Un schéma de représentation adaptatif en temps et en fréquence pour le codage audio

Gilles Gonon[1] Silvio Montrésor[2] Marc Baudry[3]
[1,3]Laboratoire d'Informatique,
[2]Laboratoire d'Acoustique, UMR CNRS 6613,
Université du Maine, 72085 Le Mans Cedex.

[1]Gilles.Gonon@univ-lemans.fr

Résumé

Cet article présente un schéma de codage adaptatif en temps et en fréquence. La segmentation temporelle est effectuée à l'aide du critère entropique local et la segmentation fréquentielle est basée sur une extension de l'algorithme de recherche de la meilleure base à partir de la décomposition en paquets d'ondelettes. L'allocation utilise des critères énergétiques et psychoacoustiques pour pallier aux problèmes de sélectivité fréquentielle des paquets d'ondelettes. Les premiers résultats subjectifs informels sont satisfaisants pour des rapports de compression de l'ordre de 10 à 15.

Mots Clef

Codage audio adaptatif, temps fréquence, paquets d'ondelettes, entropie.

1 Introduction

Le codage audio de haute qualité permet aujourd'hui d'atteindre des rapports de compression situés entre 10 et 15 pour une qualité quasi-transparente de restitution (MPEG 2 - AAC, AC-3 ou TwinVQ). Cela est en partie dû à l'amélioration de la transformation du signal, autrement dit sa représentation dans un espace concentrant son énergie. L'utilisation d'une segmentation temporelle adaptée permet par exemple de supprimer certains artefacts tels que le pré-écho par à rapport à une segmentation fixe. Toutefois rares sont les schémas de codage proposant une adaptation aussi bien en temps qu'en fréquence [1, 2].
Les normes MPEG-2 et AC-3 utilisent une transformation du signal adaptée temporellement mais n'imposent pas de méthodes de segmentation. La transformation utilise un banc de filtres dont la partition fréquentielle correspond aux bandes critiques de l'oreille humaine pour appliquer les critères psychoacoustiques. Une autre approche basée sur la décomposition en paquets d'ondelettes permet d'adapter les sous-bandes fréquentielles au contenu du signal par l'intermédiaire de l'algorithme de recherche de meilleure base. Cependant la minimisation entropique de la base est limitée d'une part par la structure dyadique de la librairie des bases mais aussi par le manque de segmentation temporelle adaptée.
Nous proposons dans cet article un schéma ouvert de codeur audio basé sur une segmentation adaptative du signal en temps et en fréquence. Les segmentations temporelles et fréquentielles sont effectuées de manière disjointe pour respecter la nature aléatoire des évènements dans les signaux audios en limitant l'augmentation d'information parallèle. Les différentes segmentations sont basées sur des mesures entropiques du signal et l'allocation des ressources utilise des critères énergétiques et psychoacoustiques pour une quantification uniforme des coefficients compressés.
La deuxième partie de cet article présente l'architecture du codeur et les choix adoptés pour effectuer chaque opération de la chaîne de codage. La troisième partie expose les segmentations temporelle et fréquentielle du signal. La quatrième partie présente le critère d'allocation mixte. Enfin, la cinquième partie discute les résultats obtenus avant de conclure et de donner quelques perspectives au codeur réalisé.

2 Schéma de codage

Dans le schéma de codage proposé, le signal est d'abord segmenté temporellement dans le but de détecter les évènements temporels au sens d'une variation de stationnarité. La segmentation fréquentielle est dissociée de la segmentation temporelle pour permettre l'utilisation d'un détecteur de ruptures précis. En effet, dans le cas d'une segmentation conjointe temps fréquence, la précision des ruptures est souvent limitée par le choix de la transformation, comme par exemple la dyadicité [1].
Le critère de segmentation utilisé est basé sur une mesure des variations d'entropie du signal comme indice de désordre spectral. Le critère entropique local fournit un vecteur de points de ruptures dans le signal utilisé pour la segmentation temporelle, partie 3.1. Dès lors, chaque tranche ainsi définie est supposée de caractère stable au sens stationnaire. En pratique la segmenta-

tion permet au choix de détecter des points de ruptures ou de délimiter des zones de transition.
Dans la suite du schéma, l'algorithme de recherche de meilleure base s-dyadique est appliqué à chaque tranche temporelle pour fournir un découpage en sous bandes adapté au signal, partie 3.2. En parallèle, les courbes de masquage de chaque tranche sont calculées pour permettre l'allocation des ressources binaires de chaque paquet d'ondelettes, ou sous bande fréquentielle, partie 4.
D'autre part, pour supprimer les effets de blocs qui apparaissent lorsque le taux de compression augmente, un recouvrement de la taille de l'ondelette analysante (longueur de la RI du filtre QMF utilisé) entre deux tranches temporelles est suffisant.
La figure ci-dessous résume les principales caractéristiques du schéma de codage utilisé au niveau de la transformation et de l'allocation. Les informations parallèles à transmettre comprennent :
- les points de ruptures temporelles, sous forme de longueur des tranches
- l'arborescence de la meilleure base,
- l'information d'allocation et de quantification des paquets

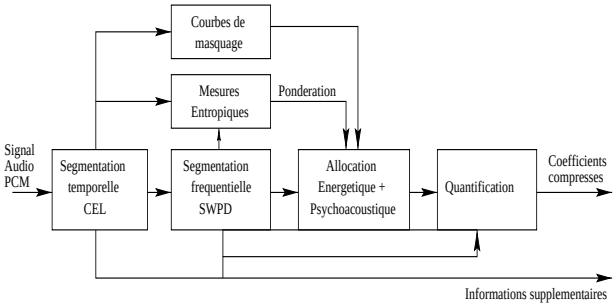

FIG. 1 – Schéma de codage adaptatif basé sur le critère entropique local et la SWPD.

Dans ce schéma, la complexité est principalement portée à l'encodage et le décodage consiste simplement en la déquantification des paquets puis en la synthèse du signal par décomposition inverse. Chaque tranche synthétisée est ajoutée à la précédente sur la zone de recouvrement. L'inverse de la décomposition en paquets d'ondelettes étant de complexité faible (inférieure à une FFT par exemple), le décodage peut être effectué en *temps réel*.

3 Segmentation adaptée

3.1 Segmentation temporelle

Des travaux récents sur la segmentation temporelle ont donné des résultats satisfaisants pour les signaux audios [3]. Le détecteur utilisé, le Critère Entropique Local, est non paramétrique et s'appuie sur une mesure relative de l'entropie du signal sur deux fenêtres glissantes. Ainsi, le CEL est un détecteur sensible aux variations de spectre à court terme du signal et plus particulièrement à la façon dont il concentre l'énergie. La mesure de concentration de l'énergie utilisée est l'entropie de Shannon appliquée à la transformée de Fourier discrète du signal.
En notant $X(k)$ la transformée de Fourier discrète normalisée d'un signal $x(n), (k,n) \in [0, N-1]$

$$X(k) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x(n) W_N^{nk} \quad \text{, où } W_N^{nk} = e^{-2j\pi \frac{nk}{N}}. \tag{1}$$

L'entropie de Shannon de x sur l'intervalle $[0, N-1]$ notée $E_{x[0,N-1]}$, est définie par

$$E_{x[0,N-1]} = -\sum_{k=0}^{N-1} |X(k)|^2 \log |X(k)|^2. \tag{2}$$

Le CEL est alors défini comme une différence relative d'entropie entre une fenêtre glissante de longueur N et ses 2 sous-fenêtres de longueur $\frac{N}{2}$. En notant respectivement E_{xc}, E_{xg} et E_{xd} les entropies de la fenêtre principale et de ses deux sous-fenêtres gauche et droite, définies par (Fig. 2)

$$\begin{aligned} E_{xc}(n) &= E_{x[n-\frac{N}{2}, n+\frac{N}{2}-1]}, \\ E_{xg}(n) &= E_{x[n-\frac{N}{2}, n-1]}, \\ E_{xd}(n) &= E_{x[n, n+\frac{N}{2}-1]}, \end{aligned} \tag{3}$$

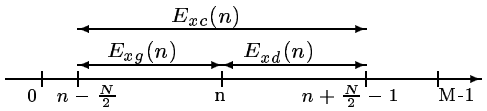

FIG. 2 – Notations des entropies pour le calcul du CEL

Le CEL est défini pour un signal de longueur M sur l'intervalle $[\frac{N}{2}, M - \frac{N}{2} - 1]$ par la formule suivante :

$$CEL_x(n) = \frac{E_{xc}(n) - (E_{xg}(n) + E_{xd}(n))}{|E_{xc}(n)|}. \tag{4}$$

La taille de la fenêtre définit le *contexte* d'analyse. Afin de réduire les variations d'entropie en fonction de la fréquence, chaque tranche est apodisée par une fenêtre de Hamming. Ceci permet de supprimer les différences de spectre causées par le calage de la fréquence sur une harmonique de la TFD qui modifie la valeur de l'entropie résultant de la somme de tous les coefficients.
Le terme de normalisation du dénominateur du CEL est choisi de telle sorte que le CEL ait une valeur proche de -1 pour un signal stationnaire monocomposante. Pour un signal nul, le CEL conduit à une forme

indéterminée du type $\frac{0}{0}$; aussi nous fixons la valeur du CEL à -1 pour un signal nul.
Le CEL est donc un estimateur temporel des variations d'entropie du signal. Les points de ruptures sont obtenus en prenant les maximas locaux du CEL supérieurs à un seuil fixé en fonction du rapport signal à bruit, typiquement -0.5. Il est de même possible d'isoler une zone de changement en plaçant des points de ruptures en début et fin de dépassement du seuil de détection.

3.2 Segmentation fréquentielle

La décomposition en paquets d'ondelettes (WPD) est un outil efficace en terme d'adaptation fréquentielle par le biais de l'algorithme de recherche de meilleure base [4]. Cet algorithme permet en effet d'obtenir un banc de filtres adapté au signal dans le sens où l'arrangement des sous-bandes est choisi pour minimiser l'entropie dans le domaine transformé. Les sous-bandes contenant du signal sont le plus étroites possibles (jusqu'à la meilleure résolution fréquentielle) et les sous bandes contenant pas ou peu de signal sont regroupées en larges bandes par le test entropique.
Cependant la minimisation entropique de la base est limitée d'une part par la structure dyadique de la librairie des bases mais aussi par le manque de segmentation temporelle.
Dans des travaux précédents [5], une extension de la recherche de meilleure base à faible coût algorithmique a été proposée, afin de pallier au problème de la structure dyadique de la WPD. Cette représentation, la SWPD, permet notamment de réduire l'entropie de la meilleure base qui est directement liée au facteur de compression dans le cas d'une allocation énergétique.

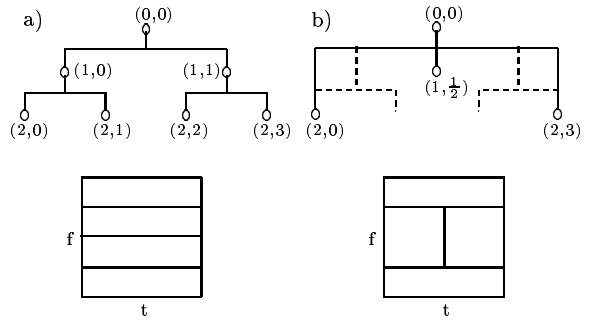

FIG. 3 – Représentation des arborescences de meilleure base et pavages du plan temps-fréquence associés : a) arbre dyadique, b) arbre $\Sigma-$dyadique .

4 Allocation

L'allocation suit un schéma original mélangeant critères énergétique et psychoacoustique. Dans notre cas, des tests subjectifs informels ont montré que l'utilisation d'une allocation uniquement psychoacoustique pouvait entraîner des artefacts caractérisés par la naissance d'une inharmonicité (effet *son de cloche*, comparable à un écho fréquentiel). D'autre part, une allocation uniquement énergétique sur des signaux de type percussif se traduit par une dégradation qualitative du signal proportionnel au facteur de compression.
Le premier effet est dû à la façon dont est effectuée l'allocation psychoacoustique de chacune des tranches. La meilleure base obtenue fournit les paquets correspondant aux sous-bandes d'une partition de l'axe fréquentiel. Ce sont les limites théoriques de cette partition qui sont utilisées pour réaliser l'allocation psychoacoustique, en fonction de la courbe de masquage calculée pour la tranche analysée (quasi stationnaire). Les ressources sont alors allouées dynamiquement aux sous-bandes ayant le plus fort rapport signal à masque, qui est mis à jour à chaque bit alloué.
Malheureusement, la partition de la meilleure base ne prend pas en compte le compromis temps fréquence effectué lors du choix de l'ondelette analysante qui se traduit par une mauvaise localisation fréquentielle lorsque la profondeur de décomposition est élevée. En d'autres termes, les limites fréquentielles utilisées pour l'allocation des paquets ne correspondent pas exactement au contenu fréquentiel des paquets. Il est donc dans ces conditions important de faire une préallocation énergétique pour allouer même faiblement les sous-bandes les plus énergétiques, car elles peuvent correspondre à une partie de signal non masqué repliée dans une autre sous-bande de la décomposition en paquets d'ondelettes. L'allocation énergétique est basée sur l'énergie des différents paquets de la meilleure base.
En résumé, la procédure d'allocation suit un algorithme classique de ditribution dynamique des ressources. Les ressources sont simplement divisées en deux groupes, le premier étant destiné aux paquets les plus énergétiques et le second servant aux paquets les plus masquants.
Du fait que les artefacts précédents dépendent fortement de la nature du signal codé (tonale ou percussive), les mesures d'entropie aux niveaux temporel et fréquentiel peuvent être utilisées lors de la phase d'allocation pour pondérer relativement l'énergie et le masquage. Aucun critère de pondération n'a encore été mis au point et les deux types d'allocations se partagent actuellement les ressources de manière fixe.

5 Résultats

Dans cette partie, nous montrons les différentes étapes du codage sur un signal de clavecin échantillonné à 44.1kHz pour un rapport de compression de 10. Le signal est constitué de 3 notes successives constiuant un accord. La figure 4 montre le signal encodé puis décodé ainsi que l'erreur de codage. La grille verticale indique les points détectés comme points de ruptures par le CEL, qui détectent précisément les occurences

des trois notes. De plus, en regardant l'erreur de codage, on s'aperçoit que la localisation précise de la première note permet de réduire l'effet de pré-écho, au sens où l'erreur n'est pas répercutée avant l'arrivée de la note comme cela serait le cas si la segmentation était mal adaptée (apparition de la note au milieu d'une fenêtre d'analyse).

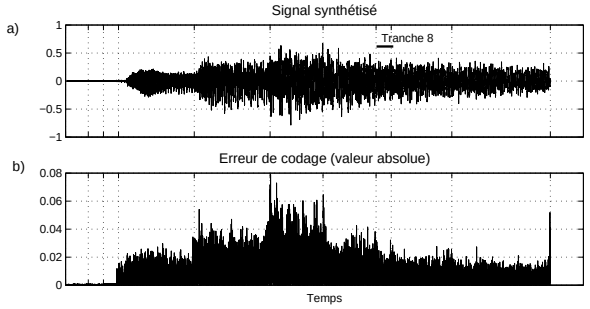

FIG. 4 – a) Signal décodé et indices de segmentation temporelle obtenus par le CEL (grille verticale). b) Erreur de codage.

La figure 5 montre le détail de l'allocation constituée des contributions énergétique et psychoacoustique d'une tranche de signal, notée Tranche 8 à la figure 4.a). La courbe du haut représente le rapport signal à masque obtenu pour la tranche et la grille verticale marque les limites correspondantes à la meilleure base issue de la décomposition en paquets d'ondelettes. Chaque paquet est ensuite quantifié uniformément.

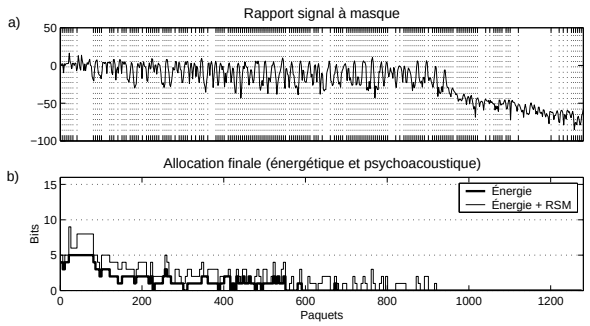

FIG. 5 – a) Rapport signal à masque et partition fréquentielle correspondant à la meilleure base. b) Allocation mixte pour la tranche, contributions énergétique et psychoacoustique.

Le codeur étant achevé depuis peu, seuls quelques tests subjectifs informels ont été effectués sur des signaux audios très variés : extraits de musique classique, variété, parole et instruments seuls dont violon, guitare, clavecin, flûte, didgeridoo, castagnettes, métronome. La qualité, notée par des auditeurs non spécialistes a été jugée très bonne pour tous les signaux pour un rapport de compression de l'ordre de 12 et bonne ou très bonne pour un rapport de compression de l'ordre de 15. La qualité reste aussi satisfaisante lorsque le taux de compression augmente.
Une démonstration comparative de quelques signaux originaux et codés/décodés est accessible ou téléchargeable depuis l'adresse :

`http://www-ic2.univ-lemans.fr/~gonon/`

6 Conclusion et perspectives

Alors que seule la transformation a été optimisée, les résultats de ce schéma de codage sont encourageants, puisque pour un facteur de compression compris entre 10 et 15, la qualité est jugée bonne (proche de la transparence), pour la majorité des signaux testés. Il reste cependant important de mener des tests subjectifs permettant de noter la qualité *absolue* mais aussi la qualité relative par rapport aux codeurs de références pour des rapports de compression comparables.
De plus, de par la nature ouverte du schéma de codage proposé, de nombreuses améliorations dans les différentes parties de la chaîne de codage sont encore possibles, notamment au niveau du modèle psychoacoustique peu adéquat pour les ondelettes. Il reste aussi important de définir un critère basé sur les différentes mesures entropiques permettant de pondérer l'allocation en faveur des critères énergétiques ou psychoacoustiques. Enfin, en raison de la teneur aléatoire des signaux temporels, il semble judicieux d'adapter l'allocation aux besoins des tranches analysées, et de mettre en place un codeur à débit variable.

Références

[1] Herley C., Kovačević J., Ramchandran K., and Vetterly M. Tilings of the time-frequency plane : Construction of arbitrary orthogonal bases and fast tiling algorithms. *IEEE Trans. on Signal Processing*, 41(12) :3341–3359, décembre 1993.

[2] Erne M. and Moschytz G.S. A bit-allocation scheme for an embedded and signal adaptative audio coder. *AES 108th Convention, Paris*, février 2000.

[3] Gilles Gonon, Silvio Montrésor, and Marc Baudry. Segmentation multibande adaptée basée sur le critère entropique local pour le codage audio. In *18^e colloque GRETSI*, septembre 2001.

[4] R.R. Coifman and M.V. Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Trans. on Information Theory*, 38(2) :713–718, March 1992.

[5] Gilles Gonon, Silvio Montrésor, and Marc Baudry. Extended best basis familly tree and entropy diminution, application to audio coding. In *International Congress of Acoustics, ICA 2001*, septembre 2001.