CODAGE EN SOUS-BANDES À ANALYSE-PAR-SYNTHÈSE

Andrei POPESCU et Nicolas MOREAU

ENST, 46 rue Barrault,
75634 PARIS Cedex 13

RÉSUMÉ

On propose un nouveau schéma de codage adapté à la compression de signaux audiofréquence. On utilise une approche de type "analyse-par-synthèse" pour quantifier les paramètres intervenant dans la modélisation des signaux de sous-bande. Pour une application en codage de parole, un prédicteur à long terme peut être facilement intégré à ce schéma. Une évaluation informelle d'un codeur de parole en bande téléphonique à un débit de 8 kbit/s donne un résultat encourageant. On obtient un retard de codage relativement faible (10...50 ms) et une complexité de traitement comparable à celle du codeur CELP.

ABSTRACT

We propose a coder structure, called SBAS ("Sub-Band Analysis-by-Synthesis"), for use in medium- and low-rate coding of speech and audio. We use an analysis-by-synthesis approach to quantize the parameters of a subband signal model. For speech coding applications, long-term prediction can be easily integrated with this structure. The evaluation of SBAS for the coding of telephone-band speech at 8 kb/s gave promising results. A relatively low coding delay (10...50 ms) is achievable with this type of coder, and its computational complexity is comparable to that of a CELP coder.

1. INTRODUCTION

Pour comprimer des signaux audiofréquence (de la parole ou de la musique), il n'existe pas un schéma de codage unique. Deux démarches, *a priori* assez distinctes, sont habituellement utilisées. Pour du signal de parole en bande téléphonique, le codeur CELP a montré sa supériorité. Il s'agit fondamentalement d'une simple quantification vectorielle de type "gain-forme" et "multi-étages" avec un dictionnaire de quantification régulièrement adapté (toutes les 20 ms environ) aux caractéristiques statistiques à court terme du signal par une opération de filtrage [6]. Cette démarche est justifiée par le fait que le signal de parole admet un modèle de production simple et efficace de type "source-filtre". Il s'agit d'un schéma "d'analyse-par-synthèse". L'exploitation sommaire d'un modèle d'audition permet une mise en forme spectrale grossière du bruit de reconstruction. Pour du signal de musique, il n'existe pas de modèle de production étant donnée la variété des sources sonores possibles; par contre un modèle d'audition élaboré est généralement utilisé. Les codeurs adaptés à de la musique, par exemple Musicam à la base du codeur MPEG-Audio [7] ou Aspec utilisé dans la troisième couche de MPEG-Audio cherchent essentiellement à réaliser régulièrement (toutes les 20 ms environ) une mise en forme spectrale du bruit de reconstruction de façon à ce que la densité spectrale de ce bruit soit, quelle que soit la fréquence et quel que soit l'instant, toujours en-dessous d'une courbe, le seuil de masquage [5].

On donne, Figure 1, le schéma de principe d'un codeur audiofréquence. La transformation T_1 symbolise le calcul d'une estimation de la densité spectrale de puissance $S_X(f)$ du signal $x(n)$, par exemple à l'aide d'un périodogramme. A partir de cette estimation un modèle d'audition fournit un seuil de masquage $S_Q(f)$. Les transformations T_2 et P_2 correspondent respectivement aux bancs de filtres d'analyse et de synthèse.

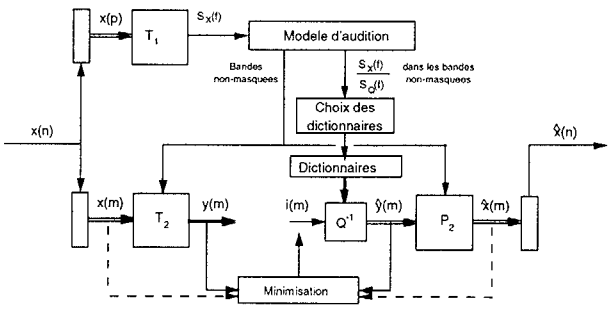

Figure 1: Schéma de principe d'un codeur audiofréquence.

Dans cet article, on rappelle d'abord quelques résultats standards relatifs à l'allocation de bits. On propose ensuite une nouvelle structure de codeur, notée par la suite codeur SBAS (SubBand Analysis-by-Synthesis coder), combinant les avantages des techniques de quantification vectorielle prédictive (le codeur CELP) et du codage par transformée (ou en sous-bandes). On montre que cette structure peut être vue comme une généralisation d'un codeur par transformée (ou en sous-bandes) et qu'elle s'interprète comme une procédure "d'analyse-par-synthèse". On propose également l'utilisation d'une quantification vectorielle "avec recouvrement". On fournit une description

(rapide) de ce codeur. On donne enfin quelques résultats expérimentaux préliminaires.

Des codeurs CELP avec des dictionnaires d'excitation composés de signaux bande étroite ont déjà été proposés [8], [9]. Le codeur SBAS présente certaines analogies avec ces codeurs mais le principe est différent. Une technique similaire à la quantification vectorielle "avec recouvrement" est décrite dans [2].

2. ALLOCATION DE BITS

On rappelle d'abord que, lorsque l'on quantifie scalairement la réalisation $x(n)$ d'une source stationnaire de puissance σ_X^2 avec la résolution b, le minimum σ_Q^2 de la puissance du bruit de quantification a pour expression

$$\sigma_Q^2 = c\ \sigma_X^2\ 2^{-2b} \tag{1}$$

où c est une constante qui ne dépend que de la densité de probabilité marginale de la source [6]. A partir de cette formule, on en déduit directement le nombre de bits minimum nécessaire pour quantifier scalairement une source avec un rapport signal sur bruit donné

$$b \geq \frac{1}{2} \log_2 c\ \frac{\sigma_X^2}{\sigma_Q^2}. \tag{2}$$

Cette formule admet une double généralisation. Si on considère que la source est avec mémoire et que l'on cherche à minimiser la puissance du bruit de quantification, on sait qu'il existe plusieurs méthodes permettant d'exploiter la corrélation (filtre blanchissant, quantification vectorielle, transformation ou bancs de filtres ...) et que, dans le cas d'un codeur idéal, on obtient une nouvelle puissance de bruit donnée par

$$\sigma_Q^2 = \frac{c\ \sigma_X^2\ 2^{-2b}}{G_p(\infty)} \tag{3}$$

où la valeur asymptotique $G_p(\infty)$ du gain de prédiction est uniquement fonction de la densité spectrale $S_X(f)$ de la source

$$G_p(\infty) = \frac{\int_{-1/2}^{1/2} S_X(f)df}{e^{\int_{-1/2}^{1/2} \log_e S_X(f)df}} \approx \frac{\frac{1}{N}\sum_{k=0}^{N-1} \hat{S}_X(k)}{(\prod_{k=0}^{N-1} \hat{S}_X(k))^{1/N}}. \tag{4}$$

L'estimation $\hat{S}_X(k) \approx S_X(f = k/N)$ fait apparaître le gain de transformation habituel comme une approximation de la valeur asymptotique du gain de prédiction. Les relations (3) et (4) permettent d'obtenir le nombre de bits minimum nécessaire pour quantifier une source corrélée

$$b \geq \frac{1}{2} \int_{-1/2}^{+1/2} \log_2 c\ \frac{S_X(f)}{\sigma_Q^2} df. \tag{5}$$

Une deuxième généralisation est nécessaire car, dans le cadre du codage de signaux audiofréquence, ce n'est pas la minimisation de la puissance du bruit de quantification qui est le problème essentiel. On cherche quel est le débit nécessaire et suffisant pour que la densité spectrale de puissance du bruit soit inférieure à une densité spectrale de puissance limite fournie par le modèle d'audition. On montre [1] que l'équation (5) se généralise sous la forme suivante

$$b \geq \frac{1}{2} \int_{-1/2}^{+1/2} \max[0, \log_2 c\ \frac{S_X(f)}{S_Q(f)}]\ df. \tag{6}$$

Intuitivement, cette formule se déduit directement de (5) ou de (2) en observant qu'il est inutile de chercher à quantifier le signal dans les bandes de fréquences où $S_X(f) \leq S_Q(f)$ et que, dans les bandes de fréquence "élémentaires" $[f, f + df]$ vérifiant $S_X(f) > S_Q(f)$, la "densité de bits" nécessaire est donnée par

$$\frac{db}{df} \geq \frac{1}{2} \log_2 c\ \frac{S_X(f)}{S_Q(f)}. \tag{7}$$

La théorie indique donc que, pour coder un signal audiofréquence, connaissant la densité spectrale de puissance du signal et du bruit tolérable, il faut déterminer les bandes de fréquence $[f_k, f_k^{'}]$ telles que $S_X(f) \geq S_Q(f)$ puis allouer les ressources binaires disponibles en fonction du rapport $S_X(f)/S_Q(f)$. On reconnait la démarche réalisée, dans une certaine mesure, par Musicam ou par Aspec. Le signal est d'abord décomposé en signaux de sous-bande (ou en coefficients d'une transformée). L'allocation de bits est faite en fonction des rapports signal sur bruit dans chaque sous-bande. Les sous-bandes qui ne sont pas suffisamment énergétiques ne se voient attribuer aucun bit (les raies masquées sont éliminées). Chaque signal de sous-bande sous-échantillonné à la fréquence critique (ou chaque coefficient de la transformée) est ensuite quantifié en utilisant des dictionnaires scalaires ou vectoriels dont le nombre d'éléments est fonction de l'allocation de bits. Le signal est enfin reconstruit par sur-échantillonnage puis filtrage (par transformation inverse).

Une autre démarche conforme au développement théorique précédent est envisageable.

3. PRINCIPE DU CODEUR SBAS

Considérons le schéma de la Figure 1. On suppose que les vecteurs $\underline{x}(m)$ et $\underline{\hat{x}}(m)$ sont des vecteurs (colonnes) de dimension N, que les vecteurs $\underline{y}(m)$ et $\underline{\hat{y}}(m)$ sont des vecteurs de dimension M et que la matrice de dimension $N \times M$ caractérisant la transformation (ou le banc de filtres de synthèse) P_2 est composée des vecteurs $\{\underline{f}^0 \cdots \underline{f}^{M-1}\}$. Ce sont les réponses impulsionnelles (inversées) des filtres de synthèse. On appelle $[\hat{y}_0(m) \cdots \hat{y}_{M-1}(m)]$ les composantes du vecteur $\underline{\hat{y}}(m)$. Ce sont les échantillons quantifiés des signaux de sous-bande. On suppose qu'à chaque signal de sous-bande $\hat{y}_k(m)$ est associé un ensemble de quantificateurs, c'est à dire un ensemble de dictionnaires correspondant à différents débits.

Dans le schéma de quantification standard "décomposition en sous bandes, quantification, synthèse", on décompose d'abord le signal en sous-bandes. On quantifie ensuite $y_k(m)$ en utilisant le dictionnaire tel que le

nombre de scalaires (ou de vecteurs) le composant soit égal à 2^{b_k} avec b_k donné par l'allocation de bits, en minimisant une distorsion entre $y_k(m)$ et $\hat{y}_k(m)$. Enfin, on sur-échantillonne $y_k(m)$ et on réalise une opération de filtrage par un filtre interpolateur. Le bruit de quantification dans la sous-bande k est réparti dans toute la bande par sur-échantillonnage puis filtré par le k-ème filtre de synthèse. Comme ce filtre ne peut pas avoïr une réponse en fréquence strictement égale à une constante dans la sous-bande considérée et nulle à l'extérieur, le bruit de quantification se retrouve dans les sous-bandes voisines. Il est possible de gérer ce phénomène parasite mais c'est assez compliqué.

Dans une fenêtre d'analyse, le signal reconstruit $\underline{\hat{x}}(m)$ s'interprète comme un vecteur dont les composantes suivant la base formée par les réponses impulsionnelles des filtres de synthèse sont les signaux de sous-bande quantifiés. En supposant $N = M$ pour éviter d'alourdir l'expression, on obtient

$$\underline{\hat{x}}(m) = \sum_{k=0}^{M-1} \hat{y}_k(m)\underline{f}^k. \qquad (8)$$

On reconnaît l'expression du résultat d'une quantification vectorielle appliquée au vecteur $\underline{x}(m)$ si cette quantification vectorielle est de type "gain-forme" et de type "multi-étages" [6]. La forme est le vecteur $\underline{f}^k$ et le gain est le scalaire $\hat{y}_k(m)$. Dans le schéma de codage standard, le nombre "d'étages" est égal à M. Le nombre de bits b_k alloué à la quantification du k-ème signal de sous-bande est totalement consacré à $\hat{y}_k(m)$. Il n'y a pas de dictionnaire pour les "formes". Si on décide de répartir le nombre de bits b_k à la fois sur le gain et sur la forme comme dans le codeur CELP, on obtient un nouveau schéma de codage que l'on appellera le codeur SBAS. On donnera, par la suite, une solution pour construire ces dictionnaires spécifiques des formes. Le schéma de codage standard apparaît comme un cas particulier du codeur SBAS puisque l'ensemble des dictionnaires spécifique des formes dans la k-ème sous-bande se réduit alors à un unique dictionnaire composé de l'unique vecteur $\underline{f}^k$. L'allocation des bits est faite directement à partir de $S_X(f)$ et de $S_Q(f)$. Il n'est plus nécessaire de calculer explicitement les signaux de sous-bande. Le banc de filtres d'analyse disparaît de notre schéma de codage. On obtient un schéma de codage de type "analyse-par-synthèse".

Considérons maintenant le cas plus général où $N > M$. La i-ème composante polyphasée de $\hat{x}(n)$ admet comme expression

$$\hat{x}(mM+i) = \sum_{j=0}^{N/M-1} \sum_{k=0}^{M-1} \hat{y}_k(m-j) f^k(jM+i). \qquad (9)$$

Dans le schéma de codage standard, la transformation devient une transformation avec recouvrement. On rappelle que l'usage d'une transformation avec recouvrement est obligatoire en codage audio pour atténuer "l'effet de bloc". De façon équivalente dans notre schéma de codage, la quantification vectorielle devient une quantification vectorielle avec recouvrement (OVL-VQ). Le principe est donné Figure 2 lorsque $N/M = 2$. Comme dans le codeur CELP, on est obligé d'enlever

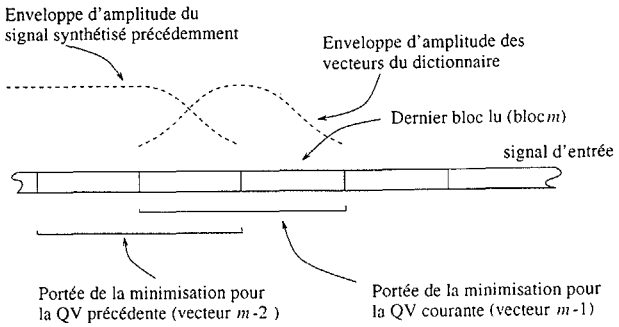

Figure 2: Principe de la quantification vectorielle avec recouvrement.

d'abord au vecteur $\underline{x}(m)$ le "ringing", la contribution due aux fenêtres d'analyse précédentes, avant de rechercher un vecteur dans un dictionnaire.

4. DESCRIPTION DU CODEUR

Le schéma du codeur SBAS est montré Figure 3. Une analyse spectrale du signal d'entrée est réalisée

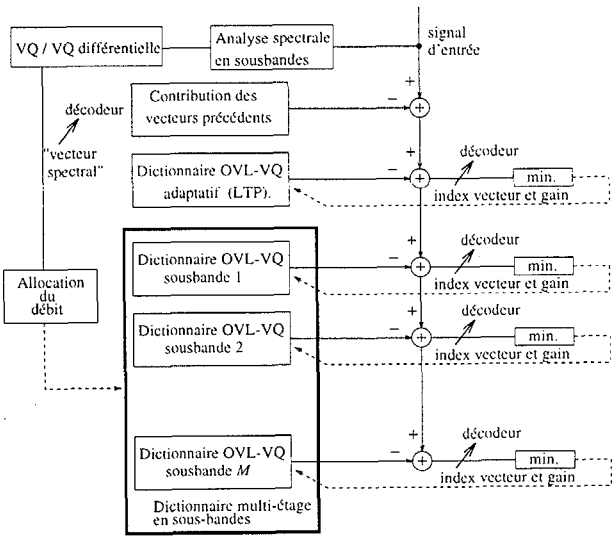

Figure 3: Schéma de principe du codeur SBAS.

périodiquement (pas forcément tous les M échantillons). Elle permet d'en déduire l'allocation de bits qui est donc dynamique. Cette information doit être transmise au décodeur. Comme dans le codeur Aspec, on réserve un certain nombre de bits pour spécifier, non pas l'allocation elle-méme, mais plutôt une "description du spectre", par exemple par les coefficients LSP. Les indices et les gains des vecteurs de sous-bande sont calculés successivement, avec une procédure d'analyse-par-synthèse, dans une structure de type multi-étages. L'ordre des étages est donné par la puissance des sous-bandes: on quantifie d'abord la plus puissante, puis la

deuxième plus puissante, etc. Le critère pour déterminer un vecteur à chaque étage est la minimisation de la distance euclidienne par rapport au vecteur "cible". Chaque vecteur de signal, à la synthèse, est la somme "à recouvrement" des contributions de plusieurs vecteurs successifs des sous-bandes où on alloue un nombre non-nul de bits.

Dans la version actuelle du codeur, les vecteurs contenus dans les dictionnaires de sous-bande sont des signaux à bande étroite; on a sélectionné des dictionnaires composés soit de sinusoïdes modulées par la réponse impulsionnelle d'un filtre passe-bas dont un exemple est donné Figure 4, soit de bruits bande étroite. Ces dictionnaires pourraient être construits à partir d'une base d'apprentissage par un algorithme de type LBG.

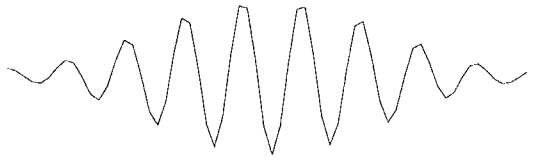

Figure 4: Exemple d'un vecteur élément d'un dictionnaire de sous-bande.

Si ce codeur est utilisé pour coder du signal de parole, son adaptation à ce type de signal est réalisée de la façon suivante. Comme le montre le schéma de la Figure 3, on rajoute un étage de quantification avec un dictionnaire "adaptatif" permettant d'exploiter la corrélation à long terme. On choisit, comme dictionnaires de sous-bande, des dictionnaires composés de sinusoïdes modulées ou de bruits bande étroite suivant une décision voisé/non-voisé comme dans le codeur MBE [4]. Les spectres du signal synthétique et du bruit de reconstruction sont déterminés dans le codeur SBAS par l'allocation de bits; l'utilisation d'un filtre perceptuel n'est donc pas nécessaire. Dans les sous-bandes où on n'alloue aucun bit, le signal synthétique a une puissance très faible, et par conséquent, dans ces sous-bandes le bruit est quasiment égal au signal. Contrairement au codeur CELP où on met l'accent sur le codage des zones spectrales inter-formantiques au détriment des formants, dans le codeur SBAS on concentre l'effort de quantification sur les formants, tout en gardant le niveau du signal synthétique très faible dans les zones spectrales où on n'alloue pas de bits.

5. SIMULATIONS ET PERFORMANCES

On a réalisé quelques simulations préliminaires concernant un codeur de parole en bande téléphonique à un débit égal à 8 kbit/s. On a sélectionné les valeurs numériques suivantes : blocs de 40 échantillons, dimension des vecteurs composant le dictionnaire adaptatif ou de sous-bande égale à $N = 120$, $M = 16$ sous-bandes réalisant une partition non-uniforme de l'axe des fréquences (8 sous-bandes de largeur de bande 125 Hz entre 0 et 1 kHz, 4 sous-bandes de largeur de bande 250 Hz entre 1 et 2 kHz, 4 sous-bandes de largeur de bande 500 Hz entre 2 et 4 kHz) correspondant approximativement aux bandes critiques. L'analyse spectrale est faite tous les 160 échantillons et réclame 28 bits. L'allocation de bits est réalisée suivant l'algorithme décrit dans [3, page 234]. L'indice et le gain spécifiques du dictionnaire adaptatif sont déterminés tous les 40 échantillons. Ils réclament respectivement 7 et 2 bits. On utilise 12 bits pour les indices (la forme) et 12 bits pour les gains pour coder l'ensemble des sous-bandes.

La complexité de cette méthode est limitée parce qu'on utilise une quantification vectorielle multi-étages avec des dictionnaires de taille relativement réduite et parce que le dictionnaire de quantification est fixe (pas de filtrage de dictionnaires). Le retard de codage "end-to-end" d'un tel codeur (avec les paramètres définis ci-dessus) est également faible, de l'ordre de 25 ms.

Des test d'écoute informels ont montré que la performance du codeur reste raisonnable dans le cadre d'une étude préliminaire et que cette méthode semble prometteuse.

6. REFERENCES

[1] T. Berger. *Rate-distortion theory : A mathematical basis for data compression.* Prentice-Hall, 1971.

[2] E.B. George and M.J.T. Smith. Generalized overlap-add sinusoidal modeling applied to quasi-harmonic tone synthesis. *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 1993.

[3] A. Gersho and R.M. Gray. *Vector quantization and signal compression.* Kluwer Academic Publishers, 1992.

[4] D. W. Griffin and J.S. Lim. Multiband excitation vocoder. *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 36, No 8:1223–1235, August 1988.

[5] N. Jayant, J. Johnston, and R. Safranek. Signal compression based on models of human perception. *Proceedings of the IEEE*, Vol. 81, No. 10:1385–1422, October 1993.

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

[7] 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.

[8] G. Yang. *Speech coding at low bit rates.* PhD thesis, Mons Faculty of Technology, Belgium, September 1993.

[9] M. Yong and A. Gersho. Subband vector excitation coding with adaptive bit-allocation. *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pages 743–746, 1989.