits de quantification différent, choisi de telle manière que:

  - le nombre de bits total soit constant,

  - le bruit de quantification suive bien le seuil de masquage.

Ce principe général de codage semble donc bien établi. Il faut noter toutefois que le seuil de masquage fréquentiel ainsi calculé, n'est valable en toute rigueur que pour un bruit à bande étroite (voire une composante harmonique) que l'on essaie de masquer dans le signal original. Autrement dit, ce *seuil de masquage fréquentiel* indique que, si une composante harmonique est superposée au signal original, et si son niveau est supérieur au seuil de masquage pour la fréquence considérée, cette composante a de bonnes chances d'être détectée; si, au contraire, son niveau est inférieur au seuil de masquage, elle ne sera pas détectée.

Dans le cas du codage cependant, le bruit de quantification ne se présente pas sous forme d'un signal harmonique, ni même d'un bruit à bande étroite isolé, mais plutôt sous la forme d'un bruit à bande large dont on peut modifier le spectre d'énergie. Le principe général de codage que nous venons de décrire repose donc sur l'hypothèse suivante:

Si on décompose le bruit de quantification (à bande large) en composantes constituant des bandes fréquentielles étroites, et si chacune de ces composantes de bruit adjacentes a un niveau inférieur au seuil de masquage pour la fréquence considérée (chacune de ces composantes de bruit présentée isolément ne serait pas détectée), le bruit total (constitué de la superposition de toutes ces composantes de bruit à bande étroite) ne sera pas détecté. Autrement dit, le bruit ne sera pas détecté si son spectre d'énergie est sous le seuil de masquage pour toutes les fréquences.

Or cette hypothèse est infirmée par l'expérience, ainsi que par notre modèle d'audition. Un bruit peut parfaitement avoir un spectre d'énergie en tous points sous le seuil de masquage (seuil de masquage calculé pour un son masqué à bande étroite isolé) et être globalement détectable. L'exemple le plus évident est le cas de la détection d'un bruit à bande large dans le silence (masqueur nul). En effet dans la référence [4] de l'annexe 2, Hellman indique qu'un bruit à large bande (75-9600 hz) est détectable dès que son niveau d'énergie total atteint 15 db. Cela correspond à un niveau de densité spectrale constant de -25 db sur la bande spectrale de 75 à 9600 hz, et donc, est sans aucun doute considérablement sous le seuil d'audition absolu de l'auditeur (seuil de masquage pour un masqueur nul) pour toute la bande spectrale considérée.

Remettre en question cette hypothèse de base, revient bien évidement à remettre en question la stratégie de codage qui consiste à faire en sorte que le spectre d'énergie du bruit suive le seuil de masquage.

Bien entendu, remettre en cause l'optimalité de cette stratégie ne nous amène pas de meilleure solution immédiate. On peut toutefois noter que, moyennant certaines approximations, la solution optimale pourrait être de faire en sorte que le spectre d'énergie du bruit suive le spectre d'énergie du signal. Cette solution serait extrêmement simple à mettre en oeuvre.

Il est clair qu'un travail important reste à faire, de manière à valider ou à invalider la stratégie de codage communément utilisée, et éventuellement de manière à trouver d'autres stratégies plus avantageuses.

5.2 Améliorations récentes du modèle d'audition "OREILLE"

Une partie du travail effectué récemment concerne des améliorations apportées au modèle d'audition "OREILLE". Ces améliorations ont consisté principalement à mettre en oeuvre un principe de détection statistique plus réaliste, pour remplacer le principe de détection déterministe que l'on utilisait jusqu'à présent.

La description de ce nouveau principe, ainsi que les résultats qu'il a permis d'obtenir sont détaillés dans l'annexe 2 du présent document. On peut résumer ces résultats comme suit:

- Dans le cas de sons harmoniques masqués par du bruit à bande étroite ou large, les seuils de détection prévus par le modèle sont en très bonne accordance (erreur de l'ordre de 1 db) avec les résultats d'expériences réelles. Il est à noter qu'avec l'ancien principe de détection, les résultats du modèle étaient déjà excellents, aucune amélioration réelle n'est donc notée dans ce cas.

- Dans le cas de bruits à bande étroite masqués par des masqueurs harmoniques, les seuils de détection prévus par le modèle sont en très bonne accordance (différence de l'ordre de 1 db) avec les résultats d'expériences réelles. Dans ce cas, par contre, cela correspond à une nette amélioration par rapport au modèle précédent. Le modèle précédent faisait des erreurs de l'ordre de 10 db sur les seuils de détection calculés, par rapport aux résultats d'expériences réelles.

Le modèle d'audition, prévoit donc avec une très bonne précision l'asymétrie (seuils de détection différant d'environ 20 db) entre la détection d'un son harmonique masqué par un bruit, et celle d'un bruit masqué par un son harmonique.

Dans la section 4.1 (page 22), nous posions la question suivante:

faut il tenir compte explicitement du fait que les zones spectrales harmoniques du signal sont plus sensibles au bruit que les zones inharmoniques (tel que le fait Johnston aux laboratoires Bell (annexe 2 - [6])), ou bien notre modèle en tient-il compte "de lui même".

La réponse est donc que notre modèle tient compte "de lui même" de cette plus grande sensibilité au bruit des zones spectrales harmoniques du signal.

De manière annexe, nous nous sommes rendus compte que lorsqu'un bruit à bande étroite est en position de signal masqué (par un masqueur harmonique), son allure spectrale a une grande influence sur le seuil de masquage, et en particulier la pente des flancs de décroissance, de part et d'autre des fréquences de coupures. Il est apparu par exemple que des bruits tiers d'octave ayant des pentes de décroissance aussi importantes que ... 100 db/octave! ne donnaient pas les mêmes résultats pour la simulation que des bruits tiers d'octave parfaits (pentes de décroissance infiniment abruptes).

Ce travail effectué sur le modèle d'audition a eu des conséquences pour les logiciels d'évaluation objective de la qualité de codage "PERCEVAL".

Dans la première partie du rapport (rapport intérimaire) nous décrivions 2 logiciels d'évaluation perceptuelle objective : PERCEVAL_1 et PERCEVAL_2.

Il apparaissait que PERCEVAL_1 fonctionnait de manière relativement classique (similaire dans le principe au NMR de Brandenburg), si ce n'est que le *seuil de masquage* était calculé en *espace continu* grâce à notre modèle d'audition. Par contre PERCEVAL_2 fonctionnait de manière beaucoup plus proche d'un modèle d'audition, et donnait comme résultat, l'intégrale de la différence de sensations basilaires entre le signal original et le signal bruité, en fonction du temps. Cette intégrale représentant la dégradation telle qu'elle est perçue par l'oreille.

Le nouveau principe de détection adapté au modèle d'audition a permis le développement d'une 3eme version de PERCEVAL. PERCEVAL_3 fonctionne de manière très similaire à PERCEVAL_2, si ce n'est qu'à partir de la différence de sensations basilaires entre le signal original et le signal bruité, on détermine une *probabilité de détection du bruit* en fonction du temps.

Cette 3ème version de PERCEVAL devrait être la plus représentative des 3. Il est facile en effet d'évaluer la performance du modèle d'audition de manière **quantitative et objective**, en termes de probabillités de détection du bruit, par comparaison avec des situations types de bruit et de signal, correspondant à des expériences réelles de psycho-acoustiques documentées dans la littérature. Par contre l'évaluation des résultats du modèle en termes d'importance de la dégradation perçue (résultat que donne PERCEVAL_2) ne peut se faire qu'en les confrontant à des résultats d'expériences réelles de comparaisons **subjectives**, qui sont beaucoup plus difficiles à quantifier.

Il est donc plus facile de valider les résultats de PERCVAL_3 (par validation du modèle d'audition utilisé) que de valider les résultats, plus subjectifs dans le principe, que donne PERCEVAL_2. Par contre, PERCEVAL_3 ne sera utilisable que sur des signaux codés de très bonne qualité. En effet, si le codage est de qualité moyenne, la probabilité de détection du bruit sera en tout temps (sauf durant les silences) très proche de 1. Ces résultats seront donc très difficiles à interpréter.

6. Orientations et travail à venir

6.1 Evaluation perceptuelle objective

Dans un premier temps, nous envisageons de mettre en oeuvre les 3 versions de PERCEVAL sur un ordinateur personnel PC-80386, de manière à faire une évaluation plus aisée de ces logiciels. Eventuellement, une version de ces programmes pourra être transmise au CRC pour fins de tests, et comparaisons des résultats à ceux d'écoutes subjectives formelles.

6.2 Stratégies de codage

A la section 1 du présent rapport, nous indiquions que la stratégie de codage habituelle qui consiste à faire en sorte que le bruit de quantification suive le seuil de masquage, n'est pas nécessairement optimale. Il serait intéressant de mettre en oeuvre un *banc d'essai de stratégies de codage*. Ce banc d'essai permettrait d'injecter du bruit dans le signal, de manière à simuler une quantification. Le bruit injecté pourrait avoir différentes configurations spectrales, dépendant du spectre du signal original et de la stratégie de codage simulée. Par exemple, le spectre d'énergie du bruit pourra suivre le spectre d'énergie du signal, ou bien suivre le seuil de masquage déduit du spectre d'énergie du signal, etc...

Le but principal de ces essais est d'évaluer de manière simple l'efficacité de différentes stratégies de codage, ainsi que d'estimer une limite, pour chaque stratégie de codage et dépendant du type de signal codé, à la quantité d'information nécessaire pour avoir un codage *transparent*.

6.3 Codage

Le travail efectué sur les stratégies de codage devrait avoir des retombées directes sur l'étude d'un codeur. Toutefois, dans un premier temps, il sera possible d'étudier un codeur fonctionnant suivant la stratégie habituelle de quantification (le bruit de quantification suivant le seuil de masquage). Ce seuil de masquage sera calculé en *espace continu* à partir du modèle d'audition."OREILLE".

6.4 Modèle d'audition

Pour finir, il est clair qu'un travail important reste à faire sur le modèle d'audition "OREILLLE". En particulier, la confrontation des résultats prévus par le modèle, avec les résultats d'expériences de psycho-acoustique réelles permettront d'obtenir un bon degré de confiance et de compréhension dans les prédictions du modèle, facilitant et validant son utilisation dans un codeur, ainsi que dans un logiciel d'évaluation perceptuelle objective de dégradation.

RÉFÉRENCES

1) Techniques d'analyse temps-fréquence

[1] S. Darlington, "On digital single-sideband modulators" IEEE Trans on Circuit Theory, Vol. CT-17, pp 409-414 Aug 1970

[2] C.R. Galand , D. Esteban, "Application of quadrature mirror filters to split band voice coding schemes" Proc. of 1977 ICASSP, pp.191-195.

[3] G. Bonnerot, M. Coudreuse, M. G. Bellanger, "Digital processing techniques in the 60 channel transmultiplexer", IEEE Trans on Communications, Vol. Com-26, No5, pp 698-706, May 78.

[4] J.D. Johnston, "A filter family designed for use in quadrature mirror filter banks", Proc. of 1980 ICASSP, pp. 291-294

[5] C.R. Galand, D. Esteban, "Design and evaluation of parallel quadrature mirror filters", Proc. of 1983 ICASSP, pp. 224-227.

[6] H.J. Nussbaumer, "Complex quadrature mirror filters", Proc. of 1983 ICASSP, pp. 221-223.

[7] V.K. Jain, R.E. Crochiere, "A novel approach to the design of analysis/synthesis filter banks", Proc. of 1983 ICASSP, pp. 228-231.

[8] J. H. Rothweiler, "Polyphase quadrature filters - a new subband coding technique". Proc. of 1983 ICASSP, pp 1280-1283.

[9] C.R. Galand, H.J. Nussbaumer, "New quadrature mirror filter structures", IEEE trans. on ASSP, ASSP-32, #3, June 1984, pp. 522-530

[10] M.J.T. Smith, T.P. Barnwell, "A procedure for designing exact reconstruction filter banks fo tree structured subband coders", Proc. of 1984 ICASSP, pp. 27.1.1-27.1.4.

[11] G. Wackersreuther, "Some new aspects of filters for filter banks", IEEE trans on ASSP, October 1986, pp. 1182-1200.

[12] M. Vetterli, "A Theory of Multirate filter banks", IEEE trans. on ASSP, ASSP-35, #3, March 1987.

[13] M.J.T. Smith, T.P. Barnwell, "A new filter bank theory for time-frequency representation", IEEE trans. on ASSP, ASSP-35, #3, March 1987.

[14] B. Paillard, J. Soumagne, P. Mabilleau, S. Morissette, "Filters for subband coding , analytical approach", Proc. of 1987 ICASSP, pp. 50.2.1- 50.2.4.

[15] J. P. Princen, A. W. Johnson, A. B. Bradley, "Subband/transform coding using filter banks based on time domain aliasing cancellation". Proc. of 1987 ICASSP, pp 50.1.1-50.1.4.

[16] P.P. Vaidyanathan, P. Hoang, "The perfect reconstruction QMF bank : New architectures, solutions and optimization strategies", Proc. of 1987 ICASSP, pp. 50.3.1-50.3.4.

[17] P.P. Vaidyanathan, "Theory and design of M channel maximally decimated QMF with arbitrary M, having perfect reconstruction property", IEEE trans on ASSP, April 1987, pp. 476-492.

[18] B. Paillard, J. Soumagne, P. Mabilleau, S. Morissette, "Subband decomposition : A new analytical approach", Traitement du signal, Vol. 5, #3, 1988, pp. 133-141.

[19] P.P. Vaidyanathan, P. Hoang, "Lattice structures for optimal design and robust implementation of two-channel perfect-reconstruction QMF banks", IEEE trans on ASSP, Vol. 36, #1, January 1988.

[20] P.P. Vaidyanathan, S.K. Mitra, "Polyphase networks, Block digital filtering, LPTV systems and alias free QMF banks : A unified approach based on pseudocirculants", IEEE trans on ASSP, Vol. 36, #3, March 1988.

[21] H. S. Malvar, D. H. Staelin, "Transform coding without blocking effects". IEEE trans on ASSP, Vol. 37, No 4, April 89

[22] T.Q. Nguyen, P.P. Vaidyanathan, "Maximally decimated perfect-reconstruction FIR filter banks with pairwise mirror-image analysis (and synthesis) frequency responses", IEEE trans on ASSP, Vol. 36, #5, May 1988.

[23] P.P. Vaidyanathan, V.C. Liu, "Classical sampling theorems in the context of multirate and polyphase digital filter bank structures", IEEE trans on ASSP, Vol 36, #9, September 1988.

[24] Z. Doganata, P.P. Vaidyanathan, T.Q. Nguyen, "General synthesis procedures for FIR lossless transfer matrices, for perfect-reconstruction multirate filter bank applications".IEEE trans on ASSP, Vol. 36, #10, October 1988

[25] M. Vetterli, D. Le Gall, "Perfect Reconstruction FIR Filter Banks : Lapped Transforms, Pseudo QMF's and Paraunitary Matrices" Proc. of 1988 ISCAS, pp 2249 - 2253.

[26] B. Paillard, J. Soumagne, P. Mabilleau, S. Morissette, "Subband Decomposition, An LMS-Based Algorithm to Approximate the perfect reconstruction bank in the general case", July 1989, C.R.C.S. internal report.

[27] B. Paillard, J. Soumagne, P. Mabilleau, S. Morissette, "Subband Decomposition and Vector Filtering", July 1989, C.R.C.S. Internal report.

2) Perception et quantification

[28] G. Theile, M. Link, G. Stoll, "Low-bit rate coding of high quality audio signals", March 87, 82nd convention of AES, Preprint 2432 (C-1).

[29] G. Stoll, M. Link, G. Theile,"Masking-pattern adapted subband coding: use of the dynamic bit rate margin", March 88, 84th convention of AES, Preprint 2585 (D-5).

[30] G. Theile, M. Link, G. Stoll, "Low-bit rate coding of high quality audio signals An introduction to the Mascam system", Aug. 88, EBU review-technical, no230.

[31] G. Theile, "IRT-proposal of low bit-rate audio coding for ATV to the FCC Advisory Comittee".

[32] K. Brandenburg, "OCF - A new coding algorithm for high quality sound signals", Proc. of 1987 ICASSP pp 141-144.

[33] K. Brandenburg, "High quality coding at 2.5 bit/sample", March 88, 84th convention of AES, Preprint 2582 (D-2).

[34] D. Seitzer, K. Brandenburg, R. Kappust, E. Eberlein, H. Gerhauser, S. Krageloh, H. Schott. "DSP based real time implementation of an advanced analysis tool for audio channels", Presented at 89 ICASSP.

[35] K. Brandenburg, D. Seitzer. "Low bit rate coding of high quality digital audio : Algorithms and evaluation of quality". May 89, AES 7th international conference Audio in Digital Times.

[36] J. D. Johnston, "Transform Coding of Audio Signals Using Perceptual Noise Criteria", IEEE journal on selected areas in communications, Vol. 6, No 2, Feb. 88.

[37] J. D. Johnston. "Estimation of perceptual entropy using noise masking criteria", Proc. of 1988 ICASSP pp 2524-2527

[38] J. D. Johnston. "Perceptual transform coding of wideband stereo signals", Presented at 89 ICASSP.

[39] Petit, Charbonnier. "Subband ADPCM coding for high quality audio signals", Proc. of 1988 ICASSP, pp 2540-2543

[40] Y. Mahieux, J. P. Petit, A. Charbonnier. "Transform coding of audio signals using correlation between successive transform blocks" Presented at 1989 ICASSP.

[41] Y. Mahieux. "Algorithmes de codage par transformée pour la réduction de bruit des voies son haute qualité", Thèse de l'université de Rennes (France) - 22 mars 89

[42] B. Paillard, "Etude d'un modèlele d'audition et application d'un critère de distance perceptuelle au développement d'un codeur en sous bandes" Mémoire de maitrise de université de Sherbrooke, Dec. 87

ANNEXE 1

DESCRIPTION DU MODELE D'AUDITION "OREILLE"

Mars 1990

DESCRIPTION DU MODELE D'AUDITION "OREILLE"

Le modèle utilisé est un modèle fréquentiel. Il permet, à partir d'une densité spectrale d'énergie (ici représentée sur 20 000 composantes de 0 à 20 Khz) d'obtenir la densité basilaire d'énergie correspondante (ici représentée sur 2500 composantes de 0 à 2500 Mel). On peut ensuite, si on le désire, ajouter l'énergie de bruit et la fonction de sensibilité logarithmique des détecteurs de la membrane basilaire afin d'obtenir un vecteur de "sensation basilaire".

Transformation d'énergie fréquentielle à basilaire

La transformation **linéaire** permettant de passer de la densité fréquentielle d'énergie à la densité basilaire d'énergie, s'effectue en trois étapes. Les résultats de ces étapes sont représentés figures 1 à 4 pour une excitation d'entrée par des raies harmoniques de niveau 40 db (d'énergie 10 000) équidistantes de 500 hz; puis figures 6 à 9 pour des bruits a large bande (largeur de bande 2Khz), de niveau 40 db, espacés de 2 Khz.

- 1ère étape : La densité fréquentielle d'énergie est multipliée par le spectre d'énergie de la fonction de transfer du conduit auditif et de l'oreille moyenne.

- 2ième étape : La densité fréquentielle d'énergie est transformée en une densité basilaire localisée d'énergie : on reporte la contribution énergétique de chaque composante fréquentielle à la position correspondante sur la membrane basilaire. Cette transformation fait donc intervenir la loi de conversion non linéaire fréquence > lieu basilaire "Tonie".

- 3ième étape : La densité basilaire localisée d'énergie est "étalée" par convolution avec la fonction de dispersion. Cette fonction est une double exponentielle décroissante (-0.1 db/Mel vers les hautes fréquences et 0.27 db/Mel vers les basses fréquences).

Sensation basilaire

L'énergie de bruit de chacun des détecteurs de la membrane basilaire est ajoutée à la densité basilaire d'énergie, puis on applique à chaque composante de cette densité

basilaire une loi logarithmique (loi de sensibillité des détecteurs de la membrane basilaire). Les valeurs des énergies de bruit des détecteurs ont été choisies de telle manière que le seuil d'audition absolu (dans le silence) soit conforme aux courbes données dans [1]. En particulier, une raie harmonique d'énergie 1 (0 db) à 1 Khz est à peine audible.

Les figures 5 et 10 représentent l'allure de ces "sensations basilaires" pour les 2 exemples précédents.

Principe de détection

La sensation basilaire représente donc la valeur "mesurée" par chacune des cellules détectrices de la membrane basilaire.

2 densités fréquentielles d'énergie (par exemple la densité correspondant à un signal original et celle du même signal bruité) donnent lieu à 2 sensations basilaires distinctes.

La probabilité de détection de la différence entre les 2 signaux est calculée pour chaque cellule détectrice, en fonction de la valeur absolue de la différence entre les 2 sensations basilaires, au point considéré.

Les probabilités de détection sont supposées **indépendantes** pour chaque cellule détectrice, de sorte que la probabilité globale de non-détection du bruit est égale au produit des probabilités de non-détection individuelles des cellules détectrices.

Mise en place du modèle et hypothèses de fond

Les hypothèses de départ sont au nombre de 3:

1) Chaque point de la membrane basilaire (donc chaque détecteur) a une réponse impulsionnelle donnée. Cette réponse impulsionnelle varie continuement avec l'abscisse b, tout au long de la membrane basilaire. Autrement dit, chaque point d'abscisse b de la membrane basilaire reçoit le signal présent au tympan, filtré par un filtre H(b), qui est propre au point b en question.

   Cette hypothèse est très peu restrictive, elle rend uniquement compte de la linéarité des phénomènes mécaniques dans l'oreille interne.

2) Chaque cellule détectrice de la membrane basilaire est sensible à l'énergie qui est présente en ce point (avec une loi de sensibilité logarithmique). Le fait que la sensibilité soit logarithmique n'intervient pas tout de suite pour la mise en place du modèle, ce qui compte surtout est que ce soit à l'énergie que la cellule est sensible.

   Cette hypothèse est, on le voit, un peu plus restrictive.

3) La troisième hypothèse faite est que le temps de réponse des cellules détectrices est assez long.

   Comme on le voit, cette hypothèse est très relative, c'est elle, en fait, qui détermine la limite de validité d'un modèle purement fréquentiel. Par exemple, un modèle fréquentiel comme le nôtre ne tient pas compte des phénomènes temporels tels que les battements lorsque 2 fréquences sont très proches. Ceci est dû au fait que ce modèle suppose que le temps d'intégration des détecteurs de la membrane basilaire est infiniment long; et donc suppose que ces détecteurs ne peuvent détecter aucune variation temporelle de l'énergie sur la membrane basilaire. Ce qui bien entendu n'est vrai que dans une plage limitée de validité.

En résumé, ces 3 hypothèses définissent un modèle schématique de l'oreille tel que décrit à la figure 11.

Les hypothèses 2) et 3) stipulent que les détecteurs de la membrane basilaire sont sensibles à l'intégrale sur un temps très long de l'énergie présente sur ce détecteur. Or, si on fait abstraction du temps, il est possible de calculer de manière très simple la densité basilaire d'énergie (intégrée entre les temps $-\infty$ et $+\infty$)en réponse à une densité fréquentielle d'énergie excitatrice (elle aussi intégrée, en toute rigueur, sur un temps infini).

Comme il est représenté sur la figure 12, chaque point d'abscisse b de la membrane basilaire reçoit entre les temps $-\infty$ et $+\infty$, une quantité d'énergie égale à l'intégrale sur l'axe fréquentiel de la densité spectrale d'énergie du signal d'excitation (obtenue en prenant la norme de la transformée de Fourier du signal d'excitation) multipliée par la densité spectrale du filtre correspondant au point b (obtenue en prenant la norme de la transformée de Fourier de la réponse impulsionnelle du filtre).

Si on décrit cette opération avec un formalisme d'espaces de Hilbert (et en utilisant les notations d'algèbre linéaire qui sont plus parlantes), il s'agit en fait du produit

scalaire du vecteur "densité spectrale d'énergie du signal" par le vecteur "densité spectrale d'énergie du filtre".

Il faut répéter cette opération pour chaque point b de la membrane basilaire pour obtenir le vecteur "densité basilaire d'énergie".

Il est possible de représenter cette infinité de produits scalaires (un pour chaque point b de la membrane basilaire) par une opération matricielle (figure 13). Le vecteur continu "densité spectrale d'énergie" est en entrée, la sortie est le vecteur continu "densité basilaire d'énergie. Entre les 2, la matrice T effectue la transformation souhaitée.

A noter que la matrice T est continue (donc de dimensions $\infty x \infty$), ce qui n'est pas facile à représenter graphiquement ...

En toute rigueur, on devrait dire que T est un *opérateur linéaire* (au sens des espaces de Hilbert). Si on observe une ligne de cette matrice, les valeurs (toujours positives) de la matrice sur cette ligne représentent la densité spectrale d'énergie d'un filtre H(b) en un point b de la membrane basilaire. De même, si on observe une colonne de T, les valeurs (toujours positives) de T sur cette colonne représentent la densité basilaire d'énergie dûe à une excitation par une raie harmonique à la fréquence f. Ces 2 descriptions de la matrice (par lignes ou par colonnes) constituent 2 façons de la représenter (matrice directe ou matrice transposée), et donc ces 2 descriptions sont parfaitement équivalentes. Il revient au même de connaitre les densités basilaires d'énergie résultant d'une excitation harmonique à la fréquence f, pour toutes les fréquences, que de connaitre les densités spectrales d'énergie des filtres correspondant à tous les points b de la membrane basilaire. La figure 14 donne une idée de l'allure de la matrice T^T.

Ceci dit, le résultat le plus important est bien l'existance (liée aux hypothèses 1) à 3)) d'une transformation **linéaire** permettant d'obtenir une densité basilaire d'énergie à partir d'une densité spectrale d'énergie d'excitation.

La linéarité de la transformation implique en particulier l'**additivité** des énergies sur la membrane basilaire:

Si $B_1(b)$ est la densité basilaire d'énergie résultant d'une excitation par la densité spectrale d'énergie $F_1(f)$

$$B_1(b) = T.F_1(f) \qquad (1)$$

$B_2(b)$ est la densité basilaire d'énergie résultant d'une excitation par la densité spectrale d'énergie $F_2(f)$

$$B_2(b) = T.F_2(f) \tag{2}$$

alors, la densité basilaire d'énergie résultant de la somme des 2 excitations est égale à la somme des 2 densités basilaires:

$$B_1(b) + B_2(b) = T.\left(F_1(f) + F_2(f)\right) \tag{3}$$

Si, en plus de la linéarité de T, on remarque que les densités basilaires d'énergie en réponse à des excitations harmoniques de fréquences différentes sont identiques, et seulement décalées sur l'axe basilaire les unes par rapport aux autres (invariance spatiale sur l'axe basilaire), on est conduit tout naturellement à décomposer la transformation T en:

- Une localisation de l'énergie spectrale sur l'axe basilaire, cette localisation devant s'effectuer à intégrale conservative.

- Une dispersion de l'énergie basilaire localisée par **convolution** avec une fonction de dispersion. Cette fonction de dispersion étant par ailleurs la densité basilaire d'énergie résultant d'une excitation harmonique pure. Dans notre référence [1], Zwicker et Feldtkeller montrent cette densité basilaire d'énergie en réponse à une excitation harmonique, comme un triangle à double pente (en échelle logarithmique). Nous avons donc pris comme fonction de dispersion (en échelle linéaire) une double exponentielle décroissante de constantes d'espace -0.1 db/Mel vers les hautes fréquences et 0.27 db/Mel vers les basses fréquences.

Il ne reste (si nécessaire) qu'à appliquer en chaque point de la densité basilaire d'énergie la loi de sensibilité (S = Log(E+e)) des détecteurs de la membrane basilaire, pour obtenir un vecteur de sensation basilaire d'où on pourra par exemple déduire la "sonie" de l'excitation.

[1] E. Zwicker, R. Feldtkeller - *Psychoacoustique, l'oreille récepteur d'information* - traduit de l'allemand par Christel Sorin - 1981 - collection technique et scientifique des télécommunications - MASSON - ISBN: 2-225-74503-X

[2] Bruno Paillard - mémoire de maitrise - Université de Sherbrooke - Décembre 87 "Etude d'un modèle d'audition et application d'un critère de distance perceptuel au développement d''un codeur en sous-bandes"

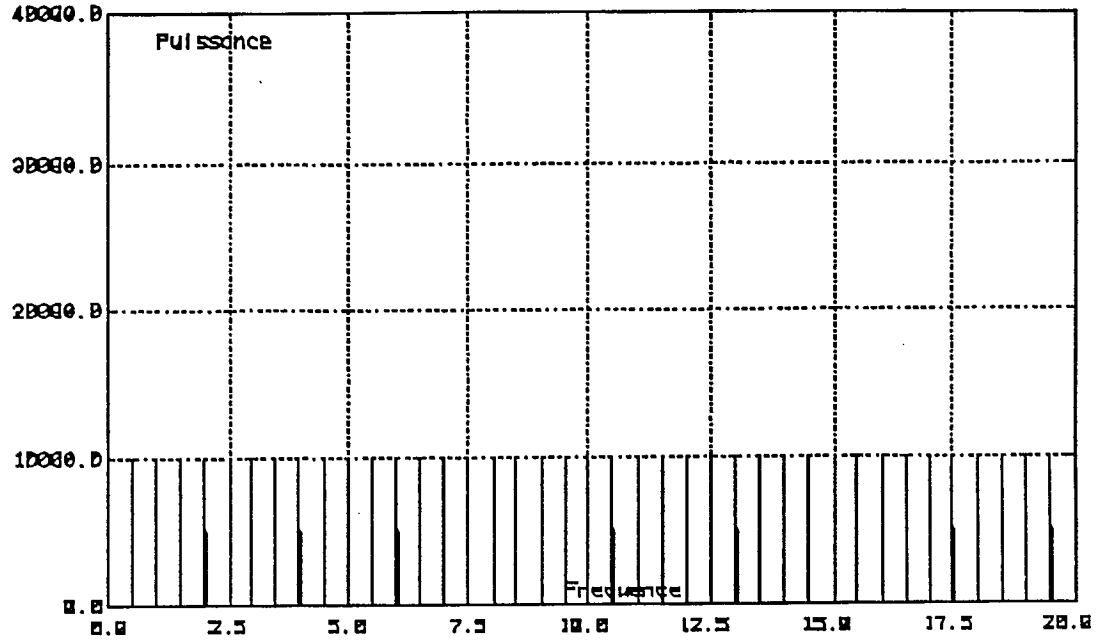


Figure 1 : répartition fréquencielle d'énergie originale

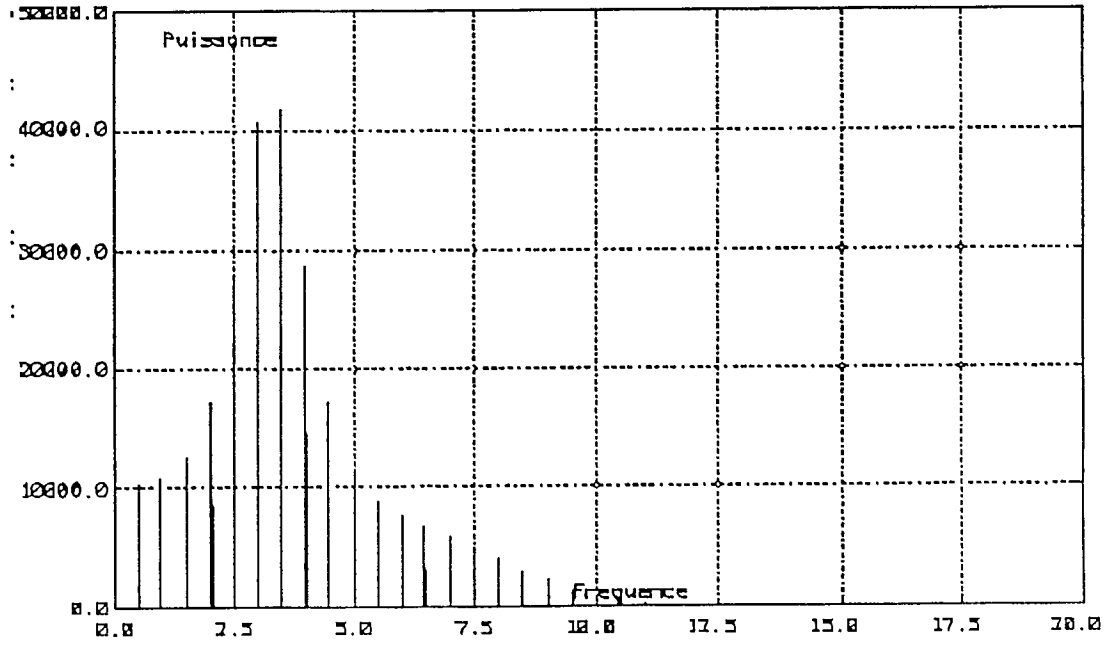


Figure 2 : Répartition fréquencielle d'énergie après atténuation de l'oreille moyenne

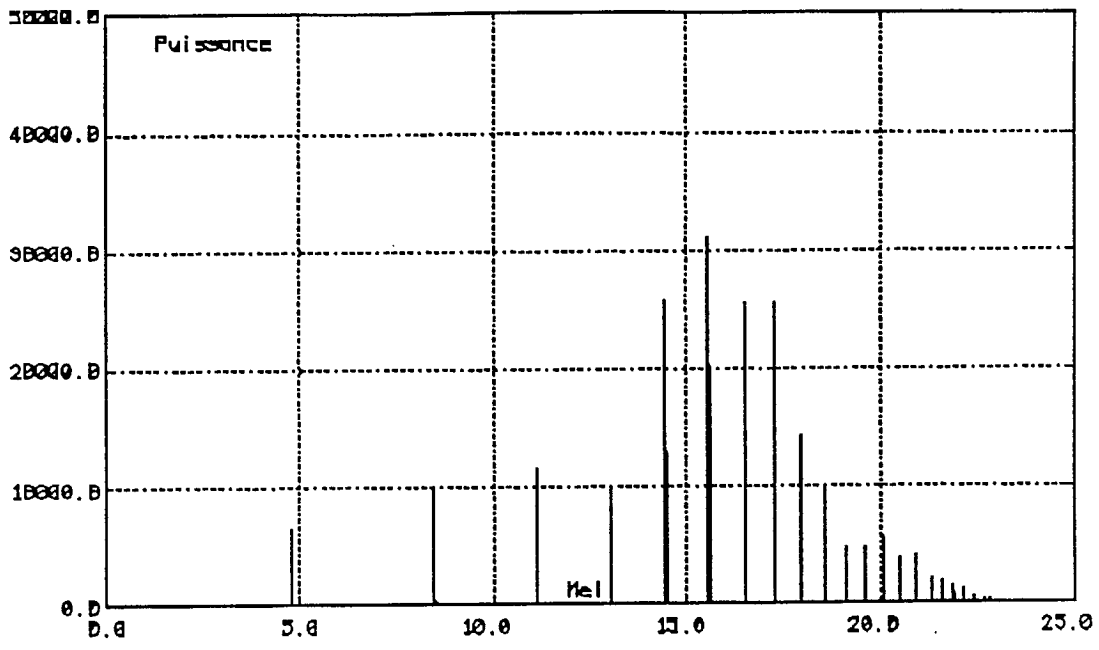


Figure 3 : Répartition basilaire localisée d'énergie

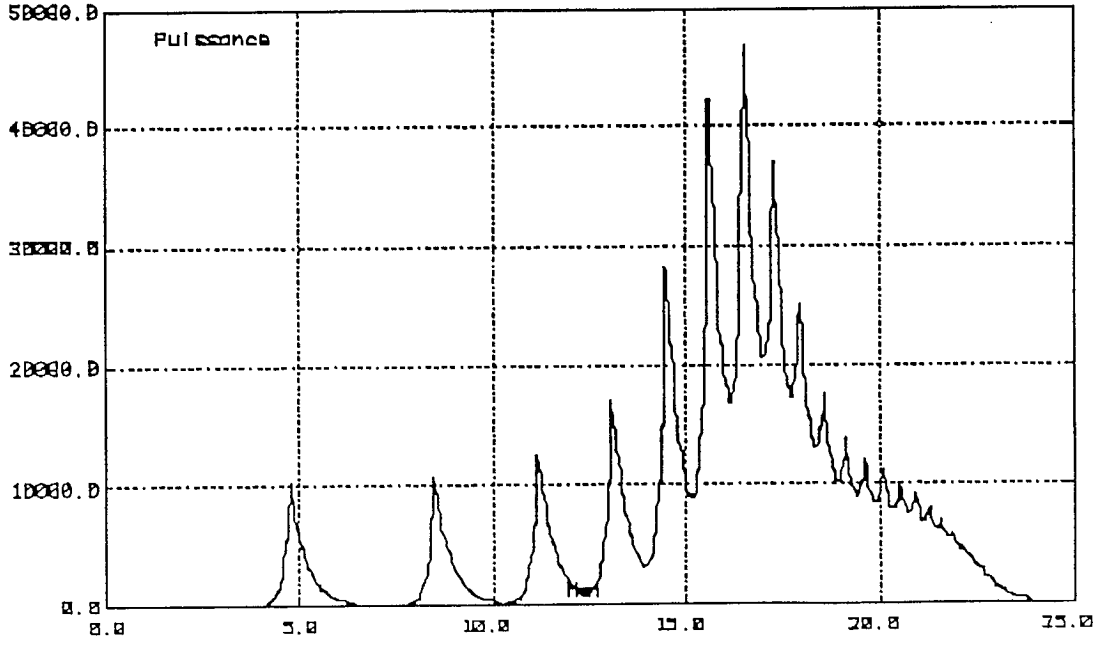


Figure 4 : Répartition basilaire dispersée d'énergie

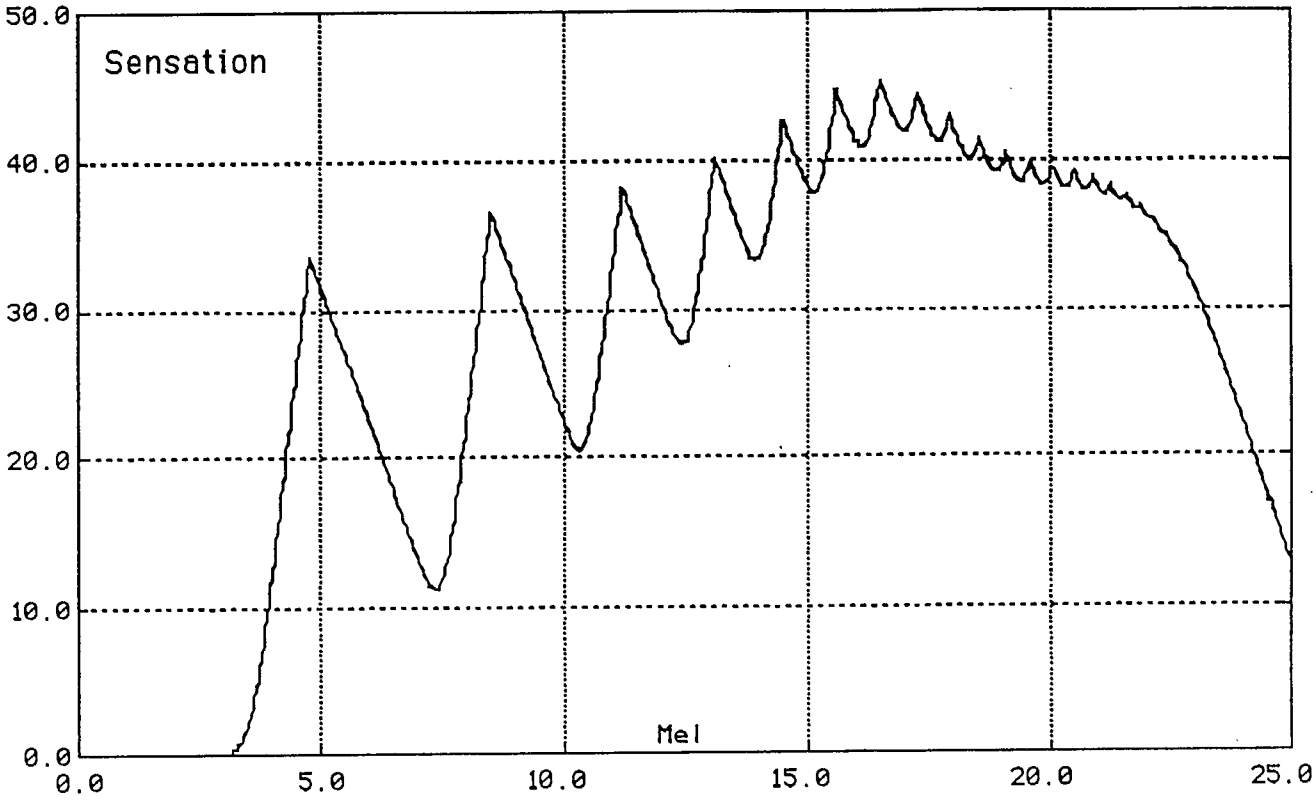


Figure 5 : Sensation basilaire

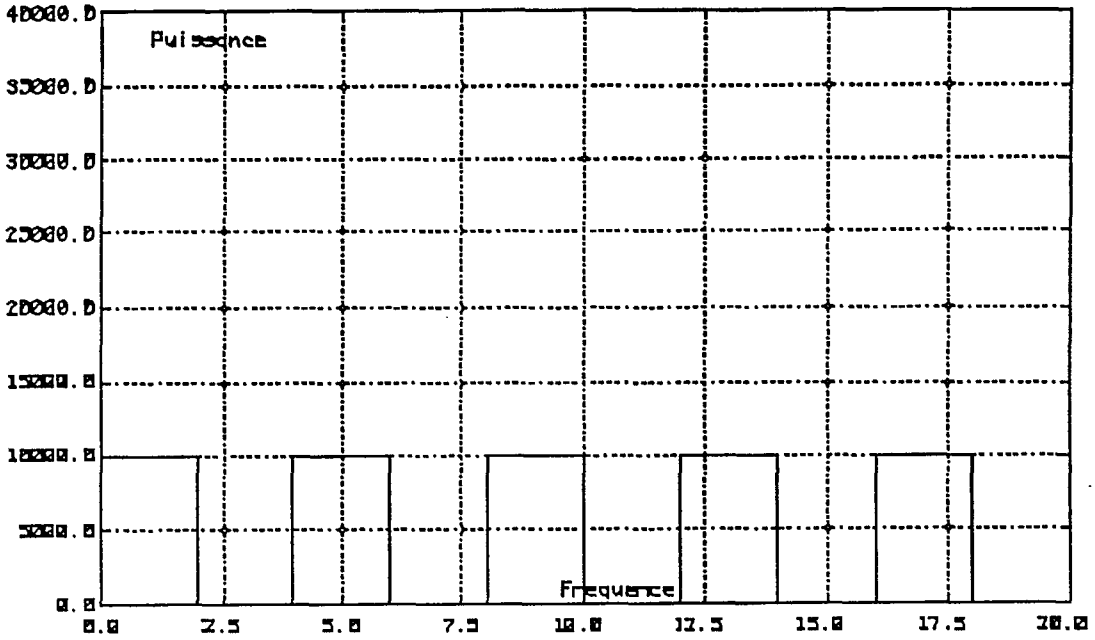


Figure 6 : Répartition fréquencielle d'énergie originale

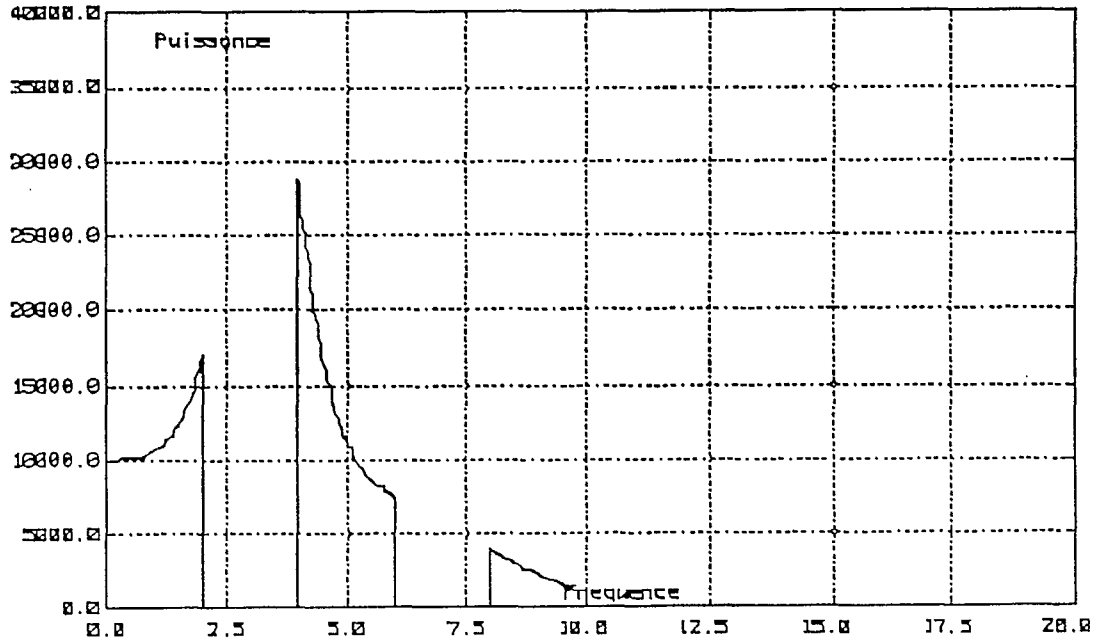


Figure 7 : Répartition fréquencielle d'énergie après atténuation de l'oreille moyenne

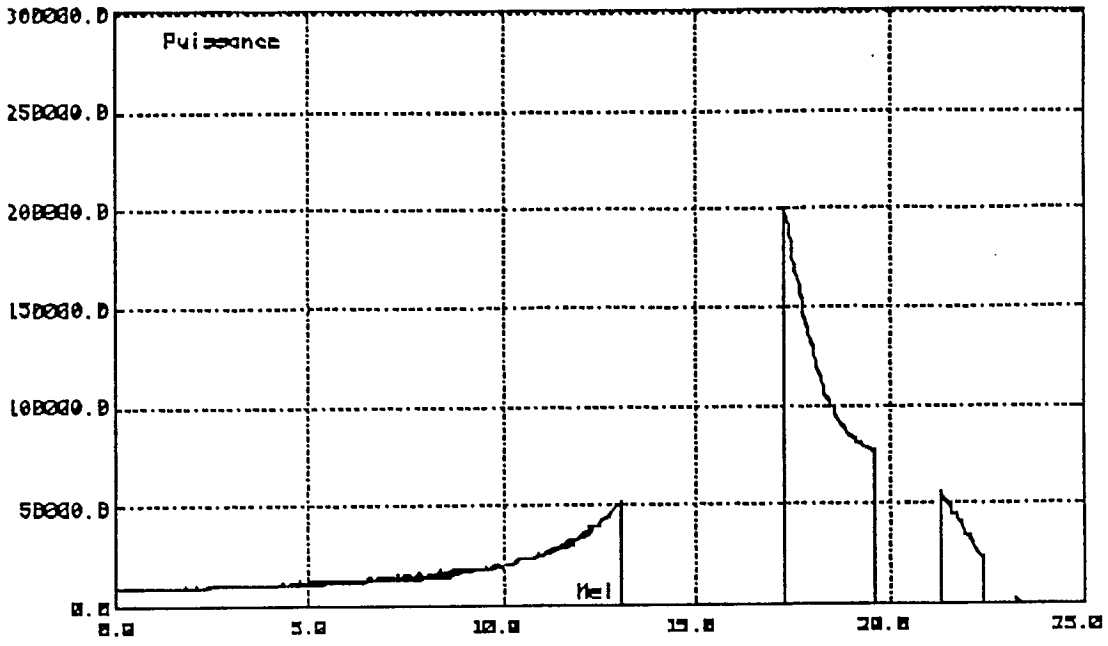


Figure 8 : Répartition basilaire localisée d'énergie

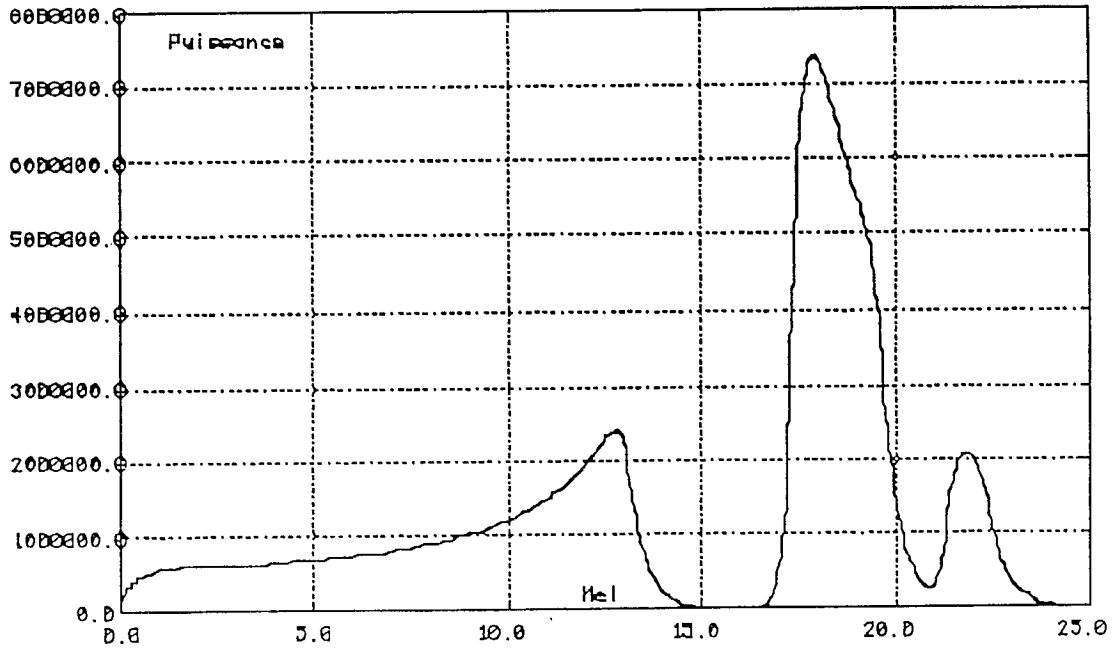


Figure 9 : répartition basilaire dispersée d'énergie

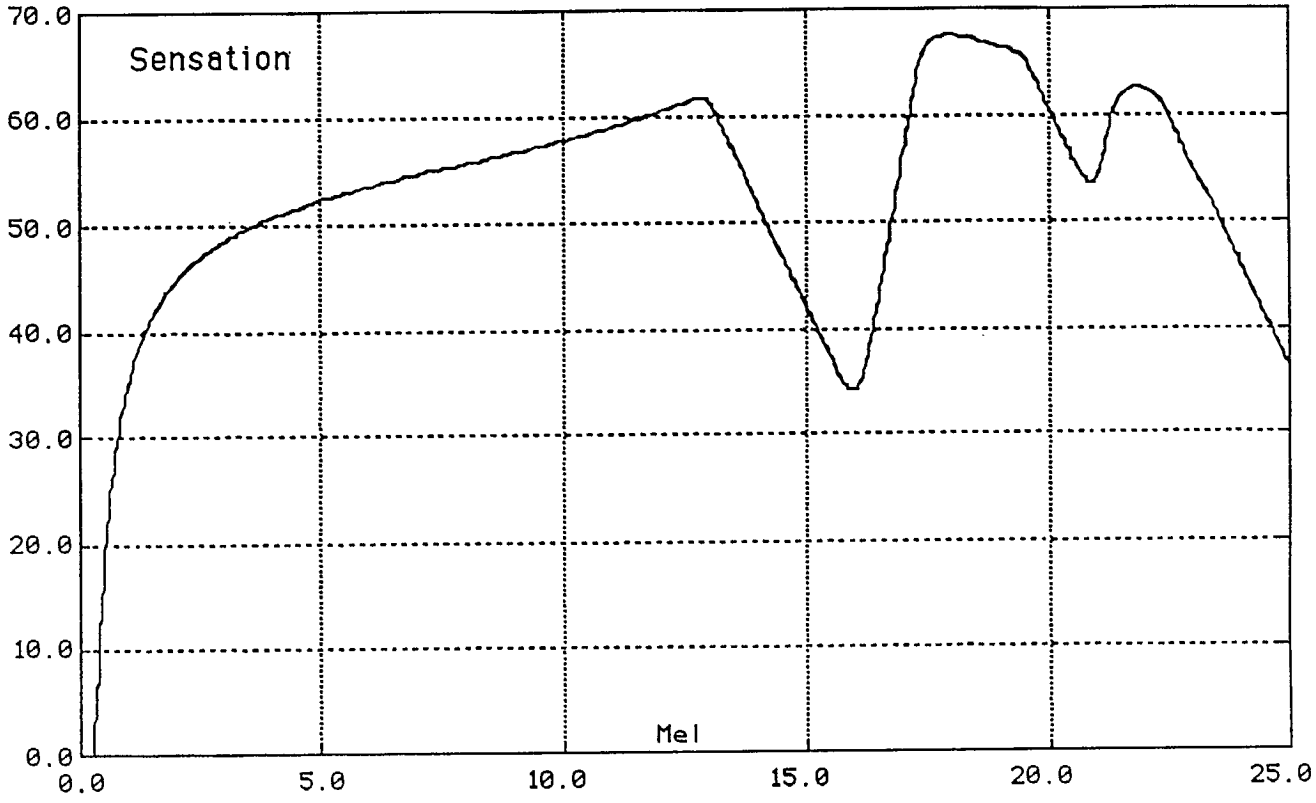
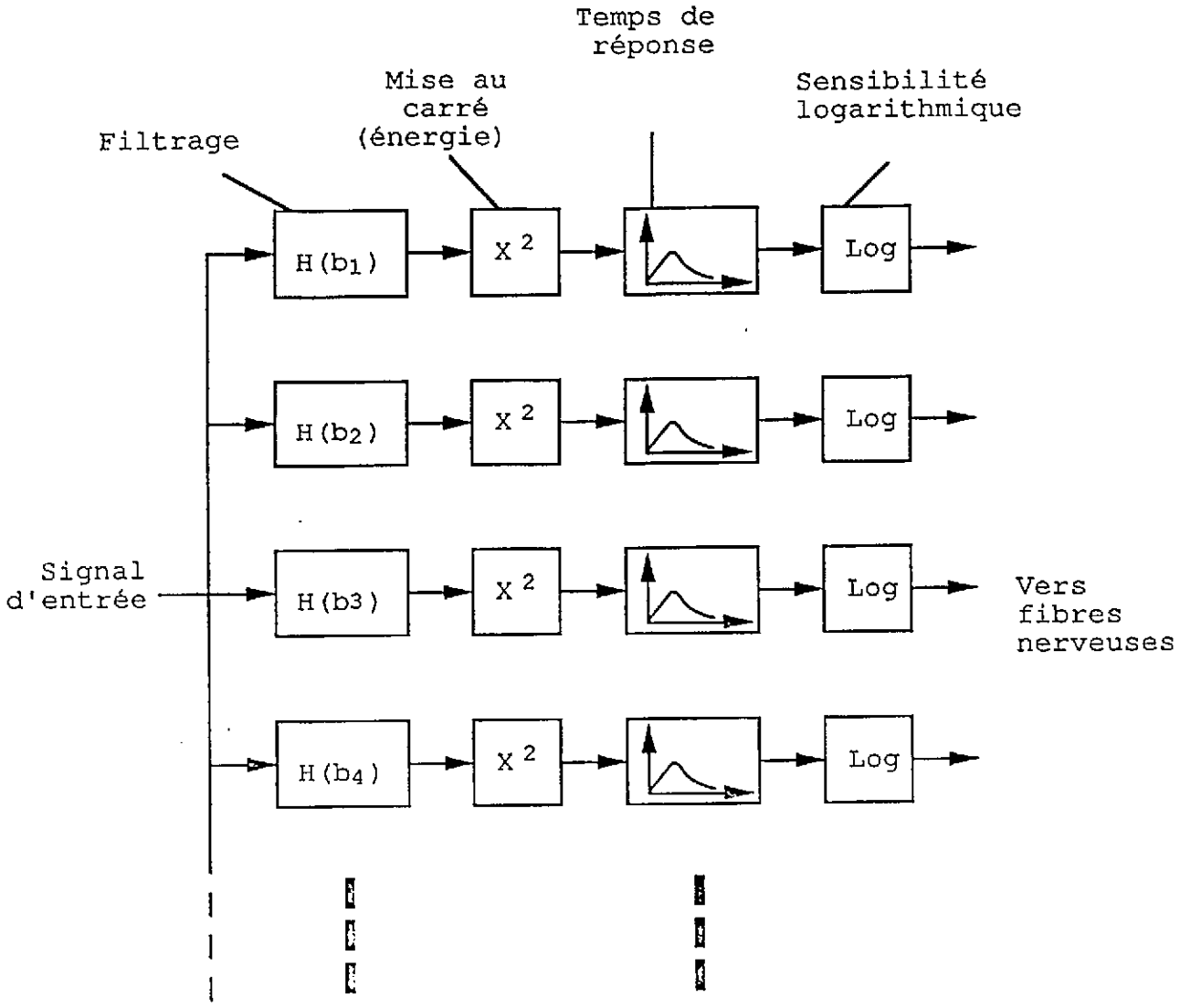
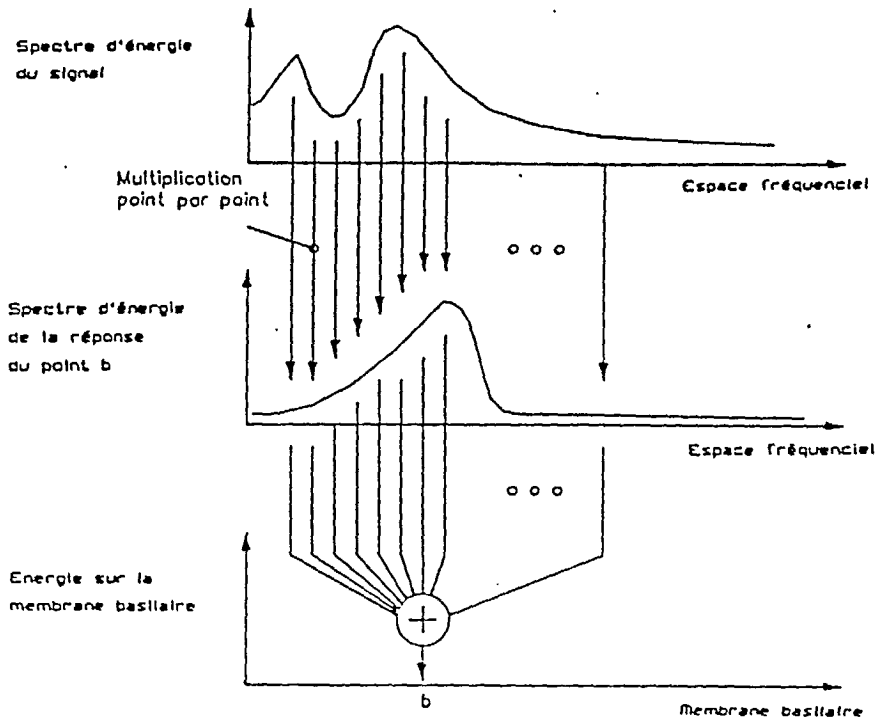
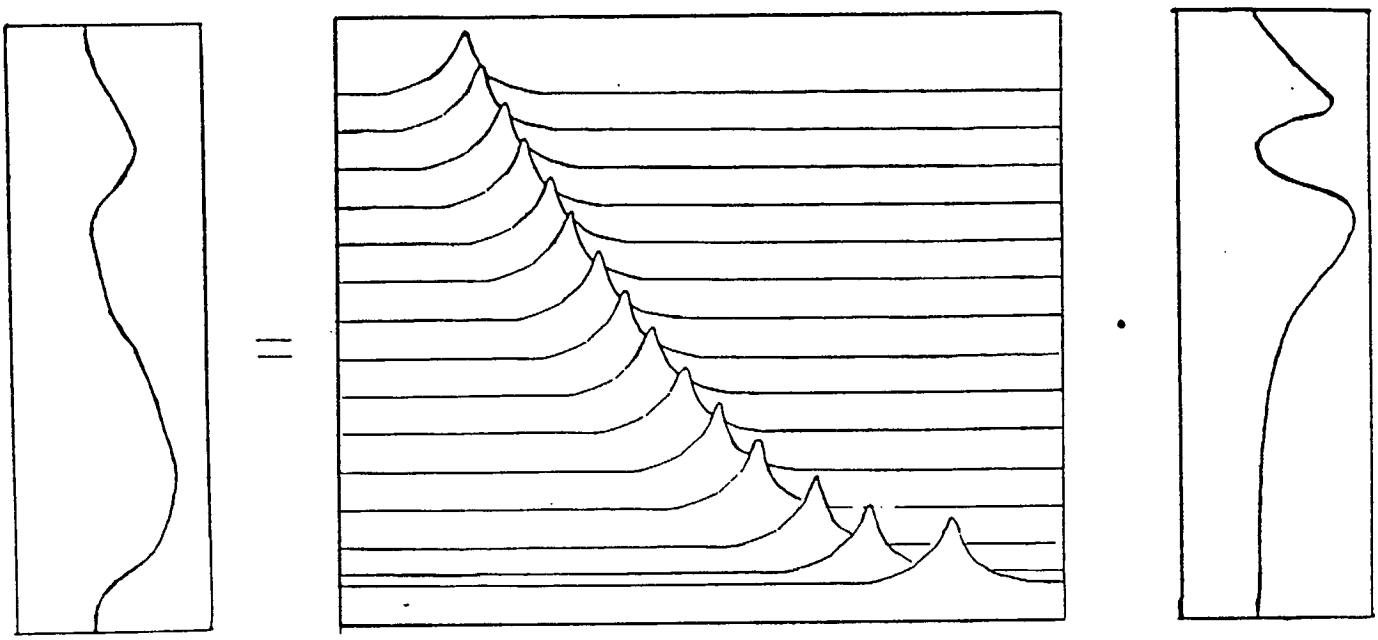


Figure 10 : Sensation basilaire



Modèle schématique de
l'oreille

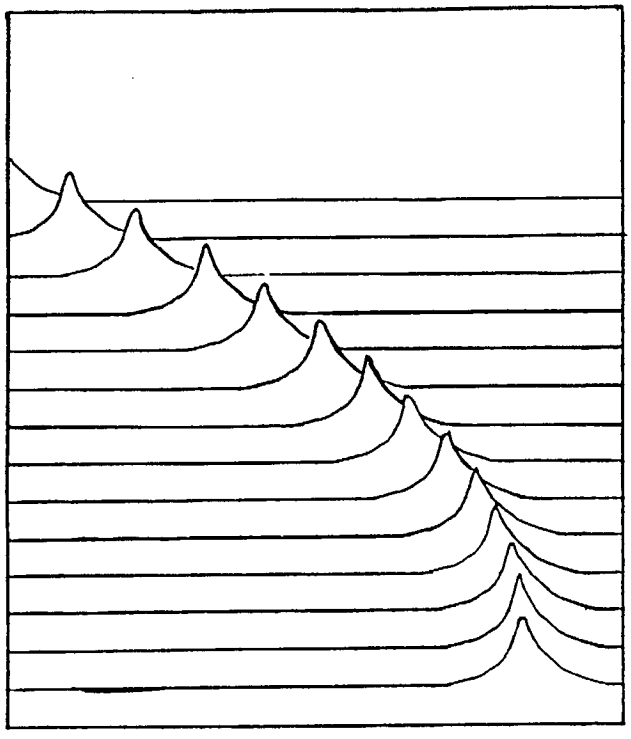

Spectre d'énergie
du signal
Espace fréquenciel
Multiplication
point par point
Spectre d'énergie
de la réponse
du point b
Espace fréquenciel
Energie sur la
membrane basilaire
b
Membrane basilaire



Matrice T

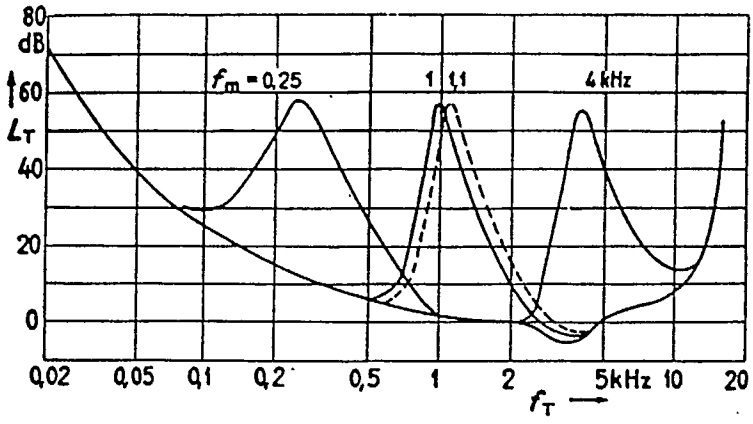
Vecteur de densité basilaire d'énergie

Vecteur de densité fréquencielle d'énergie

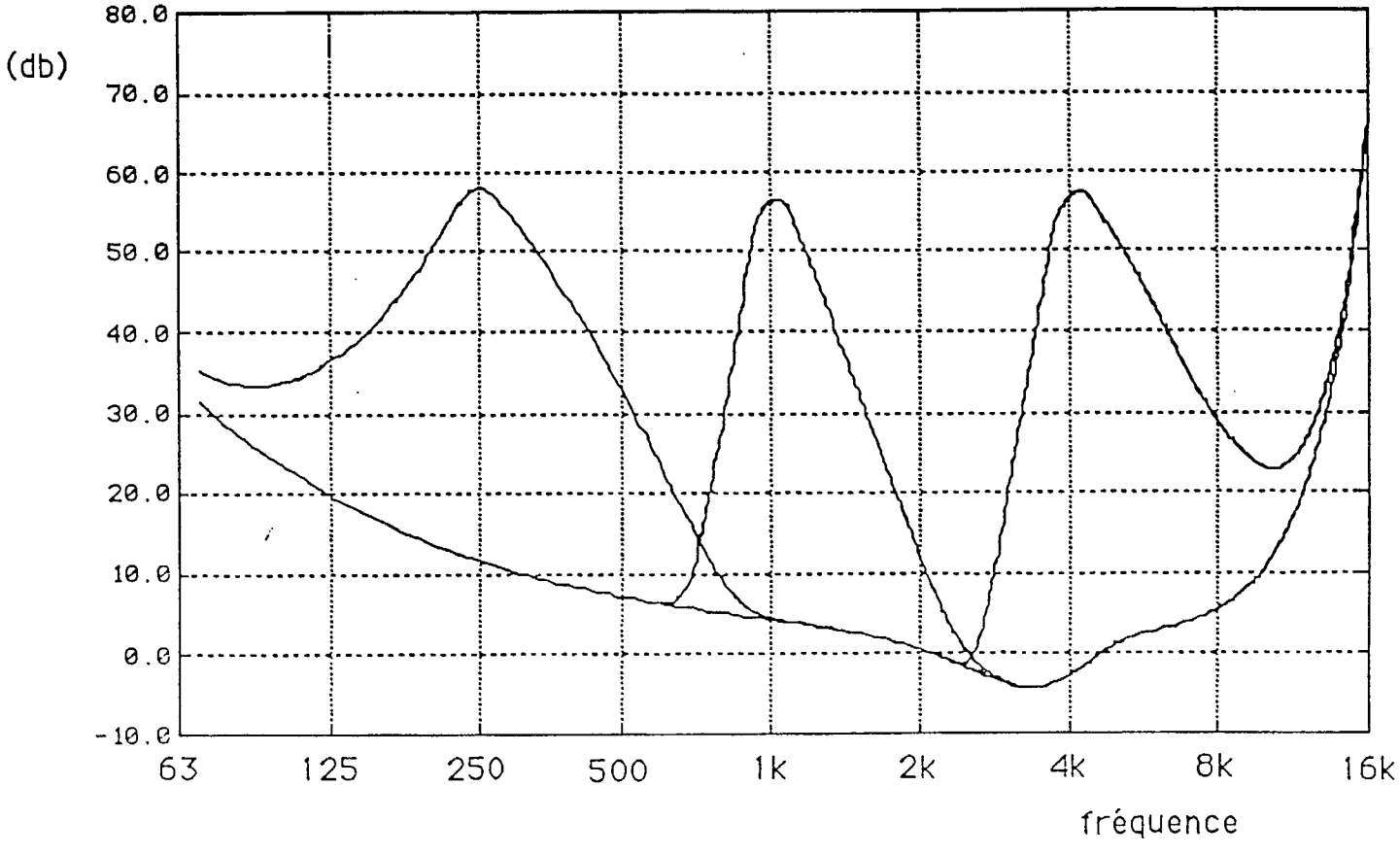


Matrice T^T
(chaque ligne représente
la densité basilaire résultant d'une excitation
fréquencielle à la fréquence f.

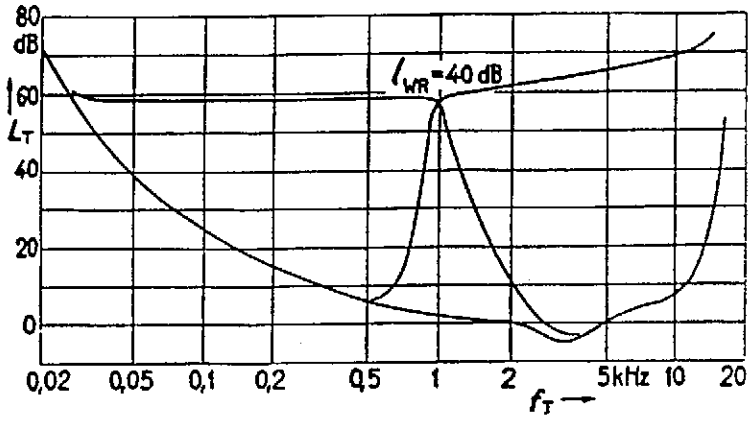
Quelques résultats



Courbes d'effet de masque de bruits à bande étroite de niveaux 60 db de fréquences centrales 250, 1000 et 4000 hz
(tiré de "L'oreille récepteur d'informations" E. Zwicker, R. Feldtkeller)

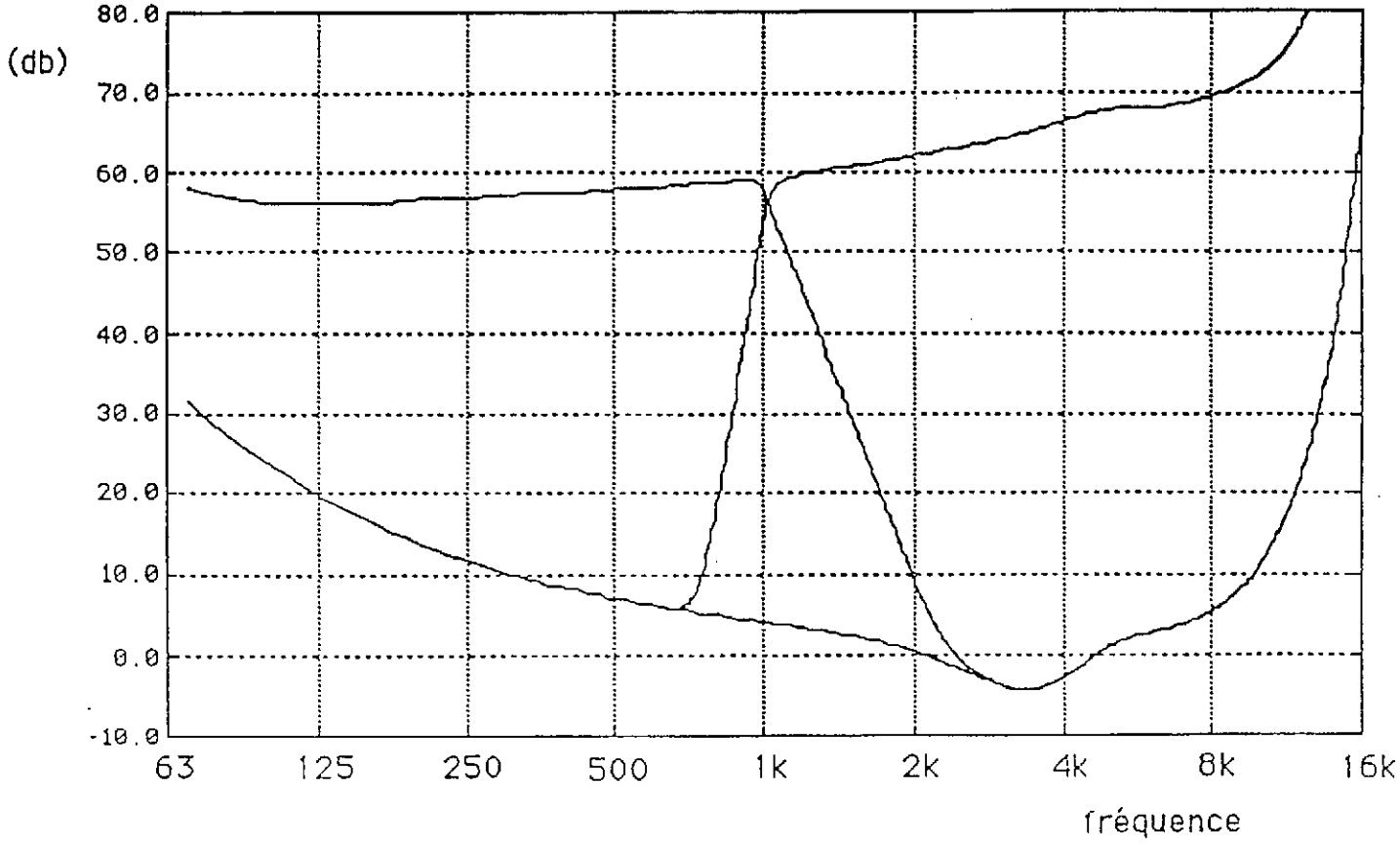


Courbes d'effet de masque de bruits à bande étroite de niveaux 60 db de fréquences centrales 250, 1000 et 4000 hz (calculés par OREILLE)

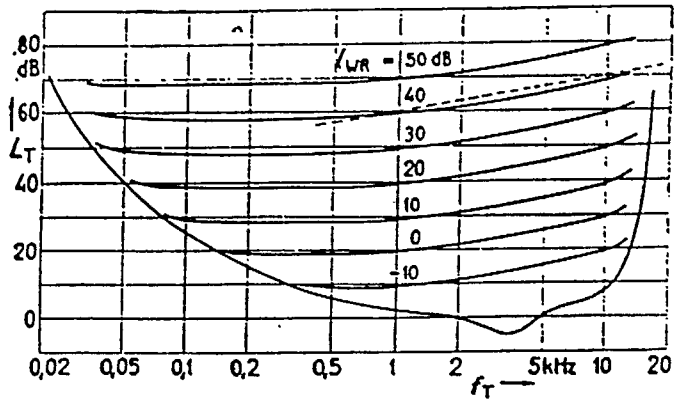


Courbes d'effet de masque de bruits passe bas et passe haut

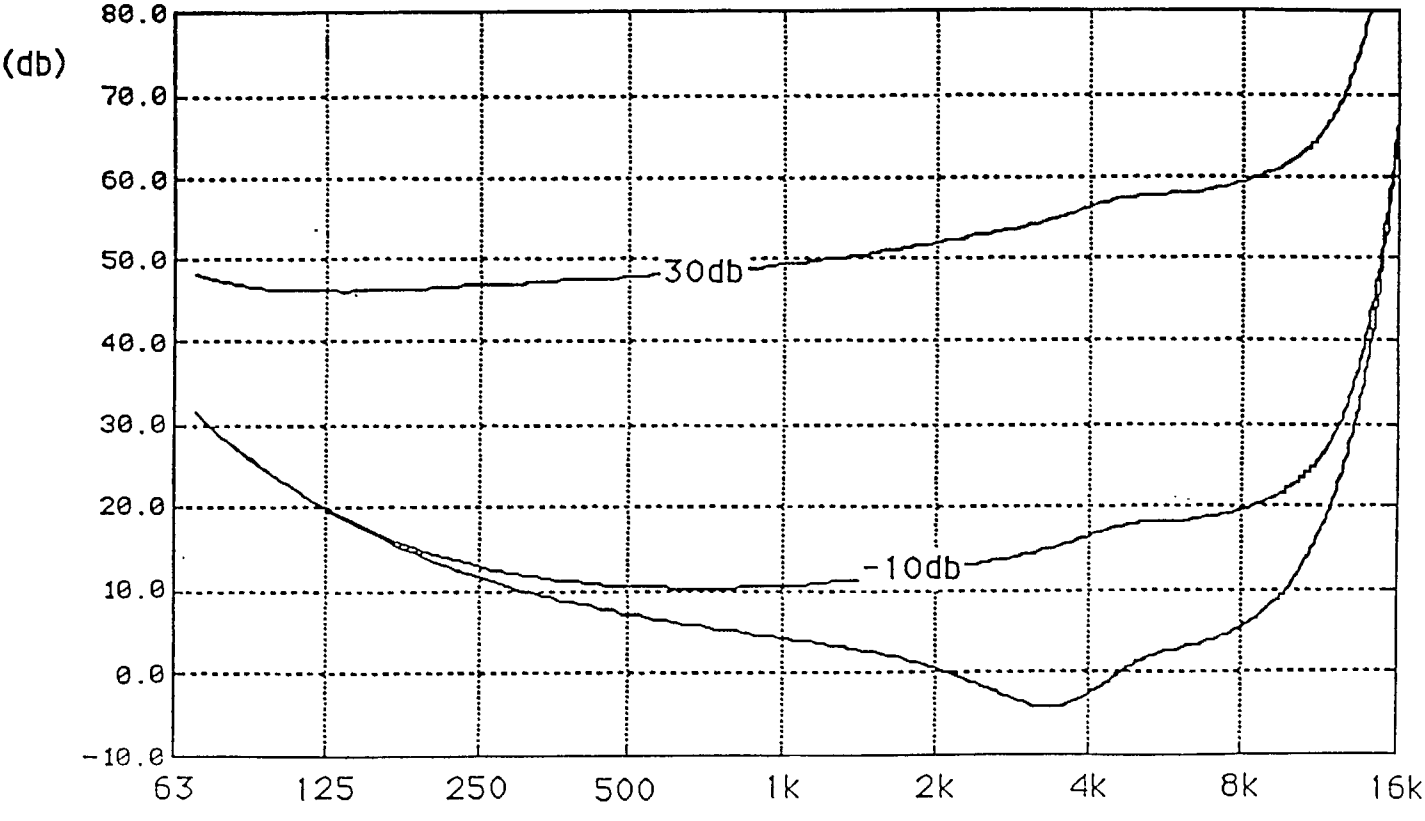
(tiré de "L'oreille récepteur d'informations" E. Zwicker, R. Feldtkeller)



Courbes d'effet de masque de bruits passe bas et passe haut
(calculés par OREILLE)



Seuils d'audition en présence de bruit blanc masquant de divers niveaux
(tiré de "L'oreille récepteur d'informations" E. Zwicker, R. Feldtkeller)



Seuils d'audition en présence de bruit blanc masquant de divers niveaux
(calculé par OREILLE)

Annexe 2

OREILLE : PRINCIPE DE DÉTECTION

OREILLE: PRINCIPE DE DÉTECTION

1 Introduction

Jusqu'à présent, le principe de détection que nous utilisions pour notre modèle d'audition était assez simple. Une différence entre 2 spectres d'énergie était détectée si la différence entre les 2 sensations basilaires correspondantes dépassait un certain seuil (de l'ordre de 1 à 2 db) sur un, au moins, des quelques miliers de détecteurs que compte la membrane basilaire.

Rappelons que la *sensation basilaire*, qui représente la grandeur *mesurée* par les détecteurs de la membrane basilaire, est obtenue en ajoutant en chaque point de la densité basilaire d'énergie, une très faible énergie de repos (qui rend compte du seuil d'audition absolu), et en prenant le logarithme de la densité basilaire d'énergie résultante (c.f. [2]). Autrement dit, pour chaque point "i" de la membrane basilaire, la sensation S_i est :

$$S_i = \mathrm{Log}\left(E_i + \varepsilon_i\right) \tag{1}$$

où:

E_i est l'énergie au point "i" de la membrane basilaire.

e_i est l'énergie de repos du détecteur "i"

Ce principe de détection déterministe, associé au modèle d'audition décrit dans [2], prédit avec une très bonne précision (de l'ordre de 1 db) les seuils de masquage de raies harmoniques masquées dans du bruit à bande étroite ou à large bande (c.f. [2]). Il prédit aussi, et ceci bien que le modèle travaille en espace continu, les propriétés macroscopiques de l'audition que constituent les *bandes critiques* (c.f. [3]).

Par contre, ce principe de détection déterministe ne prédit pas avec une bonne précision les masquages de bruit à bande large, ou étroite, par des signaux harmoniques.

C'est d'autant plus dommage que, dans le cas du codage, on est bien dans cette situation d'un bruit à bande large qu'on tente de masquer, le plus souvent, dans un signal à spectre harmonique.

Hellman [4], ou Schroeder et al. [5], notent une assymétrie entre le masquage d'un son harmonique par un bruit, et le masquage d'un bruit par un son harmonique. On peut résumer les résultats qu'ils obtiennent comme suit:

- Le seuil de détection d'une raie harmonique, masquée par du bruit à bande étroite (tiers d'octave) centré autour de celle-ci, est environ 4 db sous le niveau du bruit masqueur.

- Le seuil de détection d'un bruit tiers d'octave, masqué par une raie harmonique à la fréquence centrale du bruit, est environ 25 db sous le niveau du masqueur harmonique.

Notre modèle prévoit avec une bonne précision le seuil de masquage (niveau du masqueur moins 4 db), dans le cas où le masqueur est un bruit tiers d'octave (ou d'ailleurs toute autre sorte de bruit) et le signal masqué est une raie harmonique. Il prévoit dans le cas inverse d'un bruit tiers d'octave masqué par une raie harmonique, un seuil de masquage égal au niveau du masqueur moins 10 db. Ce résultat va bien dans le sens de l'assymétrie notée par Hellman, ou Schroeder et al., mais est insuffisant d'environ 15 db.

C'est ce genre de problèmes qui poussent certaines équipes (Johnston aux laboratoires Bell, par exemple [6]), à tenir compte de l'aspect spectral du masqueur (ici un signal à coder), et à baisser artificiellement de 10 à 15 db le seuil de masquage calculé pour ce masqueur , si celui-ci est de type harmonique.

Un principe de détection plus réaliste, basé sur une approche statistique de la détection, permet de rendre compte de cette assymétrie plus importante que prévue. L'effet qui est à l'oeuvre dans ce nouveau principe de détection peut se résumer comme suit

> *Si c'est un nombre important de détecteurs basilaires qui prennent part à la détection (comme c'est le cas pour la détection d'un bruit masqué par un son harmonique), il suffit, pour avoir détection, d'une différence de sensation plus faible sur chaque détecteur, que dans le cas où c'est un nombre réduit de détecteurs qui interviennent.*

Des problèmes d'expérimentation sont aussi à mettre en cause pour expliquer l'ampleur de l'assymétrie, notament l'allure spectrale des bruits tiers d'octave utilisés.

Comme nous allons le voir, lorsqu'ils sont en position de signal masqué (par un masqueur harmonique), des bruits tiers d'octave ayant des pentes de décroissance aussi fortes que ...100 db/octave! ne peuvent pas être considérés comme parfaits (le modèle prévoit des résultats notablement différents de ceux prévus pour des bruits tiers d'octave parfaits).

2 Principe de détection

Imaginons 2 signaux, par exemple l'un correspond à un signal original, et l'autre au même signal, où on a ajouté du bruit. A chacun de ces signaux correspond une sensation basilaire qui peut être calculée par le modèle d'audition (c.f. [2]).

Appelons **S** la sensation basilaire originale, et $\mathbf{S}^*$ celle qui correspond au signal bruité.

Chaque détecteur basilaire "i" mesure une valeur S_i dans un cas, et S^*_i dans l'autre.

Jusqu'à maintenant, nous considérions que si, pour le détecteur S_i, la différence de sensation $|S^*_i - S_i|$ était supérieure à un certain seuil (de l'ordre de 1 à 2 db), ce détecteur notait la différence. Globalement, une différence (un bruit) était détectée par l'auditeur, si elle était détectée par un détecteur basilaire au moins.

Le nouveau principe de détection utilisé est une généralisation du précédent:

Chaque détecteur a une *probabilité de détection* dépendant de la différence de sensation $|S^*_i - S_i|$ qu'il enregistre. Bien entendu, cette probabilité de détection tend vers 1 si cette différence tend vers l'infini, et elle tend vers 0 si la différence tend elle même vers 0.

La figure 1.b montre l'allure de cette probabilité de détection en fonction de la différence $|S^*_i - S_i|$. La figure 1.a montre cette même probabilité de détection pour le principe de détection déterministe précédent.

Les probabilités de détection sont supposées indépendantes pour chaque détecteur, et globalement il y a détection si un détecteur au moins note une différence de sensation. De cette façon, on peut calculer la probabilité de non-détection globale comme le **produit** des probabilités de non-détection individuelles des détecteurs basilaires.

Notons que cette approche statistique correspond mieux aussi à la réalité des phénomènes de détection, ainsi qu'aux techniques d'expérimentation en psycho-

acoustique. En général, on ne définit les seuils de détection qu'en termes de *probabilités de détection*. Ces seuils sont obtenus pour une probabilité de détection de 50 %.

Ayant décidé de ce principe de détection, il nous reste à obtenir la fonction donnant la probabilité de détection en fonction de la différence de sensation $|S^*_i - S_i|$.

2.1 L'expérience de Buus et al

Dans [7], Buus et al. étudient plusieurs principes de détection, dont celui des *canaux indépendants* que nous venons de décrire. Cette étude est faite à la lumière d'expériences de détection de raies harmoniques dans du bruit uniformément masquant.

L'expérience principale se déroule comme suit:

En présence de bruit uniformément masquant (de densité spectrale 25 db autour de 1100 hz), on mesure tout d'abord les probabilités de détection de divers harmoniques, présentés isolément, en fonction de leur niveau. On note bien que ces probabilités sont presque identiques pour chacun des harmoniques présentés. Le bruit est uniformément masquant, donc le seuil de masquage correspondant est le même, soit environ 44 db, quelle que soit la fréquence de l'harmonique à détecter.

On présente ensuite, toujours en présence de bruit uniformément masquant, un son composé de 18 harmoniques de même niveau, et on mesure la probabilité de détection en fonction de ce niveau.

La figure 2 représente les résultats de Buus et al., tirés de [7]. On note que les probabilités indiquées varient entre 0.5 et 1. Ceci est dû à la procédure *2 alternatives à choix forcé* utilisée pour l'expérimentation. Ces probabilités représentent donc les probabilités de se tromper lors du choix (dans le pire des cas on a 1 chance sur 2 de se tromper). A partir de ces probabilités d'erreur, on peut facilement obtenir les probabilités de détection (variant bien de 0 à 1), par simple changement d'échelle et recadrage.

Ces résultats amènent plusieurs conclusions :

- D'une part le niveau sur chaque harmonique, nécessaire pour détecter le son complexe, est inférieur d'environ 6 db au niveau nécessaire à la détection d'une seule composante présentée isolément. Ceci invalide bien le principe de détection qu'on utilisait jusqu'ici, qui aurait prévu le même seuil de détection, que le son soit simple ou complexe.

- D'autre part, on remarque que les courbes de *probabilités versus niveau* sont toutes parallèles entre elles. A partir de cette propriété, Buus et al. isolent une famille de fonctions possibles, donnant la probabilité de détection en fonction du rapport *signal masqué sur signal masqueur* pour un détecteur. Ils vont même plus loin, et calculent le paramètre particulier qui, appliqué à cette famille de fonctions, donne des prévisions cohérentes avec les résulats de l'expérience.

Toutefois, même si nous conservons la famille de fonctions proposée par Buus et al., la valeur de ce paramètre ne nous est d'aucune utilité. En effet, les auteurs supposent que la détection est effectuée par autant de *canaux indépendants* qu'il y a de composantes harmoniques dans le signal a détecter. Chaque harmonique excite donc un canal auditif et un seul. Pour nous, cette notion de canal auditif a une signification quelque peu différente, puisque ces canaux sont en fait les détecteurs de la membrane basilaire. Un harmonique unique excite donc en fait (à des degrés divers) un grand nombre de *canaux* à la fois.

Nous avons dû faire une autre modification (mineure) à la fonction proposée par Buus et al., puisque pour chaque canal, ces derniers travaillent à partir du rapport *signal masqué sur signal masqueur*, et que nous travaillons à partir de la différence de sensation basilaire $|S^*_i - S_i|$. Il y a en fait une relation univoque entre les 2 grandeurs puisque:

$$\left|S^*_i - S_i\right| = \left|\mathrm{Log}\left(E^*_i - \varepsilon_i\right) - \mathrm{Log}\left(E_i - \varepsilon_i\right)\right| \qquad (2)$$

Si on appelle **B** la densité basilaire d'énergie du signal masqué, on obtient :

$$\left|S^*_i - S_i\right| = \left|\mathrm{Log}\left(E_i + B_i - \varepsilon_i\right) - \mathrm{Log}\left(E_i - \varepsilon_i\right)\right|$$
$$\left|S^*_i - S_i\right| = \left|\mathrm{Log}\left(1 + \frac{B_i}{E_i + \varepsilon_i}\right)\right| \qquad (3)$$

Où $\dfrac{B_i}{E_i + \varepsilon_i}$ est la quantité utilisée par Buus et al.

2.2 Ajustement du paramètre pour la famille de fonctions proposées par Buus et al

Pour ajuster le paramètre de cette fonction *probabilité de détection versus différence de sensation*, il nous a paru plus simple d'étudier une expérience où la différence de sensations dues à chacun des 2 signaux (original et bruité) est la même pour chacun des 2500 détecteurs de notre modèle. Nous nous sommes donc basés sur une expérience de détection du taux de modulation d'un bruit à spectre blanc, décrite dans [1].

Pour cette expérience, on module lentement en amplitude un bruit à spectre blanc, et on mesure le taux de modulation juste perceptible, en fonction du niveau moyen du bruit.

Une façon alternative d'interpréter cette expérience est de la voir comme une expérience de masquage de bruit blanc, par du bruit blanc. On part avec un certain niveau de bruit blanc (le masqueur), correspondant au niveau le plus bas du bruit blanc modulé, puis on superpose à ce bruit un autre bruit blanc (le masqué), de même contenu spectral, de telle sorte que la somme des 2 bruits ait un niveau correspondant équivalent au niveau le plus haut du bruit blanc modulé. On est bien alors dans une situation de masquage de bruit blanc par du bruit blanc. On peut déduire le rapport du bruit masqué sur le bruit masquant au seuil de détection, à partir du taux de modulation juste perceptible. Bien entendu, comme le bruit est blanc, la sensation basilaire est différente d'un détecteur basilaire à l'autre, mais comme le spectre du signal masqué est identique à celui du signal masqueur, la différence de sensation entre les 2 signaux est la même pour tous les détecteurs.

La figure 3 donne les résultats de cette expérience (tirés de [1]). On voit que dès que le niveau est suffisant pour que le seuil de détection absolu n'intervienne plus, le taux de modulation juste audible se stabilise autour de 0.03. Celà correspond à une différence de sensation de 0.52 db sur chacun des 2500 détecteurs de la membrane basilaire. Pour cette valeur de différence de sensation, la probabilité de non-détection individuelle de chaque détecteur est $2500\sqrt{0.5}$. C'est de cette façon que l'on a ajusté le paramètre de la fonction de probabilité proposée par Buus et al.

A partir de cette valeur du paramètre, et à l'aide de notre modèle, nous avons simulé l'expérience de Buus et al. Le résultat obtenu est représenté figure 4, et il peut être comparé directement aux résultats de la figure 2. Clairement, les résultats obtenus

concordent remarquablement bien avec les résultats de Buus et al., tant pour les seuils de détection, que pour la pente des courbes de probabilités.

3 Performance du nouveau modèle

Dans les situations de masquage de sons harmoniques par du bruit à bande étroite ou large, le nouveau modèle donne des résultats très proches de ceux obtenus avec l'ancien modèle (qui étaient déjà excellents). Nous n'insisterons donc pas sur ces situations. La réelle amélioration concerne les résultats obtenus dans les situations de masquage de bruit à bande étroite ou large par des signaux harmoniques. Dans ces situations, l'utilisation du nouveau principe de détection, combinée à une description spectrale plus exacte des bruits masqués, conduit à l'obtention de seuils de masquage calculés, en très bonne accordance avec l'expérience (erreur de l'ordre de 1 db).

3.1 L'expérience de Schroeder et al.

La première expérience que nous avons simulé est due à Schroeder et al. [5].

Un masqueur harmonique de fréquence variable entre 500 et 2000 hz, et de niveau 80 db, masque un bruit tiers d'octave centré autour de 1 Khz. On mesure le niveau du bruit au seuil de détection en fonction de la fréquence du masqueur.

Les résultats obtenus par Schroeder et al. sont représentés à la figure 5.

3 simulations ont été faites, dont les résultats sont représentés figure 6. La 1^{ière} simulation (a) a été faite pour un bruit masqué tiers d'octave parfait (flancs de décroissance de pentes infinies). La 2^{ième} simulation (b) a été faite pour un bruit tiers d'octave présentant des pentes de décroissance de 100 db/octave. La 3^{ième} simulation (c) a été faite pour un spectre du bruit calculé d'après la documentation de filtres tiers d'octave Brüel et Kjaer. Ces filtres sont utilisés couramment en psycho-acoustique pour obtenir des bruit tiers d'octave; ce type de filtre présente des pentes de décroissance proches de 100 db/octave près des fréquences de coupures, puis ces pentes s'adoucissent et tendent vers 18 db/octave pour des fréquences plus éloignées des fréquences de coupures.

On note que la courbe 6.c (pour le bruit tiers d'octave réaliste) est très proche du résultat de Schroeder et al. On peut noter en particulier la valeur correcte du seuil de détection lorsque le masqueur est au centre du bruit tiers d'octave (à 1000 hz). A partir d'environ 1414 hz, le seuil de détection se stabilise. Il se stabilise autour de 5 db pour nos

simulations, et autour de 20 db pour les résultats de Schroeder et al. Pour notre modèle, et certainement aussi pour l'expérience réelle, cette stabilisation est due au fait que le masqueur est alors trop loin en haute fréquence pour influencer la détection du bruit. Le seuil de détection obtenu est donc dû essentiellement au seuil d'audition absolu autour de 1000 hz. Ce seuil est d'environ 5 db pour notre modèle, et il doit être autour de 20 db pour le sujet de l'expérience.

Les seuils de masquage lorsque le masqueur est au centre du bruit masqué (à 1000 hz) sont très dépendants de l'allure spectrale du bruit simulé, en particulier des valeurs des pentes de décroissance. Il est remarquable que les résultats obtenus pour un bruit ayant des décroissances aussi raides que 100 db/octave soient notablement différents de ceux obtenus pour un bruit parfait. La différence de seuils de détection entre le bruit tiers d'octave parfait et le bruit tiers d'octave réaliste dépasse 10 db! Comme on va le voir, ces résultats sont dus au fait que, lorsque le masqueur est au centre du bruit, la détection s'effectue par les *flancs* du bruit, en particulier le flanc basse fréquence où l'effet du masqueur est beaucoup moins prononcé. La manière dont décroit l'énergie du bruit sur ce flanc basse fréquence a donc une très grande importance pour la détection.

3.2 Expérience de Hellman

Nous avons ensuite simulé une expérience (plus exactement des parties d'une expérience) due à Hellman [4].

Dans un premier temps, on mesure les seuils d'audition absolus pour:

- Une raie harmonique à 1000 hz
- Un bruit à bande étroite (925-1080 hz)
- Un bruit d'octave (600-1200 hz)
- Un bruit à large bande (75-9600 hz)

Les résultats obtenus par Hellman sont :

- Raie harmonique à 1000 hz: seuil : 6 db
- Bruit 925-1080 hz : seuil : 6 db
- Bruit 600-1200 hz : seuil : 8 db
- Bruit 75-9600 hz : seuil : 15 db

Les résultats simulés sont :

- Raie harmonique à 1000 hz: seuil : 4 db

- Bruit 925-1080 hz : seuil : 5 db
- Bruit 600-1200 hz : seuil : 7 db
- Bruit 75-9600 hz : seuil : 5 db

Les résultats sont très proches des résultats de Hellman, mis à part pour le bruit à bande large. Toutefois, il faut noter que le seuil de détection pour ce bruit fait intervenir le seuil d'audition absolu dans toute la zone 75-9600 hz, et que celui-ci n'a pas été spécifié par Hellman, et peut différer notablement de celui de notre modèle.

On mesure ensuite les seuils de détection de 2 bruits (1280-1480 hz et 1350-1450 hz), masqués par une raie harmonique de niveau 70 db à 1400 hz.

Les 2 bruits simulés ont des pentes de décroissance de 100 db/octave.

Les seuils mesurés par Hellman sont :

- Bruit 1280-1480 hz : 46 db
- Bruit 1350-1450 hz : 50 db

Les seuils simulés sont :

- Bruit 1280-1480 hz : 48 db
- Bruit 1350-1450 hz : 53 db

Ils sont donc en bonne accordance avec les résultats expérimentaux. Encore une fois, l'allure de la décroissance spectrale des bruits a une grande importance ici, et peut facilement justifier l'écart entre les seuils mesurés et les seuils simulés.

4. Différences entre la détection d'une raie harmonique dans du bruit, et la détection de bruit masqué par une raie harmonique

La figure 7 montre les probabilités de détection pour chaque détecteur basilaire:

- D'une raie harmonique à 1000 hz masquée par un bruit tiers d'octave parfait, centré autour de la raie
- d'un bruit tiers d'octave parfait centré autour de 1000 hz, masqué par une raie harmonique à 1000 hz.

Ces probabilités basilaires de détection ont été calculées au seuil de masquage : lorsque la probabilité globale de détection est 0.5.

Comme on le voit sur ces figures, la détection s'effectue de manière très différente dans les 2 cas.

- Dans le cas d'une composante harmonique masquée dans du bruit, on voit que très peu de détecteurs (seulement ceux qui sont très proche de la tonie correspondant à la fréquence à détecter) participent à la détection. Pour cette raison, tout se passe en fait comme si la détection avait lieu dès que la différence de sensations basilaires $|S^*_i - S_i|$ dépassait un seuil (de l'ordre de 1 à 2 db), sur le détecteur "i", à la tonie correspondant à la fréquence à détecter. C'est pour cette raison que le modèle déterministe que l'on avait jusqu'à présent prédisait très bien le masquage d'une composante harmonique unique masquée par du bruit.

De plus, dans ce cas, c'est uniquement la densité basilaire originale d'énergie, dans cette zone très étroite autour de la tonie de la fréquence à détecter, qui détermine le seuil de détection de cette composante harmonique. Il est donc possible de déterminer le seuil de masquage en fonction de la fréquence, pour un signal masqué harmonique, en reportant simplement dans l'espace fréquentiel la densité basilaire d'énergie, et en corrigeant pour l'atténuation sélective de l'oreille moyenne (c.f. [8]).

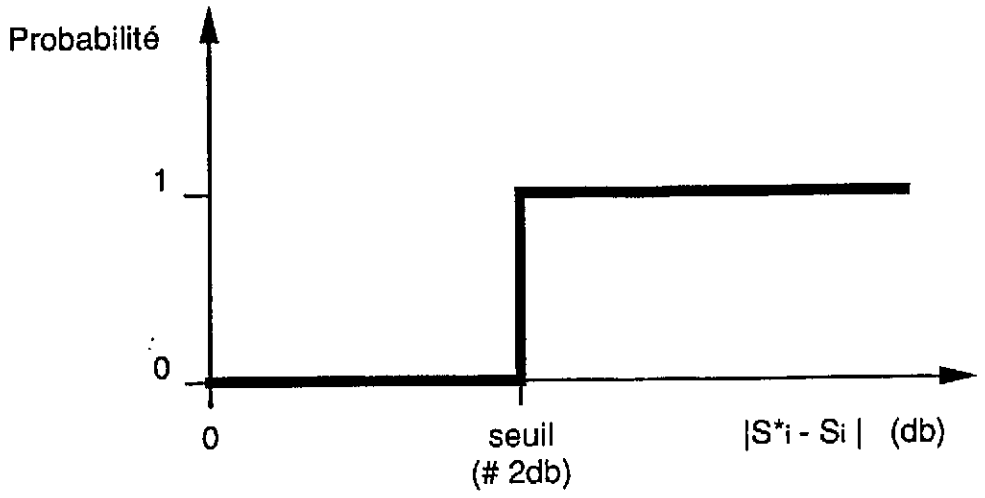
- Dans le cas d'un bruit tiers d'octave masqué par une raie harmonique, un nombre beaucoup plus important de détecteurs intervient dans la détection. Les probabilités individuelles de détection de ces détecteurs peuvent donc être beaucoup plus faibles. Contrairement au cas de la détection d'un son harmonique dans du bruit, la détection n'intervient pas essentiellement au lieu basilaire où l'énergie du signal masqué est maximum. Elle intervient ici sur les flancs du signal masqué (particulièrement le flanc basse fréquence), à un endroit où l'effet du masqueur diminue rapidement par rapport à l'effet du bruit masqué.

5. Conclusion

En conclusion, on voit que l'addition au modèle d'audition existant, de ce principe de détection statistique, permet de conserver les très bons résultats que nous avions dans le cas de sons harmoniques masqués dans du bruit. De plus, il permet de prédire avec une très bonne précision les seuils de masquage de bruits masqués par des signaux

harmoniques. Il semble donc que, armés de ce modèle, et contrairement aux solutions adoptées par d'autres laboratoires tels les laboratoires Bell, il soit inutile de tenir compte explicitement de la nature spectrale (harmonique ou non) du signal original, pour déterminer le seuil d'audition absolu. Le modèle en tient compte de lui même, implicitement.

Bien entendu, un travail important reste à faire pour valider ces résultats, mais une première retombée de ce travail devrait être la mise en oeuvre d'un logiciel d'évaluation objective de la qualité de codage : PERCEVAL_3. Ce logiciel calculerait la probabilité de détection du bruit de codage entre un signal original et un signal codé, en fonction du temps. Ce calcul serait effectué à partir des différences entre les sensations basilaires des 2 signaux, obtenues de la même façon que pour PERCEVAL_2. Ce nouveau logiciel serait particulièrement intéressant pour évaluer des signaux de haute qualité. Pour des signaux de mauvaise qualité, en effet, la probabilité de détection du bruit serait sans doute constament très proche de 1.



a) ancien modèle

b) nouveau modèle

Figure - 1- Probabilités de détection en fonction de la différence entre 2 sensations basilaires, pour une cellule détectrice.

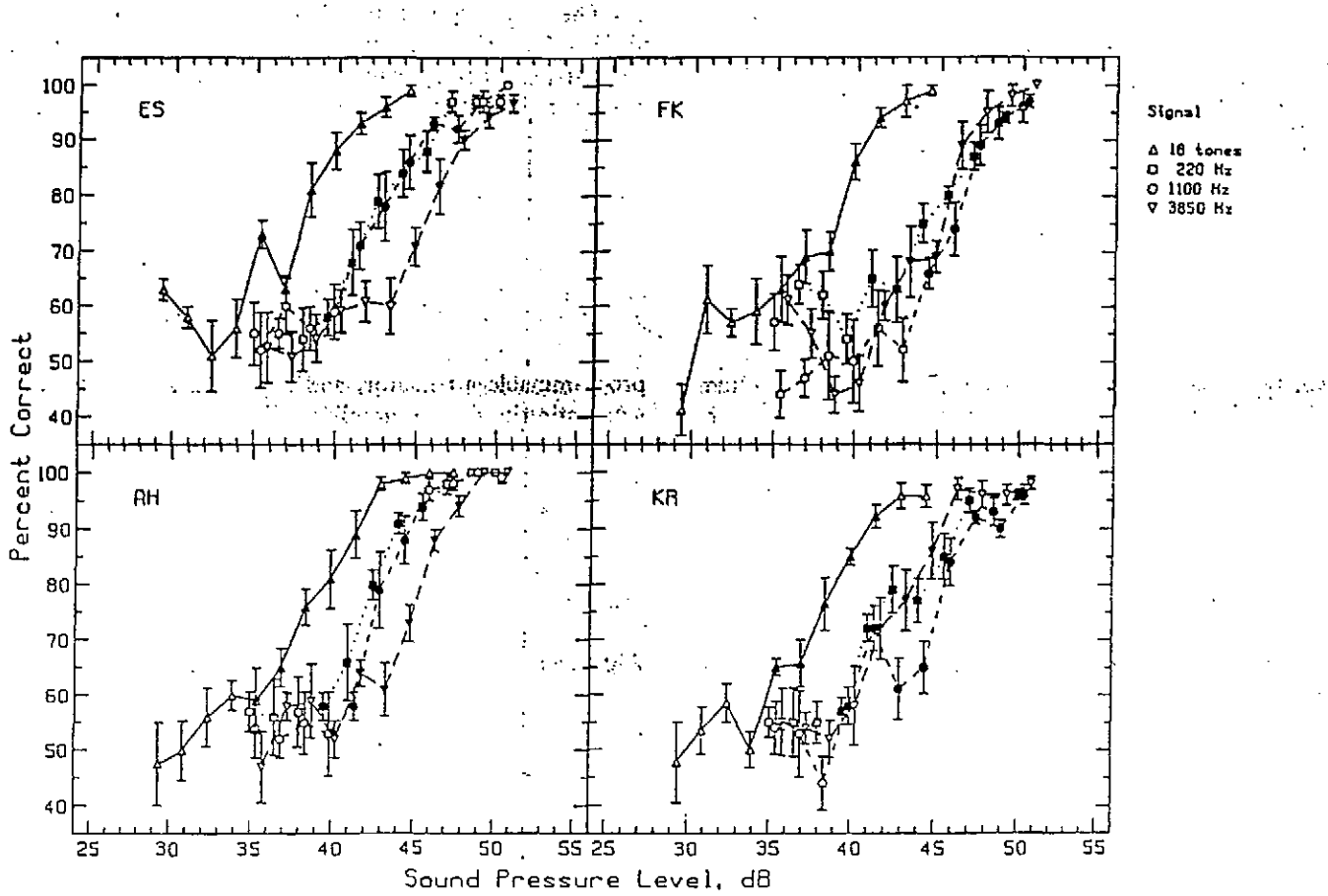


FIG. 1. Psychometric functions for pure tones at 220 Hz (□), 1100 Hz (○), 3850 Hz (▽), and an 18-tone complex (△). Percent correct obtained in a 2I, 2AFC task is plotted as a function of dB SPL per tone. Each panel shows results for one listener. The filled symbols are those included in the line fits used to summarize the data.

Figure 2 Résultats de Buus et al. (tirés de [7]

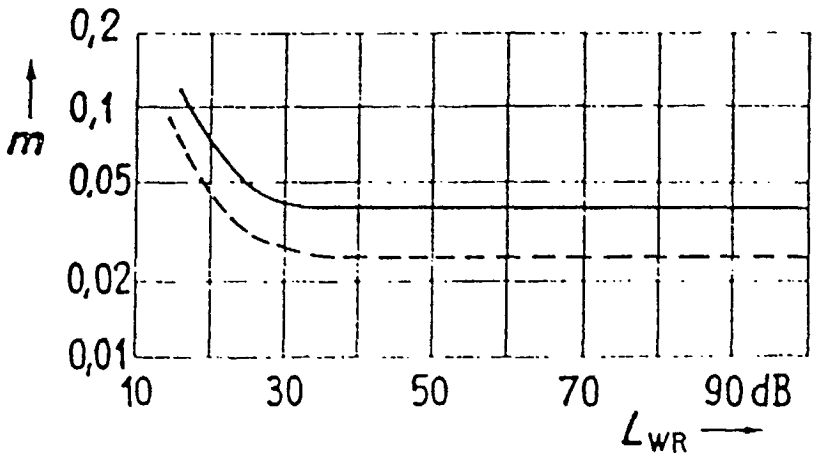


Figure 3 Taux de modulation à peine audible d'un bruit blanc, en fonction du niveau moyen.

Courbe en trait plein : modulation sinusoïdale. Courbe en pointillés : modulation carrée. Fréquence de modulation : 4 hz

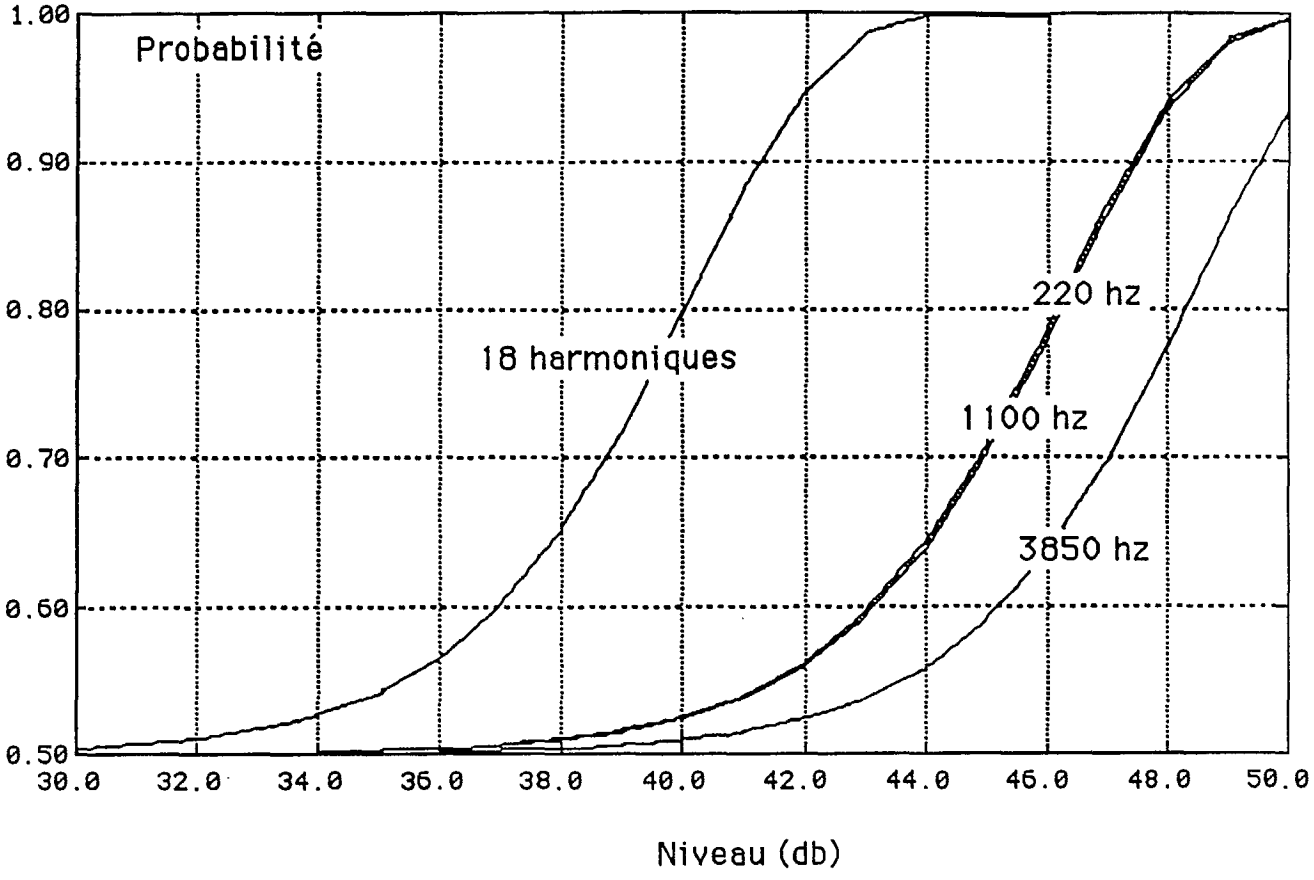


Figure -4- Simulation de l'expérience de Buus et al.
Probabilités de détection en fonction du niveau pour 3 signaux harmoniques pris isolément, et pour un signal complexe constitué de 18 harmoniques. Le masqueur est un bruit uniformément masquant de densité spectrale d'énergie 25 db autour de 1100 hz.

Note : Les résultats sont à comparer directement à ceux présentés à la figure 2.

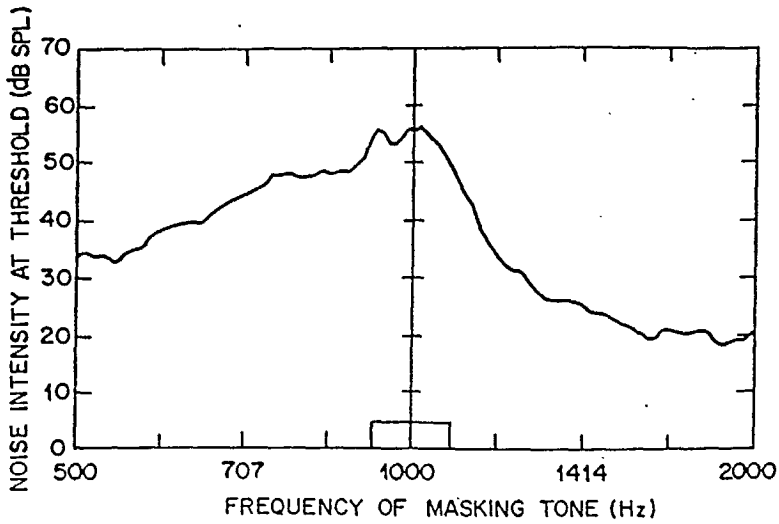


FIG. 1. Auditory threshold for a critical-band noise burst centered at 1 kHz masked by a tone of intensity 80 dB SPL. The frequency band occupied by the noise is indicated by the rectangular shaded area. Note that for a tone frequency of 1 kHz the noise intensity at threshold is 24 dB below the tone intensity. The masked threshold drops more steeply when the tone frequency is raised than when it is lowered, corresponding to the usual frequency asymmetry of auditory masking. Subject: JLH.

Figure 5 Résultats de Schroeder et al. (tirés de [5])

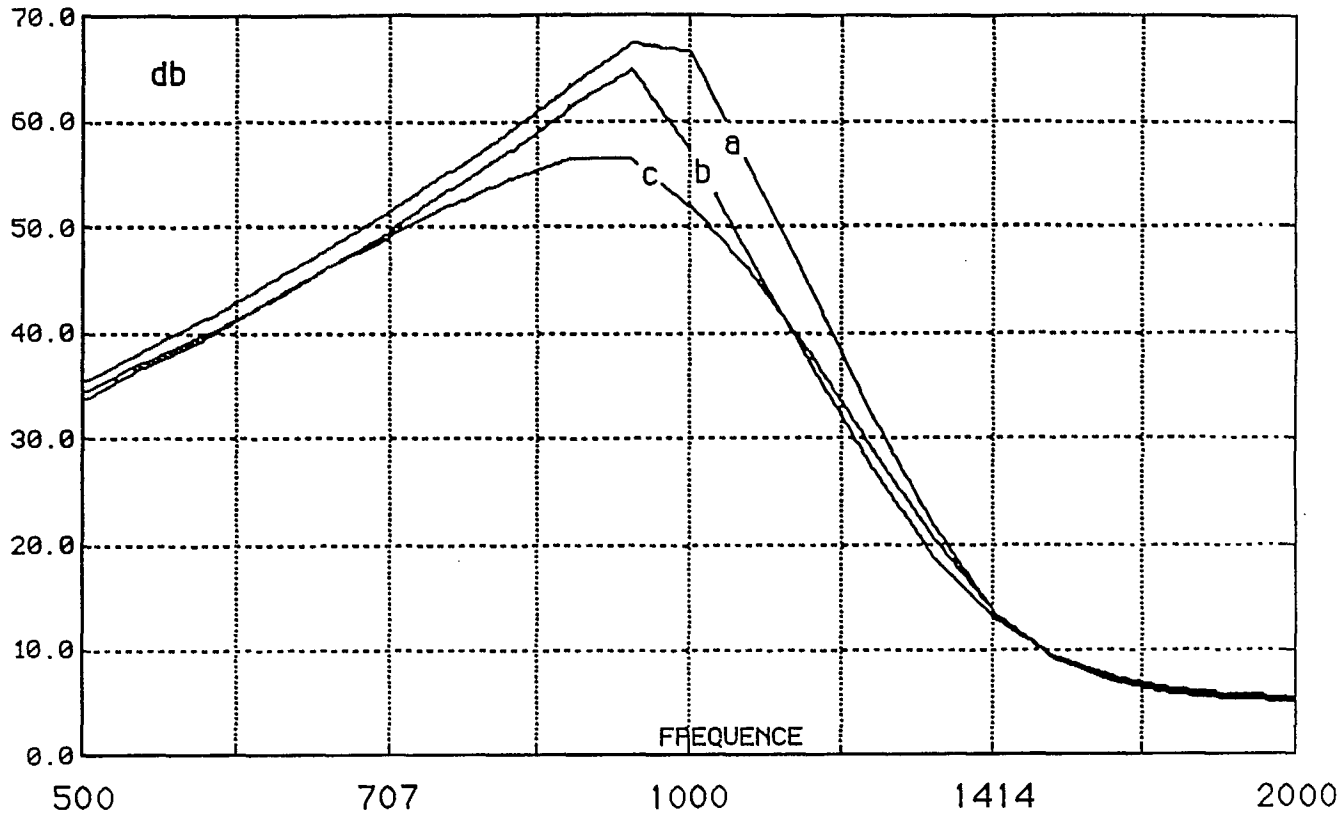


Figure -6- Simulation de l'expérience de Schroeder et al.

a) Pour un bruit tiers d'octave parfait
b) Pour un bruit tiers d'octave ayant des décroissances de 100 db/octave
c) Pour un bruit tiers d'octave réaliste (spectre calculé d'après une documentation Brüel et Kjaer)

Note : Ces résulats sont à comparer directement avec le résultat de Schroeder et al. (figure 5)

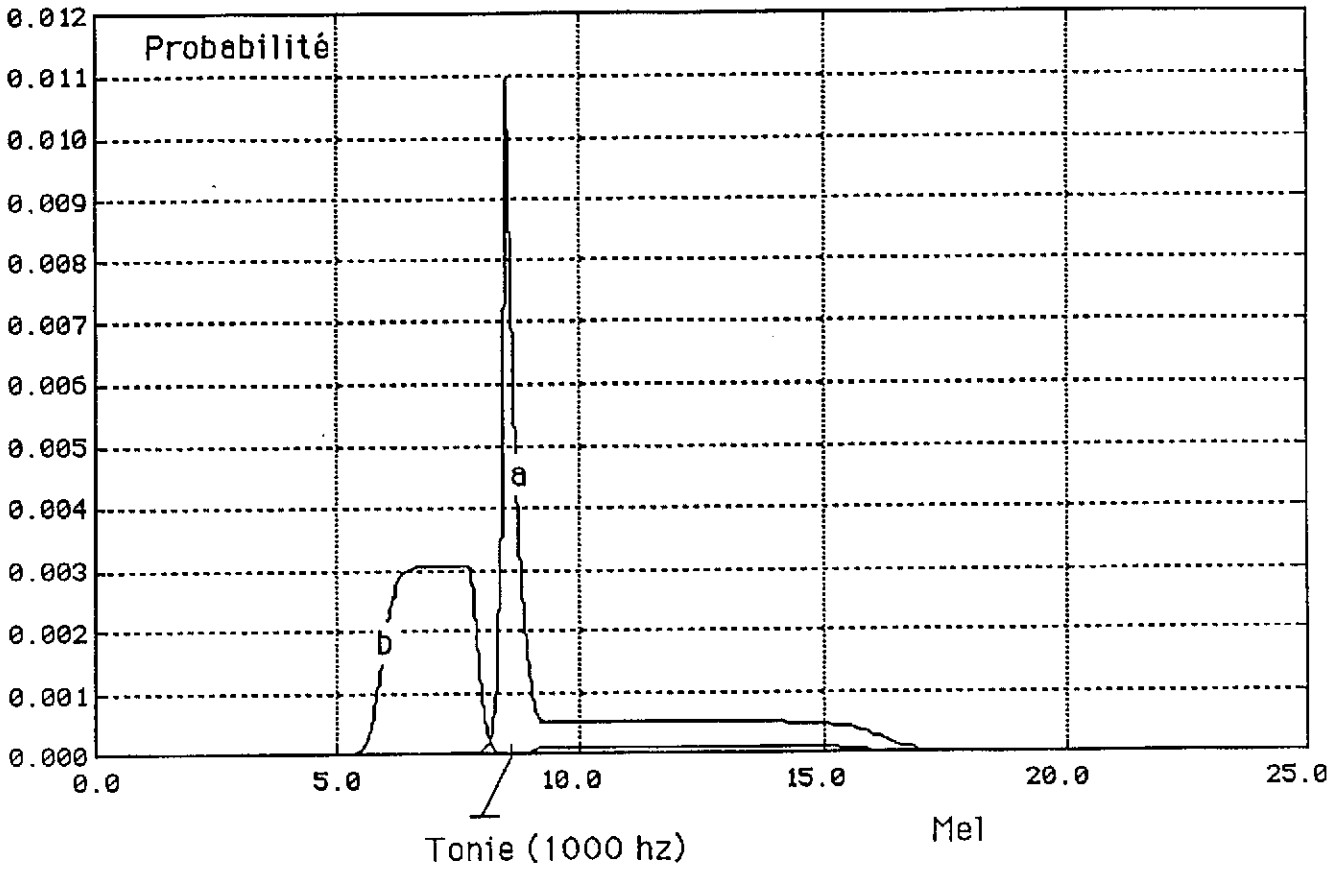


Figure 7 : Probabilités de détection en fonction du lieu basilaire au seuil de détection (lorsque la probabilité globale de détection est égale à 0.5).

a) Pour une harmonique à 1000 hz masquée par un bruit tiers d'octave centré autour de l'harmonique

b) Pour un bruit tiers d'octave centré autour de 1000 hz masqué par une harmonique à 1000 hz

Remerciements

Pour terminer, j'aimerais remercier Messieurs Raymond Hétu et Hung Tran Quoc du groupe d'acoustique de l'université de montréal (GAUM), pour les discussions très fructueuses que j'ai pu avoir avec eux, ainsi que la compétance technique inestimable et les documents qu'ils ont bien voulu mettre à ma disposition.

Références

[1] E. Zwicker, R. Feldtkeller - *Psychoacoustique, l'oreille récepteur d'information* - traduit de l'allemand par Christel Sorin - 1981 - collection technique et scientifique des télécommunications - MASSON - ISBN : 2-225-74503-X

[2] B. Paillard - *Description du modèle d'audition "OREILLE"* - CRCS rapport interne - mars 1990.

[3] B. Paillard - *Critique des bandes critiques* - CRCS rapport interne - octobre 1989.

[4] R. P. Hellman - *Asymetry of masking between noise and tone* - Perception and Psychophysics, 1972, vol 11 (3)

[5] M. R. Schroeder, B. S. Atal, J. L. Hall - *Optimizing digital speech coders by exploiting masking properties of the human ear* - Journal of the acoustical society of america 66(6), Dec. 1979

[6] J. D. Johnston - *Transform coding of audio signals using perceptual noise criteria* - IEEE journal on selected areas in communications, Vol. 6, No 2, February 1988

[7] S. Buus, E. Schorer, M. Florentine, E. Zwicker - *Decision rules in detection of simple and complex tones* - Journal of the acoustical society of america, 80(6), December 1986

[8] B. Paillard - *Mesure de l'importance d'une sous bande et attribution des bits* - CRCS rapport interne - septembre 1989