Un nouveau procédé d'analyse et de synthèse paramétriques de la parole dans le domaine fréquentiel

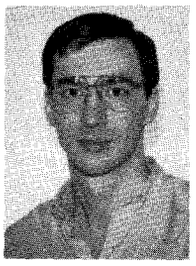
A new parametric speech analysis and synthesis technique in the frequency domain

Bruno WERY

Ingénieur de recherche du Service d'Acoustique de l'Université de Liège, Institut Montefiore, bâtiment B 28, B-4000 SART TILMAN.

Diplômé ingénieur civil électricien (électronique) en 1985 à l'Université de Liège, il occupe un poste d'ingénieur de recherche depius cette date dans le service d'acoustique de cette université.

Il y conduit un programme de recherche sous un contrat avec le Service de la Programmation de la Politique Scientifique de l'État Belge. Ce programme se situe dans le domaine du traitement numérique du signal sonore et consiste en l'élaboration de systèmes d'analyse automatique et de synthèse de la parole. Outre cette activité de recherche, il a assuré le développement d'une station de travail spécialisée pour le traitement de signal.

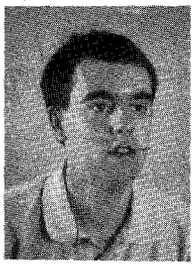


André LEROUX

Ingénieur de recherche du Service d'Acoustique de l'Université de Liège.

Diplômé ingénieur civil électricien (électronique) en 1984 à l'Université de Liège, il occupe successivement dans cette université les postes d'élève-assistant dans le service d'électronique 1982 et 1983, d'ingénieur de recherche dans le service de systèmes et automatique de 1984 à 1986 puis d'ingénieur de recherche dans le service d'acoustique depuis 1986.

Il a participé à un projet ESPRIT traitant l'interconnexion de réseaux locaux d'ordinateurs et, plus particulièrement, à la définition et l'évaluation de la liaison entre deux sous-systèmes. Il travaille actuellement sous un contrat du Service de la Programmation de la Politique Scientifique de l'État Belge dans le domaine de l'analyse et synthèse de parole par ordinateur. Il s'intéresse plus particulièrement aux problèmes d'implémentation en temps réel d'algorithmes sur DSP et à la créations d'architectures spécialisées.

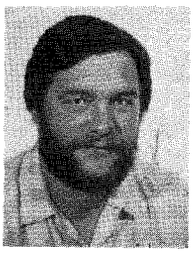


Henri-Philippe DELBROUCK

Ingénieur de recherche du Service d'Acoustique de l'Université de Liège.

Diplômé ingénieur civil électricien (électronique) en 1986 à l'Université de Liège, où il a occupé un poste d'ingénieur de recherche dans le service d'acoustique à partir de cette date. De 1988 à 1989, il a travaillé en tant que chercheur pour la chaire de télécommunication de l'École Royale Militaire Belge à Bruxelles. Monsieur Delbrouck travaille maintenant à nouveau dans le service d'acoustique de l'Université de liège.

Ses travaux ont porté sur le développement de synthétiseurs et de systèmes d'analyse basés sur le modèle de la prédiction linéaire ainsi qu'au développement de systèmes de reconnaissance du locuteur.



Jacques LECLERC

Premier Assistant du Service d'Acoustique de l'Université de Liège.

Diplômé ingénieur civil électricien (électronique) en 1978 à l'Université de Liège, il reçoit le grade de Docteur en Science Appliquées en 1984 dans la même Université. Il y occupe successivement les postes d'élève-assistant en 1977-1978, d'assistant de 1979 à 1985 et est nommé Premier Assistant en 1986 dans le service d'acoustique.

Aux côtés d'activités d'enseignement (de 2ème et 3ème cycle) et de recherches fondamentales (précision et répétitivité des mesures, caractérisation des salles de test du laboratoire), il s'est intéressé aux divers aspects de l'acoustique au cours de ses travaux : électroacoustique (modélisation des transducteurs, distortion), acoustique industrielle (tracé de cartes de bruit, modélisation des structures à baffles, efficasité des absorbants), acoustique urbaine (propagation, modélisation des écrans et des tunnels) et acoustique architecturale (logiciels d'analyse de la réponse impulsionnelle d'une salle, intelligibilité). En plus de des activités traditionnelles, il a développé un nouveau secteur de recherche dans le domaine du traitement de signal, appliqué à la parole.

RÉSUMÉ

Cet article présente une nouvelle approche de l'analyse et de la synthèse paramétriques de la parole basée sur l'élaboration d'un modèle du résidu du filtre inverse de la prédiction linéaire dans le domaine fréquentiel. Ce modèle fait intervenir la recherche d'une structure harmonique et de la période fondamentale, l'analyse en sous-bandes du signal pour la détermination de «proportions de voisement» de sorte à rendre la décision voisé-non voisé progressive et fonction de la fréquence.

Cette représentation paramétrique du signal résiduel est associée aux paramètres LPC habituels pour la transmission ou la mémorisation.

Le système ainsi créé présente la souplesse des systèmes paramétriques et est bien adapté aux problèmes de connexion entre trames intervenant dans les applications de synthèse à partir du texte.

MOTS CLÉS

Analyse-synthèse de la parole, prédiction linéaire, analyse fréquentielle, codage paramétrique, synthèse à partir du texte.

SUMMARY

This article describes a new approach for parametric analysis and synthesis of speech. It is based on the frequency domain modelling of the residual signal of an LPC analysis, including an harmonic structure and pitch extraction, and a sub-band analysis in order to determine a «voiced-unvoiced ratio» variable with frequency.

This residue parametric representation is added to the classical LPC parameters for transmission or memorisation.

Our system presents of course the flexibility of parametrical systems and is well adapted to frame to frame transition problems encountered.in text-to-speech applications.

KEY WORDS

Speech analysis-synthesis, linear prédiction, frequency domain analysis, parametric coding, text-to-speech synthesis.

Ce travail a été réalisé sous l'égide des Services de la Programmation de la Politique Scientifique de l'Etat Belge, dans le cadre de l'action concertée n° 85/90−81.

1. Introduction

Le domaine des codeurs totalement paramétriques destiné à des applications telles que la synthèse depuis le texte est encore fort restreint : en effet, depuis l'apparition du vocodeur LPC, peu de progrès ont été réalisés et peu de schémas nouveaux présentés. L'inconvénient majeur du vocodeur LPC réside dans la simplicité de la modélisation du signal d'excitation du filtre de synthèse. Si de nombreux schémas ont été proposés pour améliorer ce signal, la plupart ont pour inconvénient de renouer plus ou moins avec le codage d'onde. En effet, plutôt que d'affiner le modèle du signal d'excitation, on le remplace par une vue partielle de l'original, en le quantifiant dans les domaines temporel ou fréquentiel (méthodes par prédictions multiples, utilisation d'un résidu filtré en bande étroite, méthodes de quantification vectorielle, modélisation par des sommes de sinusoïdes et méthodes d'analyse-synthèse, LPC «multipulse»,...). Si bon nombre de ces techniques sont prometteuses et très bien adaptées à la transmission de parole à bas débit, elles sont inapplicables en synthèse à partir du texte.

Le domaine fréquentiel a été négligé jusqu'ici parce qu'il implique de grandes masses de calculs lors du passage d'un domaine à l'autre. La situation change grâce à l'apparition de processeurs capables d'effectuer ceux-ci en temps réel en utilisant des transformées de grande dimension.

La technique présentée ici requiert une masse de calculs importante, telle qu'il paraît difficile actuellement d'exécuter ce type de codage en temps réel sur des systèmes standards de coût réduit. L'entrée en production d'une nouvelle génération de processeurs de traitement de signal (TMS32030, DSP56000 et 96000) va bouleverser très rapidement cette situation.

2. Structure du système

2.1. Représentation du résidu de la prédiction linéaire

L'examen du résidu d'une prédiction linéaire conduit à la conclusion qu'il est difficile de séparer précisément le signal de parole en trames voisées et en trames non voisées. De plus, si l'on s'intéresse au spectre de ce résidu, on observe que la structure harmonique des sons voisés subit des variations importantes en fonction de la fréquence. Celle-ci disparaît parfois dans certaines bandes. Elle est souvent peu marquée au-delà de quelques harmoniques. La figure 1 présente quelques exemples de spectres dont la structure harmonique n'est pas uniforme en fonction de la fréquence. Cette constatation nous amène à penser que l'un des principaux problèmes des synthétiseurs LPC réside très probablement dans la simplicité de la décision voisé-non voisé, incapable de décrire cette situation. Nous pensons qu'une décision progressive

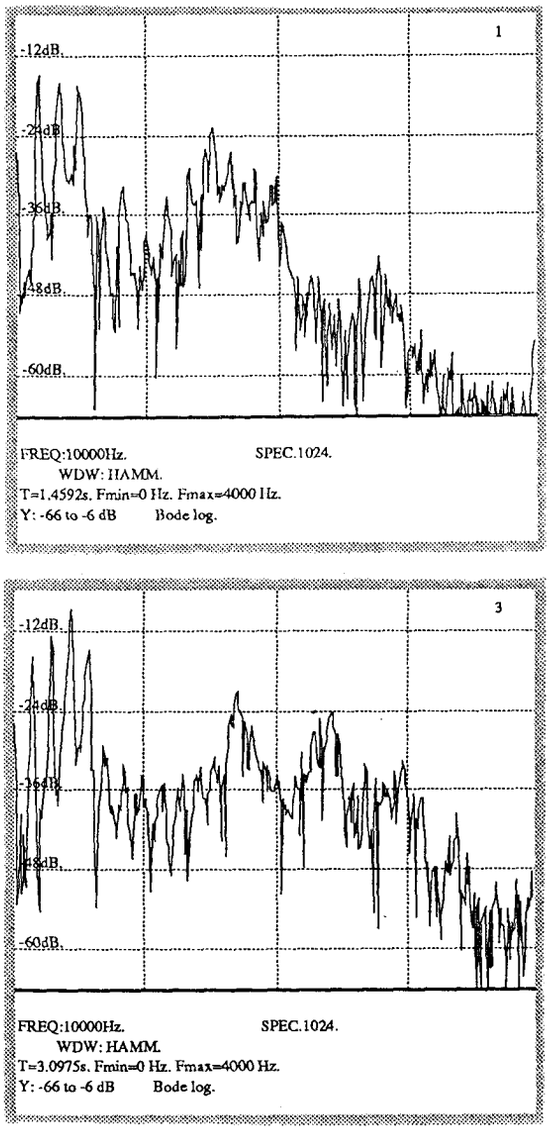


Fig. 1. – Spectres de trames de parole démontrant la variabilité du caractère harmonique des sons voisés avec la fréquence.

entre ces deux états est nécessaire et que celle-ci doit dépendre de la fréquence. Cette observation est aussi à l'origine des développements présentés par D. Griffin et J. S. Lim [10], [11], [12] et [15].

Notre but est de créer un système complètement paramétrique, capable d'évoluer au cours du temps en fonction de paramètres transmis sur base d'un découpage en trames. Le problème de raccordement entre trames est toujours aigu dans le domaine temporel et entraîne la création de systèmes adaptatifs dont le comportement est difficile à prévoir. Le fait de faire varier certains paramètres de façon «continue» au cours du temps peut résoudre le problème de génération de transitoires à la fréquence de trames, mais rend l'analyse des systèmes particulièrement ardue, sinon impossible. La théorie des filtres adaptatifs dans le domaine fréquentiel est une solution élégante à ces problèmes.

Ces considérations nous ont conduit à représenter le signal dans le domaine fréquentiel. Le principal obstacle, aujourd'hui franchi, a toujours été un manque de puissance de calcul des processeurs disponibles. Dans notre système, le résidu de la prédiction linéaire est représenté sous la forme d'une pondération entre un signal voisé et un signal non voisé. Cette pondération varie selon un découpage en bandes de fréquences. L'ensemble des opérations (à l'exception de la prédiction linéaire elle-même) s'effectuent dans le domaine fréquentiel. La réalisation des prédicteurs dans ce domaine est évidemment possible. Cette solution paraît attirante par sa capacité d'éliminer les problèmes de raccordement entre trames et de réduire le nombre de catégories d'opérations à effectuer.

Comme cela nous paraît absolument nécessaire pour des applications telles que la synthèse depuis le texte, les paramètres transmis ou stockés ont une signification physique claire, ce qui permettra l'élaboration aisée de stratégies de liaison entre éléments phonétiques. L'analyse du signal s'effectue trame par trame et le jeu de paramètres suivant est transmis pour chacune de celles-ci :

– les cœfficients PARCORS de la prédiction linéaire (une trame sur deux);

– la fréquence fondamentale;

– une proportion voisé-non voisé dans différentes bandes de fréquences (8);

– une mesure de l'énergie dans chacune de ces bandes.

La quantification de ce jeu de paramètres conduira à un débit compris entre 4800 et 9600 bits par seconde.

On remarque qu'aucune information de phase n'est transmise. Cela ne signifie pas que nous la considérions comme sans importance, mais nous ne pouvons la représenter sous la forme d'un simple jeu de paramètres. Une technique souvent rencontrée est de ne transmettre la phase que pour les harmoniques principales des sons voisés (voir, par exemple, [12]). Ici, ces harmoniques ne sont pas transmises en soi, mais seront reconstituées sur base de la période fondamentale et de ses variations. La structure des trames synthétiques ne sera certainement pas identique à celle des trames analysées, ce qui rend caduc ce type de manipulation sur la phase. En effet, sa signification physique (décalage temporel entre différentes composantes) est directement attachée à la fréquence à laquelle elle se rapporte.

2.2. Description générale du système

Les diagrammes généraux des systèmes d'analyse et de synthèse sont présentés aux figures 2a et 2b. Dans la suite de ce texte, les chiffres Romains renvoient à ces figures. Nous commençons par une brève description du système suivie d'une discussion détaillée de chaque module qui fera l'objet du chapitre suivant.

A) Analyse du signal

(I) Les deux premiers blocs symbolisent l'analyse par prédiction linéaire classique (fréquence d'échantillonnage : 10 kHz; bande passante : 4kHz). Elle est réalisée par la méthode de l'autocorrélation selon l'algorithme de SCHUR-LEROUX. Elle permet l'obtention des coefficients PARCORS, directement utilisés pour la transmission. Le filtre d'analyse

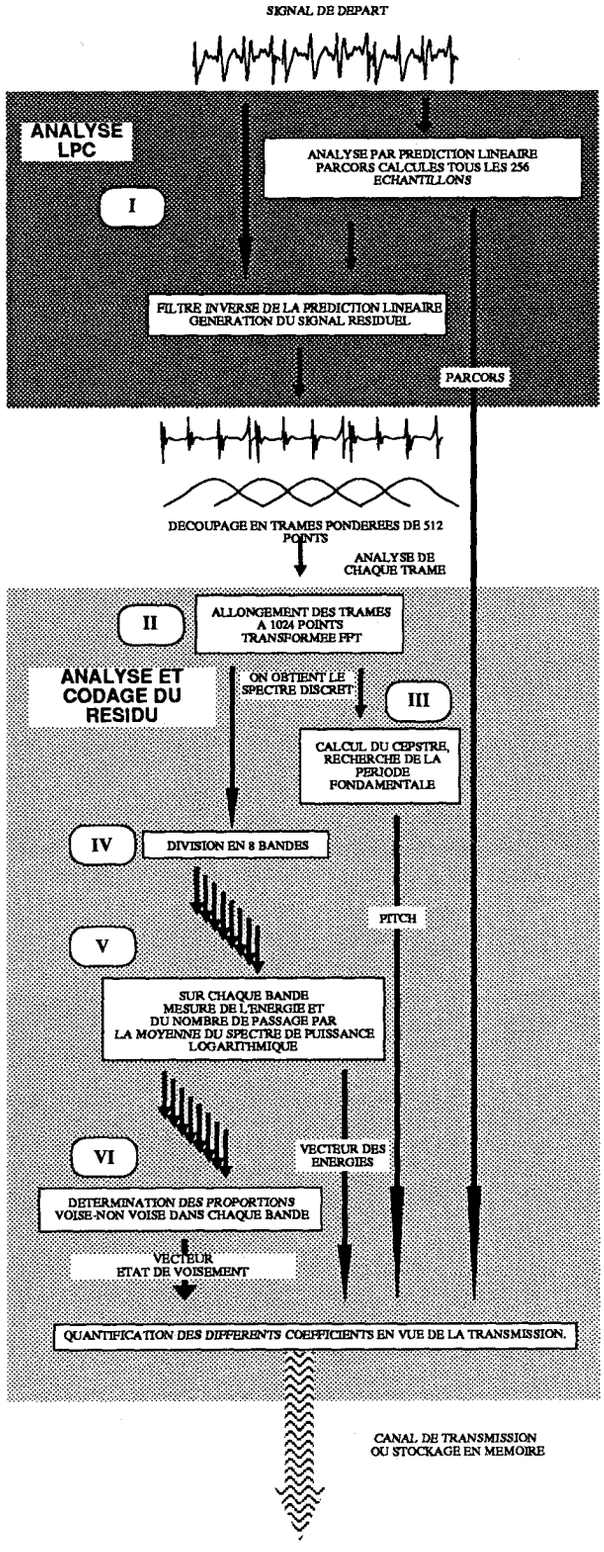


Fig. 2a. – Schéma de principe du système d'analyse du nouveau codeur.

(comme le filtre de synthèse en (XII)) a la forme en treillis conventionnelle.

(II) Le résidu de la prédiction linéaire est traité trame par trame. Après pondération par une fenêtre, une transformée de Fourier rapide est appliquée. On obtient ainsi la représentation spectrale du résidu.

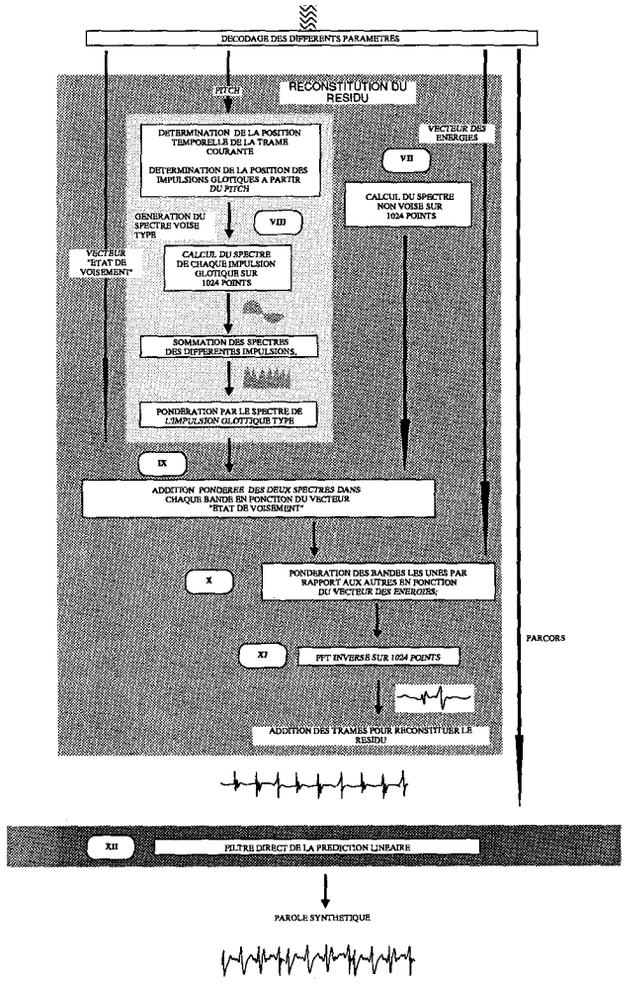


Fig. 2b. – Schéma de principe du système de synthèse du nouveau codeur.

Remarquons que la dimension des trames est liée à la dimension des FFT. Celle-ci doit être suffisante pour permettre une détection précise de la période fondamentale.

(III) La recherche de la période fondamentale s'effectue par la méthode du cepstre selon un algorithme de «poursuite» et la valeur obtenue est l'un des paramètres transmis.

(IV) Le spectre du résidu est alors séparé en sous-bandes (dans nos essais, 8 sous-bandes de 100 à 4000Hz).

(V) Chaque bande fait l'objet de deux mesures : l'énergie et le nombre de passages par sa moyenne du spectre de puissance logarithmique. Le vecteur des énergies fait partie des paramètres transmis.

(VI) On évalue une proportion entre partie voisée et non voisée dans chaque bande, à partir des mesures de passage par la moyenne et de la période fondamentale. Ces proportions constituent le dernier ensemble de paramètres à transmettre.

Ce jeu de paramètres doit être codé efficacement pour la transmission.

B) Synthèse

La première opération est la reconstitution du spectre d'une trame totalement non voisée et d'une trame totalement voisée.

(VII) La génération des trames non voisées est assez simple : il suffit de disposer d'un catalogue de spectres pondérés par une fenêtre. Il y cependant un certain nombre de précautions à prendre, pour tenir compte du fait que les trames temporelles qui en résulteront se recouvriront partiellement.

(VIII) La génération des trames voisées est une opération quelque peu plus complexe car il ne s'agit pas de calculer simplement un spectre de puissance avec une période fondamentale donnée. Il faut s'assurer qu'après reconstruction, le signal temporel généré est bien périodique à moyen terme (quelques trames) et qu'il ne comporte pas de discontinuités d'amplitude dues à des phénomènes d'interférence entre trames. Un catalogue ne peut plus suffire et un calcul spécifique de chaque spectre est nécessaire. Le spectre généré au départ est celui d'un train d'impulsions. On pourra multiplier celui-ci par le spectre d'un impulsion glottique type.

(IX et X) Les deux spectres «types» ainsi générés seront pondérés bande par bande en fonction du vecteur «état de voisement» et «énergies», afin d'obtenir le spectre du résidu synthétique de la trame.

(XI) Le résidu synthétique est alors obtenu par FFT inverse et sommation des différentes trames (recouvrement). Cette technique est dérivée de l'algorithme «Overlapp – Add» détaillé dans la référence [1]).

(XII) Le filtrage du résidu par le filtre de synthèse de la prédiction linéaire termine le processus.

3. Description détaillée du système

3.1. Analyse par prédiction linéaire

La combinaison d'une prédiction linéaire avec une analyse en sous-bandes peut paraître une redondance inutile. En effet, si notre objectif est de créer une nouvelle représentation paramétrique pour le résidu de la prédiction linéaire, il peut sembler que ce modèle serait particulièrement bien adapté à la représentation du signal lui-même. Ce serait le cas si notre hypothèse de départ durant la conception du codeur n'était pas une certaine uniformité du spectre à analyser: cette uniformité est par exemple nécessaire pour permettre la séparation voisé-non voisé selon la méthode simple du nombre de passages par sa moyenne du spectre logarithmique. Il en est de même pour résoudre le problème de quantification de l'énergie dans les bandes, qui nécessiterait un débit binaire beaucoup plus important si la prédiction linéaire n'était pas appliquée; il nous semble que celle-ci conduit à un faible accroissement du débit pour le même service. La prédiction linéaire n'est pas fondamentalement nécessaire mais permet une simplification importante du système pour une charge de calcul qui reste modeste par rapport à celle occasionnée par les opérations spectrales. Elle est également intéressante sur le plan du conditionnement numérique.

Les techniques employées sont classiques. Ces techniques sont bien développées dans la littérature, par exemple dans [4], et nous ne les détaillerons pas ici. La fréquence d'analyse et la dimension des trames sont alignées sur celles de l'analyse spectrale qui suit. Dans nos essais, nous procédons à une analyse sur 512 points pondérés par une fenêtre de Hamming. Le recouvrement entre trames est de 50 %. Le jeu de paramètres LPC n'est transmis qu'une fois par paire de trames alors que le jeu de paramètres frequentiels l'est pour chaque trame.

L'application des filtres d'analyse et de synthèse par prédiction linéaire sont les deux seules étapes du traitement du signal qui s'opèrent dans le domaine temporel. On pourrait envisager de réaliser ces filtres adaptatifs dans le domaine fréquentiel. L'analyse elle-même pourrait être conduite dans ce domaine. En effet, la relation définissant la fonction d'autocorrélation dans le domaine fréquentiel échantillonné s'écrit :

$$R(i) = \frac{1}{N}\sum_{k=0}^{N-1} S(e^{j\omega_k}).S^*(e^{j\omega_k}).(e^{j\omega_k})$$

où $S(e^{j\omega_k})$ représente la transformée de Fourier discrète du signal et R(i) les cœfficients de corrélation pour la trame considérée.

L'usage de la préaccentuation avant analyse sera discuté au paragraphe 3.7.

3.2. Dimension des trames pour la FFT

La dimension des trames a une influence capitale sur le comportement de notre codeur (débit, impératifs de l'analyse cepstrale, réponse transitoire du système).

Il faut donc choisir la plus grande longueur de trame compatible avec les hypothèses de quasi-stationnarité du signal de parole. Nous avons choisi de travailler avec des trames de 512 points, pondérées par une fenêtre de Hamming. Les trames sont allongées à 1024 points pour les opérations de FFT et le recouvrement entre trames est de 75 %; les trames se succèdent donc tous les 128 échantillons. Ces choix permettent une détection efficace de la fréquence fondamentale en dessous de 100 Hz tout en conservant au système une vitesse d'adaptation inférieureà 30 ms.

3.3. Détermination de la période fondamentale

Les critères de sélection pour la méthode de détermination de la fréquence fondamentale sont les suivants :

– la détection doit se faire à partir de la représentation spectrale du signal, afin de conserver un ensemble de paramètres cohérents, décrivant bien un même objet;

– la précision ne doit pas dépendre de la résolution fréquentielle de la transformée rapide (dans notre cas environ 10 Hz, ce qui conduirait à des erreurs de plus de 10 % pour les fondamentales basses);

– l'algorithme doit rester stable lorsque la trame est peu voisée et auraitété déclarée «non voisée» par un algorithme de détection traditionnel;

– l'algorithme ne sert pas à la détection de voisement.

Nous avons retenu la méthode du cepstre et modifié un algorithme présenté par W. Hess [3, chapitre 8, pages 399 à 409]. Notre variante ajoute au mécanisme d'asservissement une certification de la valeur obtenue à partir du spectre (figure 3).

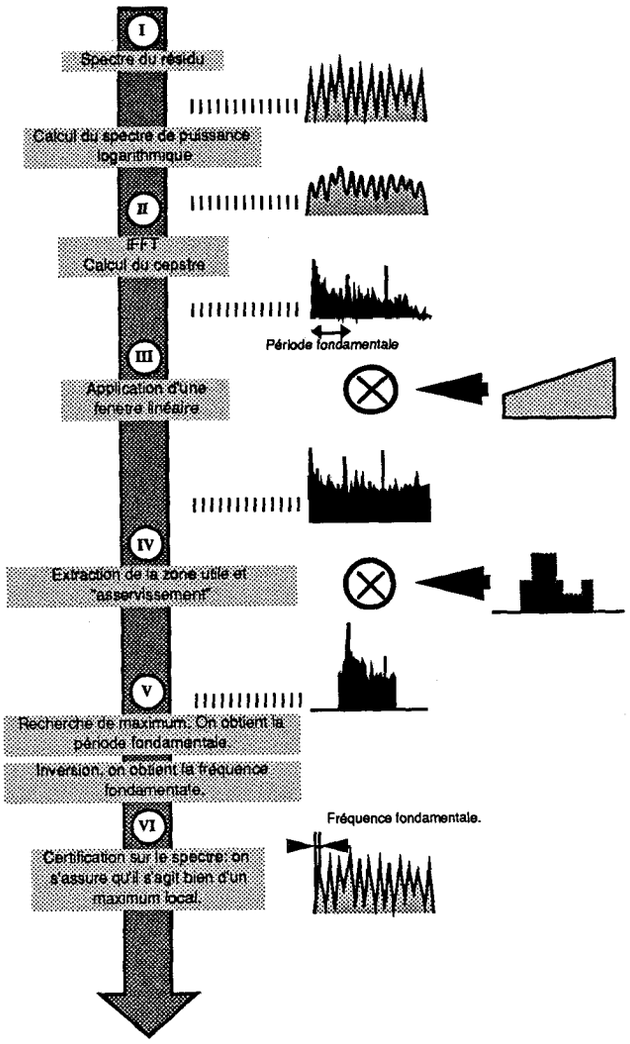


Fig. 3. – Algorithme de détermination de la fréquence fondamentale.

Le spectre est calculé de manière classique (I et II), puis est pondéré sur la zone utile par une fenêtre linéaire (III). Celle-ci est destinée à compenser la pente naturelle du cepstre et dépend, entre autres, de la largeur et de la nature de la fenêtre de pondération appliquée lors de l'analyse spectrale. Avant la recherche d'un maximum local (V), on renforce une zone autour de la valeur trouvée au cours de l'analyse précédente et on déforce une zone semblable autour du double de celle-ci (IV). Cette démarche suppose que la fréquence fondamentale ne peut pas varier rapidement et cherche à éviter les confusions avec la première harmonique cepstrale. Le maximum local correspond à la période fondamentale du signal.
Cette valeur est ensuite certifiée en s'assurant que la fréquence correspondante est bien un maximum local du spectre (VI). En pratique, cette opération doit être tolérante et est menée en comparant les amplitudes à la fréquence fondamentale supposée (F_{fond}) et aux demi-harmoniques précédente et suivante (figure 4). On vérifie que

$$S(_{fond}) > > S(F_{fond}/2) \quad \text{et} \quad S(F_{fond}) > > S(3F_{fond}/2)$$

Si cette condition n'est pas remplie, on conserve la valeur trouvée pour la trame précédente. En pratique,

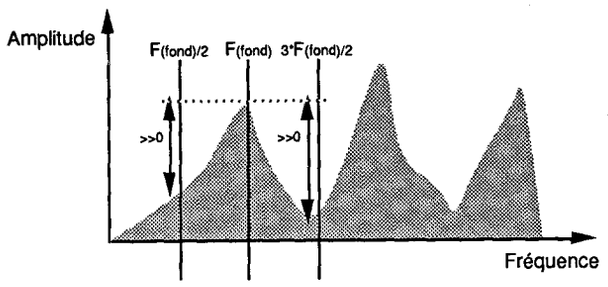


Fig. 4. – Principe de la vérification de la fréquence fondamentale.

nous avonségalement ajouté un test à une distance fixe de deux échantillons fréquentiels de part et d'autre de la fréquence fondamentale. En effet, la méthode cepstrale n'est précise que s'il existe une structure harmonique minimale. Certaines voyelles dites par certains locuteurs sont quasi sinusoïdales et nous ont posé un problème de détection, parfaitement résolu par ce mécanisme simple.
Remarquons que la charge de calcul due au cepstre se partage avec les opérations d'analyse en sous-bandes, car le spectre logarithmique y est également nécessaire.

3.4. Séparation en sous-bandes, mesure de l'état de voisement et de l'énergie

La structure harmonique du résidu n'est généralement pas uniforme en fonction de la fréquence. Par conséquent, nous pensons que l'analyse de l'état de voisement doit être non seulement progressive entre les états «voisé» et «non voisé», mais doit conduire à plusieurs décisions pour un même spectre, en fonction d'un découpage en bandes de fréquence. Deux optiques sont possibles. La plus simple consiste à diviser le spectre en sous-bandes et à prendre une décision dans chacune de celles-ci. La seconde consiste à choisir les bandes en fonction de la structure harmonique du spectre: on effectue d'abord l'analyse de voisement en fonction de la fréquence; on recherche ensuite les zones uniformes [12]. Cette seconde solution demande la transmission d'informations complémentaires concernant les fréquences charnières des bandes. Si la décision voisé-non voisé ne connaît pas d'intermédiaire, l'état de voisement alterne de bande en bande et sa valeur ne doit plus être transmise. Cette solution conduirait à un débit d'information variable.
Notre système utilise un ensemble de bandes fixes dont le choix, primordial pour un bon fonctionnement du codeur, est guidé par les considérations suivantes :
– la quantité d'informations ajoutées au jeu de paramètres traditionnels du LPC est directement proportionnelle au nombre de bandes;
– la largeur des bandes doit être suffisante par rapport à la fréquence fondamentale maximale : pour qu'une analyse en voisement y ait un sens, il faut que chaque bande soit toujours plus large qu'un pic harmonique;
– il faudrait pouvoir choisir des bandes perceptuellement équivalentes. Nous ne connaissons pas de solution au problème de l'équivalence perceptuelle de deux spectres à bande limitée. Nous avons choisi de partir des bandes «d'égale perception» définies pour

le codage en sous-bandes ([6] et [7]), légèrement modifiées en fonction des résultats de nos essais. En fait, ce choix est peu crucial pour autant que le nombre de bandes soit suffisant.

Il s'agit donc de trouver un compromis entre le débit admissible et la qualité globale du codeur. Nous travaillons pour l'instant avec 8 bandes (100-400, 400-700, 700-1000, 1000-1300 , 1300-1650, 1650-2000, 2000-3000 et 3000-4000 Hz) dont les limites ont été fixées à la suite de tests d'écoute.

L'analyse de l'état de voisement dans une bande est un problème nouveau. Un grand nombre de solutions peuvent être envisagées, avec divers degrés de complexité. Deux voies principales se dégagent :

– on peut effectuer la recherche d'une proportion qui conduirait à la synthèse du spectre le plus proche du spectre original, selon l'un ou l'autre critère. Un critère simple, basé sur le modèle de la somme d'un bruit blanc et d'un son à harmoniques constantes, serait l'égalité de la moyenne et de la variance du spectre de puissance. Des critères plus complexes pourraient être envisagés sur base du calcul d'une distance. Ces méthodes, soit nous ont paru trop complexes, soit n'ont pas donné de bons résultats pratiques ;

– on peut se baser sur des caractères «géométriques» du spectre. Ce qui nous a conduit à envisager deux méthodes simples : l'écart entre la moyenne du spectre de puissance et son enveloppe ou l'étude du nombre de passages par sa moyenne du spectre de puissance logarithmique.

Nous avons adopté cette dernière méthode.

La mesure du nombre de passages N_p par la moyenne du spectre logarithmique est effectivement un indicateur de l'état de voisement, pourvu que la bande supérieure à la fréquence fondamentale soit suffisamment étroite par rapport aux variations résiduelles de l'enveloppe spectrale du résidu. Dans le cas d'un son voisé, N_p tend approximativement vers $2(B/F_f)$ où B est la largeur de la bande et F_f la fréquence fondamentale. De même, N_p varie aux alentours de B/3 pour un son non voisé, B/2 constituant une borne supérieure de N_p (figure 5).

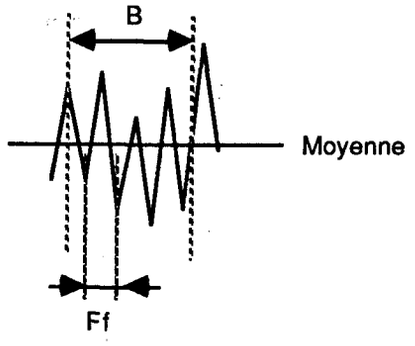


Fig. 5. – Principe de la détermination du taux de voisement.

Remarquons que cet indicateur, très simple, est d'autant plus avantageux que le spectre de puissance logarithmique est également nécessaire pour la détection de la fréquence fondamentale.

Il n'est cependant pas d'une fiabilité absolue. Nous l'avons complété par un mécanisme de décision entre bandes et entre trames capable de déclarer, selon certains critères, une trame «complètement voisée» ou «complètement non voisée». Ce mécanisme nous a permis d'obtenir une meilleure stabilité des voyelles longues, ainsi que des sifflantes. Si V_i est le taux de voisement de la bande i, c'est-à-dire le rapport entre les énergies des parties voisée et non voisée, on décide que que la trame est totalement voisée pour autant que la moyenne de V_i soit supérieure à un seuil fixé. Il en est de même pour déclarer la trame complètement non voisée. La transition d'une trame totalement voisée vers une trame totalement non voisée est interdite, ainsi que la transition inverse.

Lors de la reconstruction, il convient de synthétiser dans chaque bande, un signal qui rende la même impression perceptuelle d'intensité que celui qui y était présent lors de l'analyse. Le problème est d'autant plus aigu qu'il faut y ajouter divers mécanismes liés à la perception, tels que l'effet de masque d'une bande sur une autre. Il nous semble cependant que l'énergie du signal dans la bande est un meilleur «estimateur» que ne l'est l'enveloppe LPC. Sa mesure explicite dans chaque bande est nécessaire. Si le filtre inverse de la prédiction linéaire tend à aplatir l'enveloppe du spectre, il n'uniformise pas pour autant l'énergie par rapport à des bandes arbitraires. En particulier, dans le spectre d'un son parfaitement voisé, l'énergie est directement proportionnelle au nombre d'harmoniques présentes dans la bande, grandeur qui ne peut être conservée constante. Les travaux de Zwicker dans le domaine de l'isosonie des sons complexes ont conduit à établir une équivalence entre la perception des sons et une pondération des énergies locales du spectre par tiers d'octave [14]. Il pourrait être intéressant de se servir de ces résultats comme critère lorsque les bandes couvrent plusieurs tiers d'octave.

3.5. Génération du spectre synthétique

Le synthétiseur doit être capable, à partir du jeu de paramètres qui vient d'être défini, de générer un spectre synthétique (points VII et suivants de la figure 2). Ce spectre est ensuite utilisé comme s'il s'agissait de l'étape de reconstruction d'un filtre adaptatif par transformée de Fourier. Cette reconstruction s'effectue en trois étapes : génération d'un spectre non voisé, génération d'un spectre voisé à partir de la fréquence fondamentale et de son évolution, pondération de ces deux spectres par bandes en fonction des vecteurs «état de voisement» et «énergies». Ces constructions doivent s'effectuer en tenant compte des recouvrements entre trames et en s'assurant de la continuité des structures harmoniques de trame en trame.

3.5.1. *Génération du spectre non voisé*

La synthèse du spectre non voisé ne présente qu'une seule difficulté : il faut s'assurer qu'au moment de la reconstruction, il n'y ait pas d'interférences destructrices entre les différentes trames superposées. Pour cela, il faut que les spectres produits successivement soient une image de trames qui se recouvrent à 75 %.

De plus, pour se trouver dans les conditions de l'algorithme de filtrage adaptatif, les trames temporelles correspondantes doivent être pondérées par une fenê-

tre de Hamming. La génération dynamique de telles trames entraîne une charge inadmissible. Une solution simple consiste à réaliser un catalogue de trames calculé une fois pour toutes et conservé, par exemple, dans une mémoire morte. Le catalogue est élaboré en analysant un signal aléatoire type, en effectuant des analyses spectrales avec la fenêtre de Hamming et un recouvrement de 75 % (figure 6).

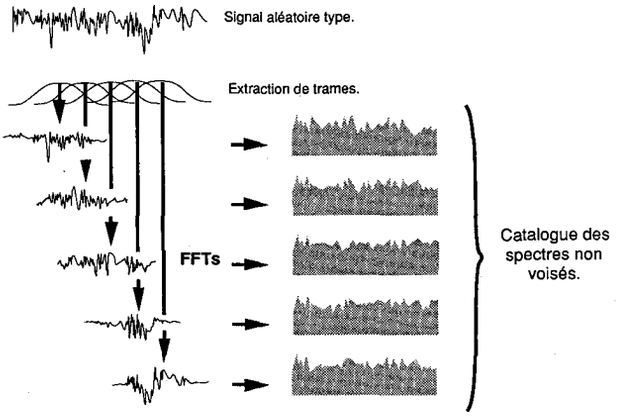


Fig. 6. – Génération des trames non voisées.

Les trames y sont choisies successivement et le catalogue peut rester limité, pourvu que les dernières trames aient été calculées en fonction d'un recouvrement avec les premières. La seule contrainte est que la séquence mémorisée soit suffisamment longue pour ne pas générer des signaux réguliers aux basses fréquences. On peut choisir, par exemple, un temps total d'une seconde, ce qui ferait un catalogue de 80000 échantillons, taille très raisonnable par rapport aux tailles des mémoires actuelles. L'espace adressable des DSPs n'est pas non plus une contrainte : on se trouve dans un cas où un mécanisme de pagination est évident.

3.5.2. *Génération des spectres voisés*

Nous nous trouvons devant un problème nettement plus épineux, car en plus des contraintes qui existent pour les spectres non voisés, il existe une profonde interaction entre les spectres successifs, liée à la continuité de la structure périodique dans le domaine temporel.

La solution qui consisterait en la reconstruction «géométrique» du spectre à partir de sa fréquence fondamentale se heurte à un certain nombre de problèmes. Pour assurer cette continuité, l'information d'amplitude ne suffit pas; il faut également l'information de phase. Celle-ci pourrait être calculée, en fonction de conditions liées à cette continuité. Nous ne connaissons cependant pas de solution acceptable à ce problème. Une autre solution consisterait à extraire cette information lors de l'analyse. Pour qu'elle soit utilisable, il faut que les positions des pics harmoniques importants soient parfaitement établies et que l'on transmette précisément cette position pour chacun de ces pics. Ce qui n'est pas possible lorsqu'on veut réduire cette information à une fréquence fondamentale! Il n'est alors plus question d'agir sur la prosodie en déplaçant les harmoniques. Enfin, les problèmes de raccord entre trames, lors de la liaison entre éléments phonétiques repose le même problème de continuité... Certains auteurs ont résolu ce problème lors de synthèses par FFT en utilisant une synthèse synchrone avec la période fondamentale et des fenêtres de longueur variable [14]. Ce type de manipulation est trop différent de ce que nous avons développé ici pour être applicable.

Il ne faut pas confondre ce problème avec celui de l'importance de la phase sur le plan perceptuel. Assurer la continuité des structures harmoniques dans le domaine temporel est absolument nécessaire et le non respect de cette règle entraîne la génération d'une parole synthétique «chuchotée» extrêmement désagréable. Le véritable problème est de savoir, lorsque cette continuité est assurée par d'autres moyens que la transmission de la phase originale, dans quelle mesure celle-ci pourrait être employée pour améliorer la qualité perceptuelle de la parole.

La solution adoptée fait implicitement appel au domaine temporel. Considérant le signal d'excitation comme une séquence d'impulsions, on calcule la position de celles-ci dans le domaine temporel à partir de la fréquence fondamentale déterminée pour chacune des trames. On sait que cette fréquence varie peu d'une trame à la suivante. En tenant compte du taux de recouvrement précité, nous avons décidé de ne faire intervenir la fréquence fondamentale déterminée pour la trame en cours d'analyse que pour le calcul de la position des impulsions situées dans le dernier quart de cette trame. Les impulsions présentes dans les trois premiers quarts sont donc issues de l'analyse des trois trames précédentes. On calcule ensuite le spectre de la séquence pondérée par la fenêtre d'analyse (figure 7).

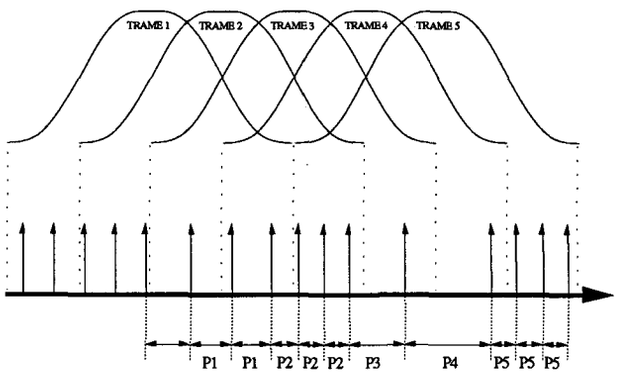


Fig. 7. – Génération du spectre voisé : calcul de la position des impulsions (les variations de fréquence fondamentale d'une trame à l'autre ont été exagérées pour la clarté de la figure).

En pratique, la position des impulsions est calculée dynamiquement en fonction de la réception des paramètres et le calcul du spectre est effectué sans qu'il soit nécessaire de reconstruire, à proprement parler, un signal temporel. Ce calcul de spectre est, de plus, très réduit car il existe rarement plus d'une dizaine d'impulsions par trame.

Cette solution génère le spectre d'un train d'impulsions, qu'il faut transformer afin d'éviter les problèmes de «métallisme» des sons voisés. Cela peut se faire aisément à ce niveau, en multipliant élément par élément le spectre par une fonction de transfert,

image d'une impulsion glottique type. Le choix d'un modèle pour cette impulsion glottique a une influence directe sur l'aspect subjectif de la parole produite. En effet, nous avons constaté (ce qui n'est pas surprenant) une très nette amélioration du timbre si une excitation complexe est employée au lieu de l'impulsion traditionnelle.

Les deux spectres «types» ainsi générés vont être pondérés dans chaque bande en fonction du vecteur «état de voisement» (IX) et «énergies» (X) afin d'obtenir le spectre du résidu synthétique sur la trame.

Cette opération revient à faire la somme des deux spectres après les avoir pondérés par les fonctions de transfert de deux filtres, ce qui correspond à l'expression générale

$$\sum a_k \left[u(\omega - \omega_{1k}) - u(\omega - \omega_{2k}) \right]$$

où $u(\omega)$ est la fonction de Heaviside ;
a_k est un cœfficient fonction des bandes ;
ω_1 et ω_2 représentent les bornes de chaque bande d'analyse.

Une telle fonction de transfert ne correspond évidemment pas à une réponse impulsionnelle finie inférieure au quart de la longueur de la fenêtre (moins la longueur de l'impulsion glottique). On risque donc d'introduire de l'«aliasing temporel». On peut résoudre ce problème en utilisant des fenêtres correspondant à chaque bande et dont la réponse impulsionelle serait finie. Celles-ci seront sommées après multiplication par les facteurs de pondération de leurs bandes respectives, afin d'obtenir les fonctions de pondération globales pour chaque spectre.

La reconstitution du résidu temporel et la prédiction directe n'appellent pas de commentaires additionnels.

3.6. Résultats pratiques

L'étude des qualités perceptuelles d'un codeur ne peut se faire qu'à travers d'importants tests d'intelligibilité et d'agrément. Ceux-ci font l'objet d'une étude en cours. Nos premiers essais donnent une intelligibilté syllabique de l'ordre de 80 % et font ressortir une bonne qualité perceptuelle, assez naturelle.

Les principales qualités sont :

- une reproduction très fidèle des consonnes longues, des semi-voyelles, des nasale et des liquides ;
- l'absence totale d'aspect métallique de la voix ;
- une très bonne reproduction en hautes fréquences, ce qui entraîne un aspect naturel certain ;
- de très bonnes transitions entre les différents sons.

Les défauts sont :

- une mauvaise reproduction des voyelles brèves, qui tendent à disparaître au profit des consonnes voisines ;
- une légère impression d'écho sur ces mêmes voyelles brèves ;
- un aspect légèrement «bruité» de tous les sons voisés ;
- les occlusives sont légèrement estompées ;
- suite à certaines simplifications indiquées dans le paragraphe suivant, la présence d'un léger bruit de fond et de faibles transitions énergétiques à la fréquence de trame.

La principale source de ces défauts est très certainement la longueur des trames qui limite l'adaptation dynamique du système. La seconde est une certaine faiblesse de notre estimateur de l'état de voisement pour les sons très voisés.

Les figures 8 et 9 comparent des représentations temporelles globales et locales du signal vocal avant et après codage. Cette comparaison fait ressortir quelques discontinuités dans les segments voisés (pointés par une flèche). Il s'agit, dans ce cas précis, d'une erreur d'estimation de l'état de voisement.

La figure 10 reprend la version «synthétique» des spectres présentés en exemple à la figure 1 (rappelés ici pour faciliter la comparaison). Celle-ci montre la bonne reconstitution de leur structure.

Ces exemples ont été obtenus sans utiliser de préaccentuation avant analyse. Si l'on en fait usage, les résultats ne sont pas fondamentalement différents. On note favorablement la disparition de quelques composantes parasites basse fréquence. Par contre, une perte de précision dans l'évaluation des états de voisements se fait sentir. Sur le plan perceptuel, ces derniers essais sont un peu plus agréables. Ces variations sont vraisemblablement dues à des problèmes de conditionnement de l'analyse et il est trop tôt pour se prononcer définitivement sur la nécessité de faire usage de cette préaccentuation.

3.7. Évaluation de la charge de calcul

On peut tout d'abord remarquer la nécessité d'une tabulation intensive de grands ensembles de données : tables des spectres non voisés, fonctions transcendantes pour le calcul des FFTs et la reconstruction des spectres voisés. Les tables en question sont souvent redondantes et possèdent diverses symétries : on peut l'exploiter par des mécanismes d'adressage adaptés.

Moyennant cette tabulation, la charge de calcul nécessaire pour la synthèse est, pour une trame :

- N*k multiplications pour la génération des spectres voisés ;
- N multiplications pour la pondération par le spectre de l'impulsion glottique type ;
- N*(2l+1) multiplications pour la pondération en fonction des vecteurs «énergies» et «état de voisement» ;
- 3N multiplications pour le calcul des énergies ;
- N multiplications pour l'application du filtre de prédiction linéaire (dans le domaine fréquentiel) ;
- p*N multiplications pour le calcul de la fonction de transfert du filtre de prédiction linéaire si $p < N/2$, approximativement N*log(N) si p est supérieur (emploi de l'algorithme de FFT avec élimination des butterfly nulles) ;
- enfin 2N*log(N) multiplications pour la transformée de Fourier finale (algorithme avec renormalisation à chaque étage).

Dans ces évaluations, le signe multiplicatif correspond à une multiplication complexe suivie d'une addition ; k est le nombre d'impulsions glottiques intervenant dans la trame, l le nombre de bandes et p l'ordre de prédiction.

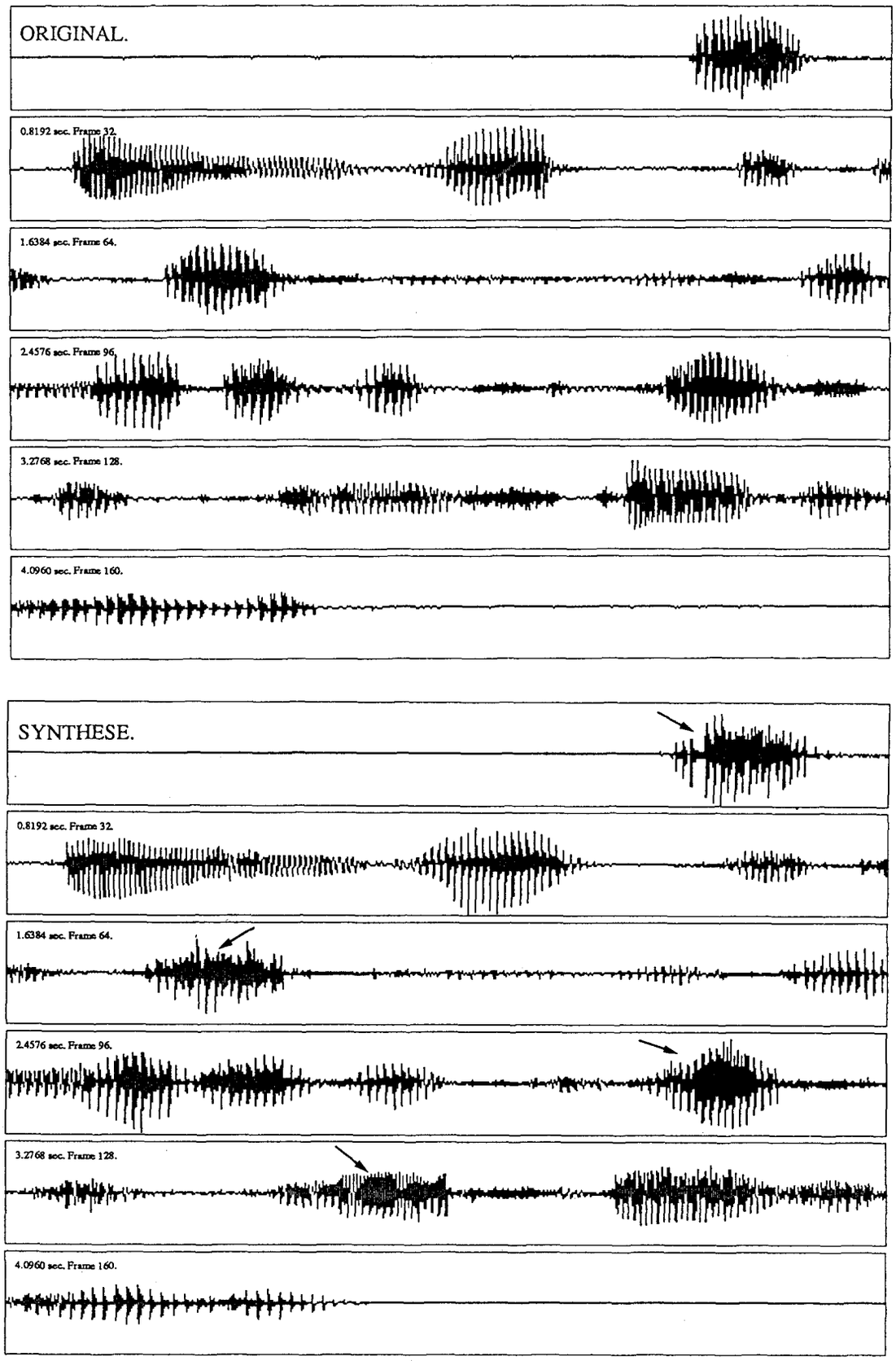


Fig. 8. – Comparaison des signaux temporels original et synthétique. Visualisation de l'enveloppe

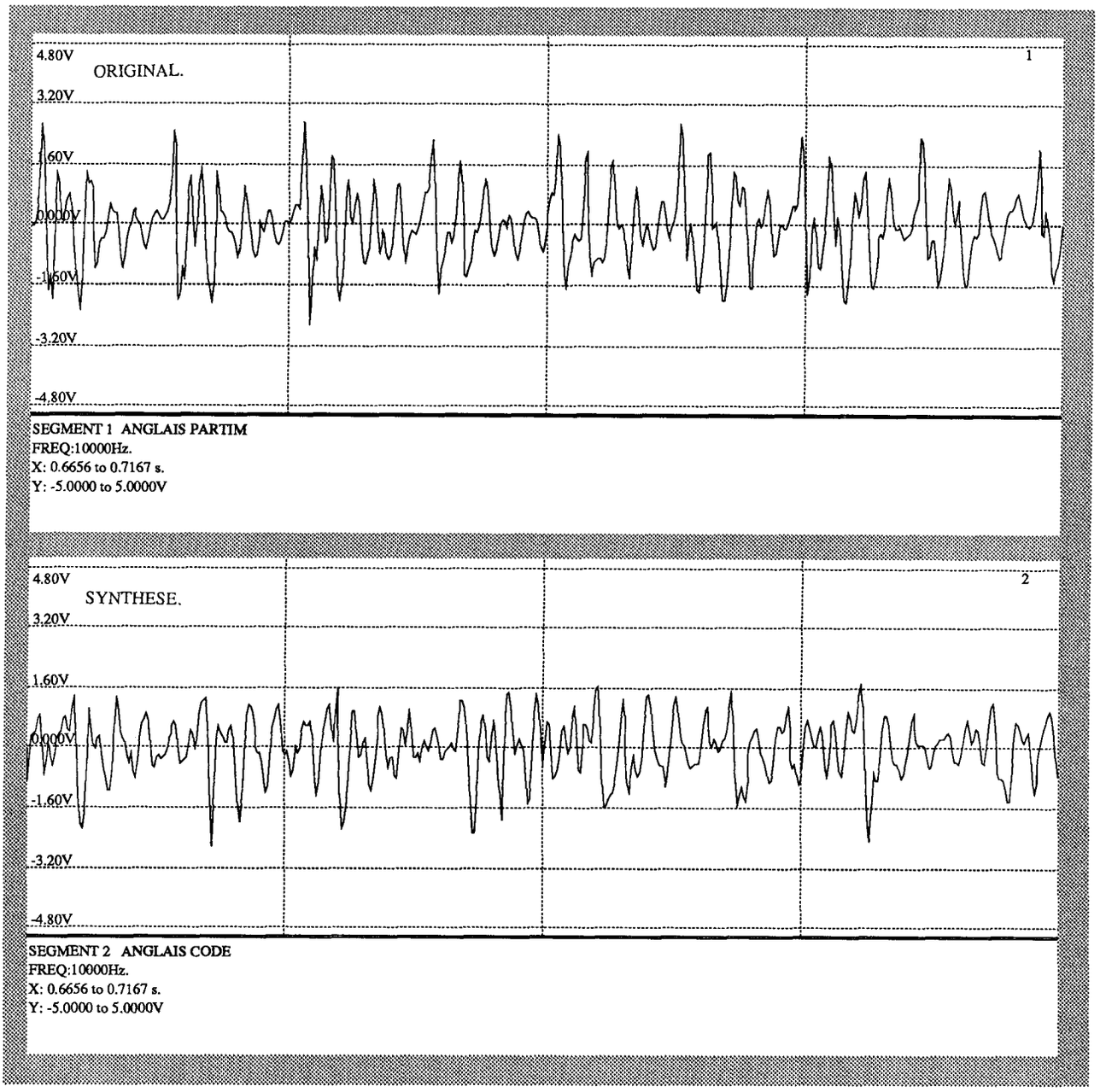


Fig. 9. – Comparaison des signaux temporels original et synthétique. Zoom sur une trame de 51.2 msec

Ceci nous donne un total de $(p+2l+k+2\log(N)+6)*N$ multiplications par trames, soit pour notre système expérimental ($k \leqslant 10$, $p=12$, $l=8$ et $N=1024$) 62464 multiplications. Il faut cependant noter que la plupart des nouveaux DSPs sont orientés vers les opérations complexes et qu'alors la charge de chaque multiplication chute à quelques instructions. La présence d'opérations très typées telles que la FFT ou la multiplication-addition de deux tableaux complexes nous fait penser qu'une solution raisonnable fera appel à un hardware spécialisé pour ces opérations.

Quelques simplifications du schéma original sont possibles afin de ramener la charge totaleà 40 Mips, voire moins si l'architecture du système est très spécifique. Des modules standards (5000$) de cette puissance apparaissent sur le marché. Il n'est donc pas utopique d'espérer réaliser un synthétiseur temps réel à moyen terme.

La charge du système d'analyse est comparable, mais pour notre application principale, c'est-à-dire la synthèse à partir du texte, l'exécution de l'algorithme d'analyse en temps réel n'a pas beaucoup d'importance. Les problèmes rencontrés sont très semblables à ceux de la synthèse. Remarquons cependant la nécessité de calculer un spectre logarithmique, ce qui peut se faire à partir de tables de logarithmes, après passage dans une représentation en virgule flottante.

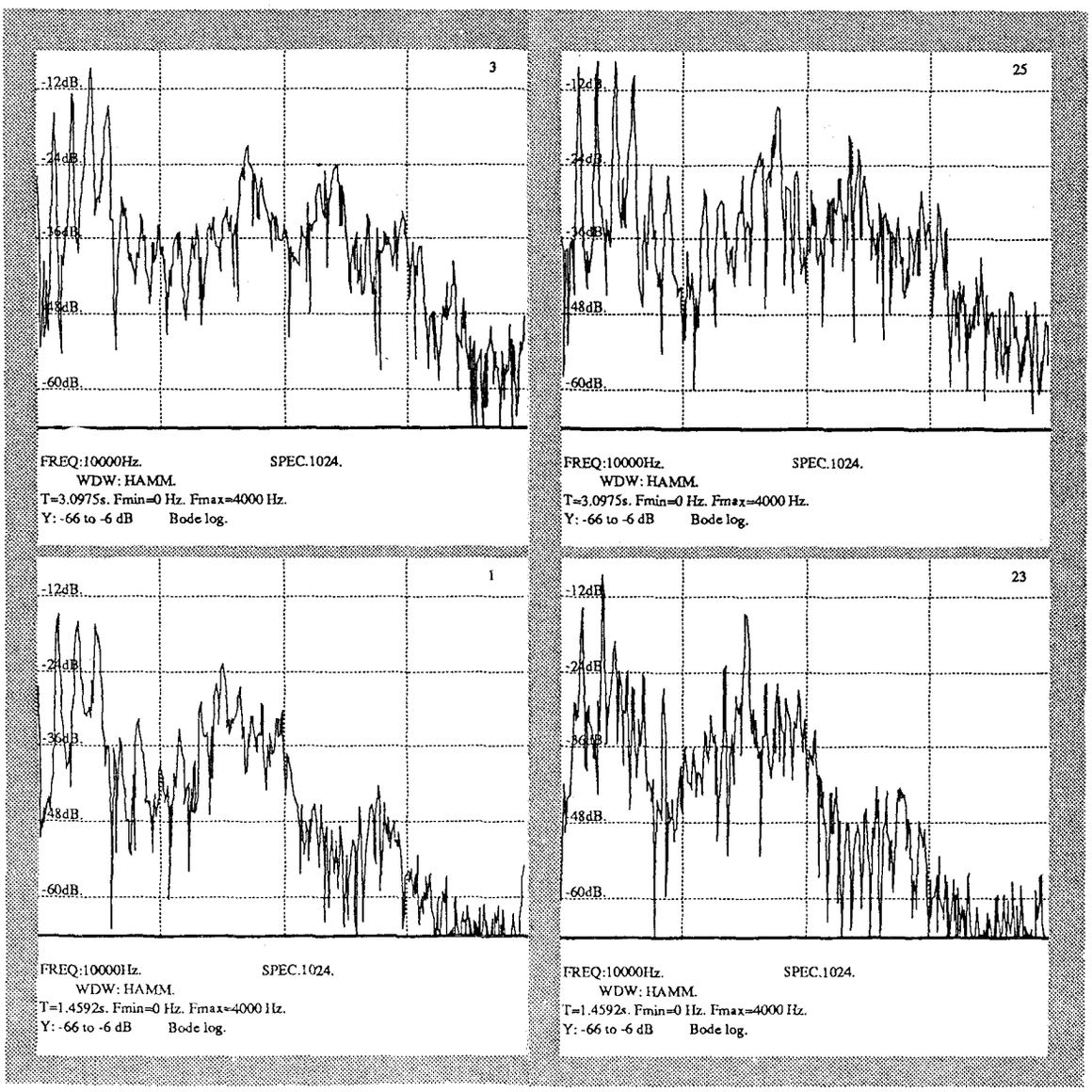


Fig. 10. – Comparaison des spectres originaux (à gauche) et synthétiques (à droite).

4. Conclusions

Nous avons développé un nouveau système de codage paramétrique faisant appel à une représentation fréquentielle du résidu de la prédiction linéaire. Ses principales particularités sont l'utilisation d'une «décision voisé-non voisé» progressive et variable en fonction de la fréquence ainsi que la reconstitution intégrale dans le domaine fréquentiel du signal résiduel par l'algorithme «overlapp-add».

La technique utilisée permet la génération d'une parole synthétique de bonne qualité, dont les avantages se distinguent assez nettement de ceux présentés par d'autres techniques visant les mêmes applications. Les résultats obtenus confirment l'intérêt de notre approche.

L'étude de l'influence perceptuelle de chaque paramètre, de son comportement au cours du temps et de la robustesse des méthodes d'analyse doit encore être approfondie. La plupart des défauts perceptuels constatés proviennent très certainement d'instabilités lors de l'analyse. Des procédures de certification et de stabilisation des valeurs obtenues doivent être élaborées.

Enfin rappelons que peu de schémas réellement paramétriques ont été proposés et que celui-ci nous paraît constituer une alternative intéressante.

Manuscrit reçu le 22 *août* 1988.

BIBLIOGRAPHIE

Ouvrages

[1] Rabiner L.R. & Schafer R.W. «Digital Processing of Speech Signal». Englewood Cliffs, N.J. : Prentice-Hall, 1978

[2] Shuzo Saito & Kazuo Nakata «Fundamentals of Speech Signal Processing». Academic Press, 1985.

[3] Hess W. «Pitch Determination of Speech Signal, Algorithms and Devices». Springer-Verlag, 1983.

Articles

[4] Makhoul J. «Spectral analysis of speech by linear prediction». IEEE Transactions AV-21, n°3, juin 1973.

[5] Makhoul J. «Linear Prediction : A Tutorial Review». Proceedings of the IEEE, vol 63, N°4, Avril 1975.

[6] Crochière R.E., Webber S.A., Flanagan J.L. «Digital Coding of Speech in Sub-bands.» Bell System Technical Journal, Octobre 1976.

[7] Crochière R.E. «On the Design of Sub-band Coders for low-bit-rate Speech Communication» - Bell System Technical Journal, mai 1977.

[8] Tribolet J.M. & Crochière R.E. «Frequency Domain Coding of Speech.», IEEE Transactions ASSP-27, octobre 1979.

[9] Griffin D. W., Lim J. S. «Signal Estimation From Modified Short-Time Fourier Transform», IEEE Transactions ASSP-32, avril 1984.

[10] Griffin D.W., Lim J. S. «A New Model-Based Speech Analysis/Synthesis System", IEEE ICASSP 1985 proceedings.

[11] Griffin D.W., Lim J. S. «A High Quality 9.6 kbps Speech Coding System», IEEE ICASSP 1986 proceedings.

[12] Hardwick J. C., Lim J. S. «A 4.8 KBPS Multi-Band Excitation Speech Coder», IEEE ICASSP 1988 proceedings.

[13] Charpentier F., Moulines E. «Text-To-Speech Algorithms based on FFT Synthesis.», IEEE ICASSP 1988 proceedings.

[14] Zwicker E., Feldtkeller R. «Psychoacoustique, l'oreille récepteur d'information.», traduit de l'allemand par C. Sorin, CNET ENST, Masson, Paris, 1981.

[15] Griffin D. W., Lim J. S. «Multiband Excitation Vocoder», IEEE Transactions ASSP-36, 1988.