APPLICATIONS DU CODAGE PREDICTIF AVEC QUANTIFICATION VECTORIELLE A LA COMPRESSION NUMERIQUE DE LA PAROLE

D.Toubol, C.Galand, M.Rosso

Centre d'Etudes et Recherches,
IBM France
06610 La Gaude

J.Menez

Lassy UA CNRS 814
41 Bd Napoleon III
06041 Nice Cedex
France

RESUME

Récemment introduite dans le codage prédictif du signal de parole, la quantification vectorielle (CELP - Code Excited Linear Predictive Coding), permet d'obtenir une qualité utilisable en téléphonie à des débits de l'ordre de 4 à 8 kbit/s. Cependant, les algorithmes proposés à l'origine [1] pour ce type de codeur sont très complexes (de l'ordre de 40 MIPS - millions d'instructions par seconde et 40 koctets - mémoire de données), et conduisent à des mises en oeuvre coûteuses et incompatibles avec les objectifs d'utilisation en radio téléphonie cellulaire où l'on doit satisfaire à des contraintes d'encombrement et de puissance dissipée. C'est pourquoi les recherches se sont rapidement orientées vers une réduction de cette complexité.

Cet article est organisé de la façon suivante: nous commençons par rappeler le principe de base des techniques CELP et par mettre en évidence leur complexité.

Puis nous proposons une technique originale de quantification vectorielle , permettant de remplir les mêmes objectifs de qualité que le CELP, tout en réduisant la complexité de mise en oeuvre microprogrammée de l'algorithme. La technique repose d'une part sur l'obtention d'un effet de masquage spectral par pré-emphase particulière du signal de parole, et non plus par filtrage de chaque séquence d'essai, et d'autre part sur l'utilisation d'un vocabulaire linéaire dont sont extraites les séquences d'essai.

Finalement, nous décrivons deux applications de notre technique à un codeur à bande complète et à un vocodeur à bande de base. Des tests de qualité sont rapportés, et démontrent l'efficacité des techniques proposées. C'est ainsi qu'on montre que la complexité de mise en oeuvre peut être réduite à 5-8 MIPS et 4 koctets sans dégradation de la qualité du signal de parole reconstruit.

SUMMARY

Combination of vector quantization with predictive coding has been recently proposed for speech compression. The so-called CELP (Code Excited Linear Predictive Coding) technique gives high quality speech at low bit rates (4 to 8 kbps), but originally led to high complexity implementations (40 MIPS - millions of instructions per second, and 40 K RAM) incompatible with the requirements of cellular telephone applications. This is why in the past few years, much work has been conducted to reduce the CELP complexity.

This paper is organized as follows: we first remind the basic CELP techniques and we outline their native complexity. Then, we propose a new vector quantization technique which allows to reach the same quality than the original CELP, with a reduction of the implementation complexity. The proposed technique is based on one hand on a noise spectral shaping effect obtained by special pre-emphasis of the speech signal, and on the other hand on the use of a linear codebook. In the third part, we describe two applications of this technique to a full-band coder and to a base-band vocoder. We report quality tests which validate the approach.

INTRODUCTION

Le CELP (Code Excited Linear Predictive Coder) est une méthode de codage du signal de parole proposée en 1985 par B.S Atal et M.R Schroeder [1], permettant d'atteindre une qualité satisfaisante sur le plan de la perception auditive, pour de faibles débits de transmission. Son principe s'appuie sur l'utilisation de techniques de prédiction linéaire et de quantification vectorielle.

Ce type de quantification nécessite un traitement vectoriel du signal de parole. La méthode d'analyse par synthèse employée dans le CELP (fig.1) consiste à produire à l'émetteur, des séquences de signal vocal de synthèse, obtenues à partir des séquences (vecteurs) d'un répertoire, afin de déterminer celle qui se rapproche le plus du signal original, au sens d'un critère pré-défini. Chaque séquence appartenant au vocabulaire est traitée par deux filtres de synthèse relatifs respectivement à la prédiction à long-terme (1/B(Z)) et à court-terme (1/A(Z)). Pour chaque segment de signal synthétique obtenu, on effectue la différence avec le signal original. Un effet de masquage spectral de l'erreur de codage est ensuite réalisé par l'intermédiaire d'un filtre de pondération W(Z). La meilleure séquence est alors déterminée en appliquant le critère des moindres carrés au signal d'erreur pondérée. Cette séquence est représentée par son indice de position I dans le répertoire et un coefficient de gain G, qui sont transmis au récepteur.

Le filtrage de chacune des séquences du répertoire rendait l'algorithme initial peu compatible avec les contraintes d'implémentation, du fait de sa complexité élevée (40 MIPS). De plus, le vocabulaire utilisé à l'origine, constitué de 1024 séquences de 40 échantillons, nécessitait une mémoire de stockage de l'ordre de 40 koctets. C'est pourquoi, le CELP a fait depuis, l'objet de nombreuses études visant à atténuer ces inconvénients.

Nous présentons dans la première partie de cet article deux modifications de l'algorithme initial du CELP:

1. une technique de masquage spectral permettant d'éviter le passage des séquences du répertoire par un filtre de pondération, réduisant ainsi la complexité de l'algorithme.

2. l'utilisation d'un répertoire linéaire adaptatif qui, outre une diminution considérable de la place mémoire, permet un accroissement de la qualité du signal de parole reconstruit.

Ces deux modifications ont mené à l'élaboration d'un codeur ACELP que nous détaillerons dans la deuxième partie. L'extension de ces techniques à un vocodeur à bande de base a permis de réduire encore la complexité de l'algoritme, sans altérer la qualité auditive des signaux vocaux synthétiques. Nous en exposons le principe de fonctionnement, puis nous présentons dans la troisième partie quelques résultats de simulation, relatifs à une application à 8.5 kbps des codeurs présentés.

1 ALGORITHME CELP

1.1 Masquage spectral

Dans le CELP (fig.1), l'effet de masquage spectral est obtenu par la pondération du signal d'erreur p(n), au moyen du filtre W(Z) défini par la relation (1).

$$(1) \qquad W(Z) = \frac{A(Z)}{A(Z/c)} \qquad 0 < c < 1 \quad (c = \text{constante})$$

A(Z) représente ici le filtre inverse du filtre modèle du conduit vocal (2). Traité par A(Z), le signal original de parole engendre un signal résiduel d'enveloppe spectrale plate.

$$(2) \qquad A(Z) = 1 - \sum_{i=1}^{p} a(i).Z^{-i}$$

La technique que nous allons exposer [2,3] repose sur la détermination d'un filtre C(Z) qui remplace le filtre de pondération W(Z) dans le schéma 1. C(Z) doit, non seulement remplir les mêmes fonctions que W(Z), mais de plus, posséder une structure permettant la simplification du schéma de base. La solution apportée consiste à donner à C(Z) la forme d'un filtre de prédiction linéaire à court-terme. La différence avec A(Z) vient de la détermination des coefficients c(i), (i= 1,p). Celle-ci est réalisée, non plus sur le signal original de parole s(n), mais sur un nouveau signal v(n) obtenu par une pré-emphase du signal s(n):

$$(3) \qquad v(n) = s(n) - u.s(n-1)$$

$$(4) \qquad u = \frac{R(1)}{R(0)}$$

où R(k) représente la k-ième valeur de la fonction d'autocorrélation du signal de parole. En utilisant la transformée en Z, on obtient:

$$(5) \qquad V(Z) = (1 - u.Z^{-1}).S(Z) = H(Z).S(Z)$$

Le signal résiduel obtenu en filtrant V(Z) par C(Z) possède une enveloppe spectrale plate. Si l'on filtre à présent le signal original S(Z) par C(Z), on obtient un signal résiduel dont le spectre a même allure que la réponse harmonique du filtre 1/H(Z). On peut également remarquer que la mise en cascade des filtres 1/A(Z) et C(Z) donne naissance à un filtre ayant lui aussi, même allure spectrale que 1/H(Z) (fig.5). En remplaçant W(Z) par C(Z) dans la figure 1, on aboutit au schéma simplifié de la figure 2 dans lequel on considère que le répertoire utilisé possède les caractéristiques spectrales de 1/H(Z). Cette structure permet une étude directe du résiduel issu du filtrage de s(n) par C(Z). Elle évite ainsi le passage des séquences du répertoire par un filtre de pondération. Après quantification, les paramètres sont transmis au récepteur. Là, le signal synthétique d'excitation est filtré par 1/C(Z) qui lui donne une enveloppe spectrale semblable à celle du signal d'origine s(n).
L'erreur additionnelle introduite par le quantificateur est supposée blanche. Ce bruit de quantification adopte donc une densité spectrale de puissance similaire à celle de 1/C(Z), lors de son filtrage par celui-ci. Ainsi que nous l'avons vu, les coefficients c(i), i= 1,p, ont été calculés sur le signal s(n) pré-emphasé grâce à H(Z). Le filtre de synthèse 1/C(Z) présente donc une densité spectrale de puissance similaire à celle du signal de parole s(n), mais avec une pente moyenne quasiment nulle.

Cette technique simplifie considérablement l'algorithme initial du CELP, tout en conservant les qualités du codeur original. Nous détaillerons dans la deuxième partie de cet article une méthode de codage s'appuyant sur ce type de masquage spectral.

1.2 Répertoire linéaire adaptatif

Les répertoires utilisés jusqu'ici dans les codeurs CELP, présentent en général deux inconvénients majeurs:

1. Leur taille, que l'on s'efforce de réduire pour des raisons de stockage, et dans le but d'atteindre des débits plus faibles.

2. Leur structure figée, qui parfois, peut s'avérer inadéquate.

Le premier point est directement conditionné par les contraintes de fonctionnement et de débit imposées au codeur. Nous allons voir qu'il est toutefois possible d'y apporter une solution grâce à l'utilisation d'un répertoire linéaire. Le second point, quant à lui, peut être amélioré par l'utilisation d'un répertoire adaptatif.

Repertoire linéaire

L'algorithme CELP proposé à l'origine [1], utilise un répertoire à deux dimensions de taille conséquente, nécessitant une importante place mémoire. Pour atténuer cet inconvénient, on peut faire usage d'un vocabulaire dit linéaire, ou vocabulaire à une dimension [2,3]. Dans un répertoire de ce type, deux séquences consécutives ne diffèrent entre elles que de i échantillons (i> 0). Elles ne présentent donc plus le caractère d'indépendance statistique présent dans un répertoire à deux dimensions. Toutefois, ce vocabulaire réduit la mémorisation des données à L+(N-1).i échantillons, N étant le nombre de mots, L leur longueur, et i le pas séparant deux séquences consécutives.

Répertoire adaptatif

Le principe du répertoire adaptatif [2] est de renouveler régulièrement une partie de son contenu, à l'aide des séquences précédemment sélectionnées. Cette technique peut être affinée en gardant fixe une partie du répertoire, qui est ainsi divisé en deux zones:
-- la première conserve le caractère stochastique original,
-- tandis que la seconde s'adapte au signal.

Cette dernière contient les caractéristiques du signal qui n'ont pu être modélisées par les techniques de prédiction linéaire. Ainsi conçu, le vocabulaire permet d'accroître les performances du codeur, en complétant efficacement l'action des prédicteurs déjà cités.

Chaque séquence de L échantillons sélectionnée est utilisée pour remplacer une séquence du répertoire.

- Dans le cas d'un vocabulaire à deux dimensions,la séquence à supprimer pose un premier problème: celui du choix. Le second inconvénient rencontré est la faible portée d'une telle méthode; l'effet du renouvellement étant limité à la seule séquence introduite.

- L'usage d'un répertoire linéaire, en revanche, accentue l'effet adaptatif. On supprime par décalage de L échantillons, la séquence la plus ancienne du répertoire. La nouvelle séquence est introduite en fin de répertoire. Elle conditionne donc les (L-1) séquences précédentes. La séquence supprimée est la première de la partie adaptative du répertoire (juste après la partie fixe). Sa disparition affecte donc L autres séquences. Ainsi, le renouvellement d'un mot du vocabulaire entraîne la modification de 2.L mots, tout en conservant une certaine continuité dans le répertoire.

2 APPLICATIONS

2.1 Codeur ACELP

La structure du codeur ACELP [2] (Adaptive Code Excited Linear Predictive Coder) est donnée à la figure 3. Elle se caractérise essentiellement par l'usage de la technique de masquage spectral décrite en 1.1, ainsi que d'un répertoire linéaire adaptatif du même type que celui présenté en 1.2.
Le signal de parole original s(n) est divisé en fenêtres (ou blocs) de durée égale à 20 ms. Pour chaque bloc, on calcule les coefficients du filtre C(Z) par lequel le bloc est ensuite filtré. Le segment de signal résiduel obtenu x(n), est alors analysé afin de déterminer

les paramètres d'un filtre de prédiction à long-terme B(Z). Ce calcul requiert l'utilisation d'une mémoire contenant des blocs d'échantillons du signal résiduel passé. Notons qu'il est possible d'utiliser le signal résiduel synthétique pour remplir cette mémoire, plutôt que le signal original, permettant ainsi d'accroître légèrement les performances subjectives du codeur. Le signal prédit est alors calculé à partir des échantillons passés reconstruits, puis retranché à x(n) pour donner un signal résiduel à long-terme y(n). Chaque fenêtre est découpée en séquences de traitement que l'on code par quantification vectorielle. La recherche de la séquence optimale consiste à déterminer la séquence n°k du répertoire, qui minimise l'expression:

$$(6) \qquad E(k) = \sum_{n} \left(y(n) - G.S(k,n) \right)^2$$

où S(k,n) est la séquence n°k du répertoire, et G un facteur de gain. Les indices k et les gains G correspondants sont transmis au récepteur, où les séquences synthétiques peuvent être reconstituées. Leur juxtaposion constitue une fenêtre de résiduel synthétique à long-terme. Celle-ci est filtrée par le filtre de synthèse à long-terme, puis par le filtre 1/C(Z) pour donner une fenêtre de signal vocal synthétique.

2.2 *Vocodeur en Bande de Base*

Les codeurs CELP ont dans l'ensemble, quelques difficultés à reproduire convenablement les hautes fréquences d'un signal [4]. A partir de cette constatation, nous avons cherché un traitement séparé des bandes hautes et basses fréquences du signal original; celles-ci étant au préalable, déterminées à l'aide d'un filtre passe-bas [7]. Nous voyons fig.3 que la technique ACELP entraîne à l'analyse,une déconvolution à court-terme du signal vocal. Il est alors possible de situer le filtrage passe-bas au niveau du résiduel. La zone de basses fréquences (dite bande de base) est ensuite traitée suivant les techniques décrites en 2.1. La bande complémentaire est régénérée à la synthèse avant le passage du signal d'excitation par le filtre 1/C(Z). Notons que ce concept rejoint celui d'un codeur de type RELP [5], dans lequel la bande de base est traitée par une technique ACELP.
La limitation de cette méthode à la seule région des basses fréquences permet une efficacité accrue dans la zone du spectre la plus énergétique.
De plus, elle permet un sous-échantillonnage du signal avant traitement, ce qui engendre une réduction notable de la complexité et du débit.

Il existe deux raisons majeures à la dégradation du signal dans un codeur en bande de base:
-- Le bruit de codage de la Bande de Base
-- Le bruit introduit par la régénération de la Bande Haute.

L'un et l'autre dépendent essentiellement de la largeur de la bande de base, et du facteur de décimation.
Plus la bande de base est étroite, plus le facteur de sous-échantillonnage est élevé. Il s'en suit une réduction d'autant plus accrue de la complexité et du débit de transmission. Cependant, se trouvant considérablement élargie, la bande haute requiert une technique de reconstruction d'autant plus fiable.

La régénération de la bande hautes fréquences (HFR) a déjà fait l'objet d'études [5,6] qui ont permis d'apporter des solutions variées à ce problème. Nous nous contenterons d'en présenter une qui nous a paru convenir à la structure étudiée.

ACELP-BB

De même que sur la fig.3, le signal vocal est décorrélé par le filtre C(Z), calculé à partir du signal pré-emphasé V(Z). Le signal d'excitation résiduelle obtenu x(n) est alors filtré par un filtre passe-bas de fréquence de coupure Fb = Fe/2.Nd (Fe = fréquence d'échantillonnage), pour donner le signal de bande de base. Ce signal est ensuite sous-échantillonné d'un facteur Nd, puis analysé afin de déterminer les coefficients du prédicteur à long-terme. Le signal résiduel à long-terme est enfin quantifié vectoriellement.
A la synthèse, l'indice des séquences sélectionnées et les gains correspondants, obtenus par quantification vectorielle, permettent de reconstituer un signal d'excitation en bande de base. Après filtrage par le prédicteur à long-terme, le signal de synthèse obtenu est un résiduel à court-terme de fréquence d'échantillonnage 2.Fb. On constitue alors un nouveau signal en insérant (Nd-1) zéros entre deux échantillons successifs.

Régénération de la bande haute

L'insertion des zéros duplique de façon symétrique la bande de base, Nd fois [6], permettant ainsi de reconstruire pour le signal résiduel une bande hautes fréquences. On peut utiliser ce signal comme fonction d'excitation du prédicteur à court-terme 1/C(Z). Cette méthode d'interpolation permet d'obtenir un signal vocal synthétique de fréquence d'échantillonnage égale à 8000 Hz. Bien que de qualité satisfaisante, ce signal présente une déformation de l'enveloppe spectrale qu'il est possible de rectifier. En effet, la technique de masquage spectral par pré-emphase confère au résiduel à court-terme une enveloppe spectrale similaire à la réponse en fréquence de 1/H(Z) (fig.5). Il s'ensuit que le résiduel à court-terme de synthèse doit posséder la même caractéristique, ce qui n'est pas le cas ici (fig.4). La technique de correction que nous avons utilisée est la suivante: La bande de base synthétique est d'abord interpolée par filtrage passe-bas, puis déemphasée afin de présenter une enveloppe spectrale plate. La régénération des hautes fréquences est réalisée par une technique dite de "folding" [6] (duplication avec retournement). Le signal résultant subit ensuite une emphase qui lui rend une allure spectrale similaire à celle de 1/H(Z) (fig.4). Il peut alors servir de signal d'excitation au filtre 1/C(Z). Notons que les opérations d'emphase et de déemphase nécessitent la transmission d'un coefficient, ce qui augmente légèrement le débit binaire.

Complexité de mise en oeuvre

- Le signal résiduel d'excitation étant sous-échantillonné , le décalage autorisé pour la recherche du facteur de périodicité M à long terme est divisé par Nd. Le nombre de points sur lequel s'effectue cette recherche étant également divisé par Nd, la complexité se trouve donc réduite d'un facteur égal à Nd^2
- Dans le cas d'un ACELP-BB, la recherche de la séquence optimale s'opère sur un nombre d'échantillons réduit d'un facteur égal à Nd par rapport à un ACELP; la complexité est par conséquent divisée, là aussi par Nd.

3 *RÉSULTATS DE SIMULATION*

3.1 *Description d'un exemple à 8.5 kbps*

Dans l'exemple que nous présentons, les coefficients du filtre de prédiction linéaire à court-terme C(Z) sont calculés et transmis au récepteur toutes les 20 ms. Chaque fenêtre traitée est décomposée en 4 sous-fenêtres de 5 ms, pour lesquelles sont déterminés les paramètres du filtre de prédiction à long-terme B(Z). Chacune d'elles est enfin divisée en séquences de 2.5 ms qui sont quantifiées vectoriellement. Le répertoire utilisé contient 275 échantillons (256 séquences).
Afin de pouvoir effectuer des comparaisons avec ce codeur ACELP, nous avons choisi d'utiliser pour le codeur ACELP-BB les mêmes caractéristiques ainsi qu'un répertoire de même dimension: 275 échantillons. De ce fait, le débit obtenu est presque inchangé.

		BB-ACELP	ACELP
Allocation de bits	8 LPC coeff.	30	30
	8 LTP coeff.	32	36
	8 indices	64	64
	8 gains	40	40
	1 coeff. HFR	5	

3.2 *Résultats des tests auditifs*

La qualité sur le plan de la perception auditive de ces codeurs a été évaluée au moyen de tests auditifs basés sur un jugement par paires. Les codeurs ACELP et ACELP-BB, fonctionnant à un débit binaire de 8.5 kbps ont traité 8 voix françaises (4 masculines et 4 féminines). Tous les coefficients ont été codés, hormis les facteurs de gains. Pour l'évaluation du débit de transmission, nous avons considéré un codage scalaire sur 5 bits pour chacun d'eux. Pour chaque voix originale, on a obtenu deux voix synthétiques que 8 auditeurs ont ensuite comparées, en indiquant leur préférence. Le résultat est donné à la fig.6 sous forme de pourcentage. On peut en conclure que les deux codeurs sont de qualité équivalente sur le plan de la perception auditive.

CONCLUSION

Dans cet article, nous avons proposé de nouvelles techniques permettant de réduire d'une part la place mémoire, d'autre part la complexité de l'algorithme CELP. Dans le cas d'une application du codeur ACELP à un débit binaire de 8.5 kbps, la complexité est de l'ordre de 8 MIPS (millions d'instructions par seconde). L'utilisation de ces mêmes techniques à un vocodeur à bande de base au même débit (ACELP-BB) permet de réduire la complexité tout en conservant une qualité équivalente. Celle-ci a été d'autre part évaluée à un niveau de 24 dB MALT [2]. Ceci permet d'envisager l'application de ces techniques à la radio téléphonie cellulaire [8] où actuellement la qualité de 24 dB MALT est obtenue à 13 kbps, avec une complexité de 4 MIPS.

RÉFÉRENCES

[1] M.R.Schroeder et B.S.Atal: 'Code-Excited Linear Prediction (CELP): High-quality speech at very low bit rates'. Proc.of ICASSP-85, pp 937-940.

[2] J.Menez, C.Galand, M.Rosso, F.Bottau: 'Adaptive Code Excited Linear Predictive Coder (ACELP)'. Proc.of ICASSP-89

[3] F.Bottau, C.Galand, M.Rosso, J.Menez: 'On different vector predictive coding schemes and their application to low bit rates speech coding'. EURASIP-88, Signal Processing IV, pp 871-874.

[4] P.Kroon, B.S.Atal: 'Strategies for improving the performances of CELP coders at low bit rates'. Proc.of ICASSP-88, pp 151-154.

[5] C.Arnaud: 'Nouvelles méthodes de régénération de la composante haute fréquence du signal d'excitation d'un codeur de type RELP'. Thèse de 3ème cycle, Université de Nice, 1987.

[6] J.Makhoul, M.Berouti: 'High-frequency regeneration in speech coding systems'. Proc.of ICASSP-79, pp 428-431.

[7] A.M.Kondoz, B.G.Evans: 'CELP Base-Band coder for high quality speech coding at 9.6 to 2.4 Kbps'. Proc.of ICASSP-88, pp 159-162.

[8] C.Galand and al.: 'Contribution française à la normalisation du codeur de parole du système Pan-Européen de communication radio-cellulaire'. GRETSI 89.

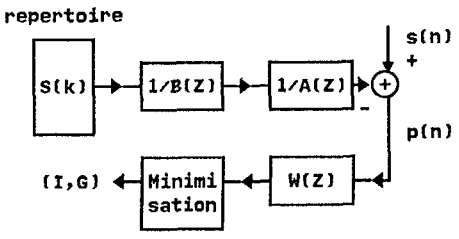


Figure 1: Principe de base du CELP original /1/

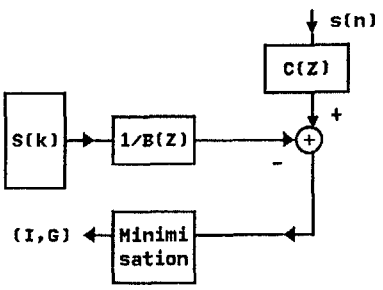


Figure 2: Principe de base du CELP modifié /2,3/

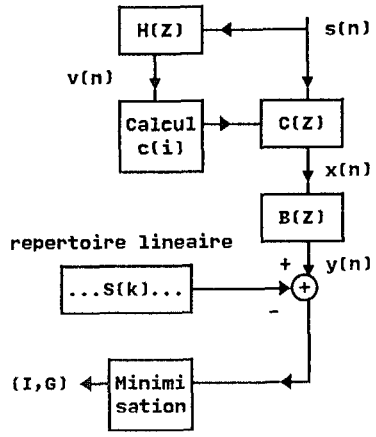


Figure 3: Principe de base de l'ACELP /2/

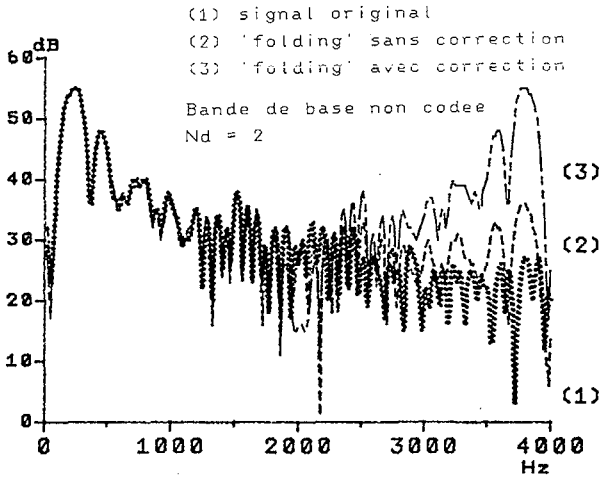
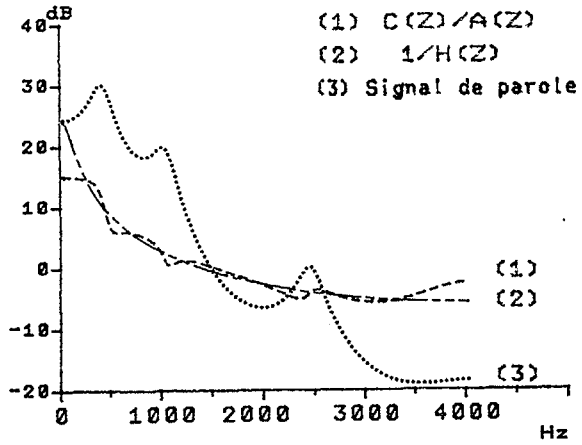


Figure 4: Densité spectrale de puissance du signal résiduel à court terme

Figure 5: Densités spectrales de puissance de:



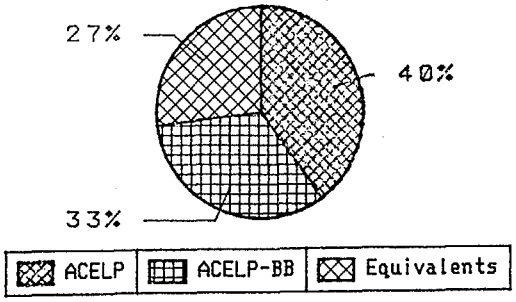


Figure 6: Résultats de simulation: