NEUVIEME COLLOQUE SUR LE TRAITEMENT DU SIGNAL ET SES APPLICATIONS

NICE du 16 au 20 MAI 1983

NOUVELLE METHODE DE SYNTHESE D'UN VOCODEUR A PREDICTION LINEAIRE

G. BRUN - E. AZIZ - J. MENEZ - J.F. GALLIANO*

LABORATOIRE DE SIGNAUX ET SYSTEMES UNIVERSITE DE NICE ERA 835 - 41 Boulevard Napoléon III - 06041 NICE CEDEX

RESUME

L'objet de cette communication est de présenter une nouvelle méthode de synthèse d'un vocodeur à prédiction linéaire permettant d'améliorer le signal de synthèse au plan de sa perception auditive, amélioration obtenue au prix d'un accroissement de la complexité du synthétiseur.

Les vocodeurs, utilisés comme systèmes de transmission de la parole, doivent fonctionner en temps réel. Ce qui impose de choisir des modèles de synthèse et des méthodes algorithmiques suffisamment simples pour satisfaire aux contraintes technologiques. Or les progrès récents en matière de technologie VLSI permettent d'accroître sensiblement la complexité des systèmes numériques sans augmentation prohibitive de leur coût de fabrication.

Ceci nous a incités à étudier un modèle de synthèse plus élaboré qui diffère de ceux préalablement proposés pour le calcul du facteur de gain qui rend au signal de synthèse une énergie proche de celle du signal original. Dans les modèles de synthèse couramment utilisés, le calcul du gain ne tient pas compte des conditions initiales lors du réajustement des coefficients du filtre modèle du conduit vocal.

Nous nous sommes intéressés à une structure à deux filtres modèles du conduit vocal identiques. Le rôle du premier est de créer la partie du signal de synthèse due au seul signal d'excitation (conditions initiales nulles au début de chaque période). Quand au second, il fournit la composante du signal de synthèse due aux seules conditions initiales qui sont réactualisées au début de chaque période.

L'originalité de ce modèle réside dans le fait que le calcul du gain est effectué pour chaque période de mélodie. Ceci conduit à un fonctionnement proche de la réalité qui se traduit par une amélioration de la qualité du signal de synthèse au plan de sa perception auditive.

SUMMARY

The purpose of this paper is to present our studies related to Linear Predictive Vocoders synthesis. The subjective quality of the synthetic speech is perceptibly better than those given by similar systems. Of course, this improvement is obtained at the expense of the increase of the implementation complexity.

The Vocoders must operate in real time, when they serve to transmit speech. Therefore the used synthesis models and algorithms are enough simple to satisfy the technological constraints. Now, the new progresses of the VLSI technology allow to greatly increase the complexity of numerical systems without damage for their fabrication cost or processing time.

This incites us to study more elaborated models. The computation of the gain factor to be applied to the synthetic speech for giving it an energy approaching the one of the original speech, has become no satisfactory solution till now. So we propose to compute a gain factor for each pitch period and to use a structure with two vocal tract model filters.

These improvements lead to a model which nearly works like reality. This signifies a better auditive perception quality of the synthetic speech.

*Monsieur J.F. GALLIANO est à présent à CIT ALCATEL La Verrière - 78320 LE MESNIL SAINT DENIS.

NOUVELLE METHODE DE SYNTHESE D'UN VOCODEUR A PREDICTION LINEAIRE

INTRODUCTION

Un vocodeur à prédiction linéaire est un système de codage de la parole caractérisé par un faible débit (1,2 à 4,8 Keb/s). Il est fondé sur l'élaboration d'un modèle de l'organe de phonation représenté par un conduit sonore à section variable dont les coefficients sont ajustés périodiquement en raison du caractère non stationnaire du signal vocal. Ces derniers sont liés à la fonction de transfert du conduit vocal et aux caractéristiques de la source d'excitation.

Au cours de cette étude, nous avons été amenés à nous intéresser à plusieurs modèles de synthèse à base de prédiction linéaire.

Le premier de ceux-ci est composé d'un filtre modèle du conduit vocal. Facile à mettre en oeuvre, il ne tient pas compte des conditions initiales lors du réajustement des coefficients.

La qualité du signal de synthèse a pu être améliorée par l'utilisation d'une structure à deux filtres modèles du conduit vocal, afin de prendre en compte séparément les effets des conditions initiales et du signal d'excitation. Dans cette structure, le gain est calculé une seule fois par fenêtre. Or ce calcul ne correspond pas à une réalité physique. Cela se traduit souvent par une modulation de l'enveloppe du signal de synthèse qui dégrade la qualité de l'audition.

Nous avons donc été amenés à proposer un calcul du gain, non plus par fenêtre de synthèse, mais par période du fondamental. Parallèlement, nous nous sommes attachés à accroître les performances des méthodes de détection du fondamental.

Il est évident que la qualité du signal synthétisé dépend aussi bien de la précision des paramètres extraits à l'analyse que de l'exactitude du modèle de synthèse. C'est pourquoi, la première partie du texte est relative à l'analyse et aux paramètres qu'elle extrait. La seconde est consacrée à la synthèse. Elle décrit la structure utilisée, la façon d'exploiter les paramètres transmis et la manière de calculer le facteur de gain.

I- L'ANALYSE : EXTRACTION DES PARAMETRES

Comme tout système de codage de la parole, le vocodeur à prédiction linéaire se compose de deux parties : l'émetteur et le récepteur.

A l'émetteur, l'analyse permet d'extraire périodiquement un ensemble de paramètres qui caractérisent le signal vocal. Ces paramètres sont relatifs au conduit vocal et à la source d'excitation. Il sont au nombre de quatre :

- la décision son voisé ou son non voisé
- la mélodie ou période du fondamental
- les coefficients PARCOR
- l'énergie du signal original

I-1 LA DECISION :

Ce paramètre, relatif à la source d'excitation, représente la nature du signal vocal : un son voisé présente un caractère périodique alors qu'un son non voisé a un caractère aléatoire.

La source d'excitation du filtre modèle du conduit vocal se compose donc de deux générateurs distincts : le premier délivre un train d'impulsions périodiques pour les sons voisés, le second fournit une séquence aléatoire engendrant un bruit blanc pour la production de sons non voisés.

La décision est prise en fonction du rapport de la longueur de la trajectoire du signal vocal à la somme des valeurs absolues des échantillons. Si ce rapport est inférieur à un seuil fixé, la séquence analysée est considérée comme voisée.

Soit d ce rapport. Il est défini par la relation suivante :

$$d = \frac{\sum_{n=1}^{N-1} | S(n+1) - S(n) |}{\sum_{n=1}^{N} | S(n) |}$$

avec N= nombre d'échantillons dans la fenêtre d'analyse

I-2 LA MELODIE :

Ce paramètre concerne également la source d'excitation et plus précisément le générateur de trains d'impulsions périodiques. La mélodie est en effet la période des impulsions. Il est indispensable de déterminer de manière aussi exacte que possible la période du fondamental puisque c'est un élément clé du système de synthèse décrit ci-dessous. C'est pourquoi, nous avons cherché à améliorer l'une des nombreuses méthodes de détection du fondamental : celle de l'AMDF normalisée. Les résultats obtenus sont aussi présentés au cours de ce Colloque /1/.
La méthode AMDF est définie par l'équation :

$$p(K) = \sum_{n=1}^{N} | S(n) - S(n-K) |$$

La période du fondamental est obtenue par la localisation de l'indice K qui correspond au minimum de cette fonction.

I-3 LES COEFFICIENTS PARCOR :

Appelés aussi coefficients de réflexion, ils définissent le filtre modèle du conduit vocal.

I-4 L'ENERGIE DU SIGNAL ORIGINAL :

Il est possible de transmettre l'énergie du signal original une fois par fenêtre d'analyse. Toutefois, afin de suivre correctement les évolutions du signal, en particulier lors des transitions entre zones voisées et non voisées, il est souhaitable de transmettre plusieurs fois l'énergie du signal. La fenêtre d'analyse est donc partitionnée en plusieurs fenêtres d'énergie. Le rapport de deux énergies consécutives permet de déterminer assez précisément l'instant de transition. Ce critère est utilisé à la synthèse pour modifier éventuellement la décision. La combinaison de ces deux méthodes donne des résultats très satisfaisants.

II- LA SYNTHESE

II-1 LE MODELE DE SYNTHESE

Ce modèle est plus élaboré que ceux couramment proposés. Il diffère de ceux-ci par la méthode de calcul du facteur de gain. Ce dernier a pour but de restituer au signal de synthèse une énergie proche de celle du signal original.

La figure 1 représente la structure qui est le plus souvent employée /2/. C'est un modèle autorégressif excité soit par un train d'impulsions périodiques, de période égale à celle du fondamental pour la création de sons voisés, soit par une séquence de bruit blanc pour la production de sons non voisés.
Ce modèle est régi par l'équation aux différences finies suivante :

$$S(n) = e(n) - \sum_{i=1}^{p} a(i).\ S(n-i)$$

avec p= ordre du modèle
e(n)= excitation
a(i)= coefficients de prédiction calculés à partir des coefficients PARCOR

En appliquant la transformée en Z aux deux membres de la relation précédente, on obtient :

$$S(Z) = H(Z) . E(Z)$$

$$\text{avec } H(Z) = \frac{1}{1 + \sum_{i=1}^{P} a_i . Z^{-i}}$$

En approximant E(Z) par une constante (impulsion ou bruit blanc), on fait l'hypothèse que le filtre du modèle est un filtre numérique linéaire récursif ne comportant que des pôles.

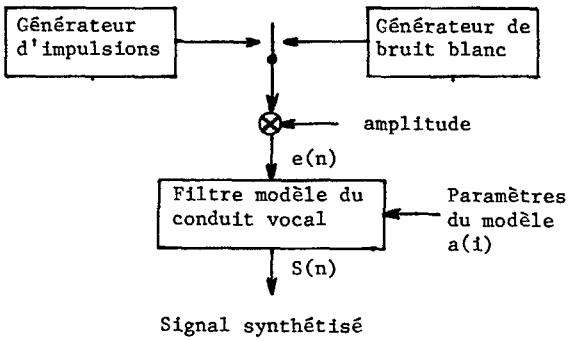


Figure 1 - Structure classique

Ce modèle a pour avantage une très grande simplicité de mise en oeuvre. Par contre, il a pour inconvénient de ne pas tenir compte des conditions initiales lors du réajustement des coefficients du filtre modèle du conduit vocal. C'est pour pallier ce défaut que nous avons choisi une structure plus complexe /3/. La complexité des systèmes n'étant plus un facteur aussi critique qu'il y a quelques années, l'amélioration de la qualité du signal de synthèse se fait à moindre coût.

Ce nouveau modèle comprend deux filtres identiques excités différemment. (Figure 2).

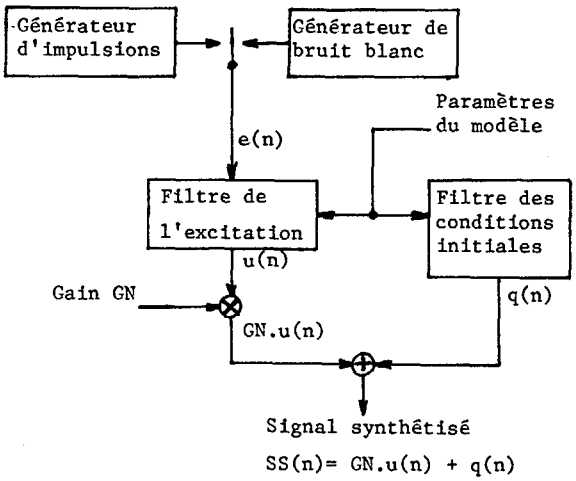


Figure 2 - Structure choisie

Un filtre engendre la composante du signal synthétisé due au seul signal d'excitation avec des conditions initiales nulles au moment du réajustement des coefficients des filtres. Le second délivre la composante du signal synthétisé résultant des seules conditions initiales.

La composante u(n) du signal synthétisé est la réponse d'un filtre transversal de coefficients a(i) à une excitation e(n). Elle est régie par la relation suivante :

$$u(n) = e(n) - \sum_{i=1}^{P} a(i+1) . Y(i)$$

avec Y(i)= mémoires du filtre

De même, la composante q(n) du signal synthétisé est la réponse d'un filtre transversal de coefficients a(i) à une excitation nulle. Elle s'écrit donc :

$$q(n) = - \sum_{i=1}^{P} a(i+1) . X(i)$$

avec X(i)= mémoires du filtre

Ce modèle a déjà été décrit en préconisant de réinitialiser les coefficients des filtres à la fin de chaque fenêtre de synthèse. Cela nous a semblé tout à fait artificiel puisque le découpage du signal en fenêtres ne présente aucune réalité physique dans un fonctionnement synchrone.

L'originalité de notre étude réside dans le fait que les coefficients des filtres sont mis à jour à la fin de chaque période de mélodie et le facteur de gain est calculé sur celle-ci. Cette façon de procéder est proche de la réalité car la période du fondamental est une entité véritablement caractéristique du signal traité. Cela se traduit pour une amélioration de la qualité du signal de synthèse au plan de sa perception auditive.

La suite de ce paragraphe est consacrée à la description du fonctionnement de ce modèle à partir des paramètres fournis pour l'analyse.

II-2 L'ENERGIE ET LA DECISION :

Dans les zones où le signal a une nature constante (soit voisé, soit non voisé), la décision utilisée est celle calculée à l'analyse. Dans les zones de transition (voisé-non voisé et non voisé-voisé), la décision de l'analyse peut être modifiée. En effet, l'analyseur délivre une décision affectée à une fenêtre d'analyse. Or cette dernière peut, dans ce cas, contenir des sons de deux types. Il faut donc pouvoir les différencier. C'est possible si plusieurs fenêtres d'énergie sont transmises à l'intérieur d'une même fenêtre d'analyse. Le rapport de deux énergies consécutives permet de connaître assez précisément l'instant de transition. Ce rapport, comparé à un seuil fixé, modifie éventuellement la décision initiale.

Ainsi des démarrages en sons voisés peuvent-ils être anticipés ou retardés par rapport au début d'une fenêtre d'analyse. Il faut alors corriger les énergies sinon une partie de l'énergie du signal voisé se reporterait sur le signal non voisé créant une discontinuité et donc un bruit désagréable à l'audition.

Tout cela est mis en évidence sur les figures suivantes. La figure 3 montre un exemple de transition non voisé-voisé. La fenêtre de synthèse est partitionnée en trois fenêtres d'énergie.

Afin de réduire autant que possible les brusques variations d'énergie, cette dernière est interpolée linéairement de milieu de fenêtre d'énergie à milieu de fenêtre d'énergie. Dans ce cas, l'énergie E(3) est très grande par rapport à l'énergie E(2) car elle représente le signal voisé. Aussi l'interpolation en reporte une grande partie sur la fenêtre non voisée comme le montre la figure 4. Ce qui a pour effet d'amplifier le bruit. Pour éliminer ce défaut, il faut procéder à une correction de l'interpolation d'énergie. E(3) est remplacée par E(2) pour maintenir l'énergie constante sur la partie non voisée. La perte d'une partie de l'énergie E(3) n'est pas préjudiciable à la qualité auditive (Figure 5).

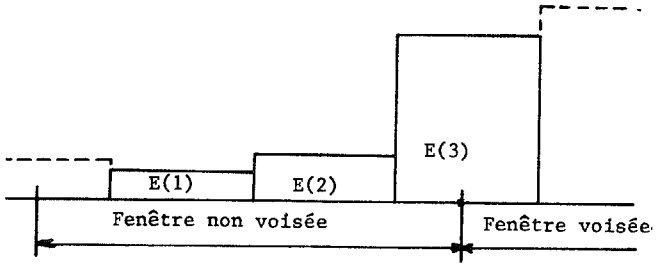


Figure 3 - Transition non voisé-voisé avec trois fenêtres d'énergie

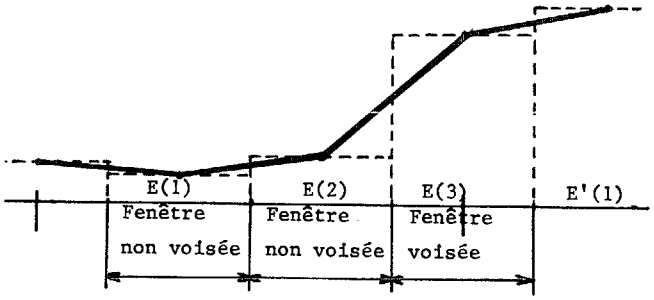


Figure 4 - Energies interpolées

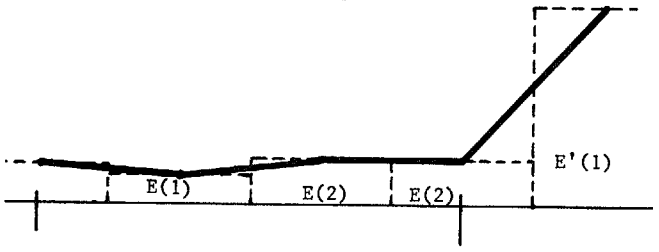


Figure 5 - Correction de l'énergie

II-3 LA MELODIE :

Il faut distinguer quatre cas correspondant aux différents types de transitions possibles :

- Durant les parties non voisées, la mélodie est fictive. Elle a été choisie égale au quart de la fenêtre de synthèse.

- Au cours des zones voisées, la mélodie est interpolée de fenêtre à fenêtre entre les valeurs fournies par l'analyse. C'est le seul cas où une interpolation est faite.

- Lors des transitions entre sons non voisés et voisés, le démarrage en voisé s'effectue au début de la fenêtre d'énergie dont la décision est voisée avec la mélodie donnée par l'analyse.

- Au moment du passage d'une zone voisée à une zone non voisée, la dernière mélodie voisée est conservée jusqu'au démarrage en non voisé .. S'il n'y a pas coïncidence entre fin de période et fin de fenêtre d'énergie, un test est réalisé pour savoir de combien déborderait une période voisée supplémentaire.

Si le débordement est inférieur à la moitié de la mélodie, la période voisée supplémentaire a lieu et le démarrage en non voisé est retardé. Dans le cas contraire, il est anticipé.

II-4 COEFFICIENTS PARCOR :

Du fait de la segmentation du signal vocal au moment de l'analyse, les coefficients peuvent subir des variations rapides d'une fenêtre à l'autre même dans les zones où les sons gardent la même nature. Cela conduit, dans certains cas, à des discontinuités pouvant se traduire par des bruits gênants. Il est possible de remédier à ce problème en faisant varier de manière continue la valeur de ces paramètres. C'est pourquoi les coefficients sont calculés en chaque point par interpolation linéaire entre les valeurs fournies par l'analyse de milieu de fenêtre à milieu de fenêtre.

Dans les zones de transition, les coefficients sont caractéristiques de sons de natures différentes. De ce fait, l'interpolation ne se justifie pas. C'est pourquoi la zone qui s'achève conserve ses coefficients constants du milieu de sa fenêtre à l'instant de démarrage. La zone qui débute le fait avec ses coefficients et les maintient inchangés jusqu'au milieu de sa fenêtre.

II-5 MEMOIRES DES FILTRES :

Celles du filtre de l'excitation sont remises à zéro au début de chaque période. Celles du filtre des conditions initiales sont également réinitialisées au début de chaque période mais avec les dernières valeurs du signal synthétisé. Elles ne sont remises à zéro que dans le cas des transitions entre sons non voisés et sons voisés puisque le signal synthétisé est alors du bruit blanc et ne présente aucune signification pour la partie voisée.

II-6 CALCUL DU GAIN :

Soit GN le gain à appliquer à la composante due à l'excitation du signal synthétisé afin de restituer à ce dernier l'énergie du signal original. Le signal synthétisé s'écrit (voir figure 2) :

$SS(n) = GN.\ u(n) + q(n)$

Son énergie sur une période est égale à $\Sigma\ SS^2(n)$

Soit ENR l'énergie du signal original sur une période

L'égalité des deux implique :

$$\Sigma\ SS^2(n) = \Sigma(GN.u(n) + q(n))^2 = ENR$$

Il faut donc résoudre l'équation du second degré en GN

$$\Sigma(u^2(n)).GN^2 + 2.\Sigma(u(n).q(n)).GN + \Sigma(q^2(n)) - ENR = 0$$

q(n) représente le signal résiduel. Physiquement, son énergie doit être inférieure ou à la limite égale à celle du signal original : $\Sigma(q^2(n)) - ENR \leqslant 0$

C'est toujours vrai sauf dans les zones de décroissance rapide de ce dernier. Les signaux produits par les filtres à ce moment-là ne meurent pas assez vite. C'est pourquoi le signal résiduel est multiplié par un coefficient d'amortissement CA tel que l'énergie du résiduel soit égale à l'énergie de l'original :

$$CA^2.\ \Sigma(q^2(n)) = ENR$$

L'équation se réduit alors à :

$$GN.(\Sigma(u^2(n)).GN + 2.\Sigma(u(n).CA.q(n))) = 0$$

Si l'équation ne peut être résolue, le gain est forcé à zéro. Dans le cas contraire, la solution positive est retenue.

Les algorithmes qui traduisent la mise en oeuvre de ces différents calculs sont présentés en annexe sous forme d'organigrammes ainsi qu'un exemple de synthèse.

CONCLUSION

Nous avons présenté une nouvelle méthode de synthèse d'un vocodeur à prédiction linéaire qui fonctionne à 3,6 Keb/s. La qualité du signal synthétique produit est meilleure que celle obtenue à l'aide des systèmes proposés jusqu'alors. Cela tient bien sûr à la méthode de synthèse mais également à des améliorations apportées à l'analyse.

En effet, nous avons augmenté le nombre d'e.b. généralement affectés au codage des coefficients PARCOR car une quantification trop faible dégrade nettement la qualité du signal synthétisé. Nous avons également codé l'énergie du signal original plusieurs fois par fenêtre d'analyse. Ces deux points expliquent le débit binaire cité plus haut. De plus, nous

avons utilisé des algorithmes plus sophistiqués pour la détection de la période du fondamental.

En ce qui concerne la synthèse proprement dite, nous proposons une structure plus élaborée qui comprend deux filtres modèles du conduit vocal et dans laquelle le calcul du facteur de gain est fait de manière synchrone à la période du fondamental.

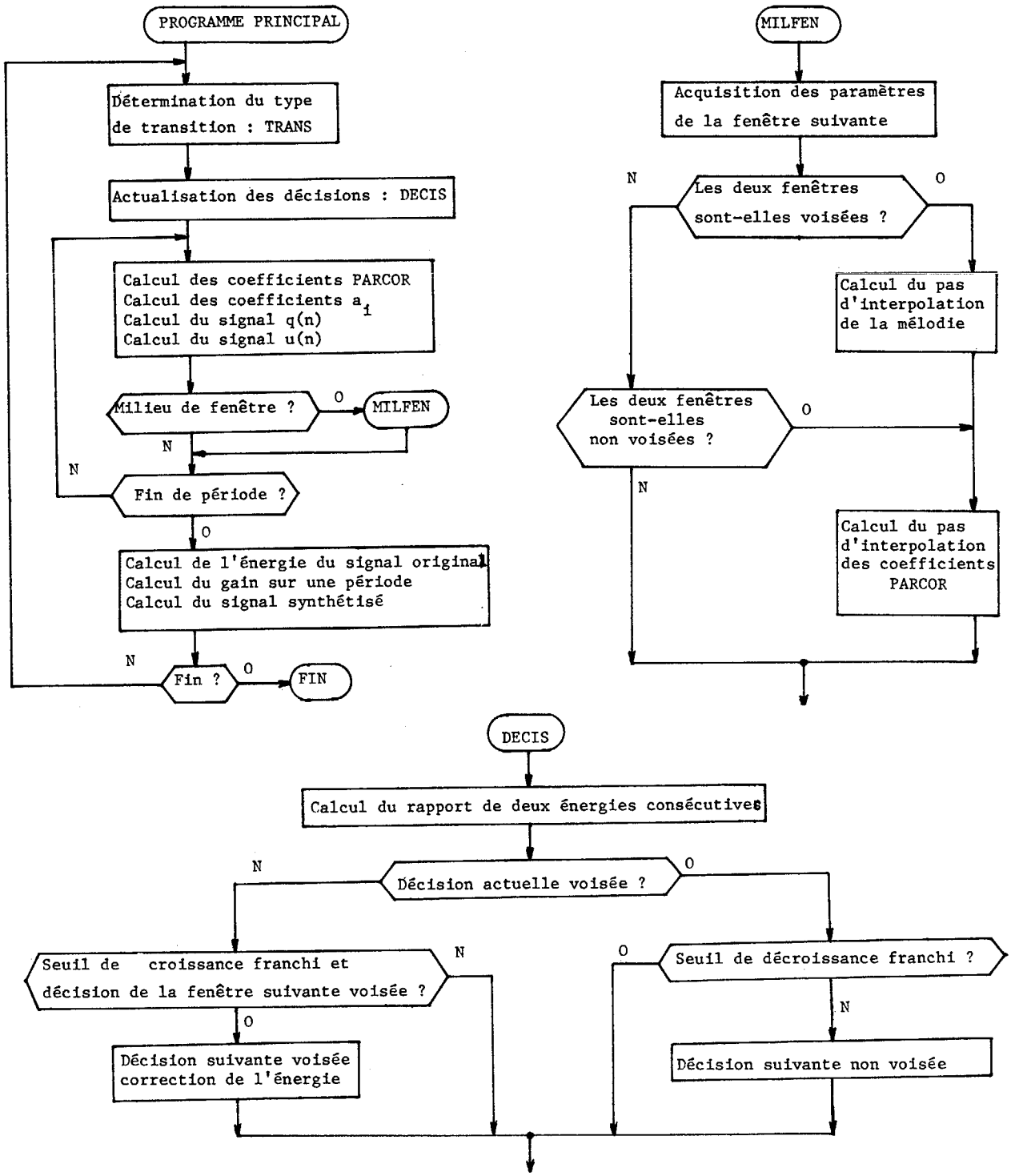
BIBLIOGRAPHIE

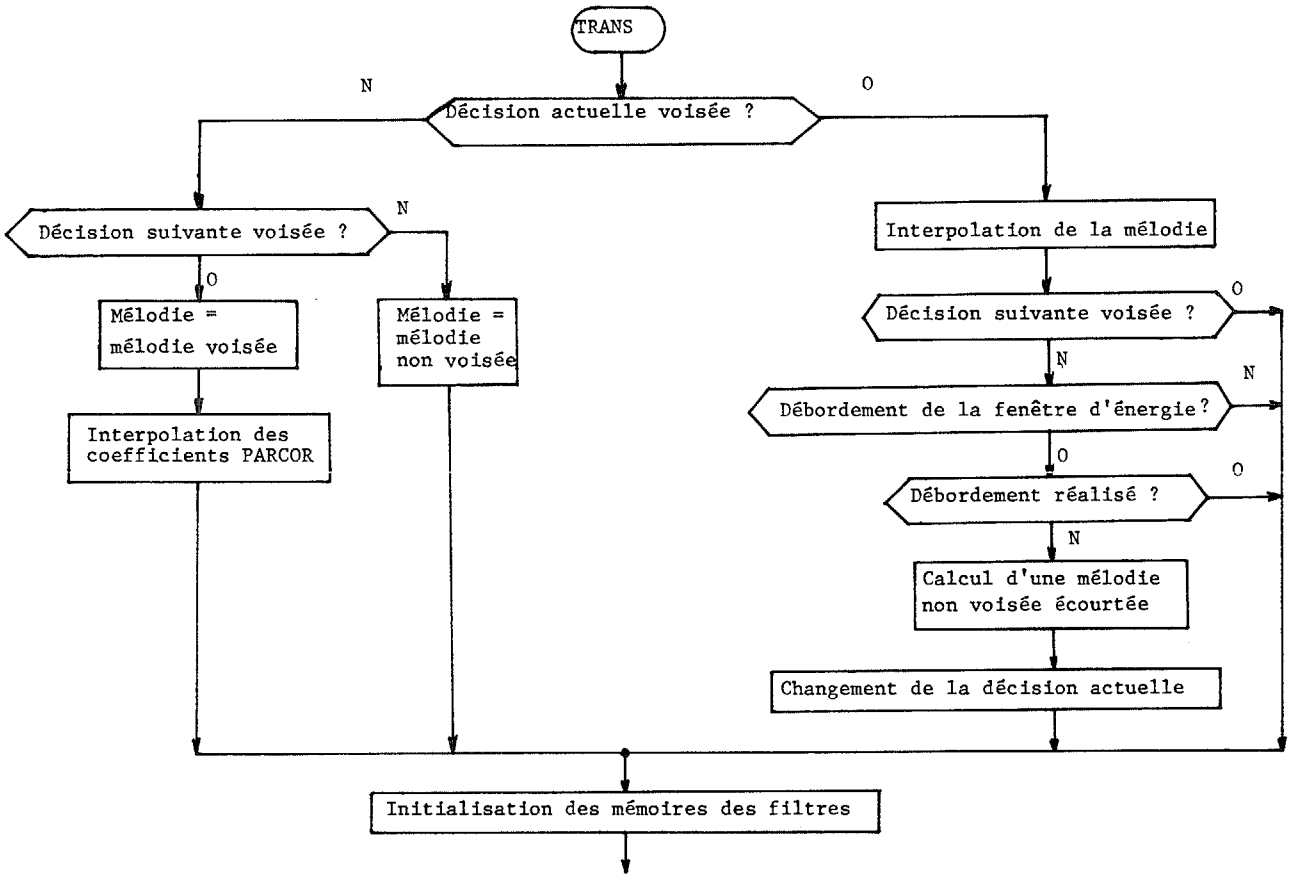
/1/ E. AZIZ, G. BRUN, J. MENEZ -
"Détection de la période du fondamental à l'aide de la fonction AMDF normalisée".
GRETSI, 1983

/2/ J.D. MARKEL, A.H. GRAY -
"Linear Prediction of Speech".
Springer Verlag, New York, 1976

/3/ B.S. ATAL, S.L. HANAUER -
"Speech analysis and synthesis by linear prediction of the speech wave".
J. Acoust. Soc. Amer. Vol 50, pp 637-655, Sept. 71.

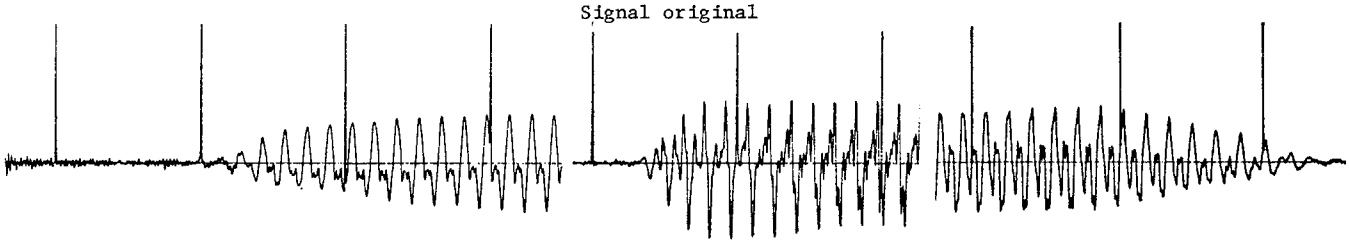
ANNEXE



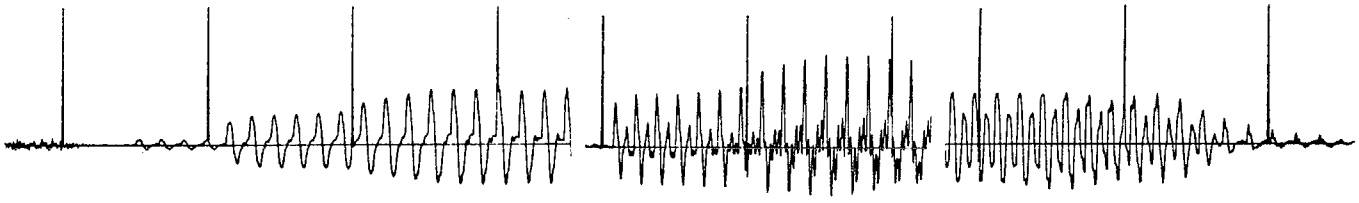


EXEMPLES DE SYNTHESE

Signal original



Signal synthétisé selon la méthode de Markel



Signal synthétisé selon notre méthode