Étude comparative de filtres perceptuels adaptés à des codeurs audio

Marcos Perreau Guimaraes[1], Nicolas Moreau[2] et Madeleine Bonnet[1]

[1]Université René Descartes-Paris V UFR de Mathématiques et Informatique
45 rue des St Pères, 75270 Paris Cedex 06
perm,bonnet@math-info.univ-paris5.fr
[2]ENST/SIG,
46 rue Barrault, 75634 Paris Cedex 13
moreau@sig.enst.fr

RÉSUMÉ

Les codeurs audio de haute qualité utilisent souvent un modèle psychoacoustique pour prendre en compte les propriétés de l'oreille. On compare des filtres perceptuels, calculés à partir d'une prédiction linéaire, avec des filtres obtenus avec des seuils de masquage utilisés dans des codeurs de musique. Nous avons remarqué que ces derniers ne donnent pas de meilleurs résultats. Si la démarche la plus naturelle consiste à définir un meilleur modèle psychoacoustique, on propose ici une méthode intermédiaire consistant à donner plus de degrés de liberté à une méthode de type standard, en traitant individuellement les zéros du filtre blanchissant.

ABSTRACT

High quality music coders commonly use auditory masked thresholds to account for the characteristics of the human ear. Perceptual filters (based upon linear signal prediction used in speech coders) are compared to filters using masked thresholds. Using listening tests, we have noticed that the second method does not provide better perceptual results. A natural way of proceeding would be to define a better psycho-acoustical model. However, an intermediate method is presented here which allows additional degrees of freedom in a standard technique. The roots of the whitening filter are treated individually.

1 Introduction

Les codeurs de parole ou de musique actuels acceptent des taux de compression de l'ordre de 8 sans perte de qualité (au lieu de 2 en exploitant simplement la redondance statistique du signal) parcequ'ils prennent en compte largement des propriétés particulières de l'oreille. Les codeurs "perceptuels" cherchent à mettre en forme spectralement le bruit de quantification de façon à ce que ce bruit soit toujours en dessous d'un "seuil de masquage" [2, 10, 7]. Les codeurs de musique en bande Hi-Fi (f_e = 44.1 kHz) et à débit élevé (96 kbit/s) utilisent un modèle d'oreille sophistiqué pour distribuer les ressources binaires disponibles là où elles sont perceptuellement les plus utiles. Les codeurs de parole en bande téléphonique (f_e = 8 kHz) et à débit faible (8 kbit/s) utilisent un filtre perceptuel, proposé par Atal [1], qui est basé sur un modèle d'audition très simplifié mais efficace.

Pour des codeurs de parole en bande élargie (f_e = 16 kHz) à des débits intermédiaires (32 kbit/s), Ordentlich et Shoham [11] ont montré que le nombre de paramètres apparaissant dans l'expression du filtre perceptuel $W(z)$ présenté dans [1] est insuffisant pour avoir une bonne approximation d'un seuil de masquage. Ils corrigent le "tilt" spectral en multipliant $W(z)$ par un polynôme $T(z)$ avec un petit nombre de coefficients. Chang et Wang [3] proposent de calculer directement le filtre $W(z)$ à partir d'un seuil de masquage. Une implémentation de cette méthode dans un codeur de parole et de musique en bande FM (f_e = 32 kHz) à des débits de 64 kbit/s, réalisée par Murgia et al. [9], a montré une amélioration de la qualité des signaux reconstruits.

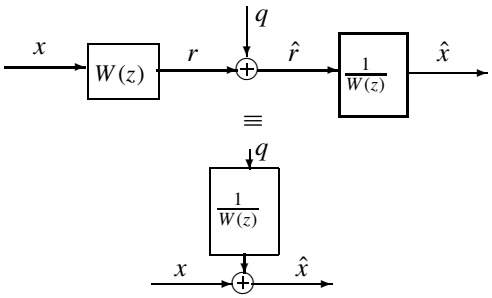

FIG. 1 — Schéma de principe

On compare ici les performances d'un certain nombre de filtres perceptuels calculés suivant différentes méthodes. Le but est de choisir un filtre perceptuel dans le cadre d'un codeur, actuellement en cours de développement, en bande FM et à débit variable (24-64 kbit/s) de type TCX [6]. Dans ce papier, on se limite à des conditions expérimentales relativement indépendantes d'un codeur. En particulier, on suppose que toutes les procédures de quantification peuvent être modélisées par un bruit blanc additif comme l'indique le schéma de la figure 1.

La section 2 rappelle les principes de bases de la psychoacoustique et décrit succinctement quatre modèles psychoacoustiques classiques. La section 3 décrit dans un premier temps les filtres perceptuels standards [1, 11] utilisés en co-

dage de la parole, puis elle décrit une méthode de calcul de filtres perceptuels à partir d'un seuil de masquage [3]. Dans la section 3.3, nous proposons une nouvelle méthode de calcul du filtre perceptuel en ajoutant des degrés de liberté au filtre perceptuel de [11]. Enfin dans la section 4 nous décrivons une procédure de comparaison subjective des différents filtres perceptuels.

2 Quelques rappels en psychoacoustique

Quand on parle de mise en forme spectrale du bruit de reconstruction, le but à atteindre est de trouver le bruit qui minimise les ressources binaires tout en restant inaudible, ou si on est à débit fixé, le bruit qui maximise la qualité subjective du signal reconstruit. On appelle seuil de masquage la densité spectrale de puissance (dsp) du bruit qui minimise le débit à la limite de la transparence. Ce seuil de masquage est le reflet des propriétés physiologiques du système auditif humain, expliquées et modélisées par la psychoacoustique.

Le calcul d'un seuil de masquage se fait en deux étapes. Le signal $x(t)$ est d'abord analysé par un banc de filtres, dits filtres cochléaires. Divers travaux [12, 5] proposent des résultats expérimentaux et des expressions analytiques pour ces filtres cochléaires. Dans une échelle de fréquences semi-logarithmique, les Bark, les modules au carré des réponses en fréquence notées $H_j(f)$ de ces filtres sont presque triangulaires (en décibels), et les fréquences centrales régulièrement disposées. Le maximum de $|H_j(f)|^2$, au sommet du triangle, vaut 1. De plus la forme de ces filtres varie avec la puissance du signal. La puissance des signaux sortant de chaque filtre cochléaire définit une analyse temps-fréquence du signal, appelée excitation, qui correspond à l'intensité des vibrations transversales le long de la membrane basilaire de l'oreille :

$$E(j) = \int |H_j(f)|^2 S_X(f) df$$

où $S_X(f)$ est la dsp du signal $x(t)$. À partir de maintenant, on adoptera des notations correspondant à une discrétisation aussi bien de l'axe des temps que de l'axe des fréquences :

$$E(j) = \sum_i |H_j(i)|^2 S_X(i)$$

sans préciser si l'indice i porte sur une échelle en Hertz ou une échelle en Bark. Dans certains modèles on utilise plutôt la fonction d'étalement, qui est la fraction de l'intensité du signal localisé à la fréquence i qui influe sur la perception à la fréquence j. Cette fonction se déduit de la réponse en fréquence des filtres cochléaires par $f_{etal}(i, j) = |H_j(i)|^2$.

Les quatre modèles psychoacoustiques utilisés dans cet article pour le calcul de filtres perceptuels ont des fonctions d'étalement très simplifiées, invariantes pour le modèle 2 de MPEG, les modèles d'ASPEC et de Mahieux et Petit. Le calcul de l'excitation se fait dans l'échelle des Bark avec une résolution de une raie par Bark dans le modèle 1 de MPEG et le modèle d'ASPEC, et avec deux raies par Bark pour le modèle 2 de MPEG. Le modèle de Mahieux et Petit fait le calcul dans l'échelle des Hertz avec 512 raies.

Cette phase d'analyse temps-fréquence est suivie d'une étape de détection, par le système nerveux, de l'information contenue dans l'onde de propagation parcourant la membrane basilaire. Ce qui nous intéresse dans le cadre du codage audio, c'est comment est perçue la différence subjective entre deux sons. La résolution de la perception de l'intensité sonore est limitée : un son x_1 (son masqué) est inaudible en présence d'un son x_2 (son masquant) si et seulement si le rapport entre leurs excitations $E_1(j)$ et $E_2(j)$ est inférieur à une certaine courbe $av(j)$, le taux de masquage [12] :

$$\frac{E_1(j)}{E_2(j)} \leqslant av(j) \quad \forall j \tag{1}$$

Des travaux récents [4] ont confirmé que le taux de masquage dépend du caractère tonal du signal. Le taux de masquage est plus faible quand le masqueur est un son pur, et plus grand si c'est un bruit à bande étroite. Pour calculer ce taux de masquage il faut donc une estimation de la tonalité du signal. Le modèle 1 de MPEG cherche les tonales dans chaque bande critique et utilise deux taux de masquage différents pour les parties tonales et non tonales. Le modèle d'ASPEC calcule une estimation globale de la tonalité du signal avec une mesure de la platitude du spectre (Spectral Flatness Measure) et calcule un taux de masquage en pondérant par cet indice de tonalité la valeur du taux pour un son pur et pour un bruit. Le modèle 2 de MPEG estime la tonalité du signal pour chaque raie du spectre en calculant le vecteur de non prédictibilité, il en déduit le taux de masquage comme dans le modèle d'ASPEC en pondérant, par les coefficients de ce vecteur, les valeurs du taux pour un son pur et pour un bruit. Le modèle de Mahieux et Petit utilise un taux de masquage constant de -30 dB.

D'après la définition que nous avons donnée du seuil de masquage, il reste à minimiser une fonction coût liée au débit avec la contrainte de l'équation (1). Les modèles utilisés dans les codeurs audio ne font pas cette optimisation qui est trop complexe. Pour simplifier le problème, ils calculent le seuil d'audition masqué, qui est le seuil en dessous duquel un bruit à bande étroite est inaudible en présence du signal original. Le problème est que le bruit de codage n'est pas en général un bruit à bande étroite, et que le seuil d'audition masqué ne donne pas de conditions sur la dsp du signal masqué. Pour en tenir compte le modèle 1 de MPEG enlève une quinzaine de dB au seuil obtenu dans le calcul des rapports signal à masque, en prenant la valeur minimum du masque et la valeur maximum de la dsp du signal dans chacune des 32 sous-bandes du codeur. Le modèle d'ASPEC et le modèle 2 de MPEG divisent le seuil d'audition obtenu par une fonction de normalisation qui correspond aux gains des fonctions d'étalement. Le modèle de Mahieux et Petit fait confiance à la faible valeur du taux de masquage choisi.

3 Différentes méthodes de calcul des filtres perceptuels

3.1 Méthodes standards

Atal et Schroeder ont proposé [1], dans le cadre d'un codeur prédictif de la parole en bande téléphonique, de mettre en

forme spectralement le bruit de reconstruction à l'aide d'un filtre dit filtre perceptuel défini par

$$W(z) = \frac{A(z)}{A(z/\gamma)} = \frac{\prod_{k=1}^{P}(1 - z_k z^{-1})}{\prod_{k=1}^{P}(1 - \gamma z_k z^{-1})}$$

où $A(z)$ est le filtre blanchissant le signal à coder, obtenu par une prédiction linéaire d'ordre P. Dans $A(z/\gamma)$, tous les zéros de $A(z)$ sont éloignés du cercle unité d'un facteur $\gamma \leqslant 1$, ce qui a pour effet de diminuer $|A(z/\gamma)|^2$ dans les régions énergétiques du signal. $|W(z)|^2$ va ainsi être inférieur à 1 aux voisinages des pics d'énergie de la dsp du signal. Plus γ est petit, plus les pics de $|1/W(z)|^2$ seront accentués. Dans la plupart des codeurs de parole en bande téléphonique, $W(z)$ est de la forme $A(z/\gamma_1)/A(z/\gamma_2)$ avec $\gamma_1 \approx 0.9$ et $\gamma_2 \approx 0.4$.

Si cette méthode est très efficace pour le codage de la parole en bande téléphonique, des améliorations ont été proposées pour son utilisation sur des bandes passantes plus larges. Ordentlich et Shoham [11], pour coder de la parole en bande élargie à 32 kbit/s, ont proposé de corriger l'allure générale de la réponse en fréquence, qu'ils appellent le "tilt", du filtre perceptuel $W(z)$ en le pondérant par un filtre $T(z)$

$$W(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)} T(z)$$

L'avantage est de pouvoir régler de manière indépendante la forme fine de $|W(z)|^2$ avec les paramètres γ_1 et γ_2, et l'allure générale avec les coefficients de $T(z)$. On obtient de bons résultats pour des signaux de parole avec $T(z)$ de la forme

$$T(z) = \frac{1}{1 + \sum_{i=1}^{2} p_i \delta^i z^{-i}}$$

où les p_i sont obtenus avec une prédiction linéaire d'ordre 2, et $\delta = 0.7$ est un paramètre servant à ajuster le "tilt".

On obtient un meilleur compromis entre la parole et la musique avec

$$T(z) = \frac{A_K(z)}{1 - \mu z^{-1}} \qquad (2)$$

où $A_K(z)$ est le filtre blanchissant d'ordre K faible, on prend $K = 2$, $|1/A_K(z)|^2$ donnant l'allure générale du spectre du signal. Le paramètre $\mu = 0.3$ permet de régler le "tilt" du filtre perceptuel par rapport à cette allure.

Tous ces filtres perceptuels n'utilisent que très grossièrement les résultats sur l'audition. Ils se contentent d'utiliser le fait qu'il vaut mieux mettre du bruit dans les régions énergétiques du signal.

3.2 Filtres perceptuels construits à partir d'un seuil de masquage

Pour mieux se servir des résultats de la psychoacoustique, Chang et Wang ont proposé, en bande élargie, de construire le filtre perceptuel à partir d'un seuil de masquage calculé à l'aide d'un modèle psychoacoustique [3]. L'idée est de considérer l'inverse du seuil d'audition comme le module au carré de la réponse en fréquence d'un filtre, puisque la dsp du bruit de reconstruction, qui est égale à $\sigma_Q^2/|W(f)|^2$, doit être inférieure au seuil de masquage.

Il existe plusieurs méthodes pour synthétiser un filtre $C(z)$, connaissant le module de sa réponse en fréquence. La méthode utilisée dans [3] consiste à considérer le seuil de masquage comme la dsp d'un signal sur la fenêtre courante, et d'en déduire la fonction d'autocorrélation à l'aide de la transformée de Fourier discrète inverse. Une analyse LPC, avec l'algorithme de Levinson ou de Schur, permet d'en déduire les coefficients c_i du filtre $C(z)$. Le filtre perceptuel est alors donné par

$$W(z) = \frac{C(z)}{C(z/\gamma)}$$

où $\gamma = 0.8$. En bande FM il vaut mieux choisir $\gamma = 0.2$.

Le modèle psychoacoustique utilisé dans [3] est le modèle 1 de MPEG. Une autre implémentation a été faite dans [9] dans le cas d'un codeur à très bas délai en bande FM avec un débit de 64 kbits/s, avec le modèle psychoacoustique de Mahieux et Petit. Ici, pour étendre la comparaison, nous avons aussi implémenté cette méthode avec le modèle 2 de MPEG et le modèle d'ASPEC.

3.3 Proposition

Quand on compare, avec la procédure décrite section 4, la qualité obtenue avec cette méthode et celle obtenue avec la méthode de Ordentlich et Shoham en utilisant (2), on est en fait assez surpris de constater que la deuxième méthode donne d'aussi bons, sinon de meilleurs résultats, surtout en bande FM.

Si la méthode de type Ordentlich et Shoham permet de bien régler l'allure générale, le "tilt", de la réponse en fréquence du filtre perceptuel, elle traite globalement la hauteur des pics de la réponse en fréquence et la hauteur des creux de $1/W(z)$ comme avec la méthode d'Atal. Pour avoir des formes intermédiaires entre les filtres de type standard et ceux obtenus à partir d'un seuil de masquage, il faut pouvoir choisir, individuellement et indépendamment des creux, la hauteur des pics de la réponse en fréquence de $1/W(z)$. On va, au lieu de multiplier tous les zéros de $A(z)$ par le même coefficient γ, multiplier chaque paire de zéros complexes conjugués, par un coefficient γ_k différent. On définit donc les coefficients γ_k en fonction des phases des zéros correspondants :

$$\gamma_k = F_{att}(arg(z_k))$$

Pour avoir une réponse en fréquence qui s'adoucit vers les hautes fréquences, la fonction $F_{att}(arg(z_k))$ doit décroître de 1 pour les basses fréquences jusqu'à une valeur légèrement inférieure pour les hautes fréquences. On choisit une fonction linéaire par morceaux, égale à 1 jusqu'à $f_c \in [0, f_s/2]$, et décroissant jusqu'à $att_{min} < 1$ après.

4 Résultats expérimentaux

Pour simuler le bruit de reconstruction d'un codeur utilisant un filtre perceptuel $W(z)$, on découpe le signal en fenêtres de N échantillons, avec un recouvrement de $N/2$ échantillons pour éviter des artefacts dus au découpage en fenêtres. Pour chaque fenêtre m, un bruit blanc de puissance $\sigma_Q^2(m)$, filtré

par le filtre $1/W(z)$, simule le bruit de reconstruction $r(m)$. Le bruit de quantification $q(m)$ dépend du type de quantificateur utilisé, mais on fait l'hypothèse qu'il est blanc et de puissance

$$\sigma_Q^2(m) = c\ \sigma_X^2(m)\ 2^{-2b}$$

avec l'hypothèse habituelle où la résolution b est élevée [8]. On fixe le rapport signal sur bruit à une valeur de 15 dB. Cette procédure est exécutée sur une base de fichiers sonores, échantillonnés à 16 kHz et 32 kHz, comprenant des échantillons de parole avec des locuteurs féminins et masculins, et des échantillons de musique pop chantée et de musique classique avec différents instruments dominants, la durée étant fixée à 8 s. Les fichiers sont écoutés avec un casque et les fichiers dégradés sont comparés au fichier original.

La méthode d'Atal engendre beaucoup de bruit dans les hautes fréquences, ce qui confirme que le "tilt" est insuffisant avec des bandes plus larges que la bande téléphonique. Avec la méthode de Wang et Chang il est préférable d'utiliser le modèle psychoacoustique 2 de MPEG. La méthode de Ordentlich et Shoham donne des résultats assez comparables à ceux obtenus avec le modèle numéro 2 de MPEG, un peu supérieurs avec les morceaux très harmoniques. Avec le filtre perceptuel que nous proposons les meilleurs résultats sont obtenus avec $f_c = 2\pi/3$ et $att_{min} = 0.98$. Par rapport à la méthode de Ordentlich et Shoham on diminue des sifflements hautes fréquences en rabotant les pics de la réponse en fréquence dans les hautes fréquences.

5 Conclusion

On compare des filtres perceptuels calculés à partir d'une prédiction linéaire avec des filtres obtenus avec des seuils de masquage utilisés dans des codeurs de musique. Nous avons remarqué que ces derniers ne donnent pas de meilleurs résultats. On propose ici une méthode intermédiaire consistant à donner plus de degrés de liberté à une méthode de type standard, en traitant individuellement les zéros du filtre blanchissant. Les meilleurs résultats sont obtenus avec un filtre assez proche de celui de la méthode de Ordentlich et Shoham, sauf dans les hautes fréquences, où on enlève quelques sifflements.

Références

[1] B.S. Atal and M.R. Schroeder. Predictive coding of speech signals and subjective error criteria. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-27, June 1979.

[2] K. Brandenburg, H. Herre, J. Johnston, Y. Mahieux, and E. Schroeder. ASPEC : Adaptive perceptual entropy coding of high quality music signals. *Proceedings of the 90th AES convention*, pages 1–11, 1991.

[3] W.W. Chang and C.T. Wang. Audio coding using masking-threshold adapted perceptual filter. *Proc. IEEE Workshop on Speech Coding for Telecommunications*, pages 9–10, October 1993.

[4] J. L. Hall. Asymmetry of masking revisited : Generalization of masker and probe bandwidth. *J. Acoust. Soc. Am.*, 101 :1023–1033, 1997.

[5] T. Irino and R. D. Patterson. A time domain, level dependant auditory filter : the gammachirp. *J. Acoust. Soc. Am.*, 101 :412–419, 1997.

[6] R. Lefebvre, R. Salami, C. Laflamme, and J.P. Adoul. High quality coding of wideband of wideband audio signals using transform coded excitation (TCX). *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pages I–193–196, 1994.

[7] Y. Mahieux and J.P. Petit. High-quality audio transform coding at 64 kbps. *IEEE Trans. on Communications*, Vol. 42, No. 11 :3010–3019, November 1994.

[8] N. Moreau. *Techniques de compression des signaux*. Masson, Collection technique et scientifique des télécommunications, 1995.

[9] C. Murgia, G. Feng, C. Quinquis, and A. Le Guyader. Very low delay and high quality coding of 20 Hz - 15 kHz speech at 64 kbit/s. *4th Europ. Conf. on Speech Comm. and Technol.*, pages 37–40, September 1995.

[10] Norme internationale ISO/CEI 11172. *Codage de l'image animée et du son associé pour les supports de stockage numérique jusqu'à environ 1,5 Mbit/s*, 1993.

[11] E. Ordentlich and Y. Shoham. Low-delay code-excited linear-predictive coding of wideband speech at 32 kbps. *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pages 9–12, 1991.

[12] E. Zwicker and E. Feldtkeller. *Psychoacoustique, l'oreille récepteur d'information*. Masson, Collection technique et scientifique des télécommunications, Traduit de l'allemand par C. Sorin, 1981.