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

PHASE IV

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 IV

Rapport final

Ministère des Communications du Canada
Ottawa - Canada

dans le cadre du

programme de Centres d'excellence

Contrat no. 36100-0-5022

Rédigé par Bruno Paillard

Sommaire: Les sections 1 à 4 font état des travaux effectués par le Centre de Recherche sur les Communications de l'Université de Sherbrooke suivant les 3 voies de recherches explorées: modélisation de l'audition, évaluation perceptuelle, codage

Responsables du projet:

Philippe Mabilleau
Sarto Morissette, Dir. CRCS

Sherbrooke, Québec

Le 20 mars 1991

TABLE DES MATIÈRES

1. Introduction

Le présent rapport fait le point sur l'avancement des travaux au CRCS. Rappelons que trois orientations principales ont été définies:

- Psychoacoustique: *développement et mise au point du modèle d'audition* " OREILLE "

  OREILLE constitue le coeur du logiciel d'évaluation perceptuelle " PERCEVAL ". Il sert aussi d'outil de développement et d'évaluation d'algorithmes de codage.

- Évaluation subjective de codeurs: *développement et mise au point des logiciels d'évaluation perceptuelle* " PERCEVAL "

  Ces logiciels, construits autour du modèle d'audition " OREILLE " visent à évaluer la qualité perçue des signaux codés. Cette évaluation se fait suivant plusieurs critères: probabilité de détection du bruit, importance de la dégradation, etc... Les applications particulières de ces logiciels sont multiples: elles sont discutées plus loin dans le document. Parmi elles figurent:

  - aide au développement de codeurs,
  - identification de séquences critiques.

- Codage: *développement de stratégies de codage et évaluation de ces stratégies au sein d'un codeur*

  Les stratégies de codage sont développées à l'aide de prédictions faites et vérifiées par le modèle d'audition OREILLE. Ces stratégies sont ensuite testées pour être finalement incorporées dans un codeur.

2. Où en est OREILLE?

A l'heure actuelle, le modèle d'audition OREILLE semble relativement bien défini et stabilisé. Les réglages des différents paramètres sont satisfaisants et le comportement global de OREILLE correspond au comportement observé d'une oreille réelle, avec une très bonne précision, pour toutes les expériences qui ont été simulées.

Nous passons donc d'une étape de développement de OREILLE à une étape d'exploitation.

A Noter: Récemment le problème de l'inversion de la transformation linéaire *énergie fréquentielle* → *énergie basilaire* à été résolu. Il est dorénavant possible d'obtenir un spectre d'énergie d'entrée à partir d'une sensation basilaire donnée. Cette transformation inverse n'a pas d'application immédiate; elle pourrait toutefois s'avérer intéressante pour obtenir des indices sur les solutions théoriques de certains problèmes, *quelle est la meilleure forme spectrale à donner au bruit, par exemple?*

D'autre part, les domaines d'application de OREILLE ne sont pas limités au codage. On pourrait penser par exemple à l'utiliser en reconnaissance de parole. Afin d'étudier ces applications, le logiciel OREILLE a été mis à la disposition de l'équipe de Jean Rouat à l'Université du Québec à Chicoutimi.

3. Où en est PERCEVAL?

A l'heure actuelle deux versions de PERCEVAL (PERCEVAL-3, c.f. Annexe 1, et PERCEVAL-4) ont été finalisées. Après une prévalidation ici au CRCS, ces deux versions sont en cours de validation au CRC à Shirley Bay.

Ces deux versions donnent des résultats sous des formes différentes et sont donc adaptées à différentes applications.

3.1 Logiciel d'évaluation perceptuelle PERCEVAL-3

Ce logiciel compare une séquence musicale originale et une séquence musicale dégradée, et donne en sortie la probabilité de détection de la dégradation, en fonction du temps (c.f. Annexe 1). A cause du type de résultats obtenus (probabilité de détection), PERCEVAL-3 est bien adapté pour travailler sur des signaux de très haute qualité (par exemple pour évaluer la transparence des codeurs). Si les signaux sont de qualité un peu moins bonne, la probabilité de détection du bruit est presque constamment autour de 100%, ce qui ne fournit pas une information très riche sur le bruit (à quel moment apparaît-il, dans quelles bandes de fréquence, quelle est son importance? ...)

Par contre, lorsqu'un codeur est proche de la transparence, les variations temporelles de la probabilité de détection du bruit fournissent une information beaucoup plus intéressante (le bruit est-il perçu pendant les décroissances des notes, pendant les attaques, pendant les silences? ...). Cette richesse d'information fait de PERCEVAL-3 un outil bien adapté pour faire du développement de codeur. A noter que pour ce type d'applications, il est possible d'obtenir une information concernant la localisation fréquentielle du bruit perçu.

En résumé, PERCEVAL-3 est bien adapté pour travailler sur des codeurs proches de la transparence, autant comme outil de développement de codeurs que comme outil d'évaluation de la qualité.

3.2 PERCEVAL-4

La nécessité de répondre à d'autres besoins identifiés par le CRC nous a poussés à développer une nouvelle version de PERCEVAL: " PERCEVAL-4 ". Ces besoins sont:

- nécessité de pouvoir évaluer des codeurs qui sont en deçà de la transparence (pour lesquels le bruit, bien que très léger, est perçu à coup sûr);

- possibilité d'effectuer une assistance aux écoutes subjectives et à l'identification de séquences critiques. Ce qu'on attend d'un logiciel dans ce cas là, est la possibilité de dégrossir les procédures d'écoutes subjectives, en fournissant avant les tests, un indice de la qualité des fichiers. La procédure d'écoutes subjectives proprement dite pourrait alors se limiter aux meilleures séquences, ou bien aux séquences que les codeurs ont le plus de mal à coder par exemple.

  Pour ce type d'applications, une information plus condensée est nécessaire. Cette information est donnée sous la forme d'un chiffre unique, représentant la qualité de la séquence musicale dans son ensemble.

PERCEVAL-3 a tout d'abord été modifié pour donner en sortie une grandeur représentant l'importance perçue du bruit en fonction du temps (à la place de la probabilité de détection).

Cette grandeur est ensuite pondérée par la probabilité de détection du bruit en fonction du temps, puis moyennée sur la longueur de la séquence.

Le fait de pondérer l'importance de la dégradation par sa probabilité de détection (entre 0 et 1) permet de minimiser dans la moyenne la contribution de bruits très faibles (non audibles) qui pourraient être présents sur des durées importantes.

Le type de résultat fourni par PERCEVAL-4 (une grandeur représentant l'importance de la dégradation) est beaucoup moins objectif que la probabilité de détection qui est fournie par PERCEVAL-3. La validation de PERCEVAL-4 sera donc sans doute plus délicate que celle de PERCEVAL-3.

Par contre les différences entre PERCEVAL-3 et PERCEVAL-4 sont très mineures, et donc la validation précise de PERCEVAL-3 apportera un certain niveau de confiance concernant l'efficacité de PERCEVAL-4.

3.3 Quelques difficultés d'utilisation de PERCEVAL

3.3.1 Synchronisation

A la suite de tests de validation effectués par le CRC sur des signaux de musique ayant été codés en temps réel, il est apparu que les logiciels PERCEVAL sont très sensibles à la synchronisation précise des séquences originales et codées. Des différences de synchronisation aussi faibles que un échantillon entre "original" et "codé" ont en effet un impact important sur la probabilité de détection (PERCEVAL-3) ou sur l'importance perçue de la dégradation (PERCEVAL-4).

Cette grande sensibilité à la synchronisation est due au fait que dans les premières étapes de l'algorithme, on calcule le signal de bruit par différence entre "original" et "codé". Cette différence, s'effectuant dans le domaine temporel, est bien entendue très dépendante de la synchronisation des signaux.

Des versions de PERCEVAL ne nécessitant pas le calcul explicite du signal de bruit ont été développées. Ces versions sont bien moins sensibles à la synchronisation que les précédentes, mais elles restent trop sensibles, en particulier lorsqu'on travaille avec des bruits à la limite de l'audibilité.

PERCEVAL reste un logiciel difficile à utiliser si on ne dispose pas des signaux originaux et codés sous forme de fichiers. En particulier, une utilisation sur des séquences musicales analogiques est très délicate.

3.3.2 Niveau de reproduction virtuel

Que ce soit en simulation ou en expérimentation réelle, le niveau d'écoute d'une séquence musicale a une grande influence sur le bruit qui est perçu ou non. En particulier, tous les bruits qui sont très peu (ou pas) masqués, deviennent audibles dès que le niveau de reproduction devient assez fort pour qu'ils soient au-dessus du seuil d'audition absolu.

De même qu'un système d'écoute a un réglage de niveau de reproduction, PERCEVAL a un réglage virtuel du niveau de reproduction. Ce réglage correspond simplement à un gain multiplicatif réglable, appliqué sur les fichiers d'entrée.

De manière à assurer que le comportement de PERCEVAL soit identique au comportement d'une oreille réelle, il convient de vérifier que le réglage du niveau d'écoute virtuel est identique au réglage physique du niveau d'écoute du système de reproduction. De cette façon, l'énergie acoustique d'excitation déduite du fichier musical par PERCEVAL est identique à l'énergie acoustique physiquement présente à l'oreille de l'auditeur.

Ce réglage est délicat, mais il a été simplifié par la mise au point d'une procédure de réglage simple et claire.

Sans entrer dans les détails, disons simplement que cette procédure a 2 étapes:

- d'une part l'auditeur écoute des sinusoïdes de très faible niveau sur le système d'écoute réel. Il doit s'agir d''un auditeur ayant une audition, et en particulier un seuil d'audition absolu, parfaitement normal.

  L'auditeur essaie alors d'ajuster le réglage du niveau de telle façon que les sinusoïdes soient à la limite de l'audibilité;

- d'un autre côté, ces sinusoïdes sont présentées à PERCEVAL comme "signal bruité", le signal original étant un silence.

  L'opérateur essaie alors d'ajuster le niveau virtuel de reproduction de PERCEVAL de telle manière que la probabilité de détection de la sinusoïde soit autour de 50%.

  A l'issue de ces 2 étapes, on dispose d'une référence de niveau commune aux deux systèmes.

4. Où en est le codage?

4.1 Instrumentation

Pour effectuer les enregistrements et les écoutes de séquences musicales, le CRCS s'est doté d'une carte de digitalisation et de reproduction de très haute qualité. Cette carte permet l'enregistrement et l'écoute à 32, 44.1 et 48 kHz de fréquence d'échantillonnage et les convertisseurs ont une résolution de 16 bits. Le rapport signal/bruit à la conversion est supérieur à 90 dB.

Le CRCS a ensuite développé les logiciels qui permettent d'utiliser cette carte facilement. Les essais comparatifs qui ont été effectués à la fréquence d'échantillonnage de 44.1 kHz (seule fréquence possible pour notre précédent système d'écoute) indiquent nettement la qualité supérieure de cette carte par rapport au système SONY PCMF1 ou 701 utilisé jusqu'alors. En particulier certains artefacts sur les sinusoïdes à fréquences variables étaient audibles avec le précédent système, et ne le sont plus du tout avec le nouveau.

De manière à accélérer le développement du codeur, le CRCS s'est aussi doté d'une carte de traitement de signal construite autour du processeur point flottant TMS320C30 de Texas Instruments.

Le CRCS a développé les outils logiciels pour effectuer l'interface entre la carte et la machine hôte (Macintosh II) et est actuellement en train de développer sur la carte les fonctions principales du codeur envisagé (décomposition temps/fréquence, certaines fonctions du modèle d'audition, etc...)

4.2 Algorithmes

Au plan du codage proprement dit, nous nous sommes concentrés sur le développement de seuils de masquages fréquentiels plus intéressants que ceux qui sont utilisés habituellement dans ce domaine.

Ces seuils de masquage ont un double intérêt:

- d'une part ils sont tels que si le spectre du bruit de quantification suit leur forme, un maximum d'énergie de bruit peut être injecté dans le signal, sans être audible;

- ensuite, il est assuré que si le bruit de quantification est sous le seuil de masquage en tous points du spectre, il sera inaudible.

Par contre, il n'est pas assuré que si le bruit de quantification dépasse le seuil de masquage en un ou plusieurs points du spectre, il sera audible. En pratique, il y aura donc 2 approches à l'utilisation de ces seuils:

- d'une part, on essaiera d'attribuer les bits suivant les composantes fréquentielles de telle façon que le bruit de quantification suive ce seuil. Cela permettra d'injecter le maximum de bruit dans le signal, donc d'utiliser globalement le moins de bits possible pour coder le signal, sans que ce bruit soit audible;

  ensuite, on pourra détecter les composantes fréquentielles du signal original dont l'énergie est sous le seuil de masquage, et s'abstenir de les coder purement et simplement.

4.3 Un premier seuil

Dans un premier temps nous avons évalué un seuil de masquage très simple: le seuil de masquage est simplement égal au spectre du signal, décalé de -13 dB.

Les justifications de principe qui prédisent que ce seuil très simple est efficace (en fait plus efficace que le seuil de masquage pour fréquences pures utilisé habituellement) sont exposées à l'Annexe 2.

Notons que pour utiliser pleinement l'efficacité de ce seuil et injecter le maximum de bruit dans le signal, il est nécessaire de contrôler finement le spectre du bruit; autrement on pourra se contenter de faire en sorte que le spectre de bruit est 13 dB sous le spectre du signal, en tous points du spectre.

4.3.1 Simulation

La validité des principes exposés dans l'Annexe 2 peut être vérifiée par simulation:

- on définit un signal original ayant le spectre décrit en figure 1;

- on définit un bruit ayant le spectre décrit en figure 2 (égal au spectre du signal original décalé de -13 dB);

- on calcule par " oreille " la probabilité de détection du bruit tout au long de la membrane basilaire (figure 3);

  on voit que cette probabilité de détection est très uniforme le long de la membrane basilaire, ce qui vérifie parfaitement les prédictions faites à l'Annexe 2;

  Note: le fait que la probabilité de détection soit uniforme le long de la membrane basilaire indique en fait que le bruit sera détecté en même temps dans toutes les zones spectrales. Aucune composante fréquentielle n'a donc été négligée par rapport aux autres.

4.3.2 Essai en vraie grandeur

Afin de vérifier pratiquement ces résultats, une expérience a été mise au point et conduite. Cette expérience, dont les résultats sont décrits dans l'Annexe 3 consistait en:

- décomposer des séquences musicales en 1024 composantes spectrales par MLT;

- ajouter du bruit aux composantes spectrales suivant trois stratégies et avec un rapport signal sur bruit variable:

  - bruit blanc,

  - spectre de bruit suivant le seuil de masquage pour fréquences pures utilisé habituellement,

  - spectre de bruit suivant le spectre du signal.

- reconstruire le signal temporel;

- écouter les séquences musicales bruitées et noter pour chacune des stratégies le rapport signal/bruit tel que le bruit est à la limite de l'audibilité.

Cette expérience a montré clairement que la stratégie qui consiste à donner au spectre de bruit la forme du spectre du signal permet effectivement d'injecter plus de bruit dans le signal.

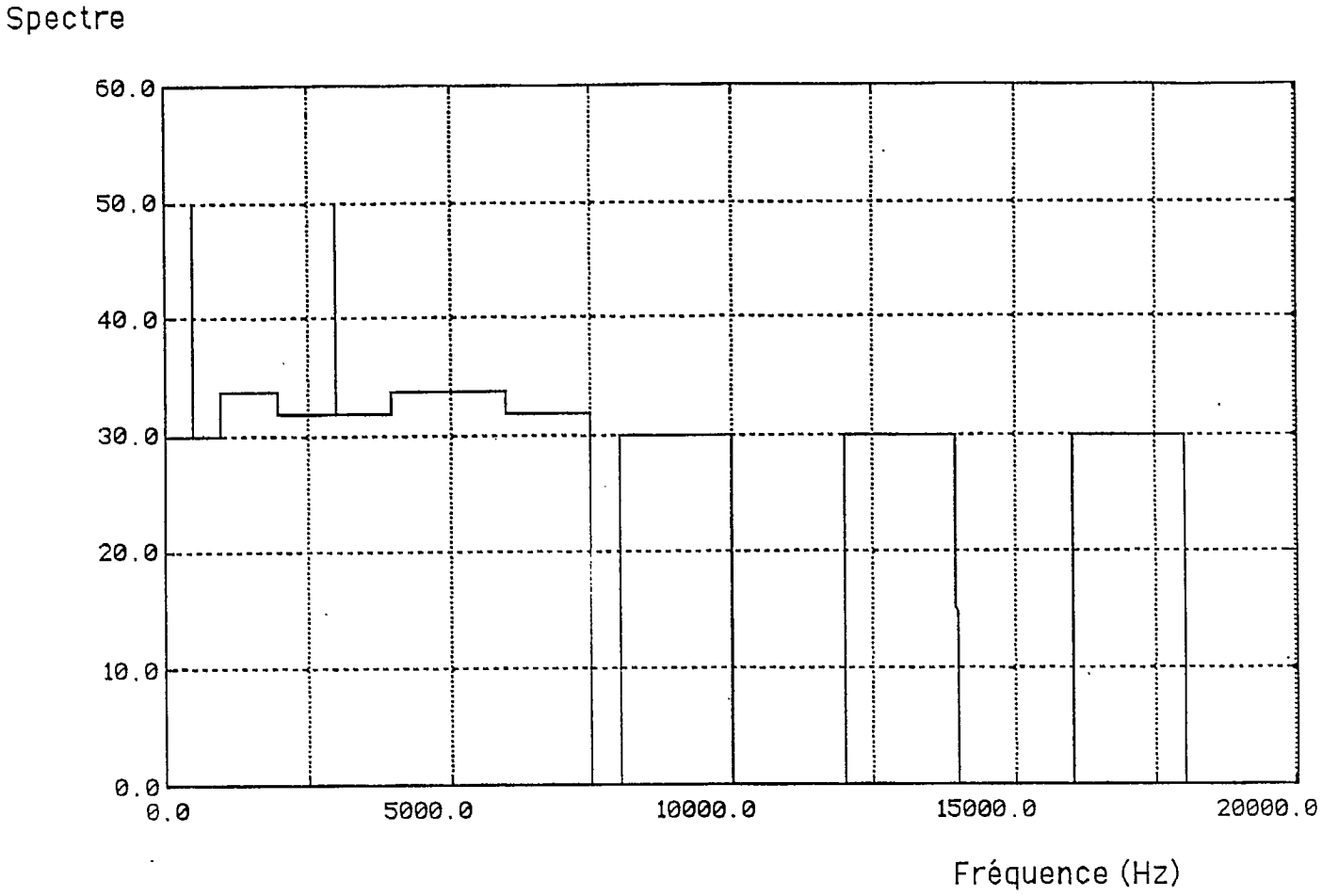


Figure 1 Spectre du signal original

Spectre

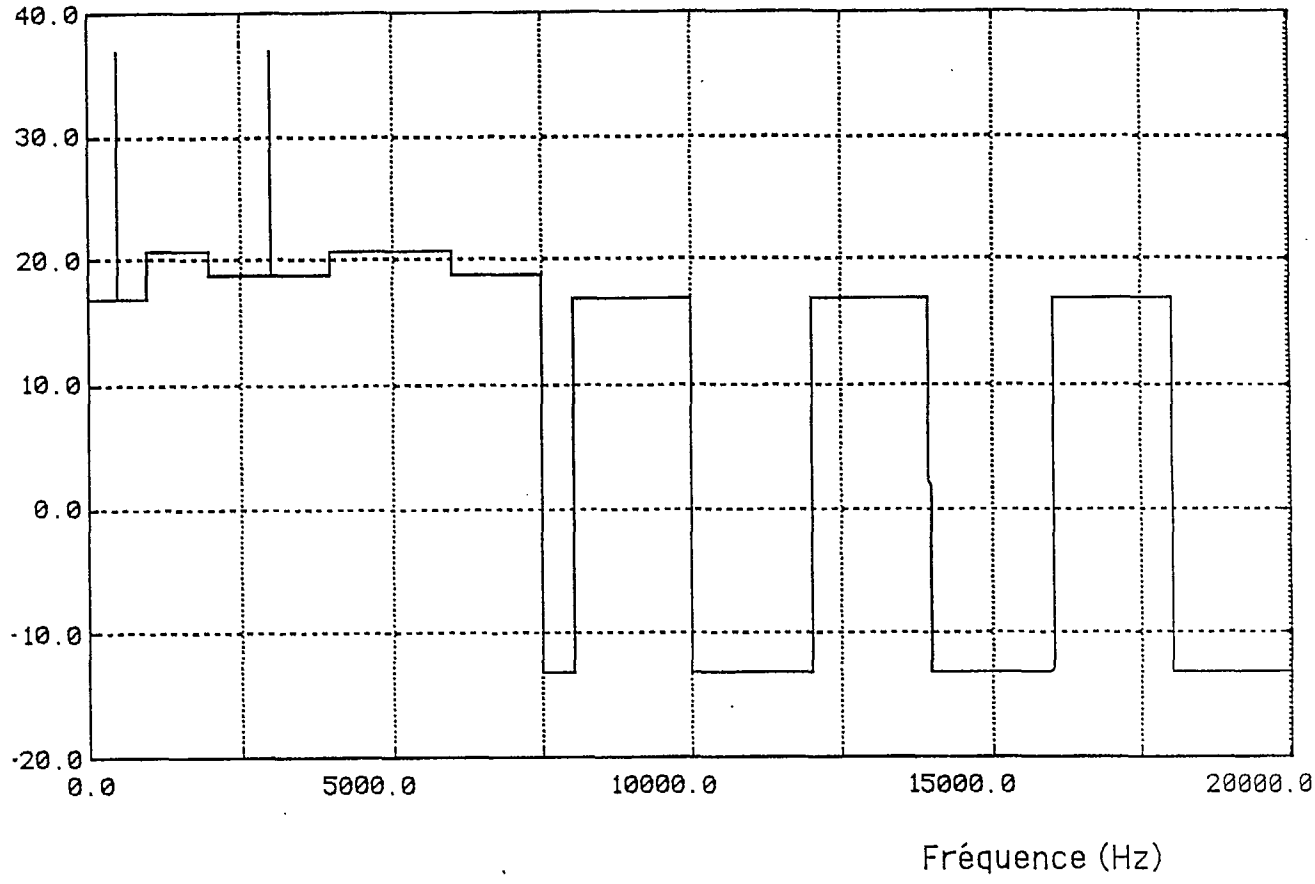


Figure 2 Spectre du bruit

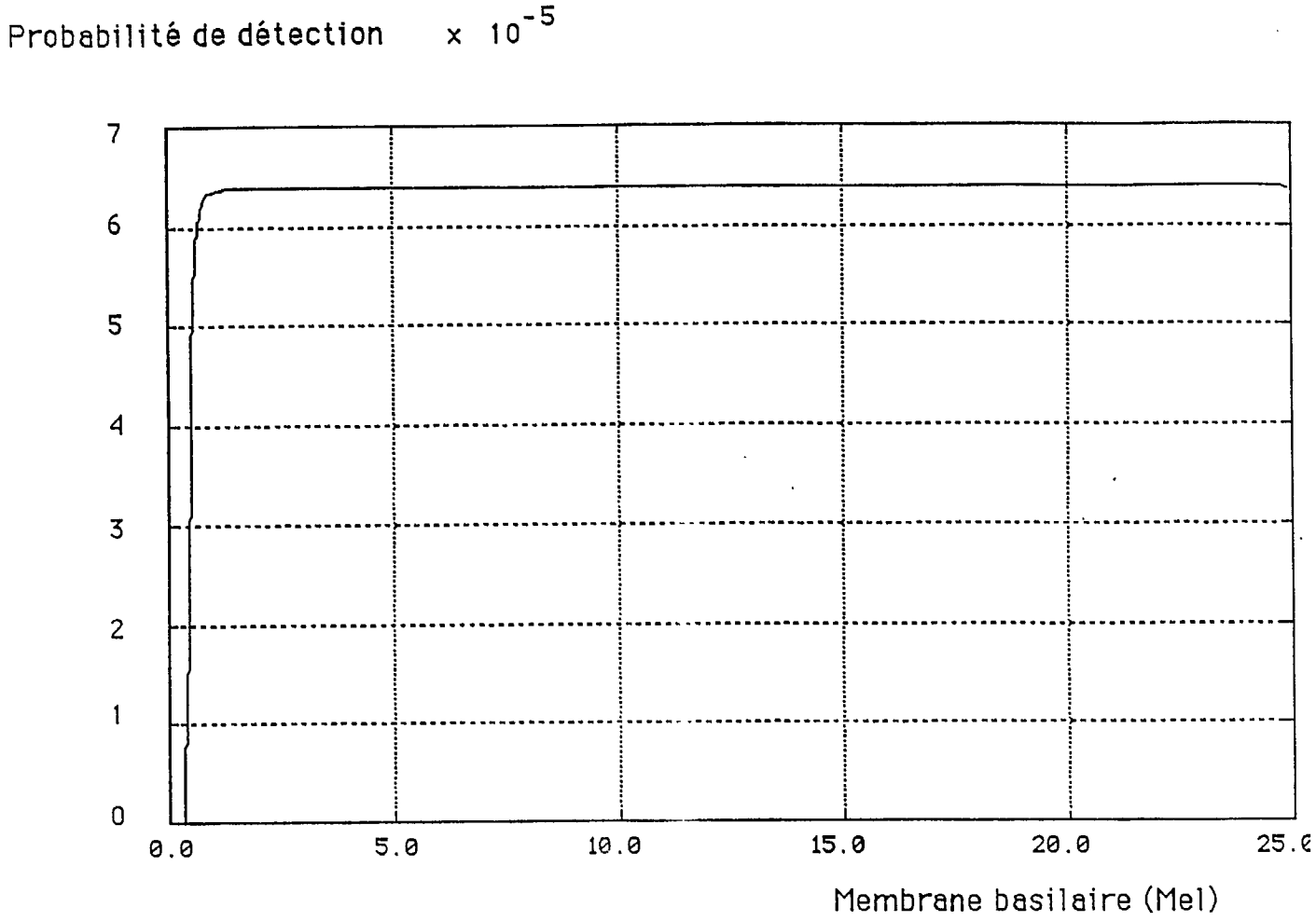


Figure 3 Probabilité de détection basilaire

4.4 Nouveau seuil de masquage

A cause de l'approximation décrite à la page 6 (ligne 32) de l'Annexe 2, le seuil de masquage suivant le spectre du signal est un seuil très conservateur. Il est en fait possible d'injecter beaucoup plus de bruit en haute fréquence que ce que ce seuil permet.

Un nouveau seuil, conservant les caractéristiques de ce dernier, mais permettant d'injecter plus de bruit en haute fréquence, peut être obtenu simplement en ajoutant au spectre d'énergie du signal, décalé de -13 dB, un seuil constant remontant en haute fréquence pour tenir compte de l'atténuation et de la moins bonne résolution spectrale de l'oreille en haute fréquence.

4.4.1 Simulation

- définit un signal original ayant le spectre décrit à la figure 4;
- on définit un bruit suivant le seuil de masquage correspondant, tel que décrit à la figure 5;

  Noter la nette remontée du seuil de masquage en haute fréquence;
- on calcule par " oreille " la probabilité de détection du bruit le long de la membrane basilaire (figure 6);
- on voit que cette probabilité de détection est très proche de la probabilité de détection obtenue avec le seuil suivant simplement le spectre du signal (figure 3);
- l'addition de cette importante énergie de bruit en haute fréquence a donc eu un effet négligeable sur la probabilité de détection; ceci est dû au fait que la remontée du seuil en haute fréquence est en fait très contrôlée, dépendant à la fois de l'atténuation et de la mauvaise résolution fréquentielle de l'oreille.

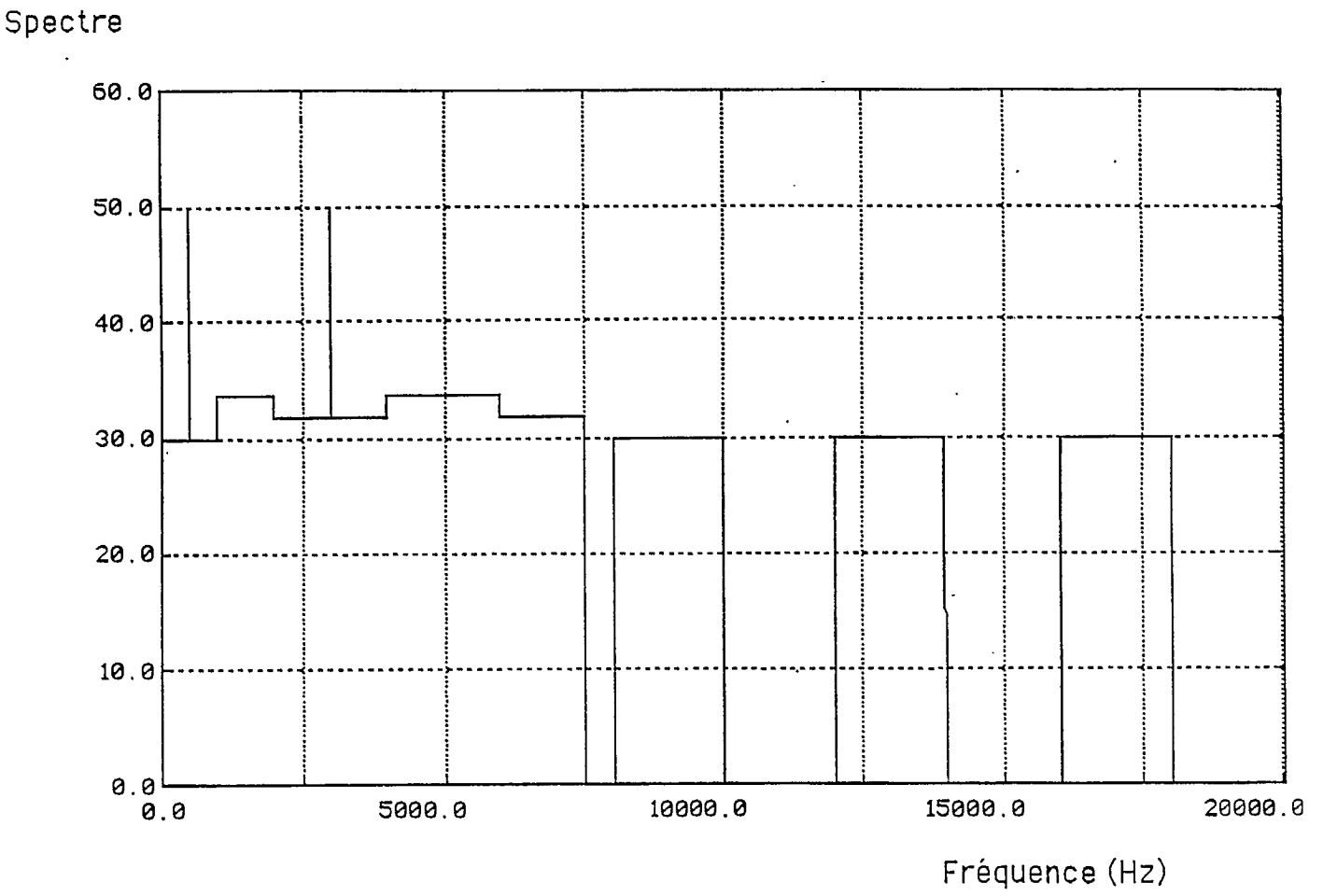


Figure 4 Spectre du signal original

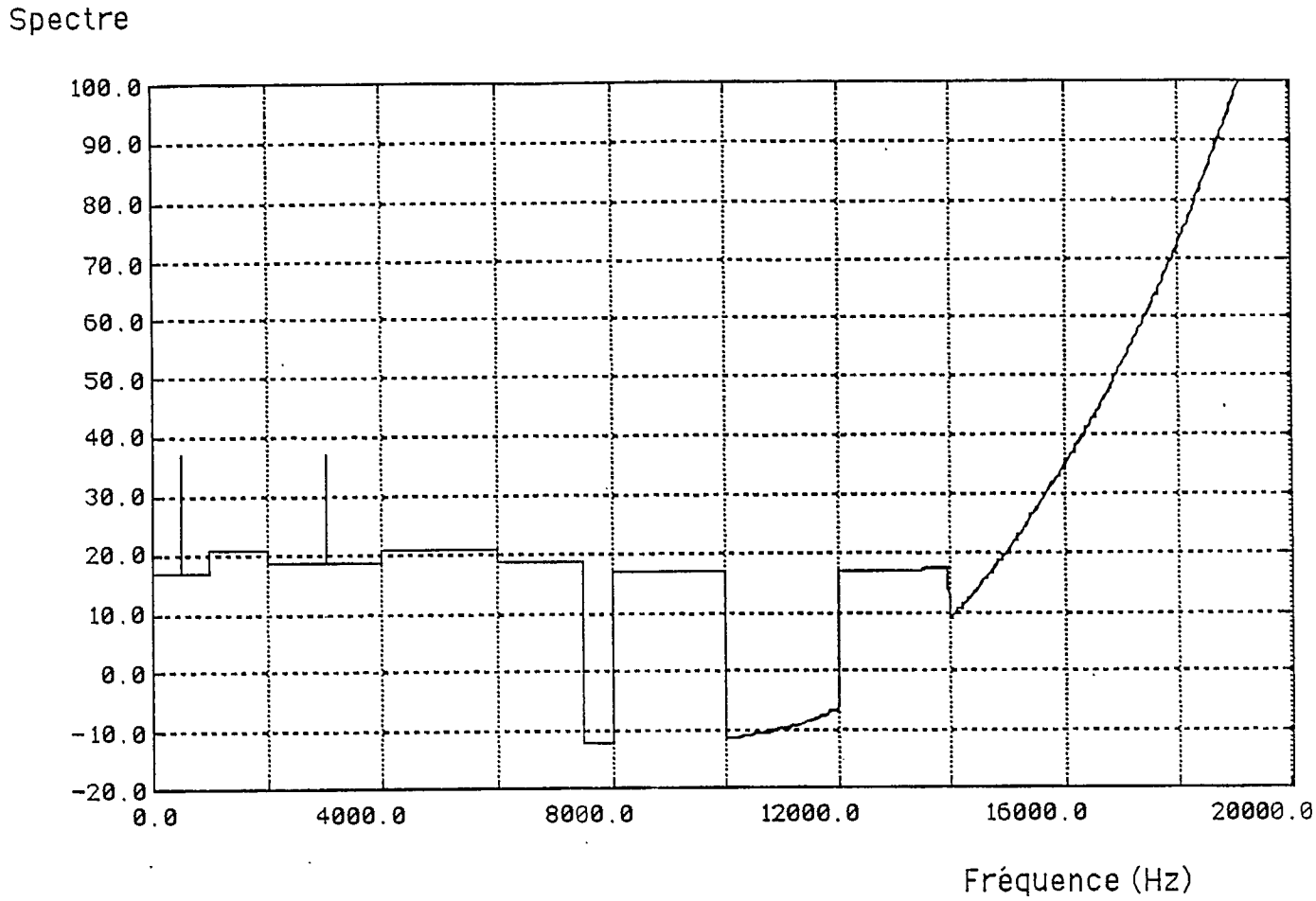


Figure 5 Spectre du bruit

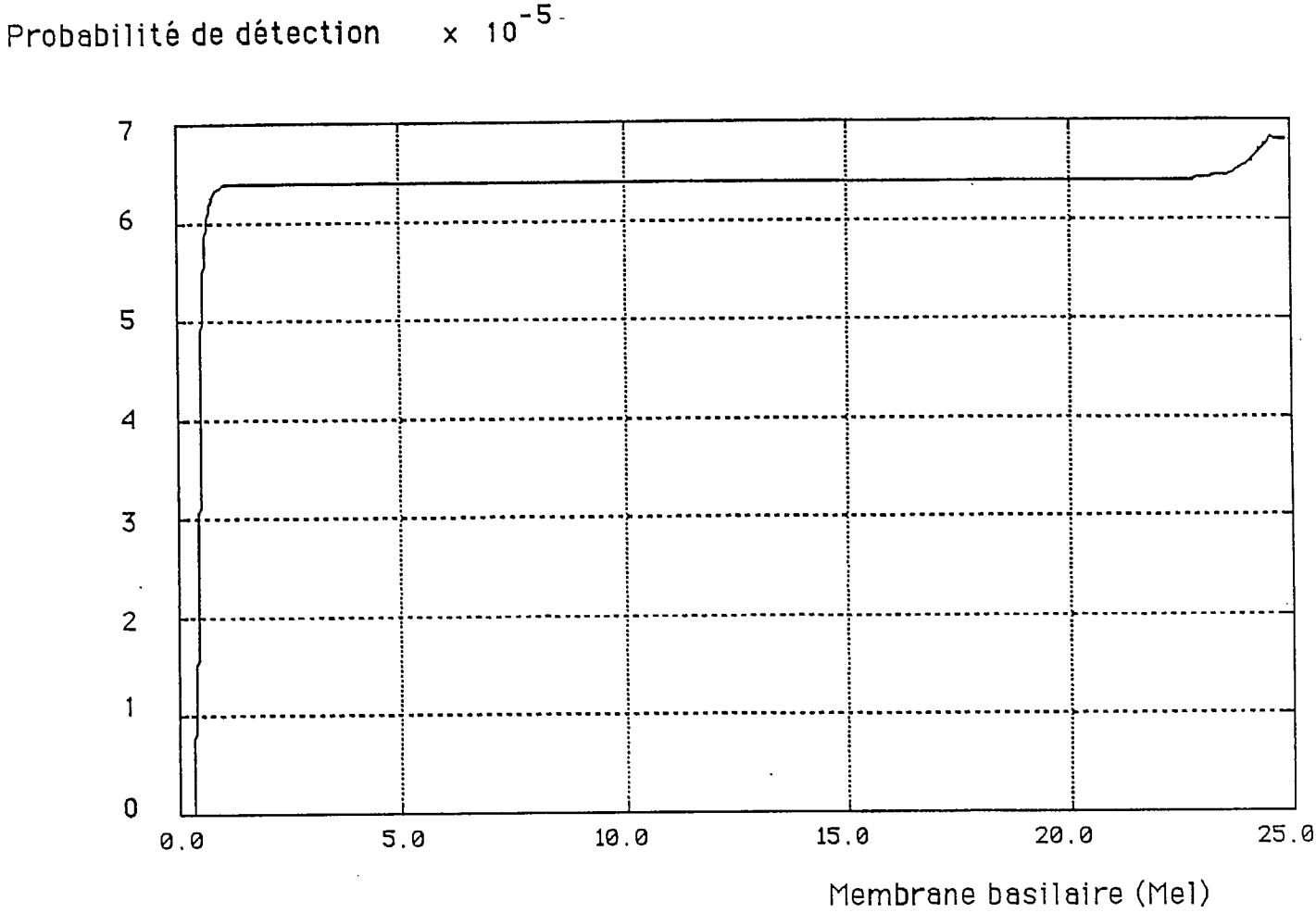


Figure 6 Probabilité de détection basilaire

5. Conclusion

Les travaux réalisés cette année ouvrent la voie au développement effectif d'un codeur utilisant au mieux les propriétés perceptuelles de l'oreille. En effet, l'étude qui a été effectuée afin de comparer les performances subjectives de différentes techniques de coloration du bruit de quantification en relation avec l'introduction d'un nouveau seuil de masquage, va permettre la définition d'une stratégie optimum pour l'allocation des bits. De plus un effort important a été réalisé pour le développement d'un environnement d'acquisition/écoute et de simulation indispensable à la mise au point d'un codeur.

ANNEXE 1

Cette annexe est un rapport qui a été préparé en anglais à la demande du CRC pour fins de présentation au comité de normalisation du CCIR.

PERCEVAL : Perceptual Evaluation of the Quality of Audio Signals

1 Introduction

Perhaps the most surprising issue in audio coding is the lack of a reliable objective quality criterion. A great deal of effort is devoted to developping better and better coders, but in the end the ultimate quality assessment must be made by listening tests which, in addition to being long and tedious, are subjective by their very nature.

Objective quality criteria do exist. For instance the signal to noise ratio (SNR) and its derivatives (frequency domain SNR, weighted frequency domain SNR...), are useful tools for the developpement of a specific coder. But since for different noise configurations, two signals with SNRs as different as 12 db and 50 db can be judged of equivalent quality in some instances (c.f. section 7), these criteria are completely inadequate to compare the quality of two different coders.

To be reliable, an objective quality criterion would have to model the hearing process.

Recently auditory models seem to have attained the required accuracy and reliability ([10]), to allow their use in objective quality criteria

The recent developpement, by our group, of a fast auditory model, led to its implementation in PERCEVAL, a software to evaluate the perceived quality of audio signals corrupted by noise.

The inputs of the program are the original signal and the noise signal. It can also be the original signal and the corrupted signal, the noise being extracted by difference.

The results are presented in the form of the detection probability of the noise versus time.

Bruno Paillard - Octobre 90

This criterion is well adapted to assess the quality of near perfect (high fidelity) signals. It cannot be used to evaluate badly distorted signals, since no indication is given of how loud the noise will sound, for noises bad enough that the detection probability is almost certain.

2 PERCEVAL : an overview

Figure 1 presents a schematic description of PERCEVAL. The two inputs are the original signal and the noise. In the case where the only two available signals are the original and the corrupted signal (for instance an original signal and its coded version), it will be necessary to compute the noise by difference. This implies that the relative temporal location of those two signals should be known precisely.

At first, a time-frequency decomposition is performed on the original signal, as well as on the noise. This decomposition is a MLT (Modulated Lapped Transform [5]), which combines the good frequency resolution of subband decompositions, and the fast algorithms of segmental orthogonal transforms (such as Discrete Cosine Transform, or Discrete Fourier Transform).

For a sampling frequency of 44.1 Khz, the decomposition is into 1024 frequency components. This gives a spectral resolution (frequency spacing between adjacent components) of 21.5 hz, and a temporal resolution (time spacing between adjacent components) of 23 ms.

From those two time-frequency representations, the energy is computed in the time-frequency space for both the original signal and the noise. This is done by performing a sliding average of the energy for each frequency component. This average is done over 3 successive samples (about 70 ms).

The results are 1024-component short term energy spectra for the original and the noise signals, obtained every 23 ms.

At this point, every 23 ms, the successive energy spectra of the original signal are passed through the auditory model which computes corresponding successive *original auditory sensations* (c.f. section 3).

The successive energy spectra for the noise are added to the corresponding energy spectra for the original signal. These *original plus*

noise energy spectra are passed through the auditory model which computes the corresponding *original plus noise auditory sensations.*

Every 23 ms, both auditory sensations are presented to a detection unit which computes the probability of detection of a difference between the two.

3 The auditory model

The auditory model around which PERCEVAL is built is a frequency model.

From a 20000-component energy spectrum (0->20 Khz), it computes a 2500-component basilar energy distribution (0->2500 Mel).

After this operation, a small intrinsic noise energy is added to each of the 2500 basilar components. This small energy limits the low-level sensitivity of the detectors, and accounts for the absolute hearing threshold.

Then, a logarithmic law is applied to each of the 2500 energy values of the basilar energy distribution, to obtain an 2500-component *auditory sensation* or *basilar sensation* vector.

This *basilar sensation vector* will be the input of the detection unit.

3.1 Hypotheses

This auditory model has been built around 3 main hypotheses :

1) *All the mechanical phenomena in the inner ear are linear and time-invariant.* In other words, the mechanical signal on each point of the basilar membrane is related to the acoustic wave picked up by the pinna, by a filtering process. The filter is different for each point on the basilar membrane (it varies continuously from one end of the membrane to the other) but all these filters include a common part consisting in the transfer function of the pinna, the ear canal and the middle ear.

2) *The detectors along the basilar membrane are quadratic (sensitive to the energy), and they have a logarithmic sensitivity function.* Which means that the sensitivity of a detector to small variations of

energy, is constant throughout the range of the detector, when the variations are expressed in db (c.f. section 4).

3) *The detectors along the basilar membrane have a long response time.* In other words, they have a poor temporal resolution.

This last hypothesis is somewhat relative. In fact, as will be seen later, this approximation enables the simplification of a *time-frequency* model (like the one described in [10] for instance), into a *pure frequency* model.

In summary, these 3 hypotheses define an auditory model such as the one described on Figure 2. Each branch corresponds to a detector on the basilar membrane.

This is a very classical model, and a lot of work has been done along those lines by the psychoacoustical community ([7]).

However, it should be noted that, in order to have a spacial resolution on the basilar membrane comparable to the resolution of the ear, several thousand detectors must be considered (in the order of 3000). This high spacial resolution is very important to model accurately certain psychoacoustical experiments, the masking of noise by tone for instance.

3.2 Simplifications

A model such as this is computationally expensive, mainly because of the several thousand filters which should be implemented.

To simplify the model, the 3rd hypothesis is used, as well as some additionnal ones.

The 3rd hypothesis indicates that the detectors are sensitive to the energy they receive, integrated over a *long* time. It is possible to compute the energy appearing on a detector, integrated over a long time, by multiplying the short term energy spectrum of the input signal by the energy spectrum of the filter corresponding to this detector, and integrating over the frequency space (Figure 3).

This operation can be seen as the inner product of the signal energy spectrum vector by the filter energy spectrum vector. If this operation is performed for every point on the basilar membrane, it is possible to represent this by a matrix operation (Figure 4).

Each line of the matrix "T" in Figure 4 represents the energy spectrum of the filters corresponding to each point on the basilar membrane.

Alternately, the columns of "T" (the lines of T^T, c.f. Figure 5) represent the distribution of energy on the basilar membrane, in response to pure tone excitations for different frequencies.

Those two descriptions of the matrix "T" (by lines or by columns) are only two different representations of the same matrix. Therefore, those two descriptions are perfectly equivalent. It is equivalent to know the energy spectra of the filters corresponding to all the points on the basilar membrane, or to know the basilar energy distributions in response to pure tones for all the frequencies on the frequency space.

The most important result is the existence (linked to the validity of the 3rd hypothesis) of a linear transformation which maps energy spectra into basilar energy distributions.

The linearity of this transformation implies the *additivity* of energies on the basilar membrane :

if

- $B_1(b)$ is the basilar energy distribution resulting from an excitation by a signal having an energy spectrum $F_1(f)$. $B_1(b) = TF_1(f)$

- $B_2(b)$ is the basilar energy distribution resulting from an excitation by a signal having an energy spectrum $F_2(f)$. $B_2(b) = TF_2(f)$

- b represents the basilar position

- f represents the frequency

then the basilar energy distribution resulting from the superposition of both excitations, $F_1(f) + F_2(f)$, is equal to the sum of the two individual basilar energy distributions $B_1(b) + B_2(b)$.

In addition to the linearity of T, we assumed a certain invariance of T^T (c.f. [6]). Basilar energy distributions in response to pure tone excitations are identical in shape, but positionned differently on the basilar membrane, depending on the frequency of the excitation.

This last assumption enables the decomposition of T into 3 steps :

- A multiplication of the energy spectrum of the signal by the *attenuation spectrum* of the ear canal and the middle ear (c.f. Figure 6).

- A localization of the *attenuated* spectral energy of the signal onto the basilar space. This localization is done according to the non-linear *frequency -> basilar position* conversion law, and conserves the integral of the energy. Each element of energy appearing on a short segment of the frequency space is added onto the basilar space, at a position corresponding to the center frequency of the spectral segment.

- A dispersion of this *localized* basilar energy distribution on the basilar membrane.
  This dispersion is done by filtering the localized basilar energy distribution with a filter whose impulse response reproduces the shape of a line of T^T (the basilar energy distribution in response to a pure tone excitation).

In [6], Zwicker and Feldtkeller show that the basilar energy distribution in response to a pure tone can be represented by a two-slope triangle on a logarithmic scale. Therefore on a linear scale, this basilar energy distribution can be represented by a double-sided decreasing exponential.

The spaces constants for these exponentials are :

- 0.27 db/Mel for the low frequency slope,
- -0.1 db/Mel for the high frequency slope

The representation of this *dispersion function* by a double-sided decreasing exponential allows the dispersion to be performed very efficiently by two 1^{st} order autoregressive filters (a causal and an anti-causal one).

Figures 7, 8, 9, 10, 12, 13, 14 and 15 show the results of these 3 steps for :

- A spectrum consisting of equally spaced 40 db harmonics, covering the frequency space from 0 to 20 Khz ,with 500 hz intervals.

- A spectrum consisting of 5 bands of noise, having 2 Khz bandwidths and 40 db power spectral densities, uniformely covering the frequency space from 0 to 20 Khz.

3.3 Sensitivity

After this transformation, the small intrinsic noise energy of the detectors is added to each component of the basilar energy distribution vector and the logarithm is taken for each of the 2500 components (c.f. Figure 2). This results in a *Basilar sensation vector* which will be one input of the detection unit. Figure 11 and 16 shows the result of this step for the first and the second example presented above.

3.4 Some results

Coupled to the statistical detection principle presented in section 4, this auditory model has been used to simulate psychoacoustical experiments which are described in the litterature. A lot of different experiments have been simulated with good results, including those aimed at describing the critical-band effects, and measuring their bandwidth.

For instance, the original experiment of Fletcher ([1]) has been simulated.

In this experiment, a noise having a variable bandwidth and a constant spectral energy density is used to mask a tone at its center frequency. The detection threshold of the tone is recorded as a function of the bandwidth of the masking noise.

The results show that for a masker bandwidth smaller than a critical value (the critical bandwidth), the detection threshold is proportionnal to the noise bandwidth. For masker bandwiths greater than this critical bandwidth, the detection threshold is roughly constant, independently of the bandwidth.

Fletcher's results are plotted in Figure 17.a. Figure 17.b shows the result of the simulation for the 1 Khz, the 2 Khz, the 4 Khz and the 8 Khz center frequencies.

As another example, Figures 18 to 20 show the performance of the model on masked detection experiments compared to results from [6].

The maskers are :

- narrow band maskers — Figure 18
- Lowpass and highpass maskers — Figure 19
- White spectrum masker — Figure 20

For all these experiments, the masked signals are pure tones. The detection threshold is defined as the energy of the masked tone necessary to obtain a 50 % chance of detection.

Section 5 will present other results for masked detection experiments where the masker is a pure tone and the masked signal is a noise.

4 Detection Unit

As can be seen from Figure 1, the detection unit compares two different basilar sensations and computes the probability of detection of the difference.

A lot of detection models (Zwicker's for instance [6]) assume that there is detection whenever the two basilar sensations differ by 1 to 2 db, or more, on one at least of the thousands of basilar detectors.

This deterministic detection principle, coupled to the auditory model described in section 3, predicts accurately the thresholds for pure tones masked by narrow or wide band noises (c.f. section 3.4). However, with this simple detection principle, the thresholds for noise masked by tones are less accurate.

This is especially unfortunate, since in the case of coding, the masked (hopefully) signal is quantization noise, and the masker is often a very harmonic signal.

A more realistic statistical detection principle allows a very accurate prediction in both cases (for noise masked by tones, and tones masked by noise).

What explains the better performance of a statistical detection principle, is that if a large number of detectors participate in the detection (as in the case of a noise masked by a tone), it allows the threshold to be reached globally for a lower difference of basilar sensations on each detector, than if only a few detectors participate in the detection. This issue is developped in section 6.

4.1 Statistical detection principle

The statistical detection principle is a generalization of the deterministic one. For each of the 2500 detectors of the basilar membrane, there is a probability of detecting a difference between two values of the energy appearing on this point. This probability depends directly on the absolute value of the difference of the two basilar sensations for this detector.

If the absolute value of the difference of the two basilar sensations is small for a detector, the detection probability tends to zero for this detector. On the contrary, as the absolute value of the difference becomes important, the detection probability tends to one for this detector.

Figure 21.b shows the schematic shape of the function *detection probability v.s. absolute value of the sensation difference*, for one detector.

Figure 21.a shows the same function for the deterministic detection principle discussed above.

The 2500 basilar detectors are supposed to be statistically independant, so that globally, the non-detection probability for the whole set of detectors is simply the product of the individual non-detection probabilities for each detector.

5 Performance of the auditory model

The combination of the auditory model described in section 3, and the statistical detection principle described in section 4 has been used to simulate a variety of experiments. This section will present the results of the simulations of two experiments which investigate the detection of noise masked by a tone.

5.1 Schroeder's experiment

The first experiment that we simulated was done by Schroeder et al. [3].

A noise, having a bandwidth of 1/3 octave and a center frequency of 1 Khz, is masked by an 80 db pure tone whose frequency is varied between 500 and 2 Khz. The detection threshold of the noise is measured as a function of the frequency of the masker.

The results obtained by Schroeder et al. are presented in Figure 22.a. The results of the simulation are presented in Figure 22.b.

They are in good accordance with the results of the real experiment. Especially, it can be seen that the slopes of the threshold function predicted by the simulation reproduce closely those measured on the curve from Schroeder et al. The maximum of the threshold function is also very close to the maximum found by Schroder et al.

One difference can be noted however. From 1414 hz to 2 Khz, the threshold stabilizes at a value around 20 db for the experiment, and around 5 db for the simulation.

For the model, it is clear that when the masker is in this frequency range, it has no masking effect on the noise which is centered around 1 Khz. Therefore the 5 db threshold depends only on (and reflects) the absolute threshold in the 1 Khz frequency range.

We assumed that this was also the case for the listener in Schroeder's experiment. Either because this listener had a high absolute threshold around 1 Khz, or because there was some ambiant noise during the experiment.

5.2 Hellman's experiment

The second experiment that we simulated was done by Hellman [2].

First of all, absolute thresholds are measured for :

- A 1 Khz pure tone

- A narrow band noise (925 - 1080 hz)
- An octave band noise (600 - 1200 hz)
- A wide band noise (75 - 9600 hz)

Hellman's results are :

- 1 Khz tone : 6 db
- Narrow band noise (925 - 1080 hz) : 6 db
- Octave band noise (600 1200 hz) : 8 db
- wide band noise (75 - 9600 hz) : 15 db

Simulation results are :

- 1 Khz tone : 4 db
- Narrow band noise (925 - 1080 hz) : 5 db
- Octave band noise (600 1200 hz) : 7 db
- wide band noise (75 - 9600 hz) : 5 db

The simulation results are in perfect accordance with the real ones, exept for the wide band noise. Here again, it should be noted that the absolute threshold for this wide band noise depends on the *absolute threshold v.s. frequency* function in the entire band from 75 hz to 9600 hz. This function was not known for Hellman's listener, and can differ notably from our model's.

In the second part of the experiment, the detection thresholds are measured for two noise signals (1280 hz - 1480 hz, and 1350 hz - 1450 hz) masked by a 1400 hz 70 db pure tone.

The thresholds measured by Hellman are :

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

The simulated thresholds are :

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

Again, the results of the simulation are very close to those of the real experiment.

6 Difference between the masking of noise by a tone and the masking of a tone by noise

Together with the statistical detection principle, this auditory model predicts accurately the detection thresholds for both a tone masked by noise, or noise masked by a tone. Therefore, it verifies the *Asymmetry of masking between noise and tone*, as described by Hellman ([2]) for instance.

The threshold level for a tone masked by a 1/3 octave noise, the tone being at the center frequency of the noise, is about 4 db below the level of the masker.

The threshold level for a 1/3 octave noise, masked by a tone at its center frequency is about 24 db below the level of the masker.

It appears then, that it is about 20 db harder to mask a 1/3 octave noise by a tone, than it is to mask a tone by a 1/3 octave noise.

Figure 23 shows the individual detection probabilities for the 2500 detectors along the basilar membrane, at threshold, in two cases :

The first case (a) represents the case of a 1 Khz tone masked by a 1/3 octave noise, centered around 1 Khz.

The second case (b) corresponds to the inverse situation of a 1/3 octave noise centered around 1 Khz, masked by a 1 Khz pure tone.

The maxima of those functions indicate the detectors for which the detection probability is highest, therefore, the width and the position of these maxima give a good indication of which detectors are most involved in the detection, and their approximate number.

Two important differences can be identified between the two cases :

- In the first case (a) the basilar detectors most likely to register the presence of the masked tone in the noise masker are precisely those detectors where the basilar excitation due to the tone alone would be maximum. In other words, the detection simply involves those detectors which are most excited by the tone.

- In contrast, in the second case (b), the detectors most involved in the detection of the noise, masked by the tone, are not the ones for which

the excitation due to the noise alone would be maximum (they are not the detectors located at the basilar position corresponding to the center frequency of the masked noise). The detection actually occurs mainly in a basilar zone which is located on the low frequency side of the masker. In this zone, the effect of the masker is notably lower than what it is at the basilar position corresponding to its frequency (1 Khz), while the effect of the masked noise (which has a wider spectrum than the masker) is still important.

This *Off frequency listening* effect accounts partly for the asymmetry of masking between noise and tone. Indeed a certain level of asymmetry (about 15 db) is predicted by the model, even with the simple deterministic detection principle described in section 4.

The second difference that can be observed between the two cases is that :

- In the first case (a) there are only a small number of detectors participating notably in the detection (mainly those few in the basilar zone corresponding to the frequency of the masked tone).

- In contrast, in the second case (b), the detection is spread over a larger number of detectors.

Therefore, in the second case the detection threshold is reached globally (for the whole set of detectors) for a lower maximum of the individual detection probabilities, than in the first case. In other words, the fact that a larger number of detectors are involved in the detection allows the detection threshold to be reached for lower individual detection probabilities. Thus, in the case of noise masked by a tone, the detection threshold is reached for lower individual sensation differences for each basilar detector, than it is in the case of a tone masked by noise.

This second effect accounts for about 5 db in the asymmetry of masking between noise and tone.

7 Performance of PERCEVAL

To evaluate the accuracy of PERCEVAL, it was enlisted in a listening test, along with 5 *real* listeners (c.f. [9]).

The experiment consisted in injecting noise in 7 different 10-second music exerpts, according to 3 noise shaping strategies.

The strategies were :

- White noise spectrum.
- Noise spectrum following the short term energy spectrum of the signal.
- Noise spectrum following the masking threshold derived from the short term energy spectrum of the signal.

In all 3 cases, the noise level was dynamically adapted to the level of the signal, to achieve a constant signal to noise ratio (SNR).

For each exerpt, and each noise shaping strategy, the noise detection threshold was measured for each of the 6 listeners (the 5 real ones and PERCEVAL).

For PERCEVAL, the detection threshold was defined as the SNR below which the detection probability was above 50 % for more than 70 ms.

The definition of the detection threshold is not really critical since the whole transition from *almost inaudible noise throughout the exerpt*, to *almost certainly perceptible noise throughout the exerpt*, generally occurs within only a 10 db SNR range.

The results are presented in table 1. The signal to noise ratios at threshold ranged from 3 db to 75 db, depending on the noise shaping strategy, and on the exerpt.

Despite this wide range, the thresholds predicted by PERCEVAL are generally in good accordance with the thresholds measured by the listeners. In the two cases of synthetic tones, and for the second and third strategies however, the predicted thresholds are somewhat underestimated (PERCEVAL seemed deaf...).

8 Conclusion

The program PERCEVAL has been designed to evaluate the perceived quality of corrupted signals. It can be used to assess the subjective quality of source coders.

The quality is represented as a detection probability of the noise, v.s. time, for a set of two input signals, the original signal and the noise signal.

The algorithm is fast so that it could probably be implemented in real time on a modern digital signal processor (TMS 320C30 for instance).

PERCEVAL is based on an auditory model which has been independentely tested and taylored on simulations of psychoacoustical experiments described in the litterature. This model shows very good results on experiments of masked detection in both cases of noise masked by tone and tone masked by noise.

A lot of work remains to be done to assess precisely the accuracy of PERCEVAL, but the first results are very encouraging.

Reférences

[1] H. Fletcher - *Auditory Patterns* - Reviews of Modern physics, January 1940, Vol. 12

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

[3] 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

[4] 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

[5] H. S. Malvar - *Lapped transforms for efficient transform/subband coding* - IEEE Trans. on ASSP, Vol. 38, No6, June 1990.

[6] 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

[7] Brian C. J. Moore. "An Introduction to the Psychology of Hearing", Academic Press, 1989, ISBN 0-12-505623-0

[8] Brian C. J. Moore. "Frequency selectivity in hearing", Academic Press, 1986, ISBN 0-12-505625-7

[9] A. Turgeon, J. Soumagne, P. Mabilleau, S. Morissette, B. Paillard. - *A study of strategies for the perceptual coding of audio signals* - To be presented at the 90th AES convention - Paris, 1991 02 19 22.

[10] R. Patterson, J. Holdsworth - *An introduction to auditory sensation processing* - MCR Applied Psychology unit, 15 Chaucer road, Cambridge CB2 2EF, England

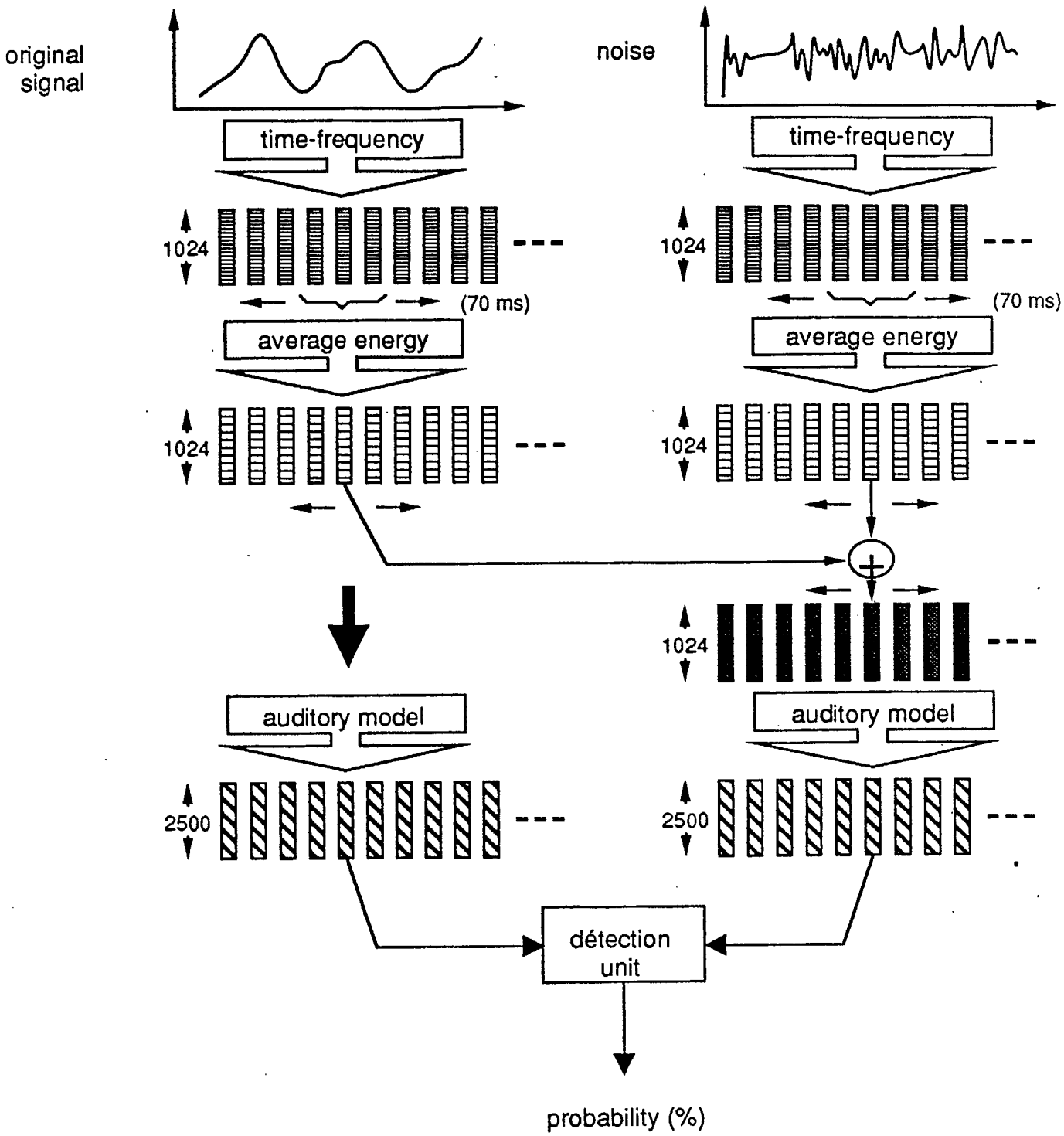


Figure 1 description of PERCEVAL

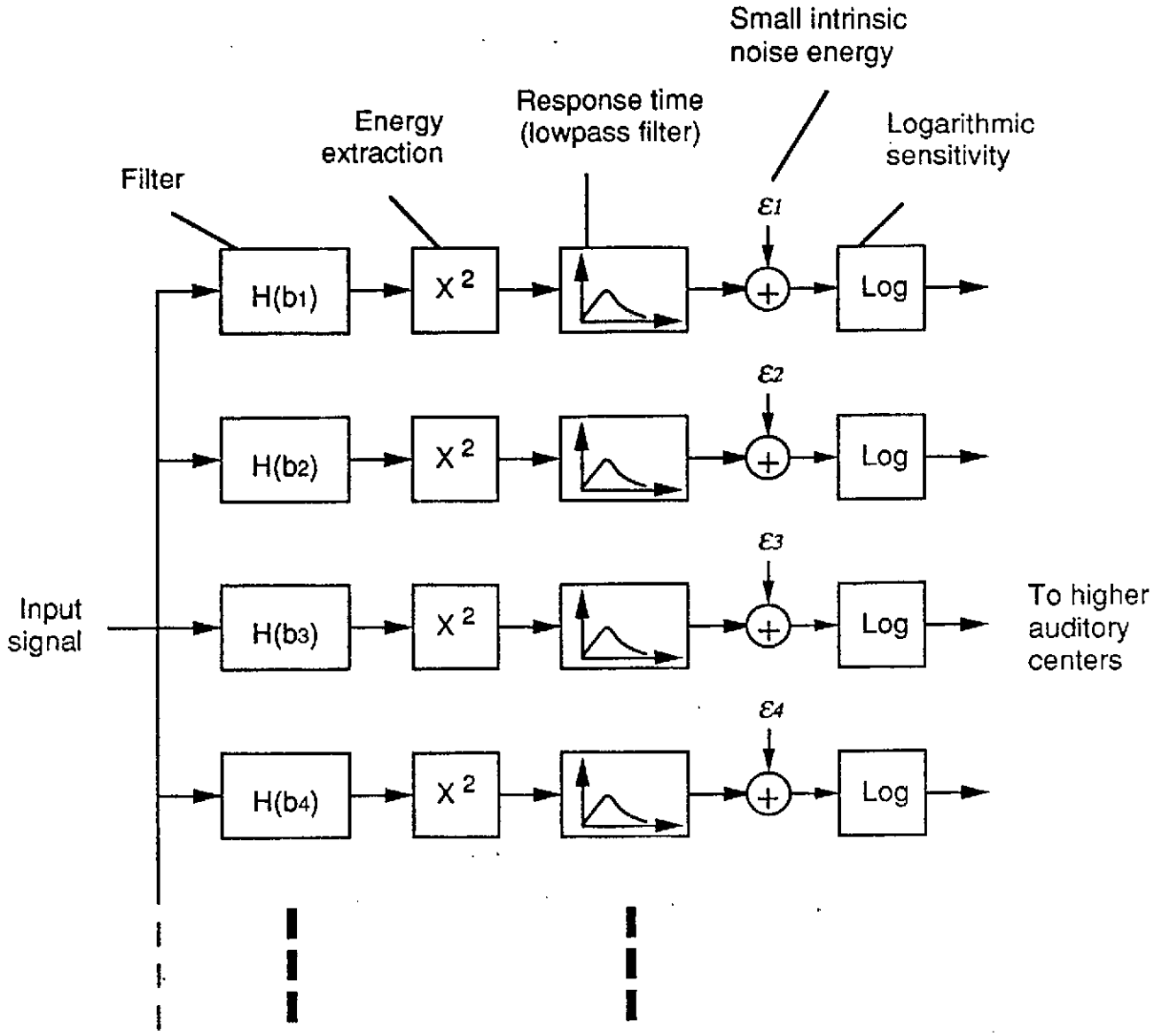


Figure 2 : Auditory model corresponding to hypotheses 1 to 3

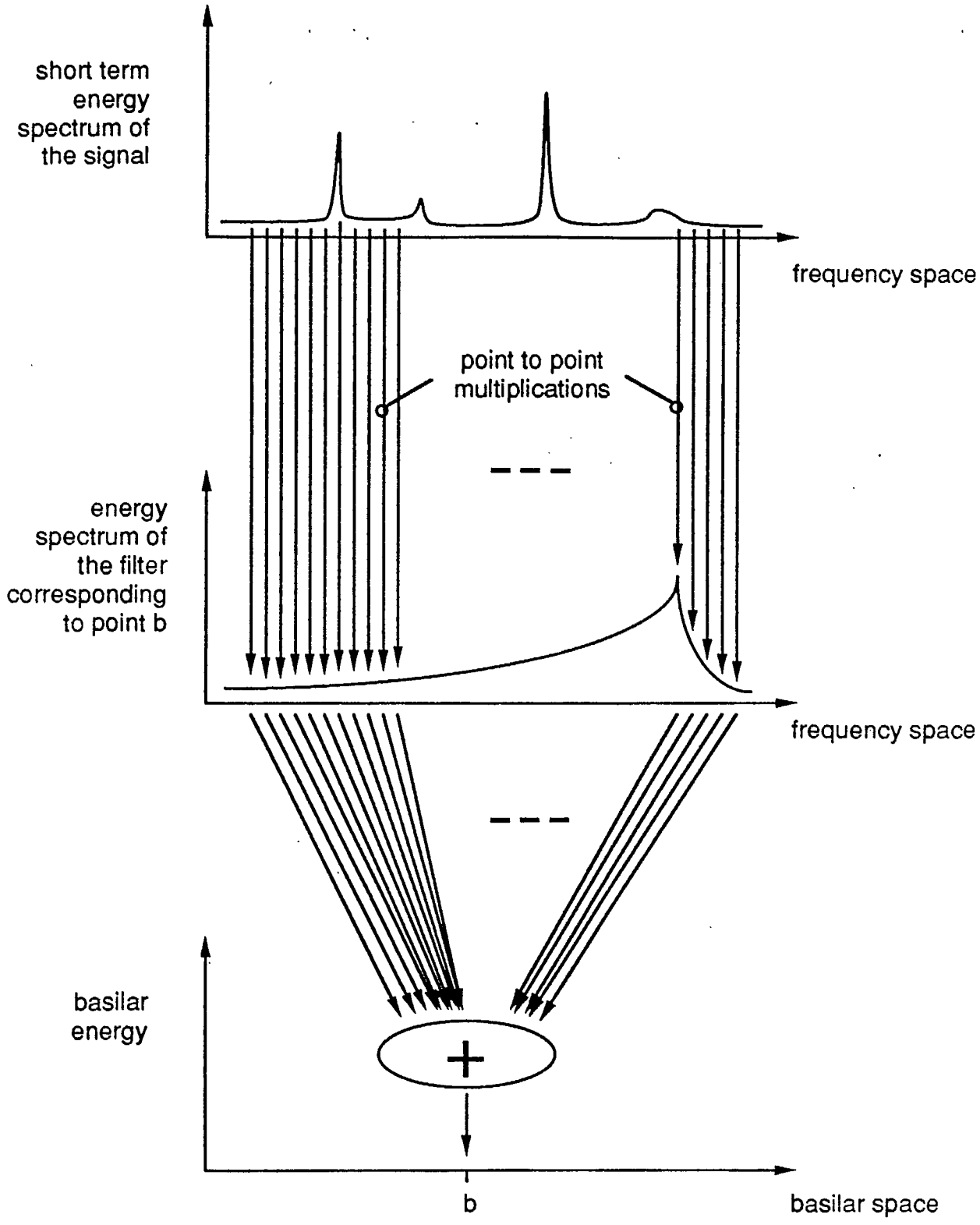


Figure 3 : Computation of the energy appearing on a point "b" on the basilar membrane

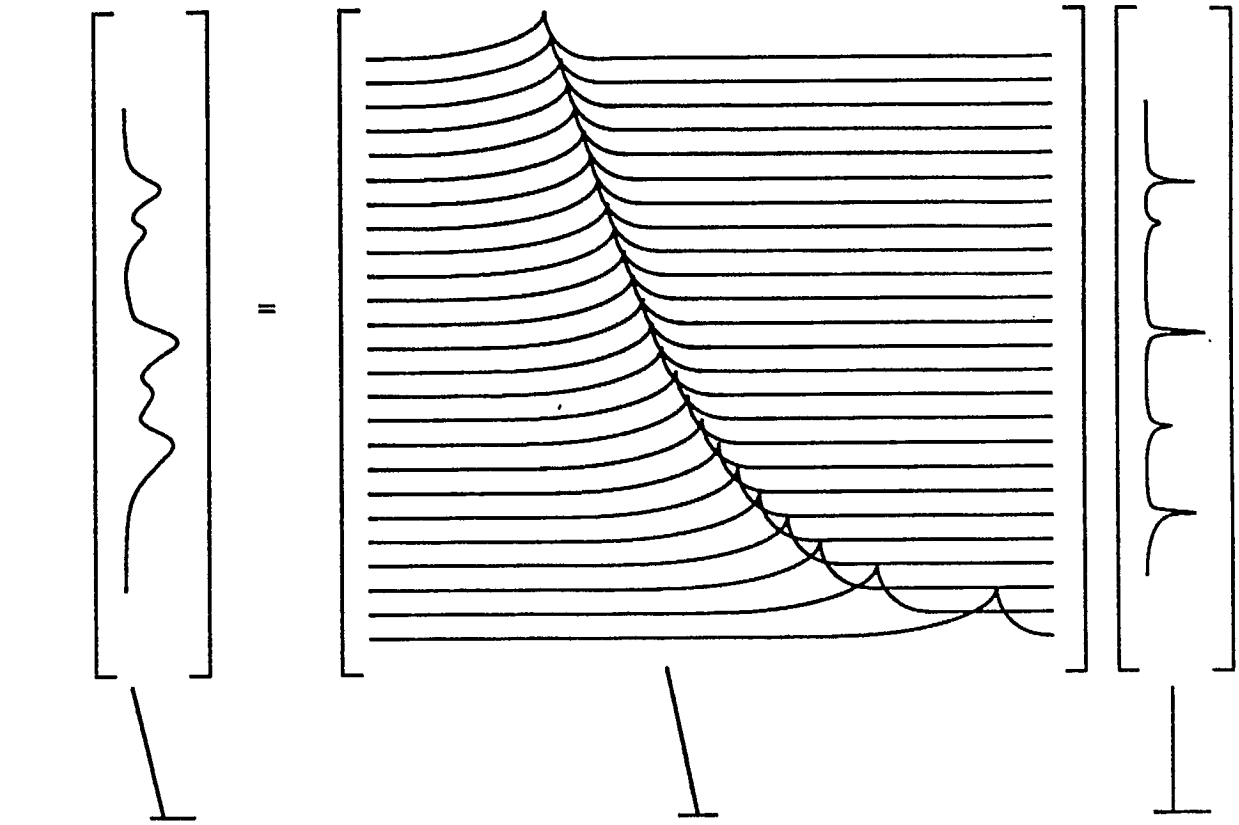
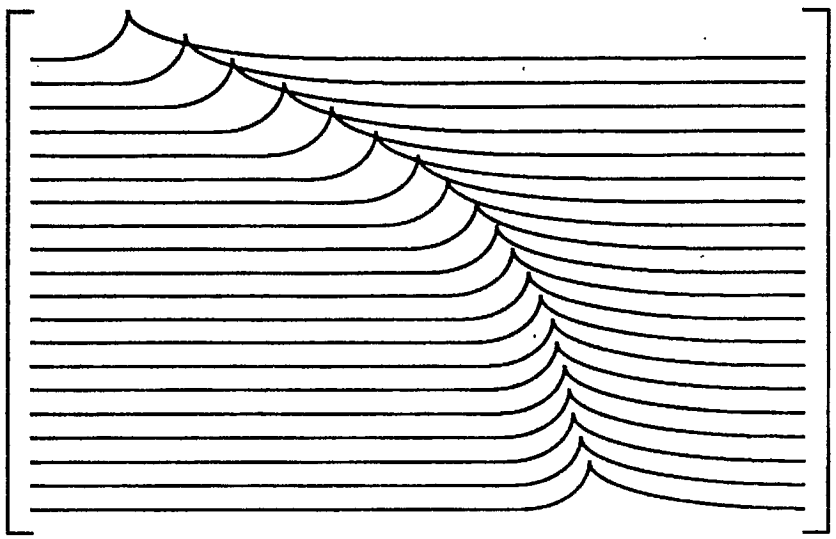


Figure 4 : transformation of the short term energy spectrum of the signal into a basilar energy distribution vector.



Matrix T^T

Figure 5 : (the lines of T^T represent the basilar energy distributions in response to pure tone excitations of different frequencies).

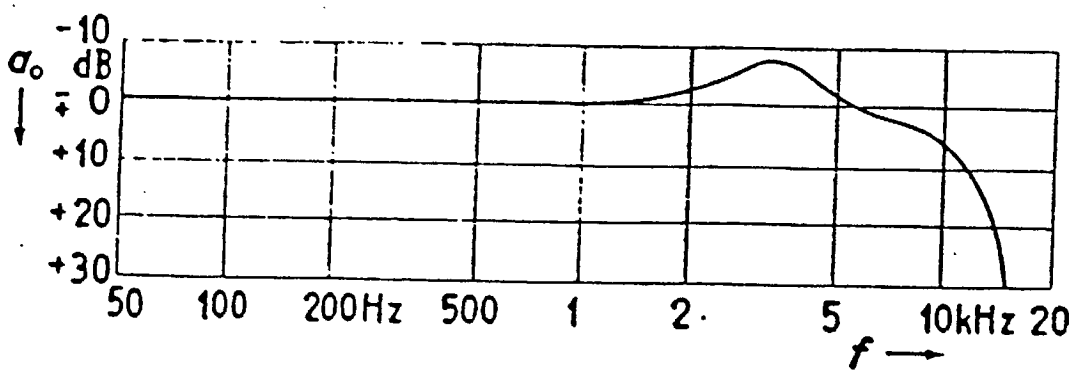


Figure 6 : Attenuation spectrum of the ear canal and middle ear

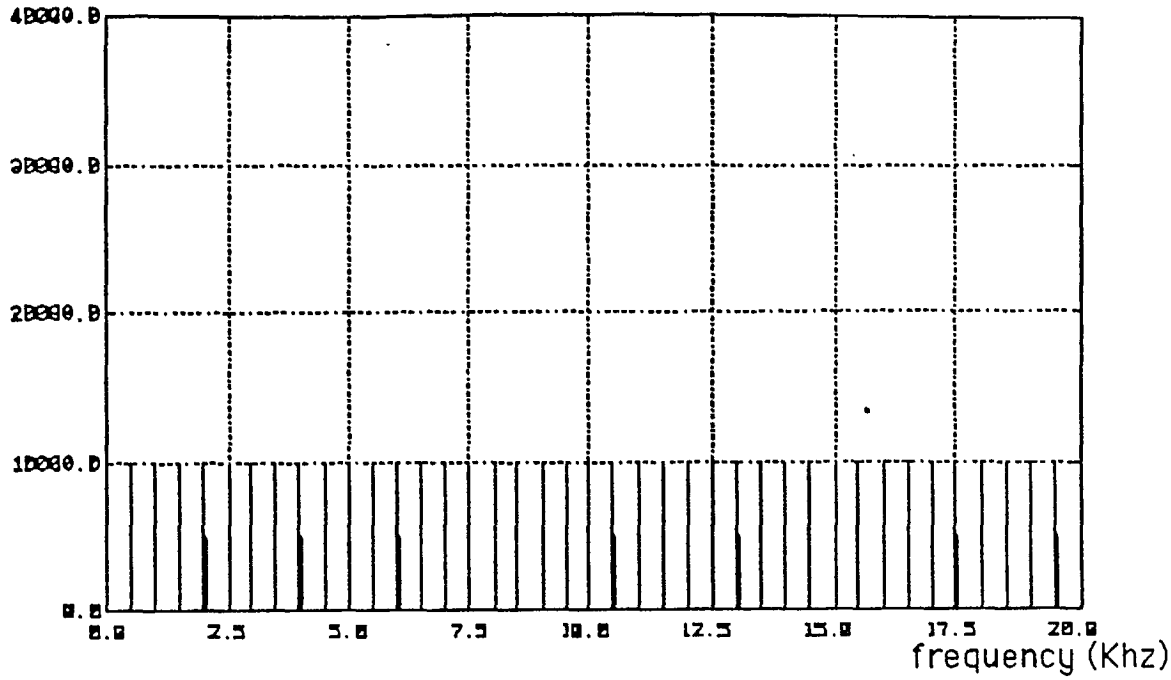


Figure 7 : original signal energy spectrum

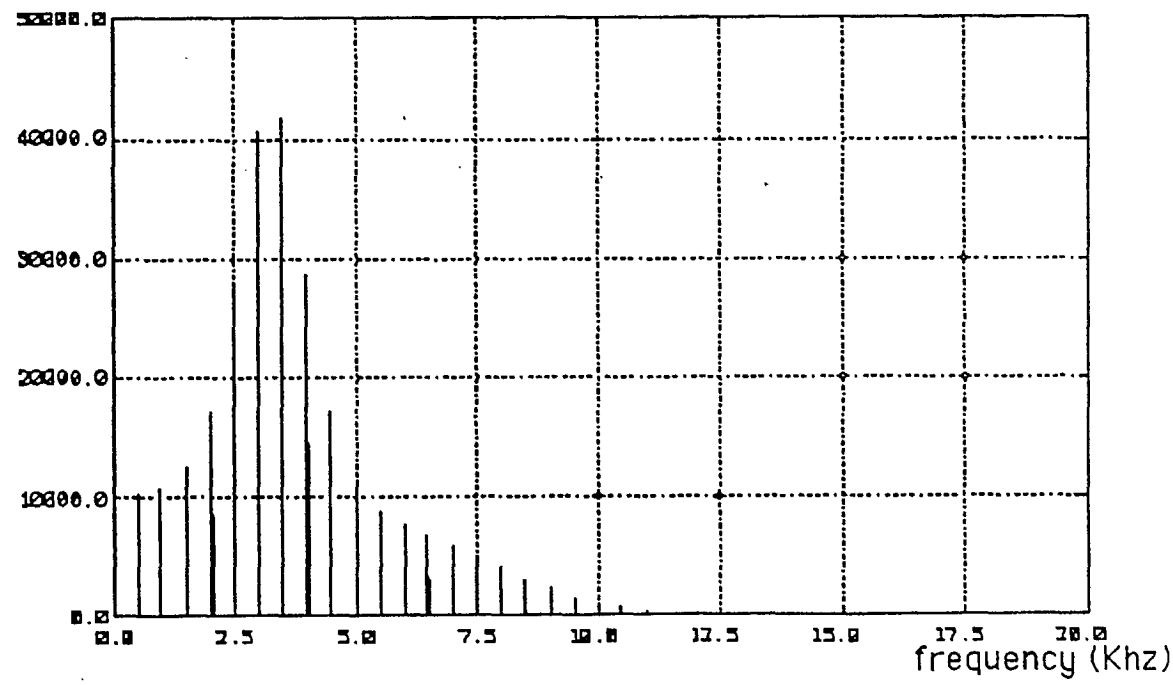


Figure 8 : Energy spectrum after attenuation by the energy spectrum of the ear canal and middle ear

energy

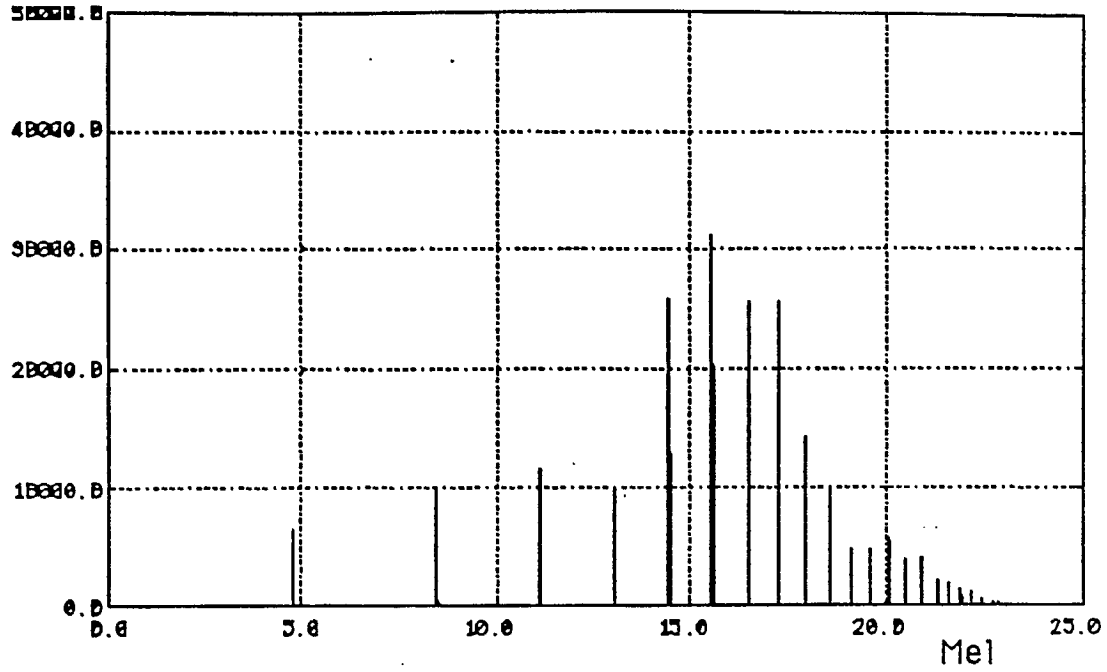


Figure 9 : localized basilar energy distribution

energy

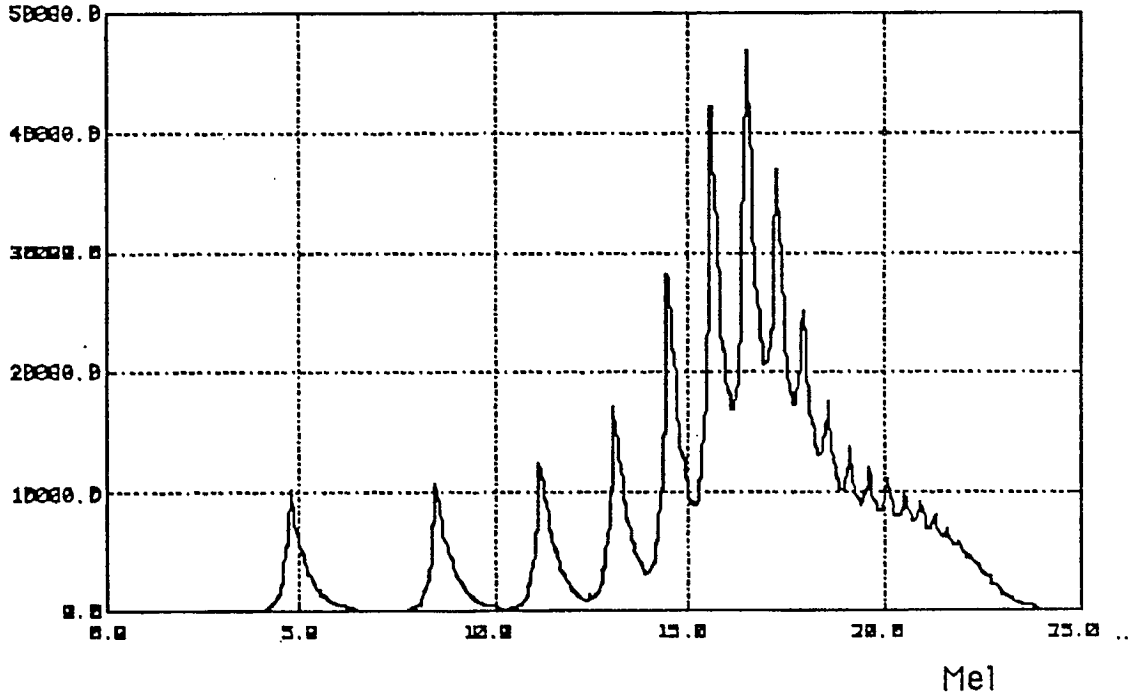


Figure 10 dispersed basilar energy distribution

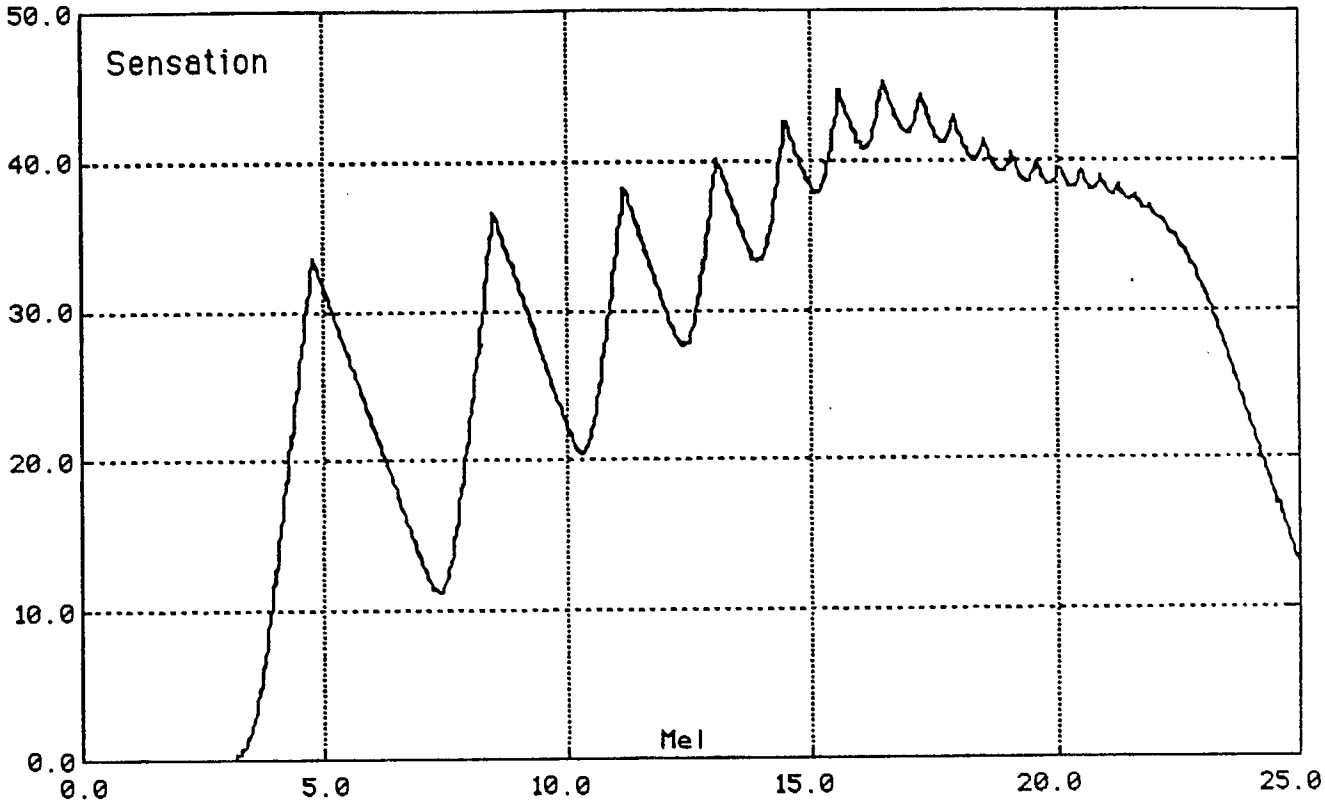


Figure 11 : Basilar sensation

energy

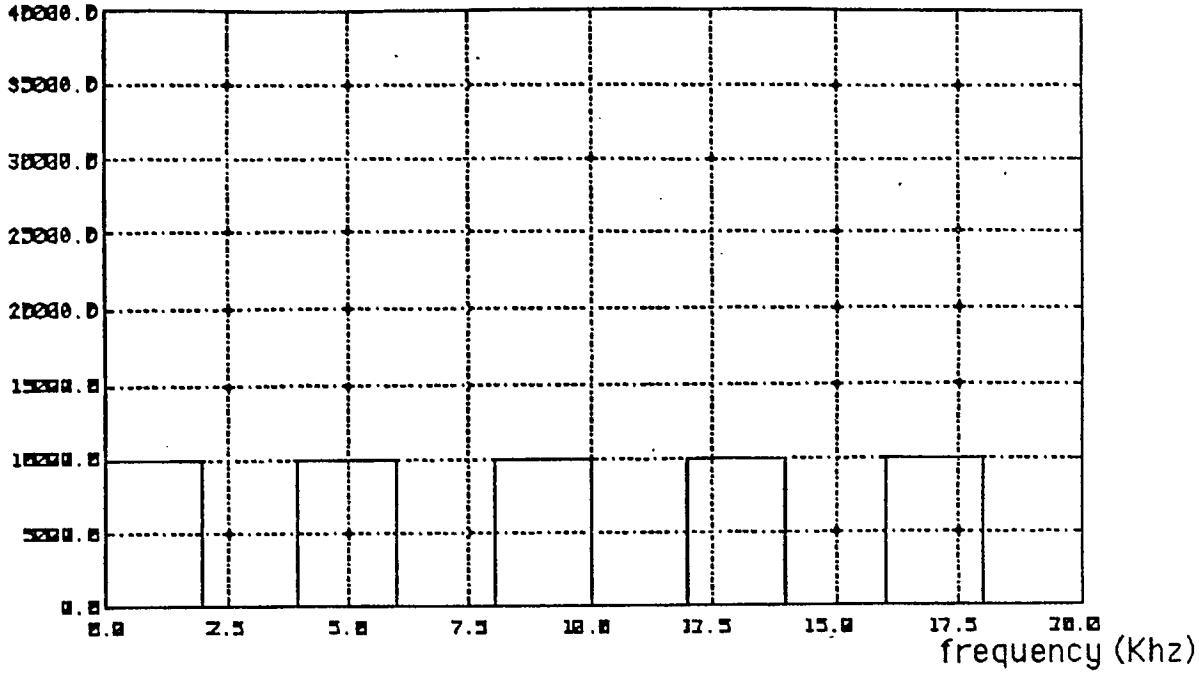


Figure12: original signal energy spectrum

energy

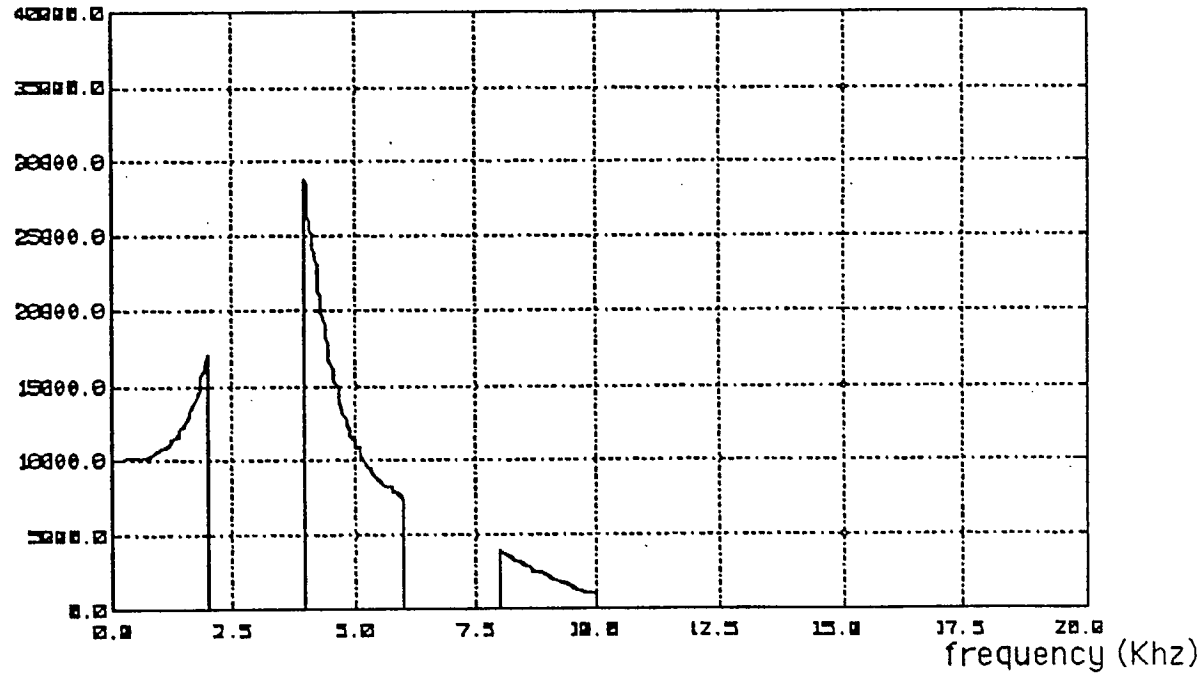


Figure13: Energy spectrum after attenuation by the energy spectrum of the ear canal and middle ear

energy

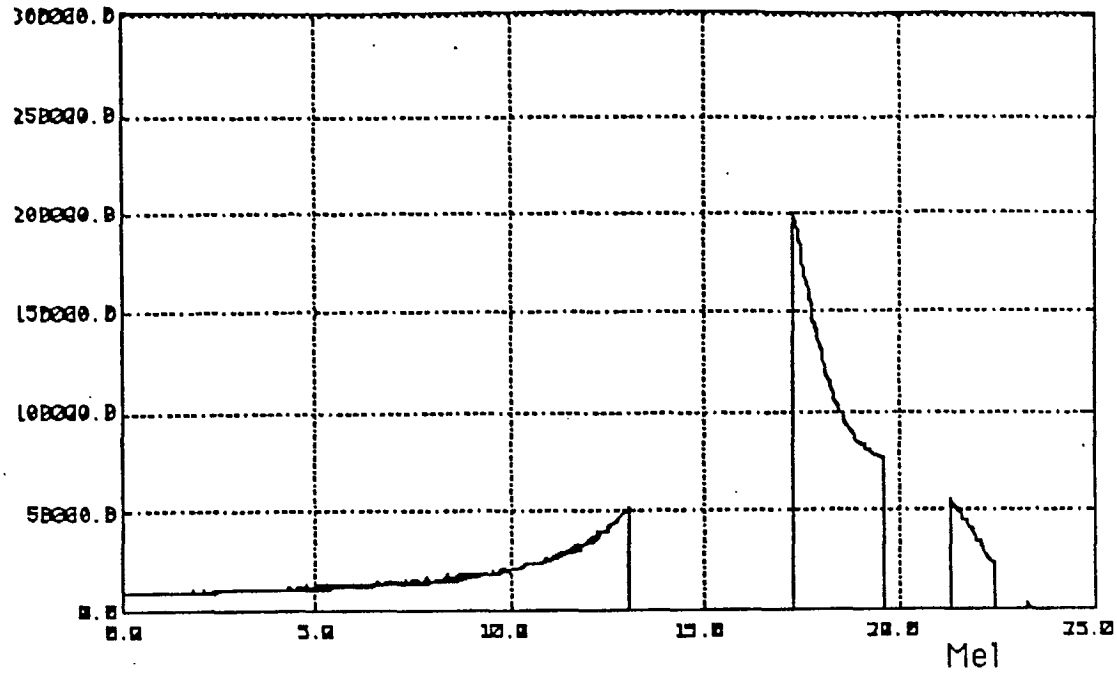


Figure14: localized basilar energy distribution

energy

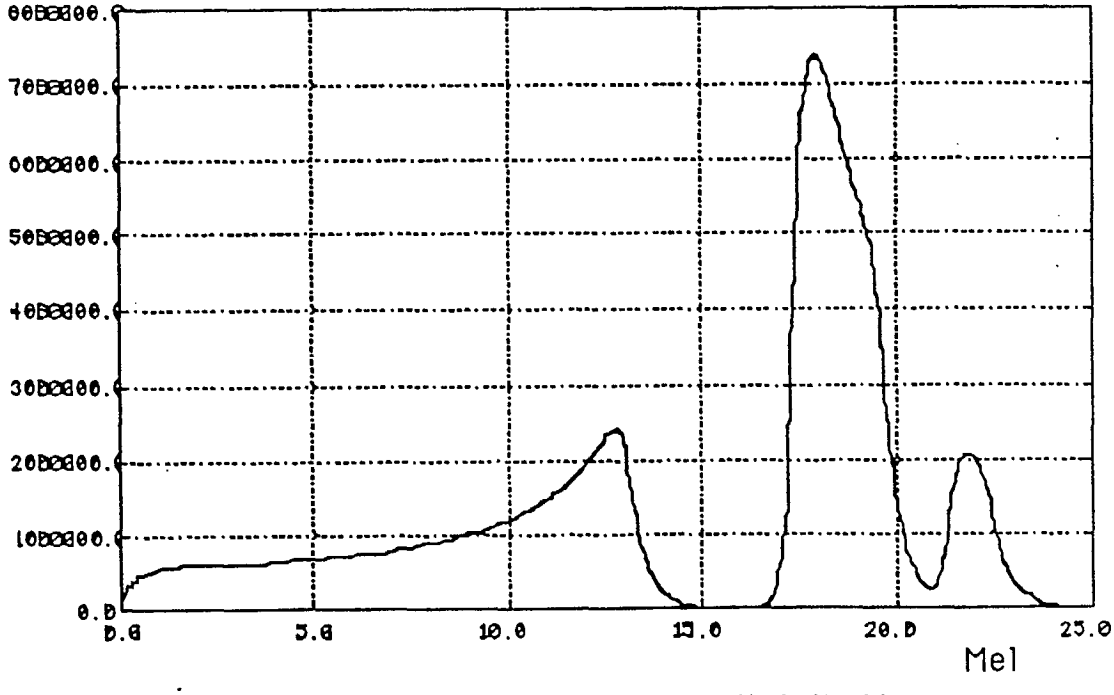


Figure15 dispersed basilar energy distribution

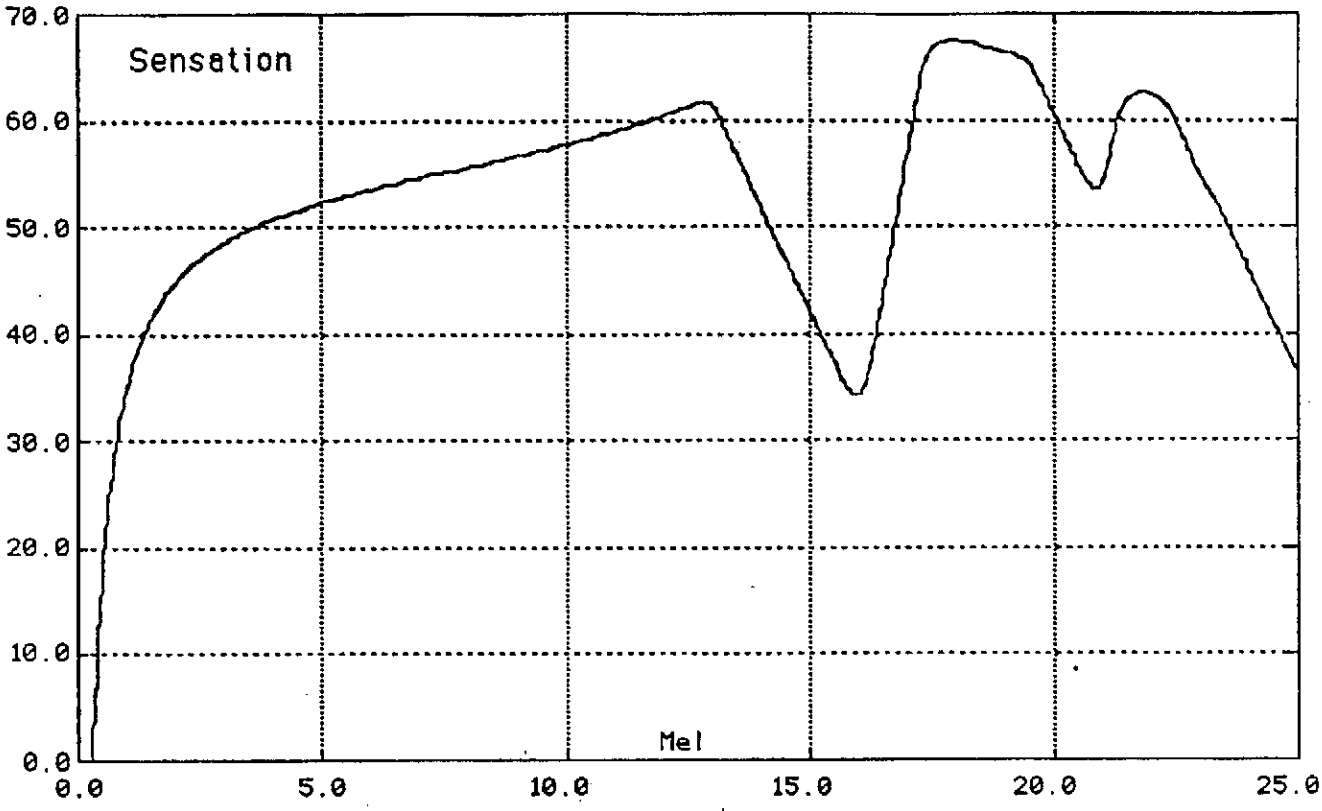
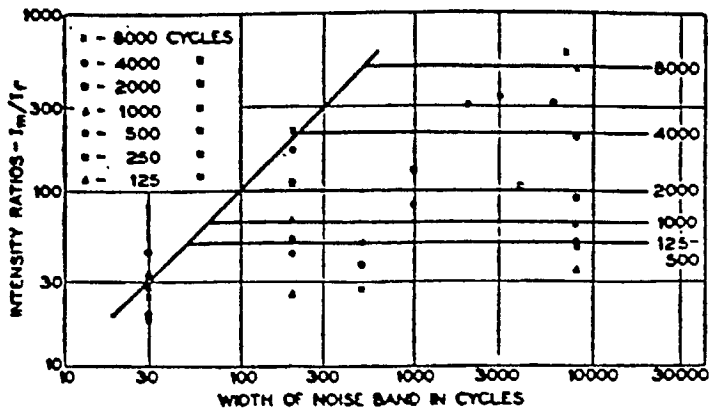
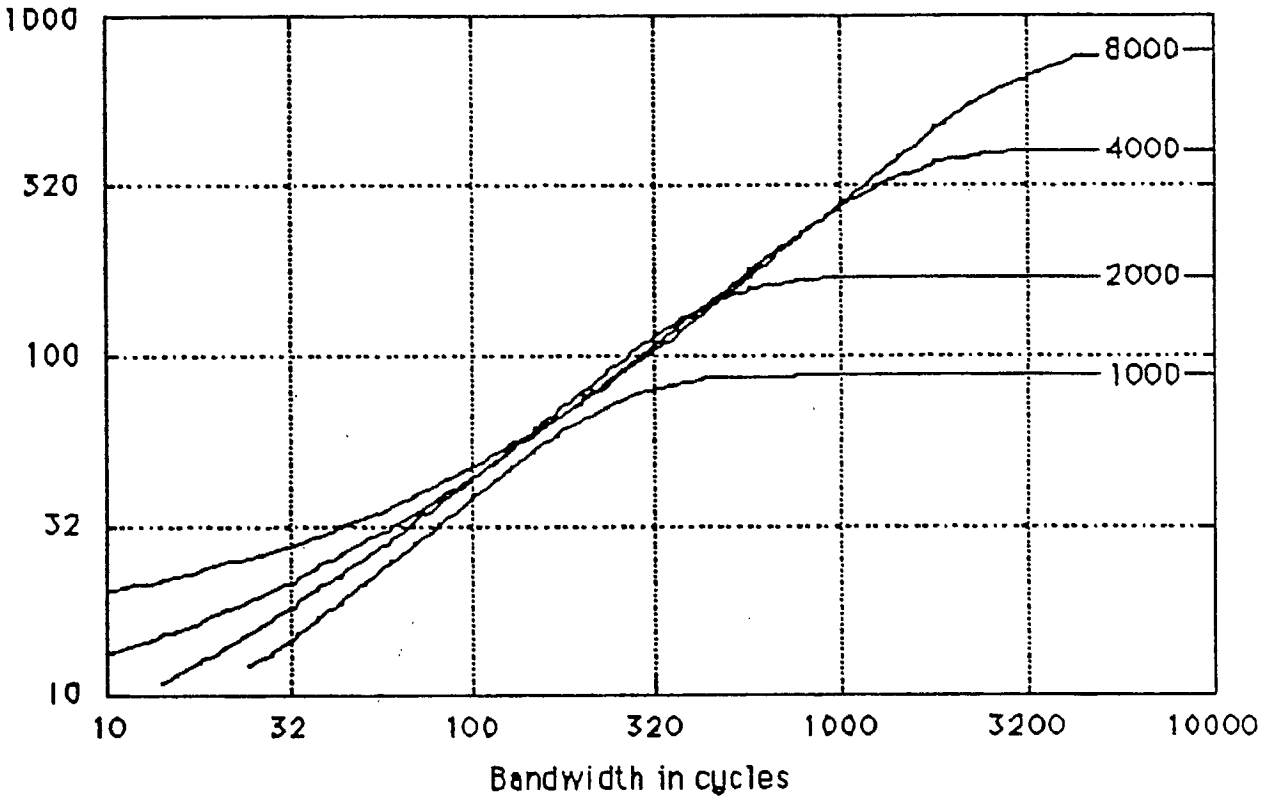


Figure 16 : Basilar sensation

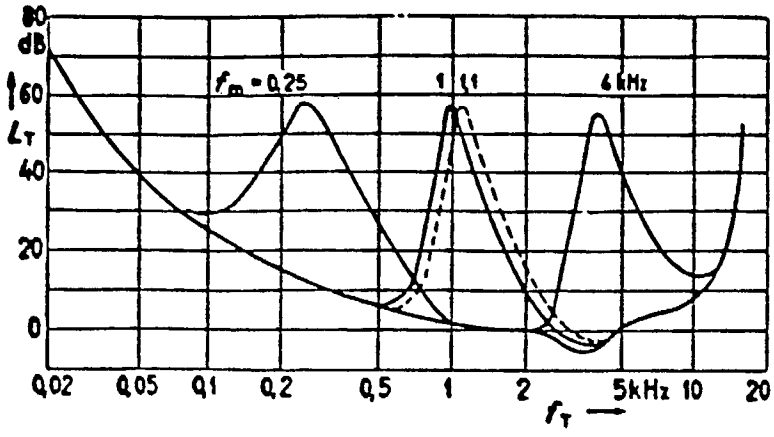


a) Fletcher's results (from [1])

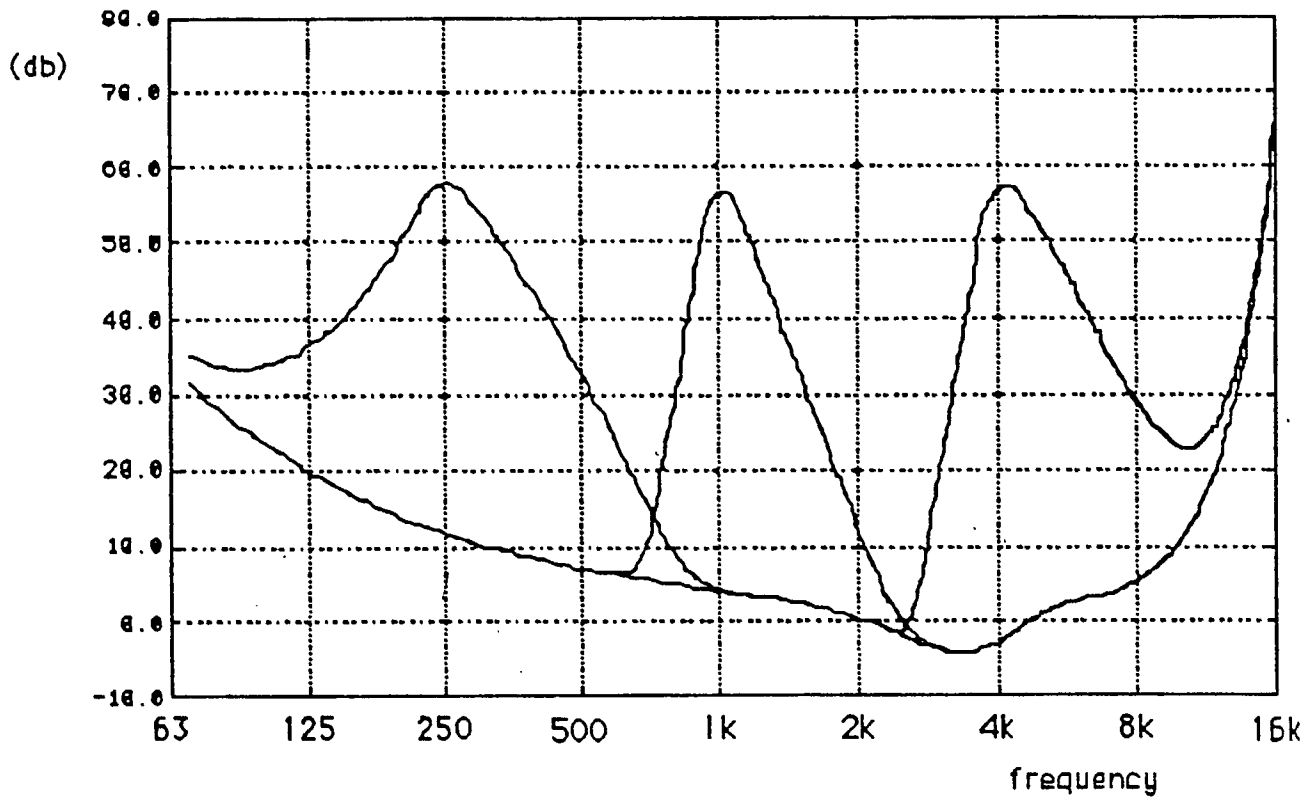


b) Simulation

Figure 17 : Fletcher's critical band experiment

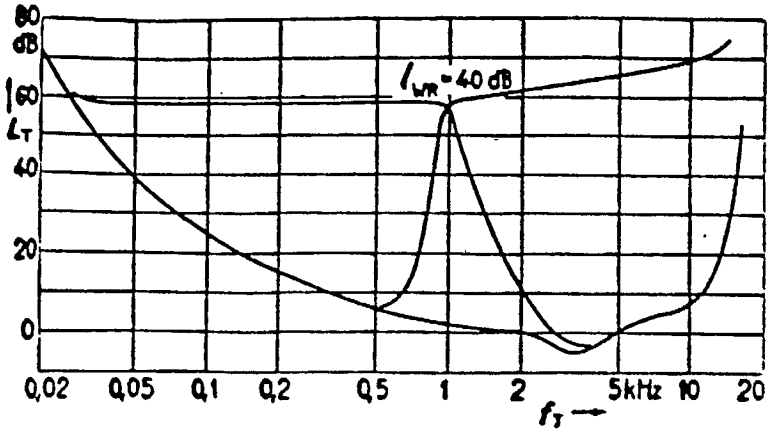


a) real results (from [6])

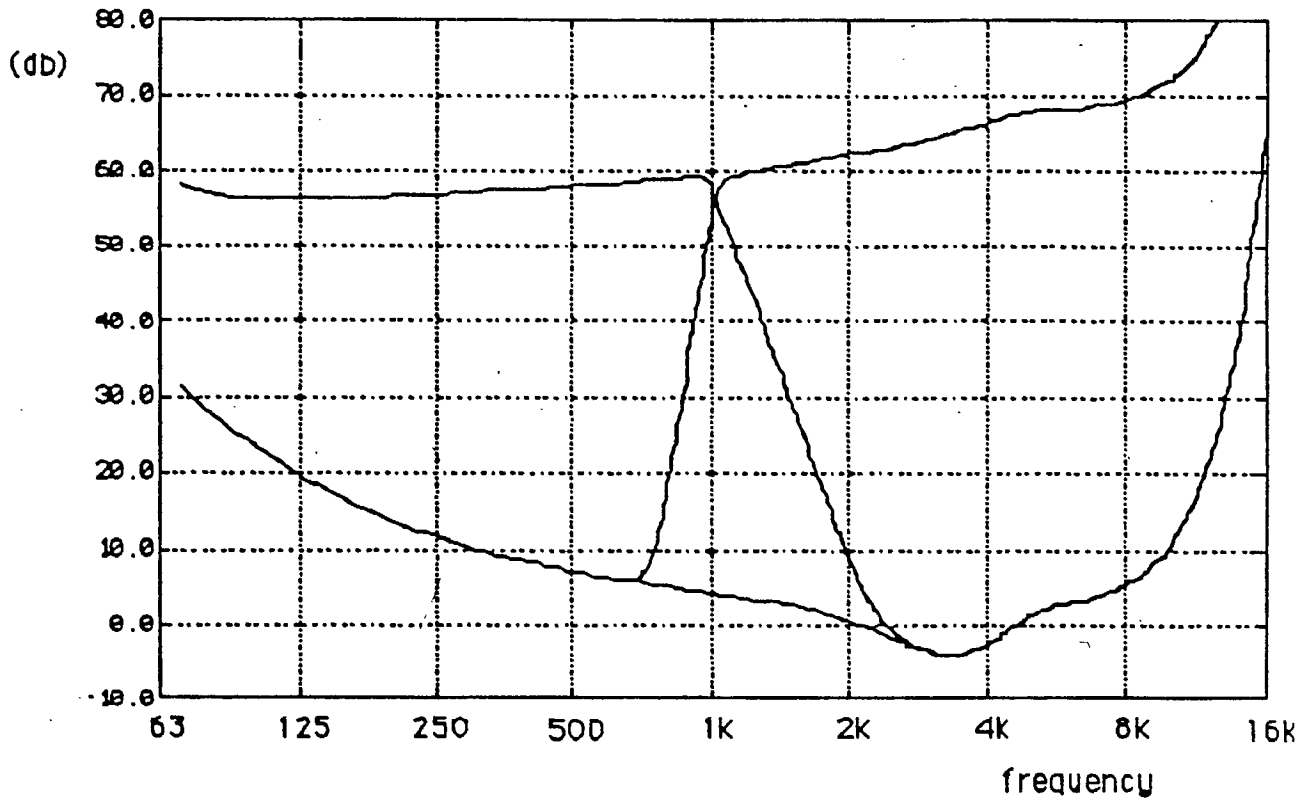


b) Simulated results

Figure 18 : detection thresholds for pure tones in the presence of 60 db 1/3 octave maskers, with 250, 1000 and 4000 hz center frequencies

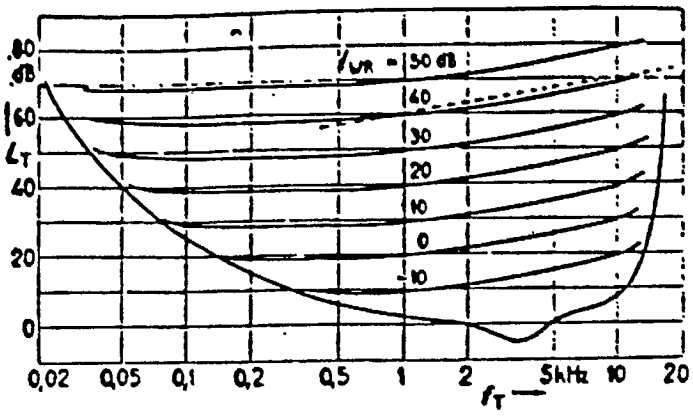


a) real results (from [6])

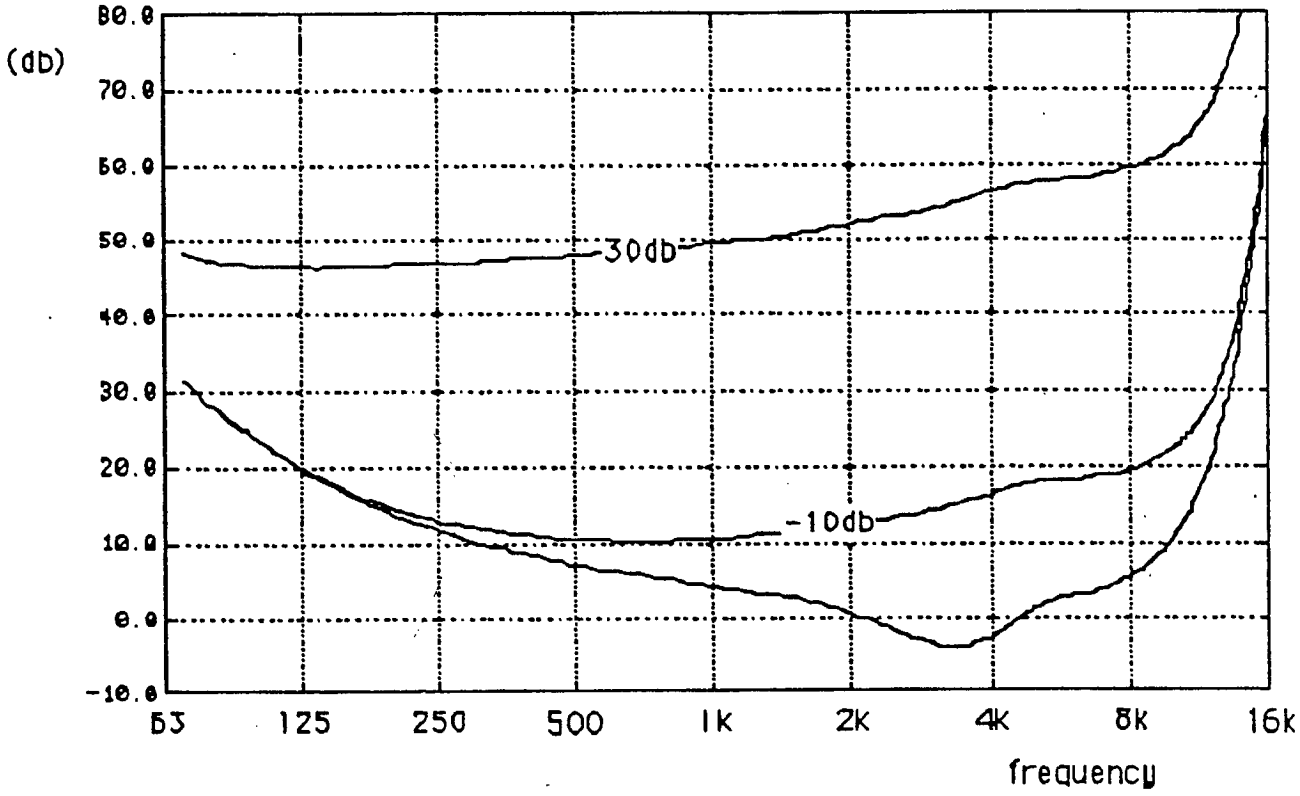


b) Simulated results

Figure 19 : detection thresholds for pure tones in the presence of lowpass and highpass noise maskers, with 1 Khz cutoff frequencies and 40 db spectral densities.

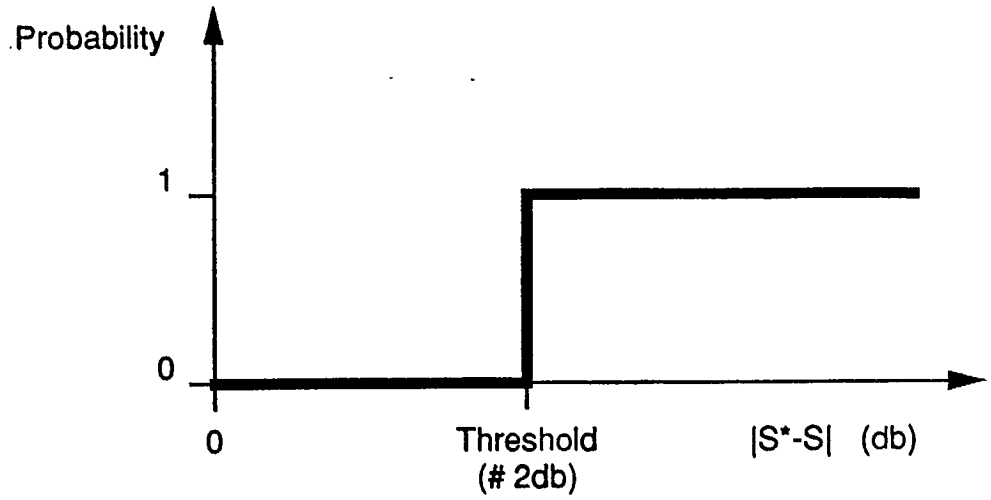


a) real results (from [6])



b) Simulated results

Figure 20 : detection thresholds for pure tones in the presence of white noise maskers, with spectral densities of -1 db and 30 db.

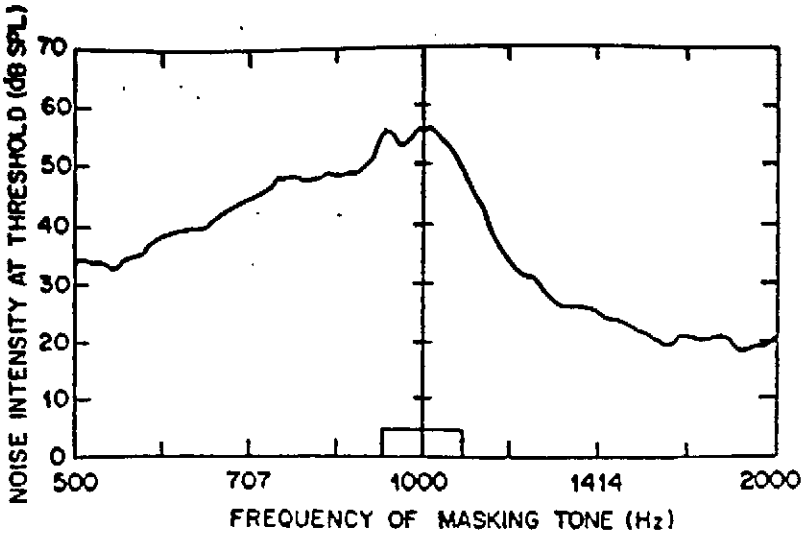


a) deterministic detection principle

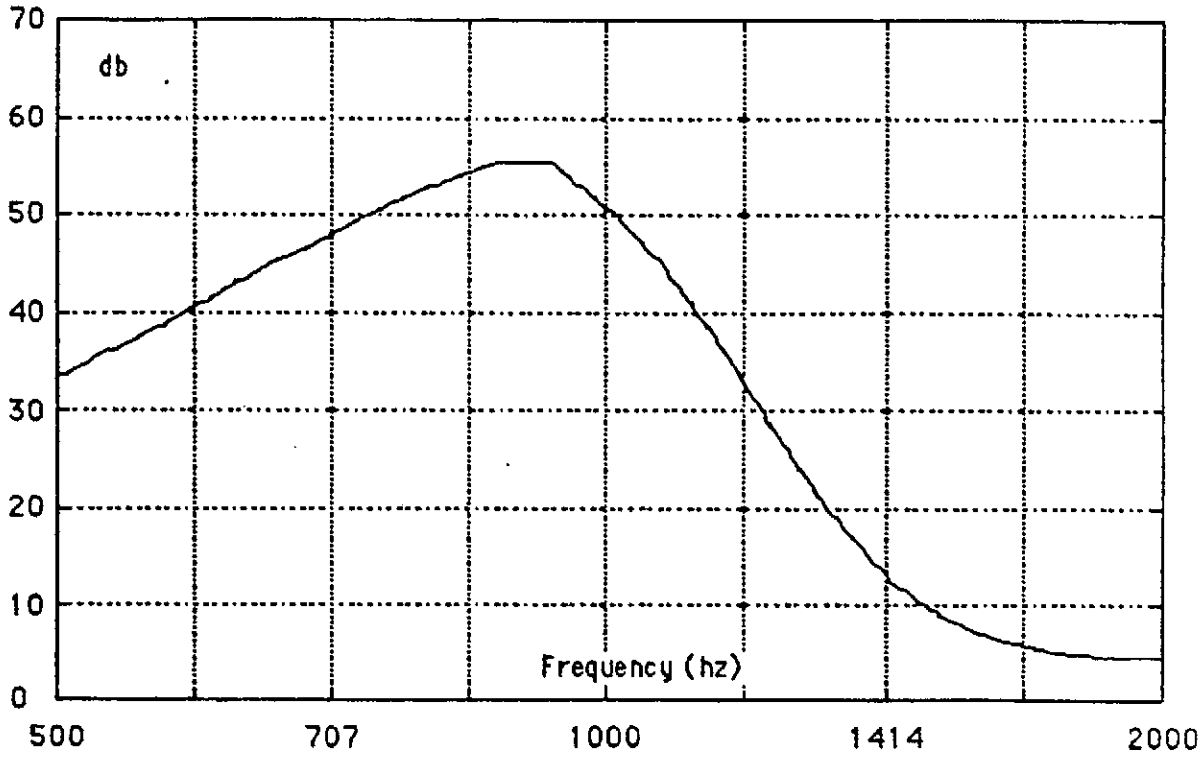
b) Statistical detection principle

Figure 21 : Detection probabilities v.s. absolute value of the difference of sensations, for one basilar detector, for the deterministic and the statistical detection principles.

S basilar sensation for the original signal
S* basilar sensation for the corrupted signal



a) results of Schroeder et al. (from [3])



b) Simulation results

Figure 22: detection threshold for a 1/3 octave noise, masked by an 80 db pure tone, as a function of the frequency of the pure tone.

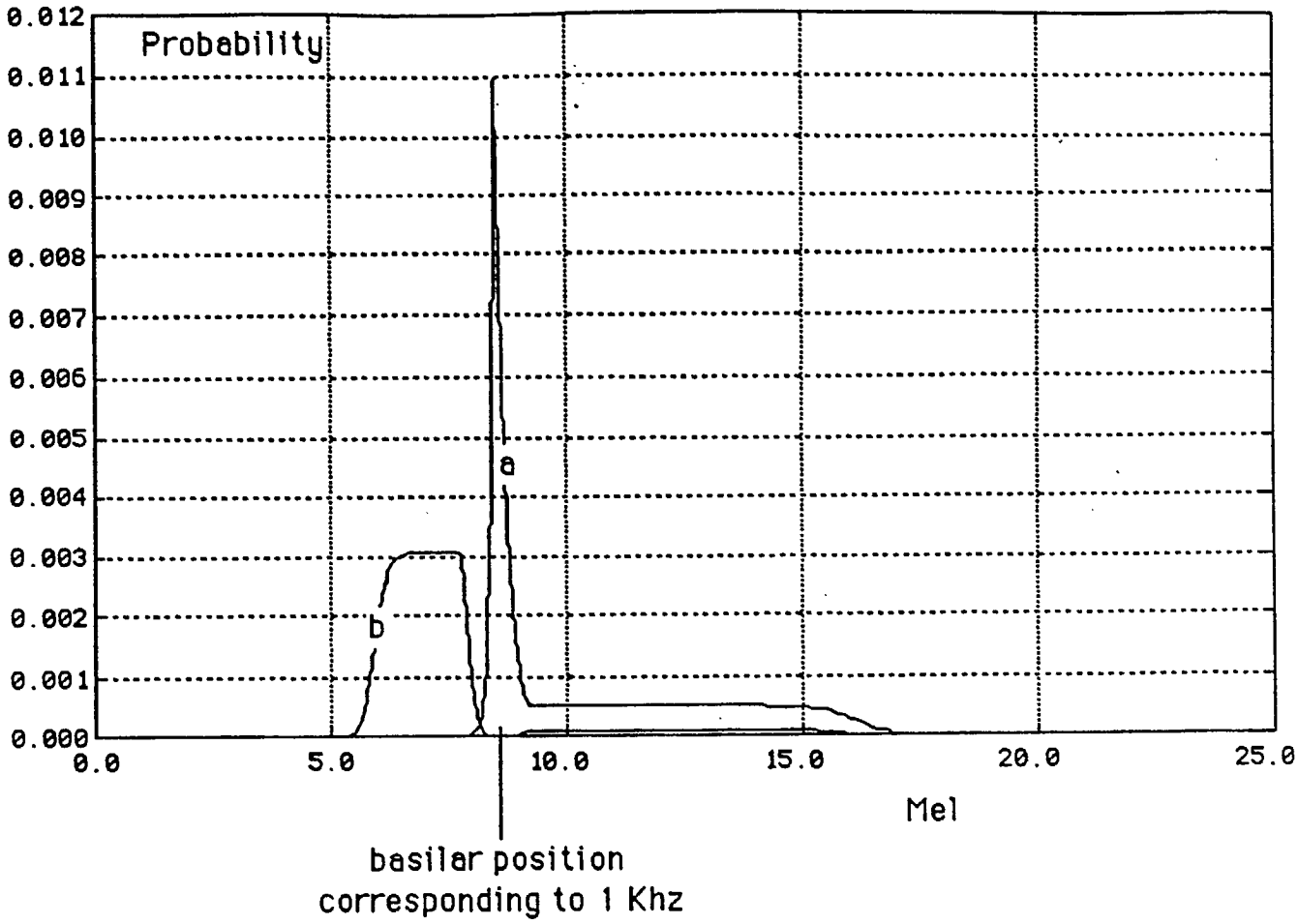


Figure 23 : detection probabilities for the basilar detectors, v.s. basilar position, at threshold (the global detection probability is 0.5)

a) for a 1 Khz pure tone masked by a 1/3 octave noise with a 1 Khz center frequency.

b) for a 1 Khz center frequency, 1/3 octave noise, masked by a 1 Khz center frequency.

	JAZZ			BACH			VOCALS			HARP		
STRATEGY	1	2	3	1	2	2	1	2	3	1	2	3
	RSB (dB)			RSB (dB)			RSB (dB)			RSB (dB)		
Philippe	40	8	20	35	5	27	45	8	30	48	10	23
Claude	30	10	20	30	3	20	48	8	23	53	5	25
François	33	6	33	38	3	35	48	3	19	60	8	38
Bruno	30	13	18	35	3	20	53	5	28	53	10	25
Alain	33	10	33	33	6	28	48	5	28	53	8	33
Perceval	33	10	18	35	10	20	45	10	18	48	12	20

	ORGAN			SYNTHETIC HF			SYNTHETIC BF		
STRATEGY	1	2	3	1	2	2	1	2	3
	RSB (dB)			RSB (dB)			RSB (dB)		
Philippe	43	13	23	63	23	23	75	18	23
Claude	40	13	30	63	23	30	73	15	25
François	48	13	35	70	23	28	73	18	20
Bruno	38	13	28	65	23	30	73	13	18
Alain	38	10	38	63	18	33	73	13	18
Perceval	40	10	20	60	10	17	68	7	15

Table 1 - Results from subjective tests.

ANNEXE 2

RAPPORT INTERNE DU CRCS

préparé dans le cadre de ce contrat

Pourquoi le spectre de bruit devrait suivre le spectre du signal ?

1 Introduction :

Comment injecter le maximum d'énergie de bruit dans un signal, de manière à ce que ce bruit reste inaudible ?

Certains groupes ([3], [4], [5], [6]) arrivent à la conclusion que le spectre d'énergie du bruit devrait pour celà suivre le *seuil de masquage fréquenciel* calculé à partir du spectre d'énergie du signal ([9]).

Il faut noter 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 (ou à la limite un bruit à bande étroite) 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.

En général cependant, dans le cas du codage par exemple, 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. La stratégie qui consiste à donner au spectre de bruit la forme du seuil de masquage 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.

Bruno Paillard - Octobre 90

Or cette hypothèse est infirmée par l'expérience ([7]), ainsi que par notre modèle d'audition ([1], [2]).

L'objet de ce mémo est de montrer qu'en faisant d'autres hypothèses et simplifications, le modèle d'audition développé au CRCS (décrit dans [1] et [2]) prédit que le spectre d'énergie du bruit devrait suivre exactement le spectre d'énergie du signal, de manière à être minimalement audible tout en transportant l'énergie maximum.

2 Démonstration :

La 1ere hypothèse faite est que le bruit est additif, et non corrélé au signal. De cette manière, le spectre d'énergie du signal bruité est égal à la somme des spectres d'énergie du signal et du bruit.

La 2eme hypothèse faite est que la détection du bruit a lieu dès que la sensation basilaire du signal bruité diffère de celle du signal original d'une valeur supérieure à un seuil constant (1 à 2 db) en un point au moins de la membrane basilaire. Cette 2eme hypothèse est une approximation qui s'avère relativement bonne ([2]).

Notations :

On désigne par

- F Les spectres d'énergie
- A Les spectres d'énergie atténués
- L Les densités basilaires d'énergie localisées
- D Les densités basilaires d'énergie dispersées
- S Les sensations basilaires

Ces différents vecteurs correspondent aux résultats d'étapes intermédiaires du modèle d'audition OREILLE, et sont décrites dans [1]

Plusieurs indices sont attribués à ces vecteurs :

- o Indique que le vecteur correspond au signal original
- b Indique que le vecteur correspond au signal de bruit
- ob Indique que le vecteur correspond au signal bruité

- L Indique que le vecteur correspond à une limite

Par exemple :

- D_{obL} est la densité basilaire limite du signal bruité
- A_o est le spectre d'énergie atténué du signal original

Comme on voit sur la figure 1, à partir de la sensation basilaire S_o dûe au signal original, on construit une sensation basilaire limite S_{obL}. Cette sensation limite est obtenue en ajoutant le seuil constant (1 à 2 db) en tous points de la sensation originale. La solution cherchée est alors le spectre de bruit F_b transportant le maximum d'énergie, tout en assurant que la sensation basilaire du signal bruité S_{ob} reste en tous points inférieure ou égale à la sensation limite S_{obL}, de manière à ce que le bruit soit indétectable.

Si on suppose que l'énergie en tous points de la membrane basilaire est grande par rapport à l'énergie interne des détecteurs basilaires, on peut négliger cette énergie interne et exprimer la relation précédente en échelle linéaire (figure 2) :

A partir de la densité basilaire d'énergie originale D_o, on construit une densité basilaire limite D_{obL}. Cette densité basilaire limite est obtenue en multipliant la densité basilaire D_o par un coefficient $k = 10^{\frac{seuil}{10}}$ où "seuil" est pris égal à 1 ou 2 db.

La solution cherchée est alors le spectre de bruit F_b transportant le maximum d'énergie, tout en assurant que la densité basilaire d'énergie du signal bruité D_{ob} reste en tous points inférieure ou égale à la densité limite D_{obL}, de manière à ce que le bruit soit indétectable.

On sait que la densité basilaire du signal bruité D_{ob} est égale à la somme de la densité basilaire du signal original D_o et de la densité basilaire du bruit seul D_b (linéarité de la transformation énergie fréquencielle -> énergie basilaire).

On sait que la densité basilaire du signal bruité D_{ob} ne doit pas dépasser une densité basilaire limite D_{obL}, obtenue en multipliant la densité basilaire originale D_o par un coefficient k.

$$D_{ob} \leq kD_o \qquad (1)$$

$$D_o + D_b \leq kD_o \quad (2)$$

$$D_b \leq (k-1)D_o \quad (3)$$

$$D_b \leq D_{bL} \quad (4)$$

La densité basilaire du bruit seul D_b doit donc, en tous points, rester inférieure à une densité basilaire limite D_{bL}, elle aussi proportionnelle à la densité basilaire originale D_o. $D_{bL} = (k-1)D_o$ (figure 3).

Bien entendu, à cause de la linéarité de la transformation énergie fréquencielle -> énergie basilaire, nous savons que le spectre d'énergie de bruit F_{bL} donnant cette densité basilaire limite D_{bL} est proportionnel (avec un coefficient de proportionnalité égal à k-1) au spectre du signal original F_o $F_{bL} = kF_o$.

Il reste à montrer que c'est bien ce spectre de bruit F_{bL} qui transporte le maximum d'énergie. Autrement dit, il faut montrer qu'il n'existe pas de spectre de bruit F_b transportant plus d'énergie que F_{bL}, qui soit tel que sa densité basilaire d'énergie D_b soit en tous points inférieure ou égale à la densité limite D_{bL}.

De manière équivalente, nous allons montrer que si un spectre d'énergie de bruit F_b donne une densité basilaire inférieure ou égale à la densité limite D_{bL}, il transporte nécessairement moins d'énergie que F_{bL}.

L'énergie totale contenue dans une densité basilaire d'énergie correspond à la composante continue du signal *densité basilaire d'énergie*

Comme la densité basilaire dispersée D est obtenue par filtrage à partir de la densité basilaire localisée L, l'énergie totale contenue dans D est proportionnelle à l'énergie totale contenue dans L, indépendamment de la forme particulière de la densité d'énergie localisée L.

L'énergie totale contenue dans la densité d'énergie localisée L est strictement égale à l'énergie totale contenue dans le spectre d'énergie atténué A (définition de l'opération de localisation).

En supposant une fonction d'atténuation constante (égale à 1 pour toutes les fréquences), on arrive à la conclusion que l'énergie totale contenue dans le spectre F est proportionnelle à l'énergie totale contenue dans la densité basilaire dispersée D, quelle que soit la forme particulière de ce spectre F.

En conséquence, l'énergie totale contenue dans D_b est proportionnelle à l'énergie totale contenue dans F_b. Donc toute densité basilaire D_b qui est en tous points inférieure ou égale à la densité limite D_{bL} (et qui donc a une énergie totale inférieure à celle de D_{bL}) a un spectre de bruit correspondant F_b qui transporte une énergie totale inférieure à celle de F_{bL}.

Le spectre d'énergie non audible qui transporte le maximum d'énergie est donc bien F_{bL}, proportionnel au spectre du signal original F_o.

3 Arguments intuitifs si on considère que la fonction d'atténuation n'est pas constante:

On peut tenter d'étendre la validité du résultat précédent au cas où la fonction d'atténuation n'est pas constante, mais est lentement variable.

Pour celà, on notera tout d'abord que l'opération de dispersion est une intégration relativement localisée (l'énergie apparaissant en un point "b" de la densité basilaire dispersée D ne fait intervenir de manière notable que les valeurs de la densité basilaire localisée L, sur un court segment autour du point."b").

En conséquence, une densité basilaire dispersée D_b en tous points inférieure ou égale à la densité limite D_{bL}, aura non seulement une densité basilaire localisée L_b d'énergie totale inférieure à celle de la densité basilaire localisée limite L_{bL} (tel que démontré au paragraphe précédent), mais de plus, cette densité basilaire localisée L_b devrait avoir une énergie *locale* (somme de l'énergie sur un court segment basilaire quelconque) toujours inférieure ou égale à l'énergie *locale* de la densité basilaire localisée limite L_{bL}.

Ce résultat peut être étendu à la densité fréquencielle atténuée A_b, puisqu'il y a une relation conservant l'énergie entre L_b et A_b :

De manière à avoir une densité basilaire dispersée D_b inférieure ou égale à la densité limite D_{bL}, la densité fréquencielle atténuée A_b devra non seulement avoir une énergie totale inférieure ou égale à celle de A_{bL}, mais de plus l'énergie locale de A_b, calculée sur de courts segments fréquenciels, devra être inférieure ou égale à l'énergie locale de A_{bL}.

Note :

Le résultat précédent serait vérifié exactement si la fonction de dispersion était une fonction rectangle (constante sur une certaine longueur basilaire, et nulle en dehors de ce segment), et en considérant des *courts segments basilaires* de longueur **supérieure ou égale** à la longueur du rectangle.

Pour que le résultat précédent soit vérifié avec une certaine précision, la longueur d'un segment basilaire considéré *court* doit donc tout de même être plus importante que la longueur sur laquelle la fonction de dispersion a des valeurs non négligeables (la longueur basilaire du sommet de la fonction de dispersion).

Dans l'espace basilaire, cette longueur minimum que doit avoir un *court segment basilaire* est donc constante et indépendante de la position basilaire. Par contre, dans l'espace fréquenciel, un *court segment fréquenciel* correspond à une longueur basilaire minimum constante, et aura donc une longueur fréquencielle proportionnelle (voire égale) à la largeur d'une bande critique en ce point (la largeur d'une bande critique est définie justement comme la plage fréquencielle correspondant à la longueur basilaire sur laquelle la fonction de dispersion a des valeurs non négligeables [10])

Etant donné que pour obtenir une densité basilaire dispersée D_b inférieure ou égale à D_{bL}, la densité fréquencielle atténuée correspondante A_b doit avoir une énergie locale (calculée sur de *courts segment fréquenciels* ayant la largeur des bandes critiques) toujours inférieure ou égale à celle de A_{bL}, et si on considère que ces *courts segments fréquenciels* sont suffisament courts pour que la fonction d'atténuation puisse être considérée comme constante sur tout le segment, on peut calculer les énergies totales de F_b et F_{bL} segment par segment, et en conclure que l'énergie totale transportée par F_b est inférieure ou égale à celle transportée par F_{bL}.

F_{bL} est donc, dans ce cas encore, le spectre d'énergie de bruit non audible, transportant le maximum d'énergie.

Attention :

Comme on vient de le voir, ce résultat repose sur l'hypothèse que sur des segments fréquenciels de longueurs correspondantes aux bandes critiques, la fonction d'atténuation peut être considérée comme constante. Cette hypothèse n'est pas vérifiée dans la zone haute fréquence du spectre,

où les bandes critiques deviennent très larges en même temps que la fonction d'atténuation varie très rapidement.

Reférences

[1] B. Paillard, "Description du modèle d'audition OREILLE" - Mars 90 - CRCS Rapport interne

[2] B. Paillard, "OREILLE : Principe de détection" - Mars 90 - CRCS Rapport interne

[3] 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.

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

[5] 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.

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

[7] Brian C. J. Moore. "An Introduction to the Psychology of Hearing", Academic Press, 1989, ISBN 0-12-505623-0

[8] Brian C. J. Moore. "Frequency selectivity in hearing", Academic Press, 1986, ISBN 0-12-505625-7

[9] E. Zwicker, R. Feldtkeiler - "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

[10] B. Paillard - "Critique des bandes critiques" - Octobre 89 - CRCS, Rapport interne

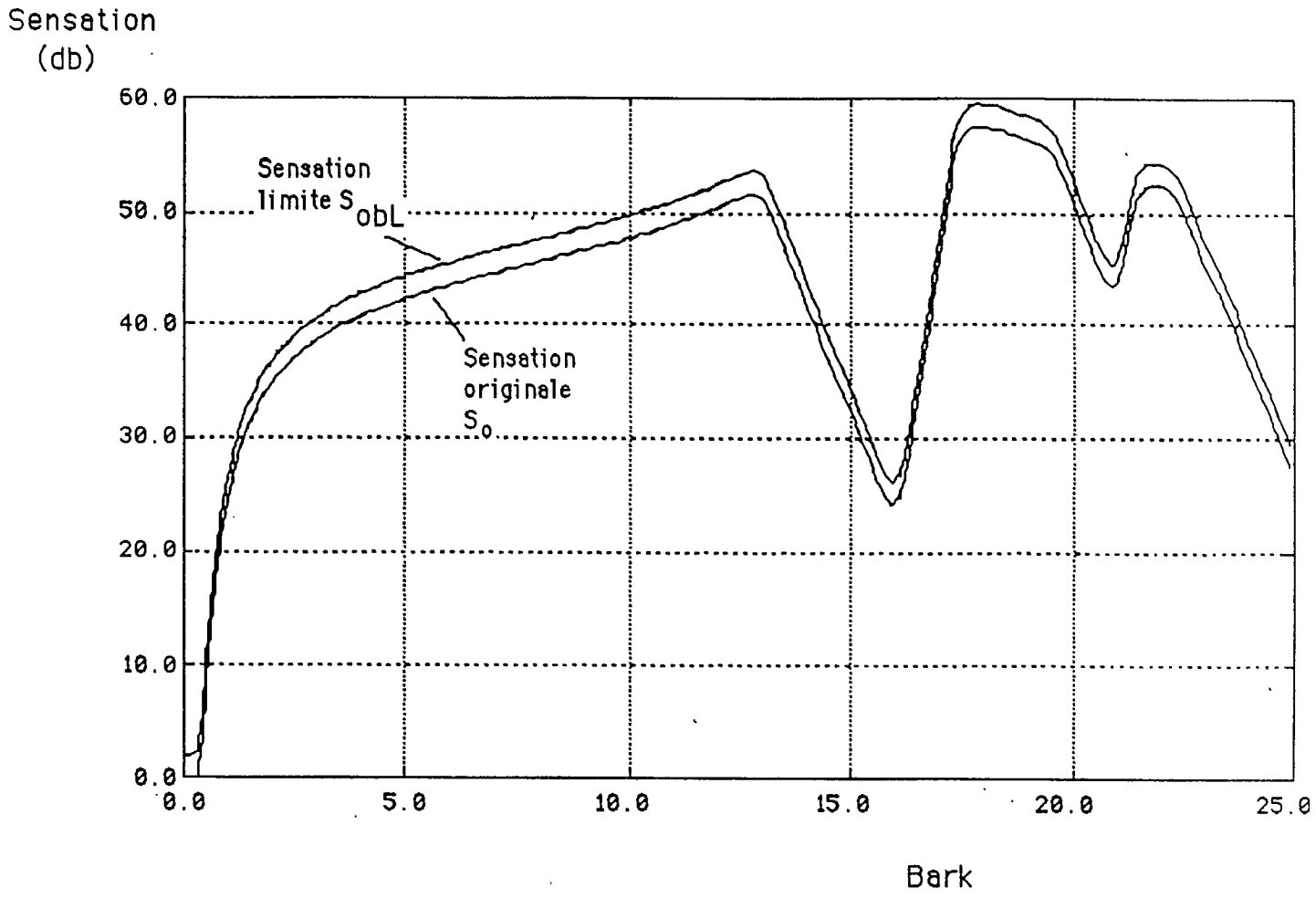


Figure 1 : Construction de la sensation limite
a partir d'une sensation originale

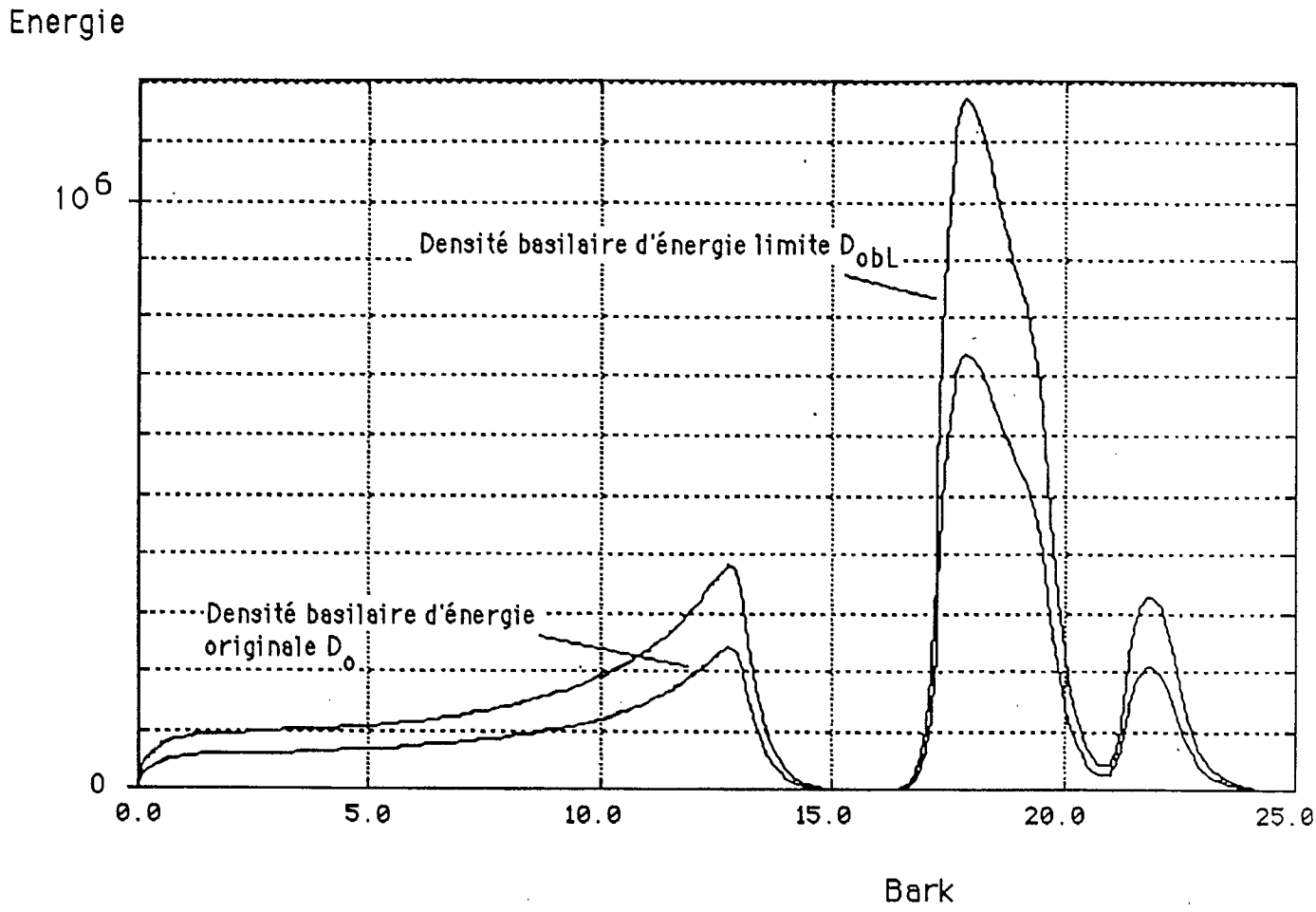


Figure 2 : Construction de la densité basilaire d'énergie limite du signal bruité, à partir de la densité basilaire d'énergie du signal original.

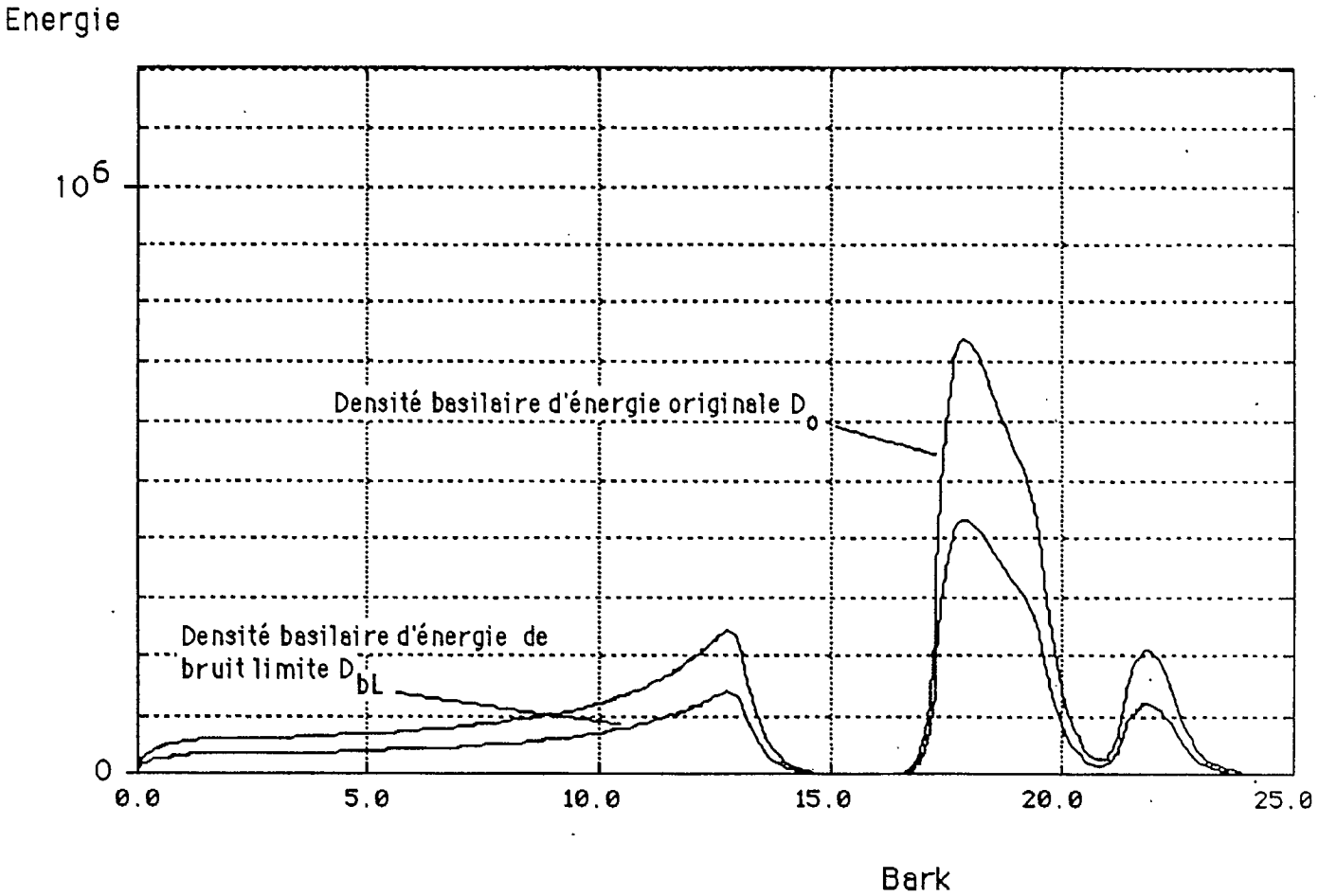


Figure 3 : Construction de la densité basilaire d'énergie limite du bruit, à partir de la densité basilaire d'énergie du signal original.

ANNEXE 3

Cette annexe présente des résultats d'expérience de coloration de bruit publiés à la 90ième Convention d'AEAS

Preprint 3010 (A-3)

A study of strategies for the perceptual coding of audio signals.

A. Turgeon, J. Soumagne, P. Mabilleau, S. Morissette, B. Paillard
University of Sherbrooke, Faculty of Applied Sciences, Electrical Engineering Dept., Sherbrooke, Canada

Presented at the 90th Convention 1991 February 19-22 Paris

AES

AN AUDIO ENGINEERING SOCIETY PREPRINT

A STUDY OF STRATEGIES FOR THE PERCEPTUAL CODING OF AUDIO SIGNALS

Alain Turgeon, J. Soumagne, P. Mabilleau, S. Morissette, B. Paillard

Sherbrooke University,

Sherbrooke, Quebec, Canada

J1K 2R1

Abstract

To take into account the frequency masking properties of the ear for the coding of audio signals, the strategy which is usually used is to shape the spectrum of the noise introduced by the coding process, according to a frequency masking threshold. The paper describes an experiment which compares different strategies, and shows that, provided the frequency resolution is adequate, it is almost always better to shape the noise spectrum according to the shape of the signal spectrum. This result had been anticipated by the behaviour of an auditory model developed by our group.

1. Introduction

In the last few years, efforts have been made to reduce the numerical rate of digital audio signals. The tendency is to take advantage of the frequency masking properties of the ear to make inaudible the noise introduced by the coding process. The usual strategy consists in shaping the noise spectrum according to a frequency masking threshold derived from a short term energy spectrum of the audio signal.

This masking threshold approach is valid for narrow bands of noise or pure tones. Indeed, the threshold indicates the energy at which a narrow band of noise is barely detectable in the audio signal. But the quantization noise is comparable to a wideband noise so that the hypothesis which is made implicitely is that the effect of a wideband noise on the ear is equivalent to the superposition of the independent effects of many adjacent narrow bands of noise.

Results from an auditory model developped by our group suggest that this is not the case, and that other much simpler coding strategies such as shaping the noise spectrum according to the signal spectrum might be better. The paper describes an experiment which shows that, provided the spectral resolution is adequate for the shaping of the noise, this simpler strategy is indeed better than the usual one.

Thias experiment consisted in injecting noise in a signal with a constant signal to noise ratio, and with a noise spectrum which was dependent on (and therefore dynamically adapted to) the short term energy spectrum of the signal.

The noise was injected in the signal according to three types of noise shaping corresponding to (simulating) three different coding strategies. To be able to control accurately the energy spectrum of the noise, we used a time/frequency decomposition of the signal (transform coding or subband decomposition) and the noise was injected in the frequency domain. It is clear then that this experiment simulates closely the behaviour and the kind of noise encountered with transform coders.

In the first section, the interest of using a time-frequency decomposition is discussed and the characteristics searched for in the use of a transform are stated. Then, three strategies for shaping the noise spectrum are described and the results of the experiments are presented.

2. Transform and subband decomposition

Transform coding is widely used for reducing the quantity of information necessary to reproduce high quality audio signals. The major aspects of this coding technique are covered in [1]. The word "transform" must be taken in a more general sense than usual since subband decomposition, polyphase filters or time

domain aliasing cancellation (TDAC) can serve as a substitute for segmentary transforms ("Discrete Fourier Transform", "Discrete Cosine Transform", etc...). Moreover, a tendency to regroup those different time/frequency decompositions into the same formal description is explored in [2], [3]and [4]. Among the coding systems existing up to now, we find the MUSICAM coding scheme [5] with Quadrature Mirror filter banks for time/frequency decomposition, the OCF system [6] with TDAC and the SEPXFM system [7] with a DFT.

Originally, the purpose of using time/frequency transforms was to concentrate the signal energy onto the minimum number of components so as to take advantage of the redundancy existing in the signal. This characteristic becomes less important however when trying to take advantage of the perceptual properties of the ear. The choice of a transform is influenced by the following characteristics:

- frequency resolution of the transform;
- possibility of an error-free reconstruction of the temporal signal from the spectral coefficients;
- possibility of a fast algorithm for realizing the analysis and synthesis filter;
- orthonormality of the transform.

The first characteristic is very important for the success of the frequency masking approach. The decomposition must have a sufficient number of frequency components to represent accurately the energy spectrum of the signal and then take advantage of the masking phenomenon. Also, the analysis filters must show an abrupt transition band and a good attenuation in the stopband. The third characteristic ensures a real time application on a digital signal processor (DSP) and the last one ensures the conservation of the energy between time and frequency domain. In particular, it ensures that the energy of the noise introduced in the frequency domain (due to a quantization of the spectral coefficients) is equal to the energy of the noise in the time domain (on the reconstructed signal).

In [8], the author suggests a transform called modulated lapped transform (MLT). This transform is a good choice in view of the characteristics that it offers. Fig. 1a shows the spectrum of a filter centered at 10 kHz and fig. 1b, the spectrum of three adjacent filters. We chose a 1024 bands decomposition for a sampling frequency of 44 kHz, thus the bandwidth of each filter is approximatively 22 Hz.

3. Strategy in shaping the noise spectrum

3.1 Structure of the noise injection system

Fig. 2 represents the block diagram showing the injection of noise in a signal. First, the signal is analyzed by an MLT (Modulated Lapped Transform) of 1024 bands. The impulse response of each filter has 2048 samples. The overlapped window is formed from 1024 new samples and 1024 samples belonging to the previous block, thus justifying the use of the word "lapped". In the second step, a short term energy spectrum is evaluated by a sliding time average on three analysis windows (70 ms for f_S= 44,056 kHz) of each spectral component. This spectrum is used to calculate the energy of the noise for a fixed signal to noise ratio (SNR). The noise added at each spectral component is derived from a random sequence of numbers whose range is determined by the noise energy. The random number generating function is described by the uniform probability law. This way, the process simulates accurately the quantization of the spectral components with uniform quantizers.

In the last step, an inverse MLT is done on the noisy components to retrieve the noisy time signal. The SNR is conserved because the transform is orthonormal.

3.2 Description of strategies

We have experimented three strategies of noise injection. In spite of an obvious poor performance, in the perceptual sense, the injection of white noise (uniform spectrum) is retained as our first strategy. It can serve as a reference for the other strategies. For the second strategy, the noise spectrum is made proportional to the signal spectrum. This would correspond to a constant bit allocation with adaptative quantizers [1]. The third strategy uses a frequency masking threshold given by an auditory model developed by our group. The noise spectrum is then proportional to this masking threshold. Fig. 3, 4 and 5 represent some examples of each strategy.

In fig. 5, the masking threshold derived from the signal spectrum is shown.

For the listening tests presented in the third section, the bandwidth of the music file is 22,028 kHz. The noise spectrum added fills all the bandwidth for the first and second strategies and is limited to the 0-15 kHz band in the case of third strategy. The reason for this limitation is that the masking threshold, and thus the energy level of the noise in the 15-22 kHz band is so high compared with the signal energy that no coding system would behave this way anyway.

4. Results of the listening tests

The listening tests are made with selected excerpts of music provided by a Sony PCM (model 701es). The sampling frequency is fixed at 44,056 kHz. For the listener, the test consists in detecting a pulsated noise in a series of musical excerpts for different signal to noise ratios of the same excerpt. Three answers were suggested to the listener, i. e. "yes, I heard the noise", "no, I don't" or "may be". The answer "may be" is used when a degradation is detected but the characteristic pulsation of the noise is not audible.

There is a practical reason to use a pulsated noise. In psychoacoustical testing, the masked noise is often pulsated to avoid an adaptation phenomenon of the ear. In our case, it is done to avoid the confusion with the noiselike parts that can be inherent in the music. For some kinds of music, it is not easy to detect a continuous noise if the signal is itself "noiselike". The pulsating noise ensures that the listener detects the noise added to the original music.

Table 1 shows the results from five listeners for different kinds of music. For the three strategies, it indicates the SNR in decibel (db) for the audibility limit of the noise, i.e. for the case where the listeners have answered "may be". The maximum uncertainty is about five db, i.e. at five db above the threshold, the noise is not detected by the listener with certainty and at five db under the threshold, the noise is detected with certainty.

The results show that for the second strategy, an audibility limit of 12 db is acceptable in general except for the files WNHF4100 and WNBF4100 where the spectrum looks like a pure tone. In fact, the performance tends to be worse in the case of pure tone. With this limit (12 db), the authors in [1] assumed that a coding can be done with 2 bits/sample (excluding side information).

The most important result is the fact that the second strategy seems to be better, in the perceptual sense, than the third one. It is also more interesting because its complexity is lower than for the third strategy, an important consideration when a real time coder has to be implemented on a DSP.

5. Conclusion

It seems that the strategy giving a noise spectrum proportional to the signal spectrum is more optimal than the one using a frequency masking threshold. This result is confirmed by subjective experiments and is valid when the time-frequency decomposition offers a good frequency resolution.

Further work is currently being carried out to develop a coder using the strategy described in this paper.

Acknowledgement

The authors would like to thank the CRSNG of Canada for the financial contribution.

References

[1] N. S. Jayant and P. Noll, *Digital Coding of Waveforms*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

[2] M. Verterli and D. LeGall, "Perfect Reconstruction FIR Filter Banks: Lapped Transforms, Pseudo QMF's and Paraunitary Matrices," *Proc. of 1988 ISCAS*, pp. 2249-2253, 1988.

[3] P. P. Vaidyanathan and 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, no. 3, March 1988.

[4] B. Paillard, J. Soumagne, P. Mabilleau et S. Morissette, "Décomposition en sous-bandes: une nouvelle approche analytique", *Traitement du signal*, vol. 5, no. 3, 1988.

[5] G. Stoll, M. Link and G. Theile, "Masking-pattern adapted subband codeing: use of the dynamic bit rate margin", *84th Convention of AES*, Preprint 2585 (D-5), March 1988.

[6] K. Branderburg, "High quality coding at 2.5 bit/sample", *84th Convention of AES*, Preprint 2582 (D-2), March 1988.

[7] J. D. Johnston, "Perceptual transform coding of wideband stereo signals", Presented at *89th ICASSP*, May 1989.

[8] H. S. Malvar, "Lapped transform for efficient transform/subband coding", *IEEE Trans. on ASSP*, vol. 38, no. 6 , pp. 969-979, June 1990.

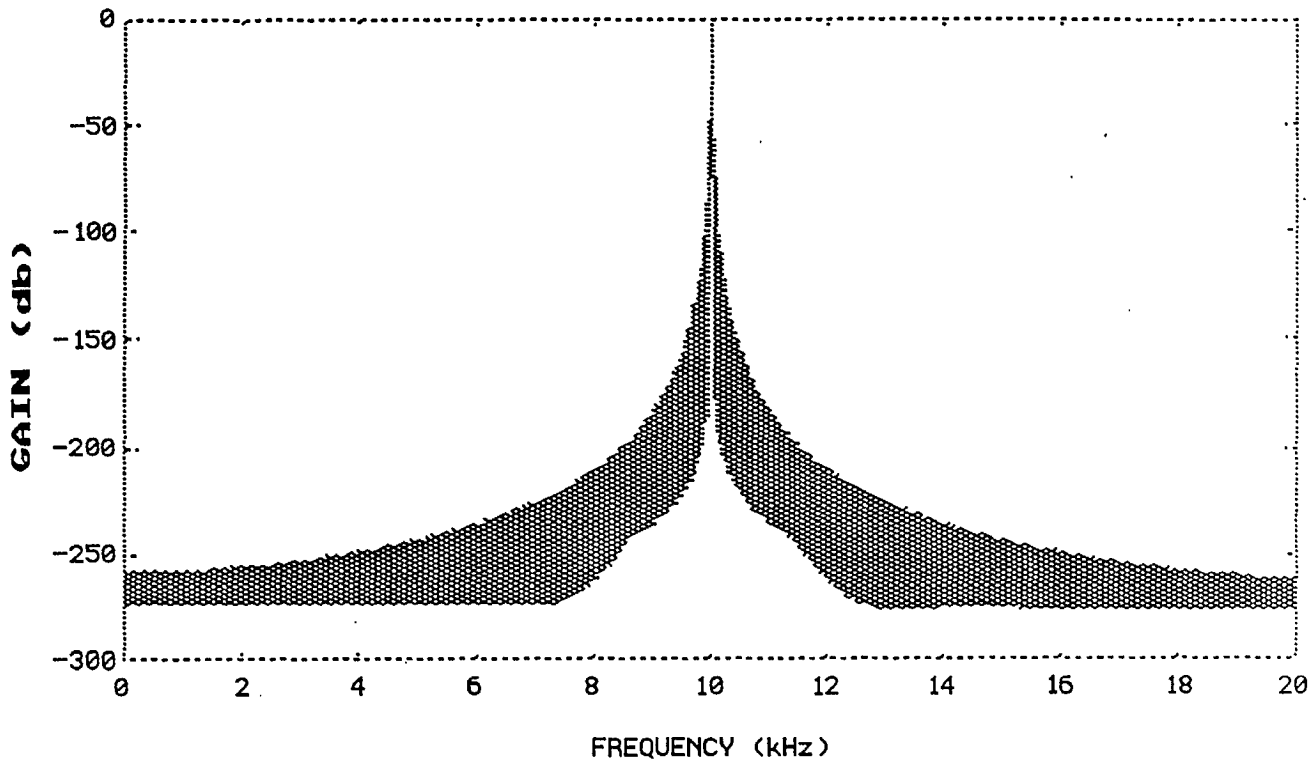


Figure 1a - Spectrum of a filter from a MLT (1024 bands).

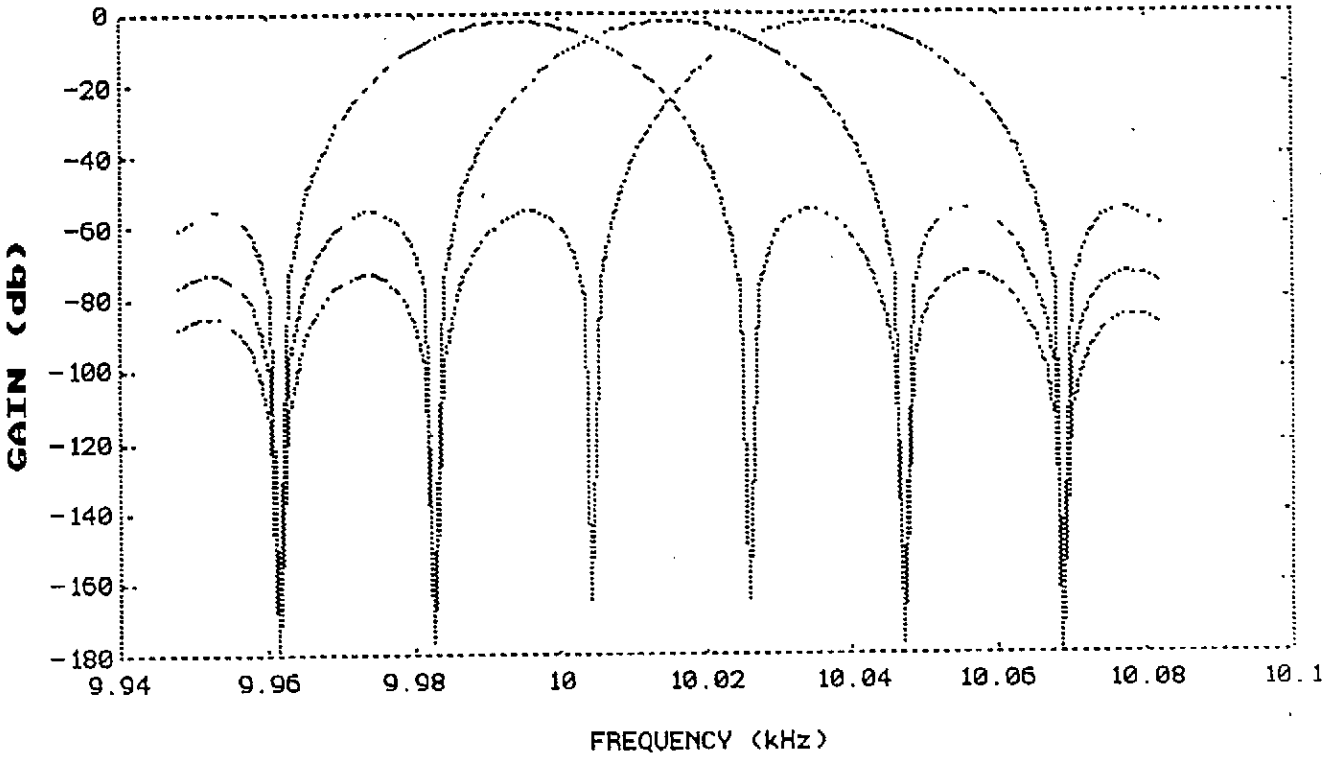


Figure 1b - Spectrum of three consecutive filters from a MLT (1024 bands).

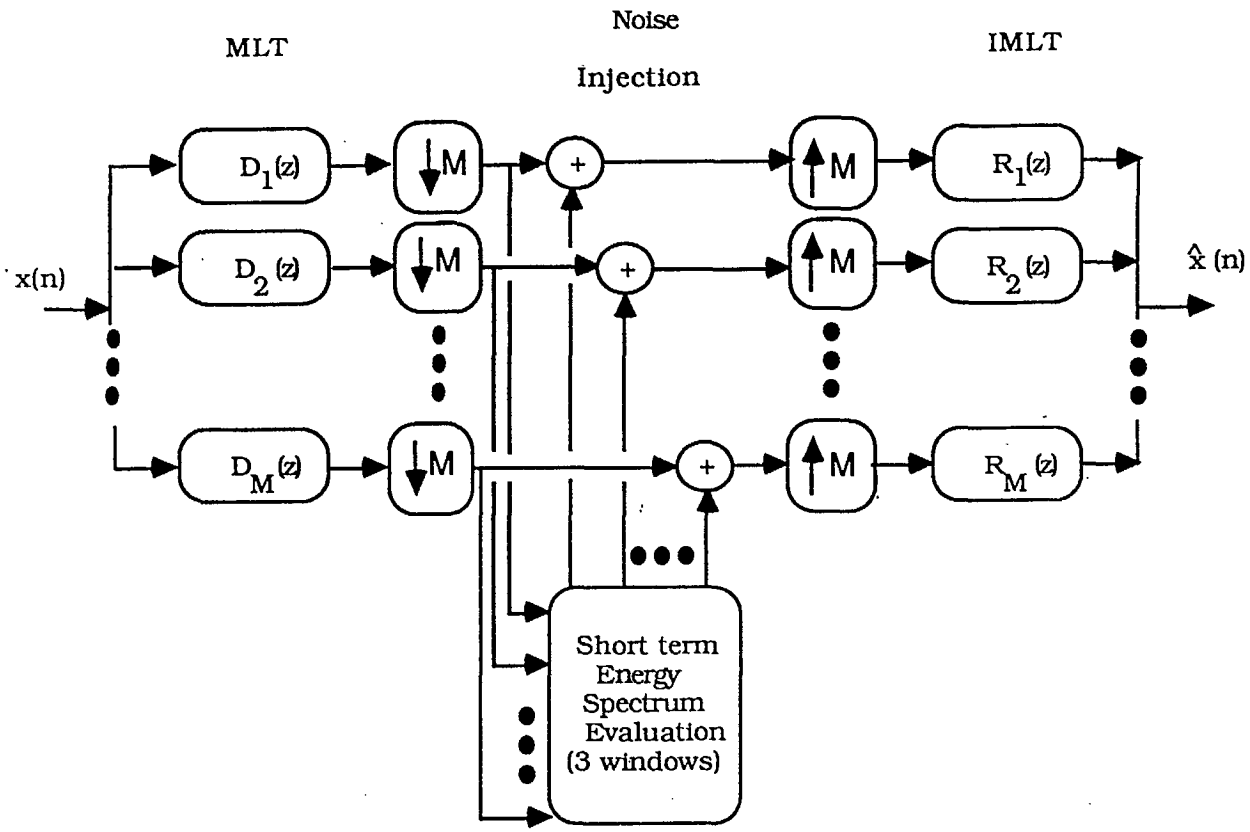


Figure 2 - Block diagram showing the injection procedure of noise in a signal.

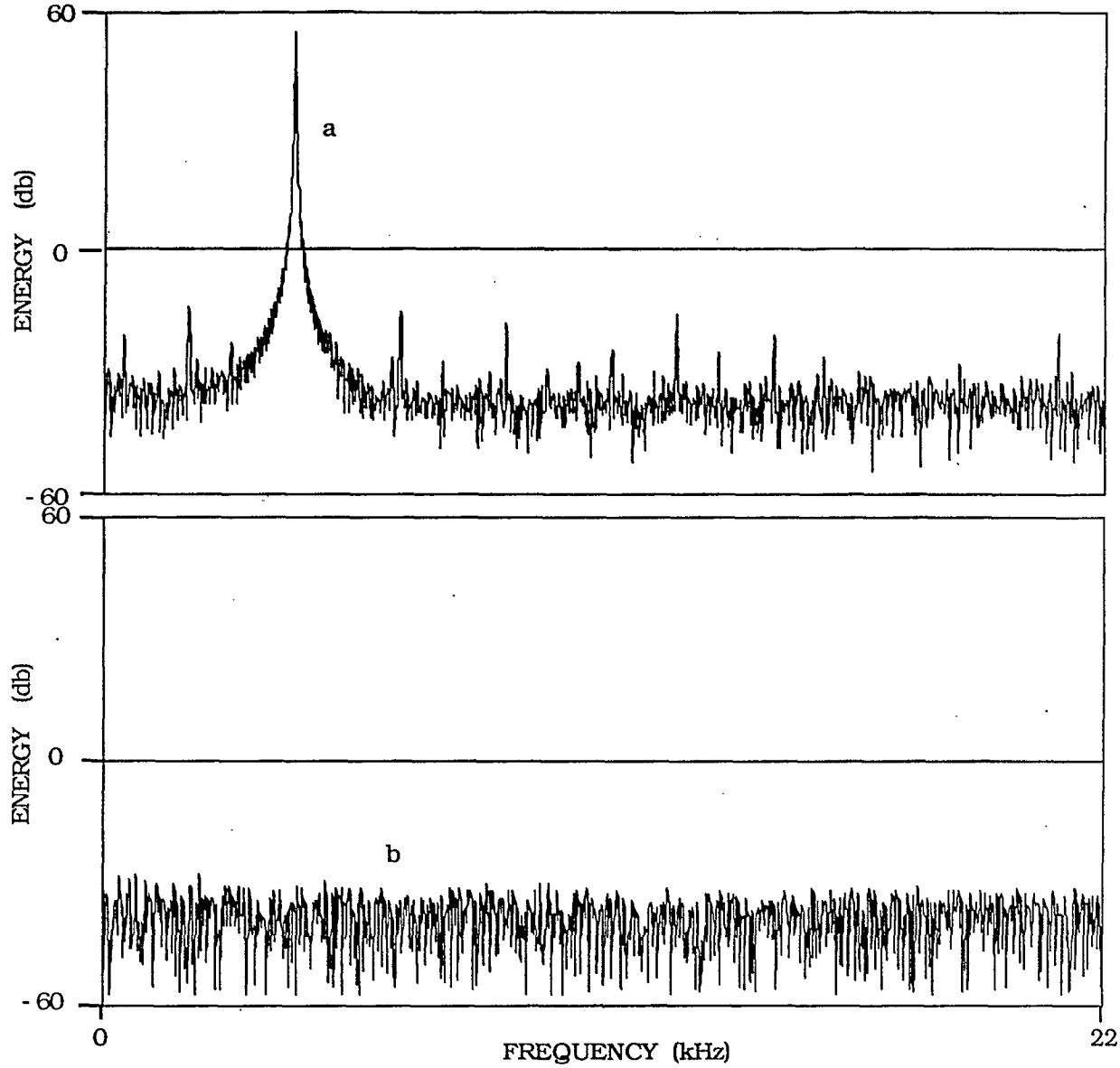


Figure 3 Energy spectrum for the first strategy (SNR = 65 db). The noise spectrum is flat.
a) Signal, b) Noise.

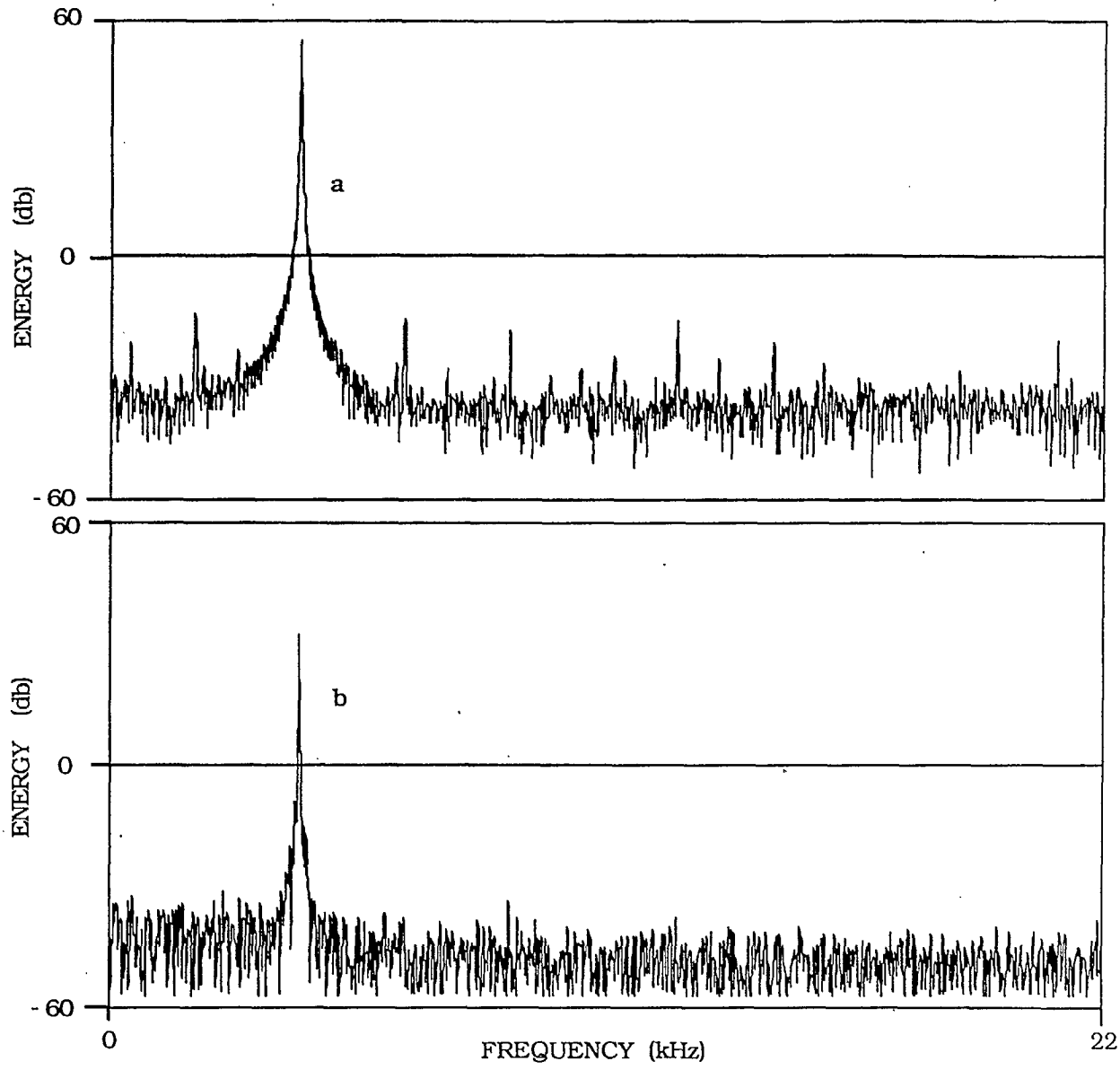


Figure 4 Energy spectrum for the second strategy (SNR = 25 db). The noise spectrum is proportional to the signal spectrum. a) Signal, b) Noise.

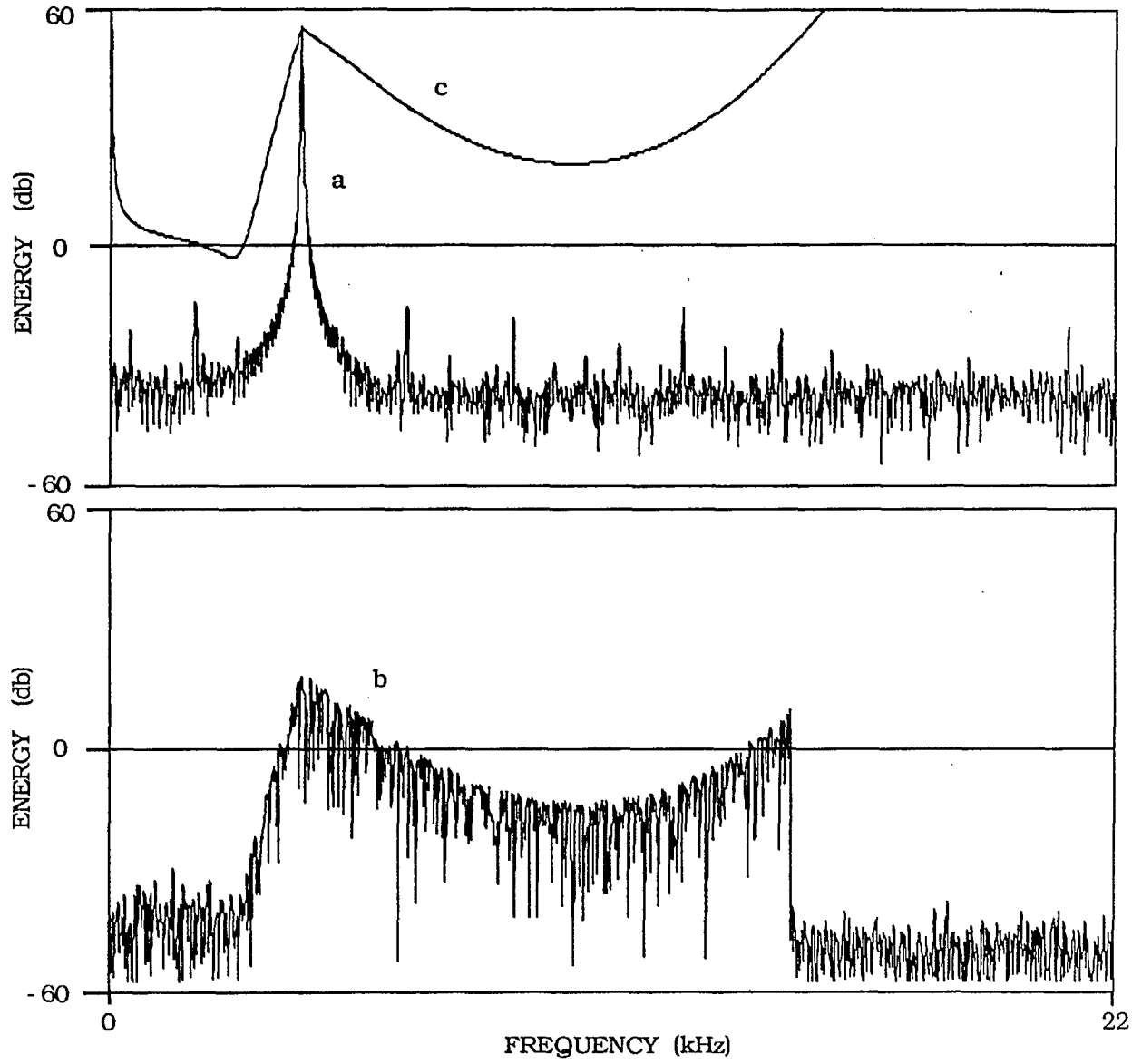


Figure 5 - Energy spectrum for the third strategy (SNR = 25 db). The noise spectrum is proportional to the frequency masking threshold. a) Signal, b) Noise, c) Frequency masking threshold.

	WJAZ4100			WBAC4100			WCHA4100			WHAR4100		
STRATEGY	1	2	3	1	2	2	1	2	3	1	2	3
	RSB (dB)			RSB (dB)			RSB (dB)			RSB (dB)		
Philippe	40	8	20	35	5	27	45	8	30	48	10	23
Claude	30	10	20	30	3	20	48	8	23	53	5	25
François	33	6	33	38	3	35	48	3	19	60	8	38
Bruno	30	13	18	35	3	20	53	5	28	53	10	25
Alain	33	10	33	33	6	28	48	5	28	53	8	33

	WORG4100			WNHF4100			WNBF4100		
STRATEGY	1	2	3	1	2	2	1	2	3
	RSB (dB)			RSB (dB)			RSB (dB)		
Philippe	43	13	23	63	23	23	75	18	23
Claude	40	13	30	63	23	30	73	15	25
François	48	13	35	70	23	28	73	18	20
Bruno	38	13	28	65	23	30	73	13	18
Alain	38	10	38	63	18	33	73	13	18

Table 1 - Results from subjective tests.