DIXIEME COLLOQUE SUR LE TRAITEMENT DU SIGNAL ET SES APPLICATIONS

NICE du 20 au 24 MAI 1985

SYNTHESE DE LA PAROLE PAR CODAGE MULTI-IMPULSIONNEL

CAELEN J., SONG J.M.

Laboratoire C.E.R.F.I.A., Université P. Sabatier, 118 Route de Narbonne, 31062 TOULOUSE CEDEX

RESUME

Le codage prédictif linéaire(LPC) est une méthode très utilisée aussi bien en analyse qu'en synthèse de la parole. Cependant, en synthèse, certains défauts introduits par le biais sur les coefficients de prédiction (ou de réflexion) ou par la mauvaise définition de la source, nuisent à la qualité de la parole produite. Le critère de minimisation habituel ne prend pas en compte les propriétés de la perception auditive: il en résulte à l'écoute, un bruit de fond perceptible en dehors des plages formantiques.

En 1982 Atal et Remede ont proposé une méthode plus élaborée, en tenant compte des critiques précédentes: le codage prédictif linéaire multi-impulsionnel (MP-LPC) qui améliore notamment la source mais au prix de calculs assez lourds. Nous proposons dans cet article un algorithme plus rapide, fondé sur une approximation a priori de la source: au lieu de chercher à déterminer les positions optimales de toutes les impulsions dans une trame, on se contente de retenir simplement les K les plus intenses mesurées sur le signal résiduel. Ce signal résiduel est obtenu d'une part par une méthode séquentielle proche de la méthode PARCOR (Partial Correlation) et d'autre part en tenant compte du masquage fréquentiel perceptif.

Ces approximations ne diminuent pas la qualité de la parole de synthèse et n'ont aucune incidence ni sur l'intelligibilité ni sur le naturel de la voix. Ceci autorise donc l'implémentation des procédures de calcul sur processeur spécialisé (SCH Novelec). Les améliorations ainsi apportées à l'algorithme de base, permettent d'atteindre le temps réel sur ce matériel.

SUMMARY

Linear-Predictice Coding (LPC) is a method frequently used both in speech analysis and synthesis. However, in synthesis, a nimber of indeirectly introduced defects, affecting predictive (or reflection) coefficients, hurt the quality of the speech output. The usual minimization criterion does not take into account aural-perception properties, and this results in a definitiley perceivable backgroung noise outside formant frequency-areas.

In 1982 Atal and Remede suggested a more sophisticated method; taking into account the above criticism: Multi-Pulse Linear-Predictive Coding (MP-LPC) which improved the source signal, but necessitated an awkward mass of computations. In this article we suggest a time-saving algorithm, based on a priori approximation at the source: instead of trying to ascertain optimal pulse positions in the chain, we merely retain the K pulses with highest intensity --measured within the residual signal. The latter is secured, on the one hand, through a sequential method akin to the PARCOR (Partial Correlation) method and, on the other hand, by taking into account perceptive frequential masking.

Such approximation do not affect either the quality of synthetized speech or voice intelligibility and naturalness. Therefore, the implementation of computation-procedures on specialized processor (SCH Novelec) is Warranted. The improvment thus conferred to the basic algorithm, makes it possible, on this hardware, to achieve on-line processing.

SYNTHESE DE LA PAROLE PAR CODAGE MULTI-IMPULSIONNEL

CAELEN J., SONG J.M.

1. INTRODUCTION

La méthode d'analyse/synthèse par codage prédictif linéaire (LPC) s'appuie sur les idées fondamentales suivantes:

(a) l'appareil de phonation peut être décomposé en deux parties indépendantes l'une de l'autre, la source et le conduit vocal,

(b) la source est une suite d'impulsions périodiques (de période $T_o=1/F_o$ avec F_o fréquence du fondamental) pour les sons voisés (voyelles par exemple) et un bruit blanc pour les sons sourds (fricative /s/ par exemple),

(c) chaque échantillon du signal de parole peut être prédit par une combinaison linéaire des M échantillons précédents, c'est-à-dire par:

$$/1/ \quad s(n) = -\sum_{i=1}^{M} a_i s(n-i)$$

La relation /1/ modélise le conduit vocal par un filtre tout pôle de coefficients a_i (coeffs. de prédiction).

En synthèse, la source permet de générer le signal excitateur qui alimente le synthétiseur proprement dit, constitué par le filtre inverse de coefficients a_i dont on modifie les valeurs à chaque fenêtre d'analyse. Depuis les années 70 beaucoup de travaux ont porté sur l'amélioration de la qualité de synthèse et surtout sur la forme de l'onde de source, trop artificielle et simplifiée telle qu'en (b), bien qu'elle présente de nombreux avantages dans une perspective de réduction de débit. (téléphone par exemple). Depuis peu, un modèle LPC-multi-impulsionnel proposé par Atal et Remede, permet d'éviter certains défauts inhérents aux approximations faites en (a) et (b) et corrigés par:

(a)' intéraction entre la source et le conduit,

(b)' source multi-impulsionnelle ne nécessitant plus la détection du voisement,

(c)' intégration de l'effet de masquage perceptuel.

Le principe du vocodeur multi-impulsionnel se fonde sur une technique d'analyse/synthèse qui permet de restituer une parole synthétique de bonne qualité en s'affranchissant essentiellement de la délicate opération de détection du voisement.

Dans cet article nous décrivons un algorithme rapide permettant l'implémentation de cette technique sur un calculateur temps réel en virgule fixe, en proposant un autre modèle de source multi-impulsionnelle.

2. L'ALGORITHME LPC-MULTI-IMPULSIONNEL

L'excitation multi-impulsionnelle est une séquence d'impulsions u(n) pouvant s'exprimer sous la forme:

$$/2/ \quad u(n) = \sum_{k=1}^{K} G_i \delta(n-n_i) \qquad 0 \leq n \leq N-1$$

où G_i est l'amplitude, n_i est la position, N est la largeur de la séquence, K est le nombre d'impulsions élémentaires δ dans la séquence. La fig. 1 donne le schéma du module d'analyse/synthèse permettant de déterminer les positions des impulsions et leurs amplitudes. Le calcul des impulsions revient à minimiser l'énergie de l'erreur perceptive e(n) entre le signal s(n) et le signal de synthèse $\hat{s}(n)$ pondérée par la fenêtre perceptuelle w(n).

$$/3/ \quad e(n) = \{s(n)-\hat{s}(n)\} * w(n) \quad (* \text{ op. de convolution})$$

Le signal de synthèse s'exprime par:

$$/4/ \quad \hat{s}(n) = \hat{s}_o(n) + u(n) * h(n) = \hat{s}_o(n) + \sum_{i=1}^{K} G_i h(n-n_i)$$

où $\hat{s}_o(n)$ est la réponse libre du synthétiseur et h(n) sa réponse impulsionnelle.

Pour trouver la source optimale au sens perceptif, on minimise l'énergie résiduelle perceptive dans l'intervalle(0,N) soit:

$$/5/ \quad E = \sum_{n=o}^{N-1} e^2(n) = \sum_{n=o}^{N-1} \{\delta_w(n) - \sum_{i=1}^{K} G_i h_w(n-n_i)\}^2$$

en posant $\delta_w(n) = \{s(n)-\hat{s}_o(n)\} * w(n)$

$h_w(n)$ correspond au filtre $1/A(z/\gamma)$, $1/A(z)$ étant celui du synthétiseur (correspondant à h(n)). Pour minimiser E par la méthode des moindres carrés on aboutit à une équation matricielle qui ne permet pas de trouver les inconnues n_i, G_i, i=1 à K de manière explicite.

$$/6/ \quad G_1 \sum_{n=o}^{N-1} h_w(n-n_1)h_w(n-n_j) + \ldots + G_K \sum_{n=o}^{N-1} h_w(n-n_K)h_w(n-n_j) = \sum_{n=o}^{N-1} \delta_w(n)h_w(n-n_j) \qquad 0 < j \leq K$$

On se contente donc d'une solution sous-optimale qui consiste à calculer les amplitudes et les positions de manière itérative. Supposant que j-1 impulsions ont été trouvées on estime la jème de la façon suivante:

$$/7/ \quad e_j(n) = e_{j-1}(n) - G_j h_w(n-n_j)$$

$e_j(n)$ erreur perceptuelle a l'itération j. On a donc:

$$/8/ \quad E_j = \sum_{n=o}^{N-1} e_j^2(n) = \sum_{n=o}^{N-1} \{e_{j-1}(n) - G_j h_w(n-n_j)\}^2$$

En annulant la dérivée de E_j par rapport à G_j on obtient la valeur de G_j sous-optimale:

$$/9/ \quad G_j = \{\sum_{n=o}^{N-1} e_{j-1}(n)h_w(n-n_j)\} / \{\sum_{n=o}^{N-1} h_w^2(n-n_j)\}$$

L'erreur perceptive totale E_j s'exprime alors par:

$$/10/ \quad E_j = E_{j-1} - G_j^2 \sum_{n=o}^{N-1} h_w^2(n-n_j)$$

Cette erreur sera minimale (et obtenue pour $\max G_j$) pour une certaine postion n_j de l'impulsion dans (O,N). Cette position ne peut être estimée qu'en procédant de manière exhaustive: cela conduit bien évidemment à des calculs longs en machine. Pour la j+1eme impulsion le calcul se poursuit de manière itérative et l'on ne s'arrêtera que lorsque l'erreur E_j sera jugée suffisamment petite. Les impulsions obtenues aux pas précédents ne sont jamais remises en cause au pas courant: c'est en ce sens que le processus est sous-óptimal. On peut améliorer légérement ceci en calculant à nouveau, par la formule /6/, les amplitudes sans modifier les positions déjà estimées.

3. MASQUAGE DE L'ERREUR

Faisant appel aux propriétés de l'audition (à partir de données psychoacoustiques), Atal et Schroeder ont montré que la pondération de l'erreur résiduelle permet d'adapter le mécanisme de masquage au bruit de codage (erreur entre le signal originel et le siganl de synthèse). L'oreille peut en effet tolérer un taux de bruit plus important dans les régions formantiques. Ceci suggère la mise en forme du bruit de codage par le filtre numérique ayant comme fonction de transfert W(z):

$$/11/ \quad W(z) = A(z)/A(z/\gamma) \quad \text{avec } o < \gamma < 1$$

γ contrôle le degré de pondération de l'erreur entre le signal originel et le signal de synthèse. Expérimentalement, la meilleure valeur se situe vers 0,8. La fig. 2 illustre le phénomène de masquage . Puisque le prédicteur 1/A(z) est une estimation optimale de l'enveloppe spectrale du signal au sens de la distance d'Itakura, on peut remplacer l'enveloppe spectrale de s(n) par celle de 1/A(z).

Pour mieux comprendre le mécanisme de masquage fréquentiel, prenons les deux cas extrêmes γ=0 et γ=1:
(a) γ=1 -il n'y a pas de masquage. Dans le plan fréquentiel le bruit de codage peut être supérieur au signal utile surtout dans les zones de non-résonances (fig. 2 a)
(b) γ=0 -ici au contraire le spectre du bruit de codage est "parallèle" à celui du signal original et donc important en basses fréquences pour lesquelles la sensibilité de l'oreille est maximum.

Il s'ensuit donc que le confort maximal pour un auditeur se situe entre 0 et 1, et expérimentalement pour γ voisin de 0.8. Il est à remarquer que pour cette valeur le rapport signal/bruit n'est pas obligatoirement optimal.

4. MISE EN OEUVRE SUR PROCESSEUR RAPIDE

Dans notre simulation du vocodeur LPC-multi-impulsionnel les conditions choisies pour l'analyse sont les suivantes:

Méthode LPC	autocorrélation
Largeur d'une trame	128 points
Longueur de la fenêtre	256 points
Echantillonnage	10 kHz
Bande-passante	300 Hz - 40000 Hz
Ordre M du prédicteur	10
Coefficient γ	0,8
Nb d'impulsions calcul.	8

TABLEAU I : Conditions d'analyse

L'ordinogramme de l'algorithme est présenté sur la fig. 3 , il contient une procédure de "raffinement" qui permet de recombiner deux impulsions superposées lors du calcul itératif. Il est à noter que ce phénomène se produit assez souvent pour certaines trames voisées.

Un des défauts principal de la méthode multi-impulsionnelle réside dans la taille des calculs, notamment pour obtenir à chaque itération la position de la meilleure impulsion. C'est pourquci nous avons songé d'implanter ce calcul sur un processeur rapide, en visant par là le temps réel (processeur SCH de Novelec France, connecté sur ordinateur hôte PDP 11/23 de D.E.C.) Comme tous les processeurs de cette gamme, SCH possède une mémoire interne limitée et une unité de calcul en arithmétique fixe. Par conséquent on établit une symbiose entre les deux calculateurs: SCH est une unité de calcul pour PDP qui est une mémoire externe pour SCH, tout ceci dans un environnement FORTRAN. Les principales caractéristiques de ce calculateur SCH sont:

1. mots de 16 bits
2. taille de la mémoire programme et données 4 kmots
3. structure pipe-line
4. temps d'exécution d'une instruction: 200 ns.

Sur SCH sont exécutées les procédures qui demandent le plus de calculs: (a) l'extraction des coefficients du synthétiseur, (b) la synthèse, (c) l'intercorrélation (formule /9/).
(a) pour extraire les coefficients du synthétiseur nous avons utilisé l'algorithme de Leroux, dont l'intérêt se trouve dans le fait que toutes les variables intermédiaires peuvent se mettre sous forme fractionnaire, ce qui facilite les manipulations en virgule fixe, simple précision. Nous avons remarqué que les paramètres ainsi calculés (les coefficients de réflexion) rendent les

coefficients de prédiction a_i sensibles au bruit de quantification. Il faut donc mettre en oeuvre une procédure de calcul pour que l'amplitude des coefficients reste grande devant le pas de quantification, tout en évitant les problèmes de dépassement de capacité (overflow) de la machine.

(b) le synthétiseur est réalisé grâce à une structure en treillis avec deux multiplieurs dans chaque cellule, construite à partir des coefficients de réflexion. Ces filtres en treillis, qui appartiennent à une classe de filtres polynomiaux orthogonaux, sont très utilisés pour l'implémentation matérielle en raison de leur facilité de mise à l'échelle et parce qu'ils sont moins sensible à l'erreur introduite par l'arithmétique de précision finie. Dans la réalisation d'un filtre en virgule fixe, le dépassement de capacité ne peut pas en principe, avoir lieu (grâce à une contrainte nécessaire mais non suffisante). A partir des théories développées par Jackson, la mise à l'échelle se fait à l'aide du rapport $1/||F||$ en supposant que l'excitation est $\delta(n)$. $||F||$ est la norme L_2 de la fonction de transfert du synthétiseur $1/A(z)$:

$$/12/ \quad ||F|| = \{ \frac{1}{2\pi} \int_0^{2\pi} |\frac{1}{A(e^{j\omega})}|^2 d\omega \}^{1/2}$$

En réalité, le calcul de $||F||$ se fait plus simplement par:

$$/13/ \quad 1/||F|| = \prod_{i=1}^{M} (1-k_i^2)$$

où les k_i sont les coefficients de réflexion du filtre. Naturellement, pour une autre excitation que $\delta(n)$, et c'est le cas ici, un autre facteur d'échelle doit être ajouté: il s'obtient par une contrainte énergétique sur le signal d'excitation. Il suffit alors d'ajuster le gain de sortie du filtre sur le gain obtenu lors de l'analyse.

(c) le calcul le plus pénalisant dans une optique temps réel se trouve dans la détermination des positions optimales des impulsions puisque la recherche se fait de manière exhaustive. Il convenait donc que ce soit une tâche laissée au calculateur SCH: un sous-programme a été développé pour calculer l'intercorrélation dans l'intervalle (O,N-1) (numérateur de l'équation /9/). Ce sous-programme donne en sortie, la valeur optimale n_j de la position de la jème impulsion à partir de laquelle on peut calculer facilement son amplitude G_j.

5. AMELIORATION DE LA SOURCE

L'excitation par impulsions multiples reste malgré tout une approximation de la source , certes plus efficace que la source classique mono-impulsionnelle pour les sons voisés. Il est donc tentant de prendre une solution intermédiaire en extrayant par exemple, les impulsions les plus amples dans le signal résiduel e(n), $e(n) = s(n) * b(n)$ (b(n) correspond à la réponse impulsionnelle du prédicteur $1/A(z)$) et de considérer ces impulsions comme réprésentatives de la source. Bien évidemment une telle solution serait acceptable si e(n) n'était pas un signal aléatoire (la fig 4 montre le cas du phonème / ɑ / pour lequel prendre les K impulsions les plus amples ne conduirait pas à un bon résultat). D'où l'idée d'utiliser une méthode séquentielle ressemblant à la méthode PARCOR (Partial Correlation) qui actualise un modèle instantané à chaque nouvel échantillon du signal s(n), ce qui permet une adaptation au cours du temps du modèle sur les variations du signal /Boll /. Le choix des K impulsions les plus amples devient alors beaucoup plus cohérent avec les critères que nous explicitons ci-après.

On sait que dans la méthode PARCOR, l'erreur de prédiction "aller" ($S_m^+(n)$) et "retour" ($S_m^-(n)$) peuvent s'écrire sous la forme récursive suivante:

$$/14a/ \quad S_m^+(n) = S_{m-1}^-(n) + k_m S_{m-1}^+(n)$$

$$/14b/ \quad S_m^-(n) = S_{m-1}^+(n) k_m - S_{m-1}^-(n)$$

pour m=1,2,...M et les conditions initiales $S_o^+(n)=s(n)$, $S_o^-(n)=s(n-1)$. La minimisation de l'erreur "aller" et "retour" conduit à l'expression:

$$/15/ \quad k_m = -E\{S_{m-1}^+(n) S_{m-1}^-(n)\} / \sqrt{E\{S_{m-1}^{+2}(n)\} E\{S_{m-1}^{-2}(n)\}}$$

k_m est au signe près, un coefficient de corrélation partielle. Ces coefficients k_m, m=1 à M, sont modifiés à la même cadence que l'échantillonnage pour obtenir un ensemble adapté au signal. En mettant l'énergie instantanée de l'erreur "aller" et "retour" sous la forme:

$$/16/ E_m = S_m^{+2}(n) + S_m^{-2}(n) = 4k_m S_{m-1}^+(n) S_{m-1}^-(n) + \{1+k_m^2\} E_{m-1}$$

puis en minimisant cette quantité E_m, on obtient:

$$/17/ \quad k_m = -2S_{m-1}^+(n) S_{m-1}^-(n) / \{S_{m-1}^{+2}(n) + S_{m-1}^{-2}(n)\}$$

Il est clair que k_m est borné, $|k_m| < 1$ et il est facile de vérifier la relation suivante:

$$/18/ \quad S_m^{+2}(n) = S_{m-1}^{+2} . (1-k_m^2)$$

qui rappelle une relation semblable dans la méthode LPC classique: $\alpha_m = \alpha_{m-1} . (1-k_m^2)$

L'erreur résiduelle est finalement: $\varepsilon(n) = S_M^+(n)$, cette erreur parait moins "aléatoire" que e(n), elle est plus "régulière" (fig.4). Il devient alors possible par ce moyen, de prélever les K impulsions les plus intenses pour représenter la source dans le cas des sons voisés. Un affinage supplémentaire peut être ajouté en considérant le masquage perceptif /6/.

SYNTHESE DE LA PAROLE PAR CODAGE MULTI-IMPULSIONNEL

CAELEN J., SONG J.M.

6. CONCLUSION

L'amélioration que nous venons d'introduire, en considérant le signal-source comme les K (K=8) plus intenses impulsions du signal résiduel $\varepsilon(n)$, diminue considérablement le nombre des calculs pour rechercher la position des impulsions optimales. Avec cette approximation il devient possible d'implémenter la procédure de résolution en temps réel sur calculateur spécialisé. La qualité de la parole de synthèse reste bonne, il n'y a aucune incidence au niveau de l'intelligibilité ni au niveau du naturel de la voix. Tous les algorithmes sont développés en virgule fixe et en simple précision (16 bits). Cela rend possible l'utilisation de cette méthode pour la transmission de la parole à bas débit avec une excellente qualité à la réception. Avec les paramètres du tableau I, ce débit est de 10 kb/s, la fréquence de répétition d'une trame étant de 78 Hz.

La Fig 5 montre que les résultats obtenus sont très proches de ceux de la méthode classique: à l'audition aucune différence n'est perceptible.

7. REFERENCES

(1) Atal B.S., Remede J.R., "A New Method of LPC Excitation for Producing Natural Sounding Speech at Low Bit Rates", Proc. ICASSP, 1982, pp. 614-617.

(2) Atal B.S., Schroeder M.R., "Predictive Coding of Speech Signals and Subjective Errors Criteria", IEEE Trans ASSP-27, 1978, pp. 247-254.

(3) Leroux J., Gueguen C., "A fixed Point Computation of Partial Correlation Coefficients", IEEE Trans ASSP-25, 1977, pp. 257-259.

(4) Viswanathan R., Makhoul J., "Quantization Properties of Transmission Parameters in Linear Predictive Systems" IEEE Trans ASSP-23, 1975, pp. 309-321.

(5) Markel J.D., Gray A.H., "Linear Prediction of Speech" Springer Verlag, New York, 1976.

(6) Singhal S., Atal B.S., "Optimizing LPC Filter Parameters for Multi-Pulse Excitation", Proc. ICASSP, 1983, pp. 781-784.

(7) Boll S.F., "Selected Methods for Improving Synthesis Speech Quality using Linear Predictive Coding. System Description, Coefficients smoothing and STREAK", UTEC-CSC-74-151 Computer Science Departement, University of Utah, 1974.

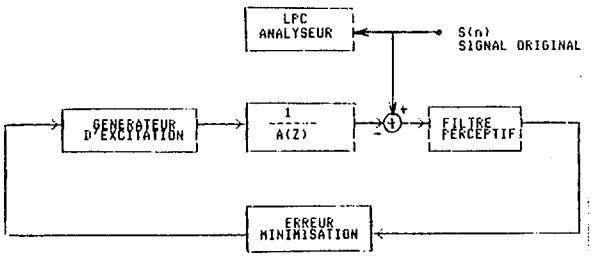

Fig. 1: Diagramme du bloc d'analyse-synthèse pour extraire les paramètres.

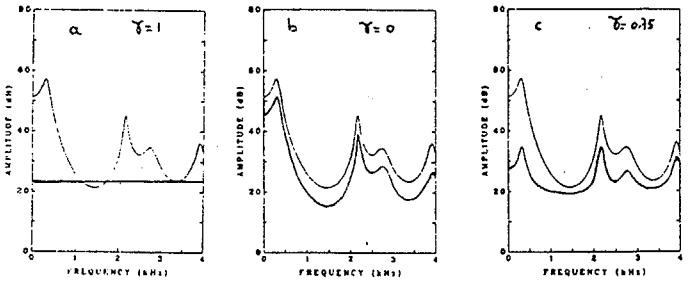

Fig. 2: Spectres du signal et du bruit de codage en fonction de γ, paramètre de mise en forme.

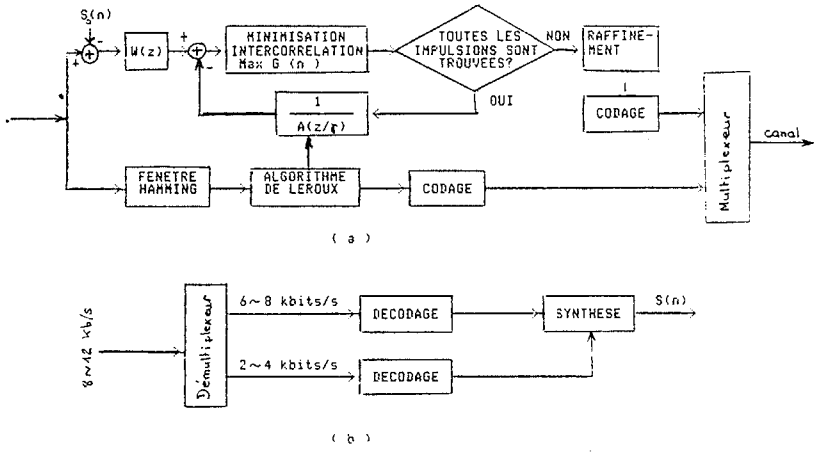

Fig. 3: Schéma du transmetteur (a) et du récepteur (b).

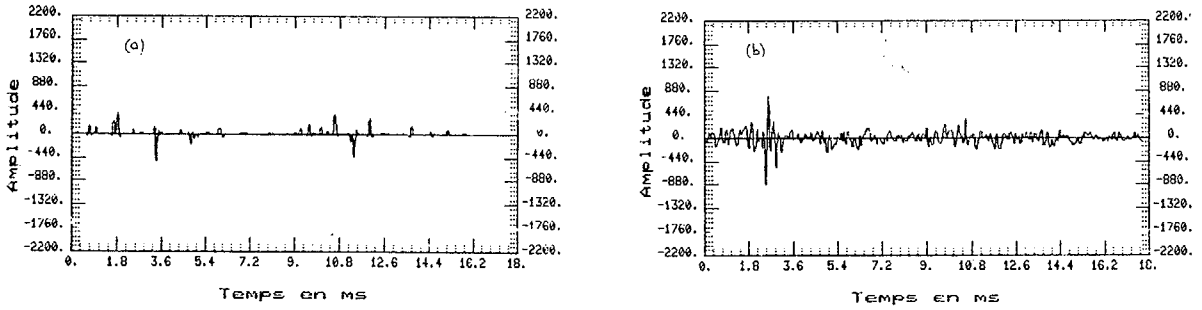

Fig. 4: Signaux d'erreur pour une trame voisée, phonème /o/,
(b) signal résiduel obtenu par la méthode classique,
(a) signal résiduel obtenu par la méthode d'adaptation instantanée.

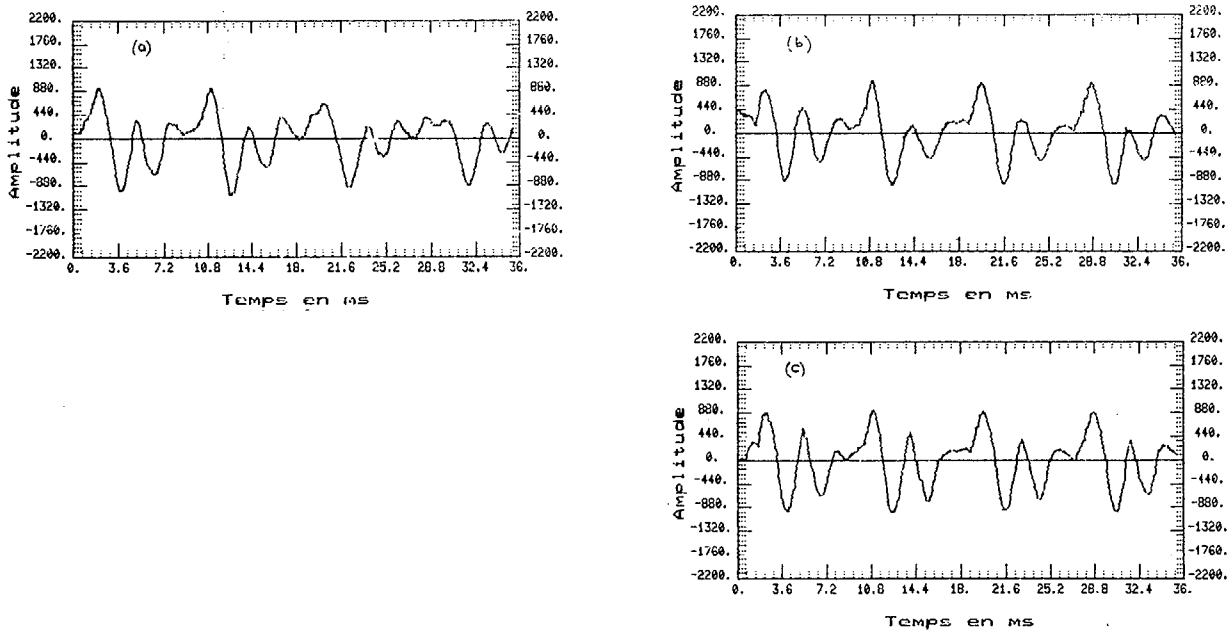

Fig. 5: Signaux d'une trame voisée, phonème /o/,
(a) signal orisinal,
(b) signal synthétisé par la nouvelle méthode,
(c) signal synthétisé par la méthode classique.