Codage de la parole basé sur la Transformation en Ondelettes Discrètes Apport des modèles psycho–acoustiques

Radwan Kastantin * , **Dan Ştefănoiu** *,** , **Gang Feng** *

* Institut de la Communication Parlée – URA CNRS 368, Équipe "Traitement et Codage de la Parole"
1180 Avenue Centrale, B.P. 25, 38040 Grenoble Cedex 9, FRANCE.

** Université "Politehnica" de Bucarest, Faculté "Automatique et Ordinateurs", Groupe "Identification des Systèmes et Traitement du Signal"
313 Splaiul Independenţei, Sector 1, 77206 – Bucarest, ROUMANIE.

@-mails: kastant@icp.grenet.fr, dan@icp.grenet.fr, dan@indinf.pub.ro, feng@icp.grenet.fr

Dans cette communication, nous présentons un nouvel algorithme de codage de la parole à débit variable, en bande téléphonique. Sa particularité réside à la fois dans l'utilisation d'une *Transformation en Ondelettes Discrètes* et dans l'intégration d'un modèle psycho–acoustique, ce qui permet d'obtenir une qualité proche du signal original, avec un débit moyen de 10 kbits/s.

In this paper, a new variable bit rate speech coding algorithm for telephonic speech domain is proposed. The properties of the *Discrete–Time Wavelet Transform* and several human psycho–acoustic features are integrated in order to obtain a high quality of coded signal, close to the original speech. This combination allows us to reduce the mean bit rate in the range of 10 kbit/s.

1. Introduction

Le codage par transformée est un outil efficace de compression de la parole ou des images [5]. Son principe est basé sur la réduction de la redondance du signal, à l'aide d'une transformation orthogonale. La transformation peut conduire à l'intégration efficace des propriétés psycho-acoustiques du système auditif humain, à condition qu'elle permette une interprétation fréquentielle de ses coefficients.

Pour comparer les performances de différentes types de transformations possible à utiliser, on dispose d'un critère objectif: le gain de codage [5]. Selon ce critère, la Transformation de Karhunen-Loève (TKL) [5] est théoriquement optimale. Cependant, la TKL ne peut pas être utilisée à cause de deux inconvenients majeurs: d'une part, la base des signaux associée dépend de la statistique du signal à coder et, d'autre part, il est difficile d'interpréter ses coefficients en fréquence. Pour cette raison, en pratique on utilise d'autres transformations, notamment la Transformation du Cosinus Discrète (TCD) [4] où la Transformation en Ondelettes Discrètes (TOD) [10], qui réalisent une segmentation en sous bandes (égales où inégales) du spectre du signal analysé. Le gain de codage de ces transformations est sous optimal (à l'égard de la TKL), mais elles permettent des implémentations efficaces.

L'algorithme de codage que nous allons présenter est basé sur la TOD, qui, par ses propriétés intéressantes, nous a permis d'intégrer un modèle du système auditif humain [11], afin de réaliser un meilleur masquage du bruit et d'améliorer le facteur de compression.

Nous précisons ensuite la manière d'élaboration du codeur, en particulier les solutions aux problèmes spécifiques suivants: le choix de la transformation, la sélection des coefficients, la quantification de ces derniers et des informations auxiliaires, ainsi que la construction des codes.

2. Description de l'algorithme de codage

2.1 Choix de la transformation

Parmi les classes des transformations orthogonales, nous avons sélectionné la TOD à support fini [10], qui possède plusieures propriétés intéressantes pour le codage de la parole. En effet, ces propriétés sont dûes à la structure de l'espace des signaux discrèts d'énergie finie, l^2, qui reflecte la structure multirésolutive de son analogue continu, L^2 [3], [2]. La TOD est orthogonale (inversable), récurrente et la régularité de ses ondelettes varie linéairement avec la taille du support. L'effet en fréquence de la TOD consiste dans une segmentation (uniforme ou non uniforme) du spectre selon un découpage souhaité. Le contrôle de la régularité aussi que de l'effet en fréquence sont des propriétés particulièrement importantes et spécifiques à cette transformation. Elles permettent l'amélioration du facteur du compression des signaux non stationnaires comme la parole.

La définition de la TOD part d'une paire des opérateurs de l^2, (F_0, F_1) et de leurs adjoints, (F_0^*, F_1^*):

$$\left[\begin{array}{lll} (F_0 s)[n] & \stackrel{déf}{=} & \sum_{p\in\mathcal{Z}} h[p-2n]\, s[p] \\ (F_1 s)[n] & \stackrel{déf}{=} & \sum_{p\in\mathcal{Z}} g[p-2n]\, s[p] \quad \forall s \in l^2 \\ (F_0^* s)[n] & \stackrel{déf}{=} & \sum_{p\in\mathcal{Z}} h[n-2p]\, s[p] \quad \forall n \in \mathcal{Z} \\ (F_1^* s)[n] & \stackrel{déf}{=} & \sum_{p\in\mathcal{Z}} g[n-2p]\, s[p] \end{array}\right. , \qquad (1)$$

où $h, g \in l^2$ sont l'ondelette-père, respectivement l'ondelette-mère discrètes (les réponses impulsionnelles d'une paire de Filtres Miroirs en Quadrature (FMQ), en effet). Par conséquent, ces opérateurs sont en quadrature:

$$F_0F_0^* \equiv F_1F_1^* \equiv \Im\,, \qquad F_0F_1^* \equiv F_1F_0^* \equiv 0\,, \qquad F_0^*F_0 + F_1^*F_1 \equiv \Im\,, \tag{2}$$

où $\Im$ est l'opérateur identité. Si le signal original (s) contient 2^L échantillons, alors la TOD directe (d'analyse) est définie par les relations récurrentes suivantes:

$$\left[\begin{array}{lcl} C^{j+1,0} & \equiv & F_0C^{j,0} \\ C^{j,1} & \equiv & F_1C^{j,0} \end{array}\right. \quad \forall j \in \overline{0, L-1} \qquad \left[\begin{array}{lcl} C^{j,2k} & \equiv & F_0C^{j,k} \\ C^{j,2k+1} & \equiv & F_1C^{j,k} \end{array}\right. \quad \begin{array}{l} \forall k \in \overline{1, 2^{L-j-1}-1} \\ \forall j \in \overline{0, L-2} \end{array} \tag{3}$$

où $C^{0,0} \equiv s$ et $C^{j,k} \stackrel{déf}{=} \{C_p^{j,k}\}_{p\in\mathcal{Z}}$ est la suite des coefficients d'ondelette correspondants à la sous bande k de l'octave j [10]. Dans ce contexte, L indique la plus grossière résolution temporelle acceptée dans l'analyse. Si tous les coefficients ci-dessus sont évalués, alors le découpage en fréquence réalisé par la TOD est uniforme. Pour obtenir un découpage non uniforme, il suffit d'évaluer seulement une partie de ces coefficients.

La TOD inverse (qui réalise la synthèse exacte) est définie d'une manière récurrente aussi:

$$\left[\begin{array}{lcl} C^{j,k} & \equiv & F_0^*C^{j,2k} + F_1^*C^{j,2k+1}\,, \\ & & \forall k \in \overline{2^{L-j-1}-1, 1},\ \forall j \in \overline{L-2, 0} \\ C^{j,0} & \equiv & F_0^*C^{j+1,0} + F_1^*C^{j,1}\,, \forall j \in \overline{L-1, 0} \end{array}\right. \tag{4}$$

Grâce aux relations récurrentes (3) et (4), l'implémentation de la TOD (analyse et synthèse) peut être réalisée à l'aide d'une paire de bancs d'opérateurs en quadrature, organisés comme des arbres binaires. La structure de l'arbre d'analyse correspond uniquement à un découpage en fréquence réalisé par la TOD. Ceci nous permet d'effectuer un choix de la TOD selon une segmentation en fréquence prédéfinie. Un découpage prédéfini peut être élaboré à partir de l'échelle des barks (la correspondance Hertz–Bark) [11], afin de prendre en compte des caractéristiques du système auditif humain. Le modèle de la perception auditive [11] exprime, en effet, la séléctivité de l'oreille humaine, au niveau des bandes critiques, à l'égard du bruit qui entache le signal utile.

À partir de ce modèle, nous avons construit l'arbre non symétrique d'analyse contenant 17 sorties et 32 blocs d'opérateurs de la Figure 1. L'étiquette 0 ou 1 associée à chaque bloc indique l'opérateur F_0, respectivement F_1. Cette structure de la TOD directe a été choisie selon les critères suivants: que le découpage fréquentiel non uniforme résulté soit tout proche de l'échelle des barks (17 bandes critiques pour le domaine téléphonique) et que la complexité de l'algorithme d'analyse-synthèse correspondant soit minimale.

En ce qui concerne les ondelettes d'analyse, nous avons choisi la classe de Daubechies [3], qui est la meilleure à l'égard d'autres classes selon le gain de codage [7].

Le signal de parole est échantilloné à 8 kHz et segmenté en trames de 32 ms avec un recouvrement de 2 ms, ce qui nous permet d'éliminer l'effet de bord entre les trames.

2.2 Sélection des coefficients d'ondelette

Grâce à la localisation fréquentielle des coefficients d'ondelettes et aux propriétés du système auditif humain, il est possible de sélectionner seulement une partie d'entre eux pour la transmission. La sélection est basée sur une classification de plus en plus raffinée des coefficients.

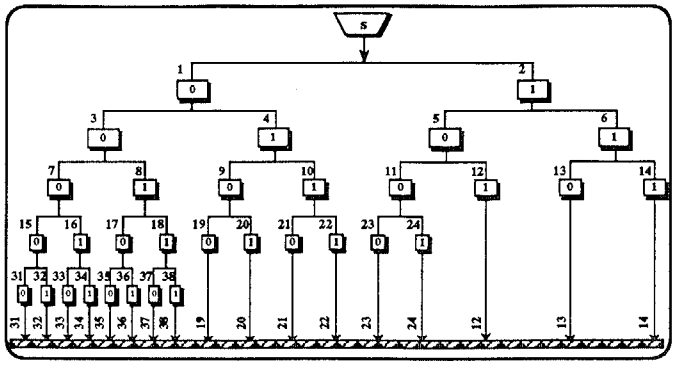

Figure 1: *Le banc d'operateurs en quadrature associé à la correspondance Hertz-Bark dans la bande téléphonique.*

D'abord, les trames de signal de parole (et les coefficients correspondants compris) sont classé en trois catégories selon le critère suivant [4]: si l'énergie d'une trame est inférieure à un certain seuil, elle constitue un morceau de silence; au cas contraire, la trame est soit non voisée (si le premier terme d'autocorrelation est inférieur à 0.6 ou si le nombre de changements de signe dans une durée de 10 ms est supérieur à 100), soit voisée.

Ensuite, on classe les coefficients de chaque trame selon un critère basé sur le phénomène de masquage fréquentiel [11], [8]. Ce phénomène exprime la capacité de l'oreille humaine de masquer les bruits selon leur intensité et leur fréquence. Plus exact, tous les bruits qui entachent le signal utile, dont le spectre est situé en desous d'une courbe dite "de masquage" sont inaudibles. Une approximation en escalier de la courbe de masquage peut être déterminée avec l'algorithme de Johnston [6]. À chaque bande critique $i \in \overline{1, 17}$ correspond un seuil de masquage constant, $\mathcal{S}[i]$. Afin de présenter la procédure de calcul des seuils de masquage, nous allons simplifier les notationsdes coefficients: C_p^i remplacera désormais $C_p^{j,k}$, étant donné que la paire (j, k) détermine uniquement une bande i de localisation. La procédure comprend 3 étapes.

1. *Le calcul du spectre énergétique*

D'abord, l'énérgie du signal répartie dans la bande $i \in \overline{1, 17}$ (avec m_i coefficients correspondants) s'exprime comme suit:

$$\mathcal{E}[i] = \sum_{p=1}^{m_i} \left|C_p^i\right|^2 \tag{5}$$

L'ensemble $\{\mathcal{E}[i]\}_{i\in\overline{1,17}}$ est une approximation en escalier du spectre original, qui s'appelle "spectre énergétique".

2. *Le calcul du spectre auditif*

Le spectre énergétique est utilisé, ensuite, dans le calcul du "spectre auditif" $\mathcal{A}$, selon la formule de convolution en fréquence suivante:

$$\mathcal{A}[i] = (\mathcal{E} \star \mathcal{B})[i] = \sum_{n=1}^{17} \mathcal{E}[n]\,\mathcal{B}[i-n]\,, \quad \forall i \in \overline{1, 17}\,. \tag{6}$$

Dans la relation (6), $\mathcal{B}$ est une approximation de la fonction d'étalement spectral caractéristique à l'oreille humaine,

qui s'exprime comme suit (en échelle linéaire) [8]:

$$\mathcal{B}[n] \stackrel{déf}{=} \begin{cases} 10^{+2.7n} & , \quad n < 0 \\ 1 & , \quad n = 0 \\ 10^{-n} & , \quad n > 0 \end{cases} \tag{7}$$

Pratiquement, le spectre auditif réel a été à son tour approximé en escalier par $\mathcal{A}$, sous l'hypothèse simplificatrice que dans chaque bande $i \in \overline{1,17}$ il n'y a qu'une seule raie d'amplitude $\mathcal{E}[i]$, centrée au milieu de la bande [6].

3. *Le calcul des seuils de masquage*

Pour obtenir l'approximation en escalier de la courbe de masquage, $\mathcal{S}$, il faut corriger le spectre auditif comme suit:

$$[\mathcal{S}[i]]_{dB} = \max\{[\mathcal{A}[i]]_{dB} - RSM[i]\,,\, \mathcal{S}_a[i]\}\ ,\ \forall i \in \overline{1,17}, \tag{8}$$

où $[x]_{dB}$ indique la valeur en dB de x, $RSM[i]$ est un offset (en dB) qui exprime le rapport signal sur masquage et $\mathcal{S}_a[i]$ constitue le seuil absolu d'audition (en dB toujours), pour chaque bande i. La valeur de RSM dépend à la fois de la nature du spectre du signal (tone ou bruit) [6] et de la bande courrante des fréquences. Pour établir cette valeur, nous avons utilisé la classification grossière des coefficients d'ondelette et les résultats de [4]. Les trames non voisées et le silence, bénéficient d'un offset constant (8dB). Pour les trames voisées, l'offset varie comme suit:

$$RSM[i] = \begin{cases} 20\,dB & , \quad i \in \overline{1,8} \\ (28 - i)\,dB & , \quad i \in \overline{9,17} \end{cases} \tag{9}$$

Les seuils absolus d'audition ont été évalués expérimentalement pour chaque bande i. Finalement, les seuils de masquage $\mathcal{S}[i]$ ont été normalisés [6].

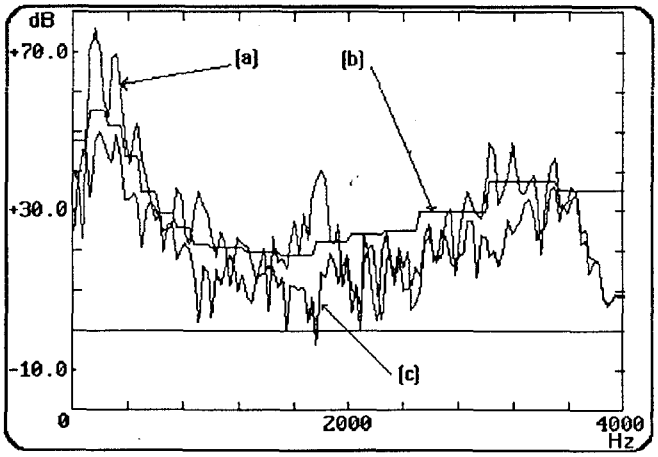

Figure 2: *Spectre d'une trame du signal (a), courbe de masquage en escalier associée (b) et bruit de quantification (c) pour l'allocation par coefficient.*

Dans la Figure 2, nous présentons un exemple de courbe de masquage en escalier, déterminée à l'aide de l'algorithme d'auparavant, pour une trame du signal original. Pratiquement, chaque seuil de masquage $\mathcal{S}[i]$ représente la quantité maximale de bruit qu'il est possible à ajouter dans la bande i, sans dégrader la qualité auditive du signal original. Pour réaliser un affinement de notre classification, nous avons supposé que la quantité $\mathcal{S}[i]$ est uniformément répartie aux m_i coefficients de la sous bande i. On obtient, alors, la quantité suivante: $\mathcal{M}[i] = \mathcal{S}[i]/m_i$, $\forall i \in \overline{1,17}$. Elle constitue, pratiquement, la variance maximale admissible du bruit induit par la quantification des coefficients.

Ces observations sont utiles dans le calcul des pas de quantification correspondants à un quantificateur uniforme:

$$\Delta[i] = \sqrt{12\,\mathcal{M}[i]}\,, \quad \forall i \in \overline{1,17}. \tag{10}$$

Pour ce type de quantificateur, tous les coefficients de la bande $i \in \overline{1,17}$ qui vérifient la contrainte suivante:

$$|C_p^i| \geq \frac{\Delta[i]}{2} \tag{11}$$

sont "non masqués" (nécessairs dans la synthèse), alors que les autres coefficients sont "masqués" (non nécessairs dans la synthèse).

Grâce aux propriétés conjointes de la TOD et du système auditif, nous avons estimé les pourcentages moyens de coefficients masqués: 40% pour les trames non voisées et 60% pour les trames voisés ou le silence.

2.3 Quantification et allocation des bits

Les coefficients non masqués sont uniformément quantifiés. En ce qui concerne l'allocation des bits, nous avons étudié deux méthodes: par coefficient et par bande critique.

1. *Allocation des bits par coefficient et codage de Huffman*

Les pas de quantification $\{\Delta[i]\}_{i\in\overline{1,17}}$ sont utilisés pour déterminer le nombre de niveaux nécessaires dans la quantification de chaque coefficient:

$$N_p^i = \left\lfloor \frac{|C_p^i|}{\Delta[i]} + \frac{1}{2} \right\rfloor\ ,\quad \forall p \in \mathcal{Z}\ ,\ \forall i \in \overline{1,17} \tag{12}$$

(où $\lfloor x \rfloor$ représente la partie entière de x). Pour chaque coefficient C_p^i, il y a deux catégories d'informations à transmettre qui déterminnent la structure des bits alloués: principale (le nombre de niveaux N_p^i) et auxiliaire (le pas de quantification $\Delta[i]$, le signe et le type de masquage).

Nous avons constaté que, pour un signal de parole de durée égale à 1 minute, prononcée par deux types de locuteurs (homme et femme), l'histogramme du pourcentage des

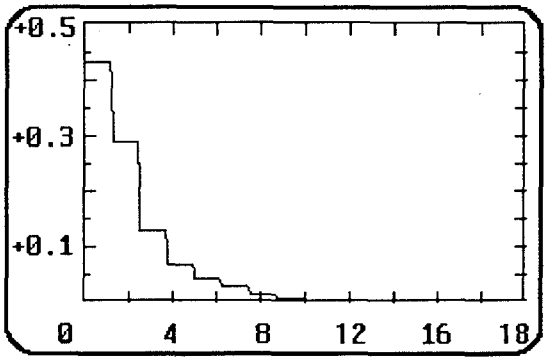

Figure 3: *Histogramme du nombre de niveaux nécessaires dans la génération des codes de Huffman.*

coefficients en fonction du nombre des niveaux de quantification montre comme dans la Figure 3. On voit facilement que plus de 40% de coefficients sont masqués et que la distribution des niveaux est non uniforme. Ceci nous amène à utiliser un codage de type entropique, notamment de Huffman [5].

Dans un code de Huffman, nous avons réservé non seulement les bits pour la valeur du nombre des niveaux de quantification et du signe, mais aussi un bit pour le type

de masquage. Pour indiquer le pas de quantification au décodeur, nous pourons envoyer un code supplémentaire à chaque bande. Hormis le pas $\Delta[i]$ (10), ce code pourrait représenter soit le seuil de masquage $\mathcal{S}[i]$ (8), soit l'énergie $\mathcal{E}[i]$ (5) où la dispersion $\sigma^2[i] = \mathcal{E}[i]/m_i$. Nous avons choisi de coder les dispersions des bandes, afin de pouvoir exploiter le silence dans la diminution du débit de transmission. Chaque dispersion a été quantifiée uniformément en échelle logarithmique, avec 4-5 bits.

Cette méthode nous a amené à un débit moyen de transmission de 18 kbits/s. Le rapport signal sur bruit (RSB) a été évalué à $20dB$ en moyenne. Le bruit de quantification induit est bien masqué, comment il le montre, par exemple, la Figure 2. Dans les tests subjectifs effectués, la plupart des auditeurs n'ont pas trouvé de différences entre le signal codé et le signal original.

2. *Allocation des bits par bande critique*

Dans cette approche, tous les coefficients localisés dans une bande $i \in \overline{1,17}$ reçoivent le même nombre de bits, $R[i]$. Il est calculé comme suit:

$$R[i] = \log_2 \frac{\varepsilon[i]\,\sigma[i]}{\sqrt{\mathcal{M}[i]}}, \qquad (13)$$

où $\varepsilon[i]$ est un facteur de performance qui dépend à la fois du quantificateur utilisé et de la dénsité de probabilité de la grandeur à quantifier [5]. Dans ce cas, nous avons utilisé un quantificateur de type Max-Lloyd optimal [9] pour quantifier les valeurs des coefficients normalisées par la dispersion de la bande de localisation correspondante. (Les dispersions ont été quantifiées comme auparavant.)

Le débit moyen de transmission est, dans ce cas, de 18 kbits/s également, et le RSB moyen de $19,2\,dB$. Le bruit de quantification induit est pratiquement masqué,

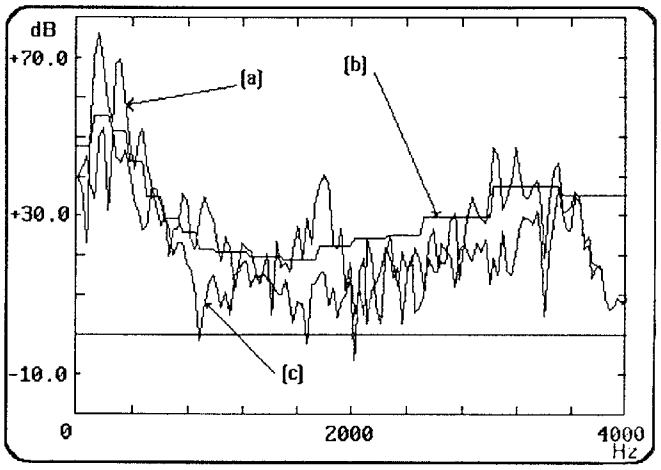

Figure 4: *Spectre d'une trame de signal (a), courbe de masquage en escalier associée (b) et bruit de quantification (c) pour l'alocation par bande.*

comment on peut le voir dans la Figure 4, où la trame du signal est identique avec celle de la Figure 2. Cependant, la qualité auditive est légèrement inférieure par rapport au cas précédent, à cause du fait que les coefficients d'amplitudes plus grandes ont été défavorisés dans la quantification.

2.4 Exploitation du silence pour diminuer le débit

Malgré la qualité auditive obtenue, le débit de transmission reste élevé. Pour cette raison, nous avons exploité l'observation que le silence occupe jusqu'à 60% du temps d'une conversation [1]. En effet, toute trame de silence contient seulement un bruit de fond et le système auditiv perçoit l'enveloppe de son spectre. Par conséquent, il n'est pas nécessaire à quantifier les coefficients correspondants avec la même fidélité que les autres. Il suffit, donc, d'envoyer au décodeur seulement les codes des dispersions des bandes. Au décodeur, le bruit de fond est généré en utilisant une source de bruit gaussien dont la variance dépend des dispersions transmises.

Nous avons ainsi obtenu un débit moyen de 10 kbits/s, sans une dégradation importante de la qualité.

3 Conclusion

Dans cette communication, nous avons présenté un nouvel algorithme de codage de la parole, à débit variable. La démarche est basée sur deux approches importantes: l'emploi de la Transformation en Ondelettes Discrètes et l'intégration des propriétés psycho–acoustiques du système auditif humain, ce qui conduit à diminution du débit moyen de transmission. Les résultats obtenus confirment l'intérêt pour l'application de la TOD dans le codage de la parole.

Références

[1] **Brady P.T.** *A statistical analysis of on–off patterns in 16 conversations.* Bell Syst. Tech. J., 47(1):25–35, January 1968.

[2] **Coifman R., Meyer Y., Quake S., Wickerhauser M.V.** *Signal Processing and Compression with Wavelet Packets.* In Proc. of the 3rd Intl. Conf. on Wavelets and Applications, 1992.

[3] **Daubechies I.** *Orthonormal Bases of Compactly Supported Wavelets.* Comm. on Pure and Appl. Math., XLI:909–996, 1988.

[4] **Dia H.** Codage par transformée de la parole à bande élargie (0-7 kHz). PhD thesis, Inst. de la Comm. Parlée, Univ. Stendhal, Inst. National Polytechnique de Grenoble, FRANCE, October 8, 1993.

[5] **Jayant N.S., Noll P.** *Digital Coding Waveforms – Principles and Applications to Speech and Video.* Prentice Hall (editor: A.V. Oppenheim), New Jersey, U.S.A., 1984.

[6] **Johnston J.D.** *Estimation of Perceptual Entropy Using Noise Masking Criteria.* I.E.E.E. Conf. on Speech Coding, 2524–2527, 1988.

[7] **Kastantin R., Ştefănoiu D., Feng G., Martin N., Mrayati M.** *Optimal wavelets for high quality speech coding.* In Proc. of the Symposium EUSIPCO '94, 399–403, 1994.

[8] **Malıieux Y.** *High Quality Audio Transform Coding at 64 kbits/s.* Annales de la Télécommunication, 47(3–4):95–106, 1992.

[9] **Max J.** *Quantizing for minimum distortion.* I.R.E. Trans. on Information Theory, I(6):7–12, 1960.

[10] **Ştefănoiu D.** *Signal analysis by the time–frequency methods.* PhD thesis, Univ. "Politehnica" of Bucharest, Dept. of Automatic Cntl and Computer Science, ROMANIA, April 1995.

[11] **Zwicker E., Feldtkeller R.** *L'oreille récepteur d'information.* Masson, 1981.