J. Acoustique 2 (1989) 163-169

Classification
Physics Abstracts
43.72

Synthèse et analyse-synthèse par fonctions d'ondes formantiques (*)

Christophe d'Alessandro ([1,2]) et Xavier Rodet ([2,3])

([1]) LIMSI, CNRS, BP 30, F-91406 Orsay Cedex, France
([2]) LAFORIA Université Paris 6, 4 Place Jussieu, F-75005, Paris, France
([3]) IRCAM, 31 rue Saint-Merri, F-75004 Paris, France

(*Reçu le 5 octobre 1987, révisé le 19 mai 1988, accepté le 14 juin 1988*)

Résumé. — La synthèse par fonctions d'onde formantique ou FOF a été introduite voici quelques années pour simuler la réponse d'un banc de filtres disposés en parallèle à un train périodique d'impulsions : il est ainsi possible de synthétiser des voyelles (parlées ou chantées), certaines consonnes et toutes sortes de sons relevant du modèle impulsion/résonance (cloches, instruments à anche, etc.). Après avoir montré que l'on peut avec cette méthode synthétiser également de la parole non voisée, nous présentons ici une nouvelle méthode permettant l'extraction automatique de formes d'ondes (ou « grains » spectro-temporels) à partir du signal de parole. Les étapes principales du processus d'analyse/synthèse sont (trame par trame) : modélisation LPC de l'enveloppe spectrale, recherche des formants, définition d'un ensemble de filtres centrés sur les régions formantiques, filtrage du signal dans les régions formantiques (par analyse/synthèse de Fourier à court terme), détection et modélisation des FOF. Cette méthode, qui conserve une bonne précision d'analyse tant du point de vue temporel que du point de vue fréquentiel, permet la manipulation de paramètres perceptivement pertinents.

Abstract. — For many years formant-wave-function synthesis has successfully been used in musical research. We present results in speech synthesis using formant-wave-functions to implement a parallel formant synthesizer. A new method for representing the speech signal as a set of formant-wave-functions is then described. The main steps of the Analysis/synthesis process are, for each frame, LPC modeling, formant tracking, filter bank definition using formant parameters, filtering in formant regions with overlap-add short-time Fourier analysis/synthesis, detection of elementary waveforms and modelisation using formant-wave-functions. This method allows to control both spectral and temporal resolution with manipulation of perceptually relevant parameters.

Introduction.

La synthèse par Fonctions d'Onde Formantique (ou FOF) est utilisée avec succès depuis plusieurs années, en particulier pour la recherche musicale (synthèse de voix chantée, de timbres instrumentaux ou imaginaires) [1]. Il s'agit de reconstruire le signal dans le domaine temporel à l'aide de fonctions élémentaires, les FOF, ayant des caractéristiques spectrales intéressantes. Les FOF ont été principalement utilisées pour simuler un synthétiseur à formants en parallèle, de façon économique en puissance de calcul. De plus, de par les fonctions utilisées, il est dans certains cas possible d'obtenir un contrôle plus fin de l'enveloppe spectrale que celui autorisé en synthèse à formants en parallèle classique. Jusqu'à présent la synthèse par FOF n'a été utilisée, en synthèse de parole, qu'avec une analyse conjointe de la fréquence de voisement et ne permettait que de générer les sons voisés : voyelles, semi-voyelles et certaines consonnes. Néanmoins il semble possible en conservant les mêmes fonctions temporelles de synthétiser également d'autres types de sons, en s'attachant à simuler non plus une source quasi périodique d'excitation, mais une source bruitée (sur toute l'étendue du spectre ou seulement dans certaines régions), une source « mixte » (bruit pulsé) ou des impulsions isolées. On peut ainsi espérer synthétiser n'importe quel segment de parole. Bien entendu il paraît nécessaire d'utiliser des procédures automatiques d'estimation des paramètres de synthèse ; notons que la plupart de ces paramètres ne sont pas spécifiques à notre méthode, mais communs à tous les systèmes de synthèse à formants.

La nature même de la méthode de synthèse présente intrinsèquement une localisation spectro-temporelle : en effet les fonctions élémentaires, calculées dans le domaine temporel, sont porteuses d'une information fréquentielle localisée dans la région d'un formant.

La nécessité d'affiner l'analyse du signal de parole en utilisant des échelles spectro-temporelles différentes —

(*) Texte issu d'une conférence présentée aux 16[es] Journées d'Etudes sur la Parole (JEP), 5-9 octobre 1987, Hammamet (Tunisie).

c'est-à-dire en utilisant des échelles de temps différentes à des fréquences différentes — apparaît dans les travaux récents de plusieurs auteurs. Citons en particulier Liénard avec l'analyse à très court terme [2], qui s'attache à décomposer le signal de parole en fonctions d'ondes élémentaires, à l'aide d'un banc de filtres dont on analyse ensuite, voie par voie le comportement temporel. D'autre part une nouvelle méthode, l'analyse en ondelettes, initialement développée par Morlet [3], fournit un cadre mathématique rigoureux pour l'analyse d'un signal sur une base de fonctions élémentaires dépendantes d'un facteur d'échelle spectro-temporelle. Il semble possible de remplacer l'analyse de Fourier par ce type d'analyse, de façon très avantageuse dans certains cas quant à la précision fréquentielle et temporelle.

Après un rappel de la définition d'une FOF et un exposé des résultats ainsi obtenus en synthèse de parole (tant voisée que non voisée), nous présenterons une méthode d'analyse-synthèse permettant la représentation du signal vocal par un ensemble de FOF.

1. Définition des FOF.

Une FOF est une fonction temporelle dont le spectre possède un maximum qui sera dénommé « formant » par analogie avec les maxima de la fonction de transfert du conduit vocal. Les FOF peuvent être calculées à l'aide d'une formule mathématique explicite ou bien être extraites de parole réelle [4]. La figure 1 présente un exemple de FOF communément utilisée : la réponse d'un résonateur du second ordre à une impulsion (approximativement en forme d'arche). On peut ainsi définir différentes FOF qui possèdent comme expression analytique le produit d'une sinusoïde et d'une fonction d'enveloppe temporelle (exponentielle décroissante, fenêtre d'analyse spectrale [5]...) dont la forme induit les propriétés d'enveloppe spectrale. La FOF de la figure 1 se déduit de la formule :

$$s(t) = A \text{ env } (t)\, p(t)$$

avec

$$p(t) = \sin (2 \pi f_c + \Phi)$$

et pour $0 \leqslant t \leqslant t_e$;

$$\text{env } (t) = 1/2\,(1 - \cos (\pi t/t_e)) \exp(-\alpha t)$$

pour $t \geqslant t_e$:

$$\text{env } (t) = \exp(-\alpha t)\,.$$

f_c : permet de fixer la fréquence centrale, ou fréquence du maximum spectral.

π : est la phase initiale de la FOF.

Les paramètres suivants concernent l'enveloppe temporelle de la FOF, dont la transformée de Fourier donnera l'enveloppe spectrale.

α : contrôle la largeur de bande spectrale à − 6 dB du sommet.

t_c : règle le temps d'excitation (temps de montée de l'enveloppe temporelle) et la largeur des « jupes » spectrales, de façon indépendante de ALPHA, ce qui est remarquable.

A : donne l'amplitude de la FOF et donc règle l'amplitude spectrale.

L'utilisation de FOF présente plusieurs avantages :

— les paramètres utilisés sont perceptivement pertinents, ce qui explique le succès de la méthode en synthèse musicale ;

— on peut obtenir une grande précision temporelle et simultanément une grande définition spectrale ;

— la méthode se prête bien à une mise en œuvre sur de petites machines (un synthétiseur FOF temps réel a été développé sur microprocesseur TMS32010) [6].

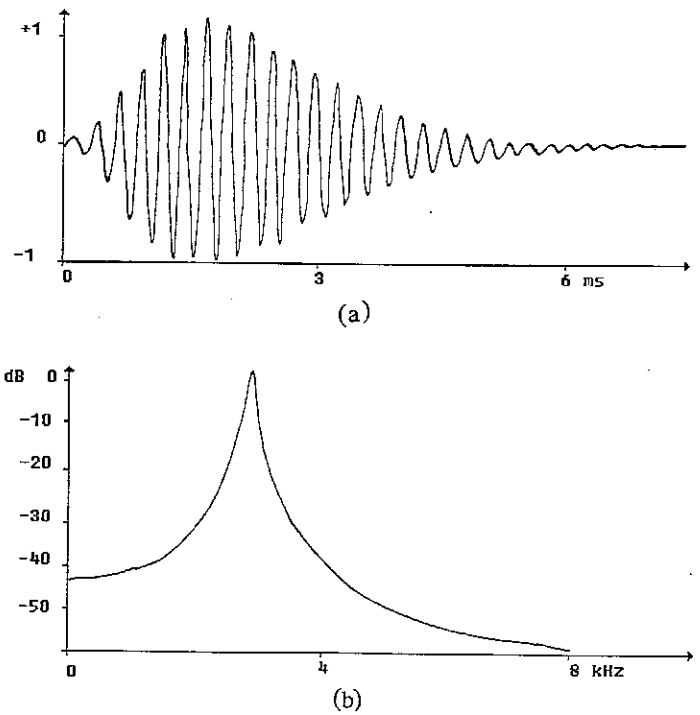

Fig. 1. — (a) Exemple de FOF : le signal temporel. (b) Exemple de FOF : le spectre d'amplitude logarithmique.

[Temporal signal of a Formant-Wave-Function. Logarithmic magnitude spectrum of a Formant-Wave-Function.]

2. Synthèse.

2.1 PAROLE VOISÉE. — La production de diverses sortes de signaux (de parole ou d'instruments musicaux par exemple) peut être représentée sous la forme d'une fonction d'excitation et d'un filtre. Ainsi les FOF ont été d'abord utilisées pour simuler un synthétiseur à formants en parallèle excité par un train d'impulsions quasi périodiques (synthèse de la partie vocalique de la parole, de voix chantée...) [7].

Bien qu'il soit envisageable de définir « manuellement » les paramètres de synthèse pour certaines applications (synthèse musicale, ou un ensemble fixé de voyelles) l'effort s'est naturellement porté vers leur estimation

automatique. Pour le synthétiseur de la figure 2 il s'agit de :

— la fréquence fondamentale de voisement ;
— les paramètres formantiques (f_c, α, t_e, A).

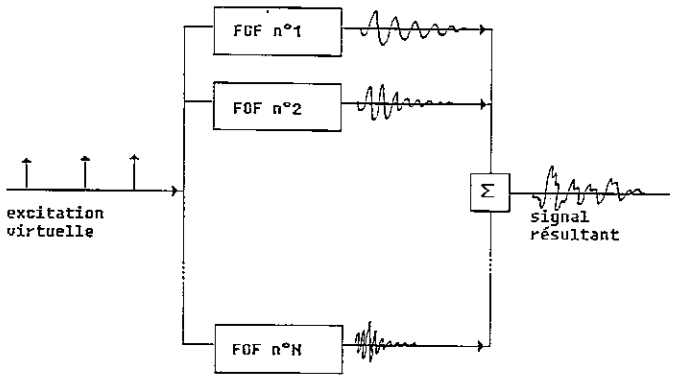

Fig. 2. — Structure d'un synthétiseur FOF parallèle.

[Structure of a Formant-Wave-Function Parallel Synthesiser.]

Pour peu que les paramètres fournis au synthétiseur soit correctement estimés, des tests informels semblent faire apparaître une qualité de synthèse (pour la parole voisée) comparable à un bon codage classique par prédiction linéaire, en conservant toutefois plein accès aux paramètres acoustiques explicites du signal synthétisé. Dans l'exemple suivant (Figs. 3 et 4) nous avons utilisé le système de détection de formants développé à l'I.R.C.A.M. [8] (joint à une détection du Fondamental).

Ce système permet la modélisation, trame par trame, de l'enveloppe pectrale du signal par prédiction linéaire (algorithme du filtre en treillis adaptatif) de façon synchrone au Fondamental. La définition temporelle est donc la même pour toute la gamme des fréquences ce qui entraîne un défaut de sonorité que l'on constate également en synthèse par prédiction linéaire classique. Ce type d'implémentation de la synthèse par FOF présente par contre l'avantage d'être simple et de demander peu de puissance de calcul pour la précision obtenue.

2.2 SYNTHÈSE DE BRUIT. — Il est également possible de simuler la sortie d'un filtre excité par un signal aléatoire, et non plus par un train quasi périodique d'impulsions, en générant des FOF de façon aléatoire — de l'ordre d'une FOF par milliseconde en moyenne. On peut ainsi simuler un bruit de friction afin de synthétiser des fricatives (Fig. 5). Ici il paraît indispensable de générer les FOF correspondant à chaque formant de façon totalement asynchrone : on utilise alors la propriété de localisation spectro-temporelle et d'indépendance des différentes FOF.

Pour obtenir une synthèse de bonne qualité il semble nécessaire de dépasser l'opposition binaire voisé/non voisé, et de définir plusieurs degrés de voisement. Ceci se vérifie tant pour les segments de parole qui relèvent directement d'un modèle de production « mixte » (un bruit fricatif se superposant aux vibrations quasi périodiques des cordes vocales) que pour les segments de parole qui paraissent « bien voisés ». Tous les degrés de mélange entre signaux quasi périodiques et signaux aléatoires coexistent dans la parole naturelle, de la voix chuchotée à la voix théâtrale. D'autre part, on constate chez certains locuteurs plusieurs excitations du conduit vocal au sein du même cycle vocalique par exemple à l'ouverture et à la fermeture des cordes vocales. Le succès du

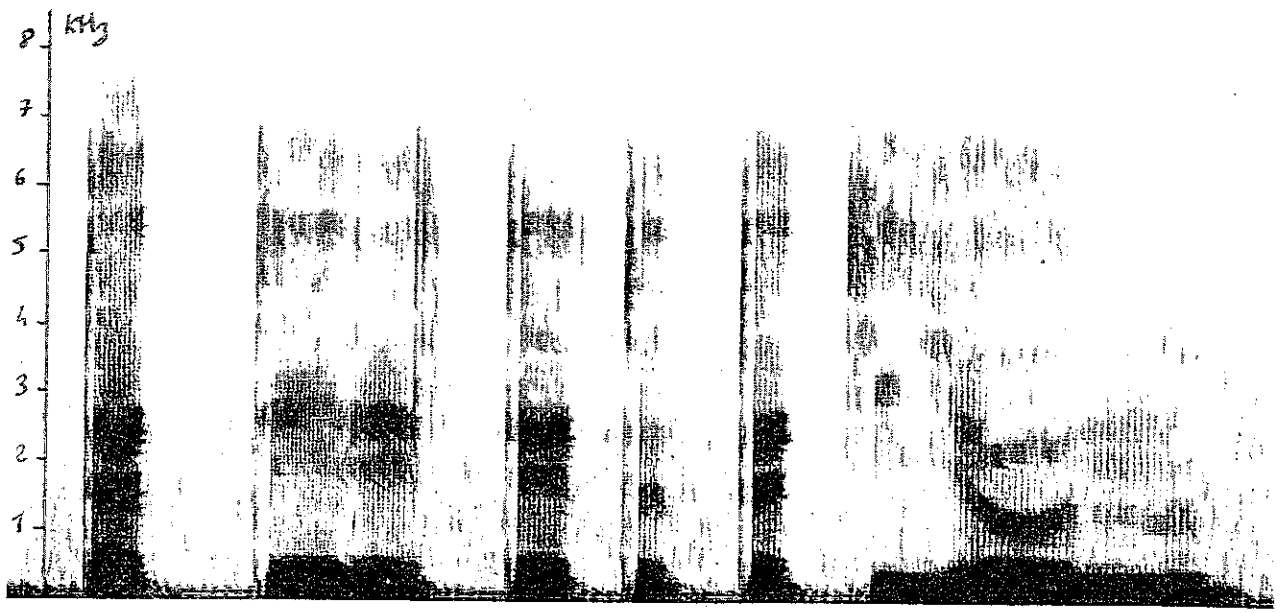

Fig. 3. — Sonagramme de la phrase « Tâter les têtes tatillonnes » parole naturelle.

[Sonagram of the French sentence « Tâter les têtes tatillonnes » natural speech.]

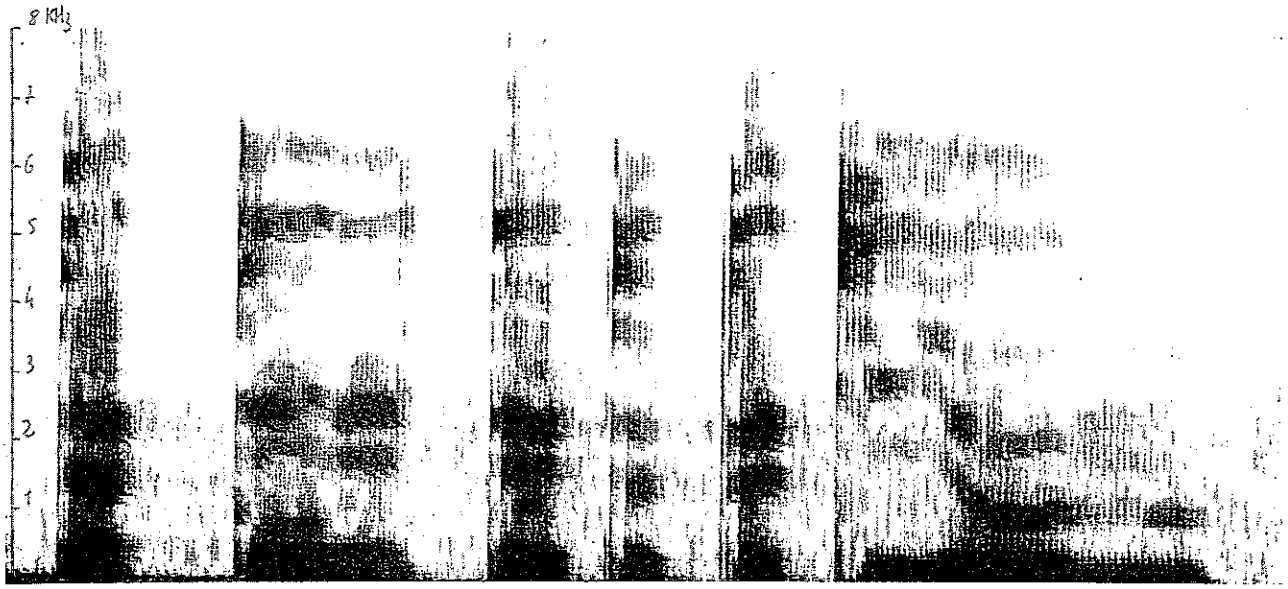

Fig. 4. — Idem : synthèse par FOF.

[Idem Formant-Wave-Function synthesis.]

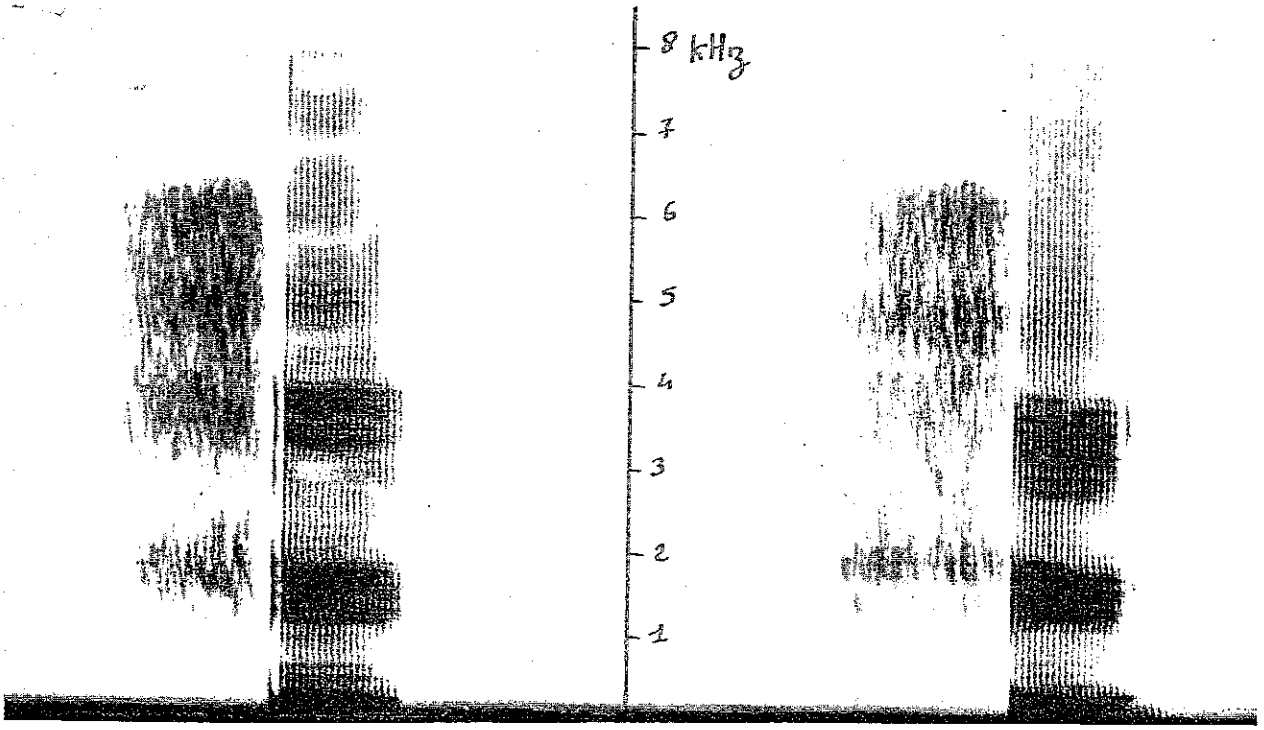

Fig. 5. — Sonagramme de la syllabe /fy/ naturelle puis synthétique (estimation manuelle des paramètres à partir du son naturel).

[Sonagram of /fy/, natural and synthetic (manual estimation of parameters).]

codage par prédiction linéaire multi-impulsionnelle [9] est révélateur à cet égard.

3. Analyse-synthèse.

Nous ne cherchons plus ici à modéliser le signal comme étant issu du filtrage d'une certaine source d'excitation, voisée, bruitée, ou « mixte », mais comme le propose Liénard [11], à le représenter par un ensemble de fonctions d'ondes élémentaires. Il faut donc rechercher ces éléments dans le signal de parole original, par segmentation spectrale et temporelle. On est ainsi conduit à élaborer un système d'analyse-synthèse automatique, valable pour tous les segments de parole, fondé

sur une représentation « granulaire » dont on peut espérer que les paramètres sont perceptivement significatifs (paramètres formantiques, enchaînements temporels des formes d'ondes). Chaque grain ou fonction d'onde, correspond à une concentration d'énergie dans une région spectro-temporelle précise, et peut être modélisé comme la réponse d'un certain filtre à une certaine excitation. Le procédé d'analyse repose sur les hypothèses suivantes :

— on peut segmenter le signal spectralement, ce qui est vérifié dans la plupart des cas puisque le signal offre un spectre comportant des formants. Néanmoins pour certaines parties du signal la segmentation spectrale peut apparaître artificielle ;

— on peut segmenter le signal temporellement : si les filtres de segmentation spectrale sont suffisamment larges, pour englober un formant, on doit retrouver des FOF dans le signal temporel. Remarquons que cela n'est pas toujours aisé, en particulier en basse fréquence ;

— les paramètres formantiques varient relativement lentement, c'est-à-dire que l'on peut les considérer comme constants pendant la durée d'une FOF ;

— la forme précise de l'enveloppe des FOF n'est pas d'une importance extrême, pour pouvoir les modéliser de façon systématique par des fonctions mathématiques simples.

L'analyse est faite selon les étapes suivantes [14].

1. *Modélisation du signal par prédiction linéaire.* Nous utilisons comme précédemment un filtre d'analyse en treillis [12], pour obtenir une estimation (non synchrone au Fondamental) de l'enveloppe spectrale. La fenêtre temporelle d'analyse est choisie assez longue pour englober une période de voisement, et le pas temporel d'analyse assez bref pour ne pas perdre les évolutions spectrales rapides.

2. *Détection des maxima spectraux, sur chaque fenêtre d'analyse,* par suivi de la courbe spectrale. On acquiert ainsi la fréquence centrale, l'amplitude et la largeur de bande de chaque formant.

3. *Définition pour chaque fenêtre d'un jeu de filtres passe-bande,* dont le gain est centré sur les fréquences centrales des formants. La méthode de filtrage que nous avons choisie, par analyse/synthèse de Fourier, autorise la définition de filtres de gains quelconques, à la fenêtre d'analyse près. Plusieurs sortes de filtres ont été mises en œuvre : rectangulaires, en arches de sinusoïdes... La somme des gains de ce jeu de filtres est partout égale à l'unité. Le but de ce filtrage est de « découper » en quelque sorte le signal correspondant à un formant.

4. *Filtrage dans chaque région spectrale* définie par le jeu de filtres précédant, et dans la région temporelle définie par l'instant d'analyse, avec une marge temporelle suffisante autour de cet instant pour la détection qui va suivre. Le filtrage est effectué par analyse/synthèse de Fourier à court terme [13], ce qui préserve les caractéristiques de phase des signaux.

5. *Détection des FOF dans chaque région spectro-temporelle,* par corrélation ou par filtrage inverse ce qui donne un signal local d'excitation. Cette détection temporelle peut également se réaliser par simple calcul de l'enveloppe dans chaque bande d'analyse et découpage temporel des formes d'ondes, qui sont ainsi centrées sur les maxima.

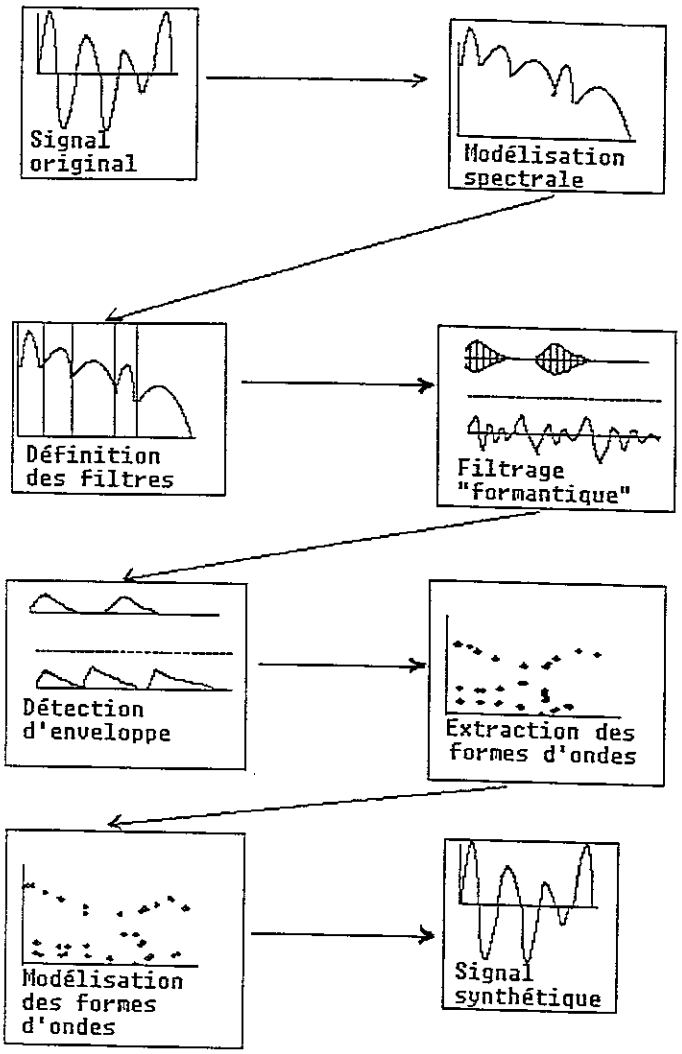

Fig. 6. — Processus d'analyse-synthèse.

[Analysis-synthesis process.]

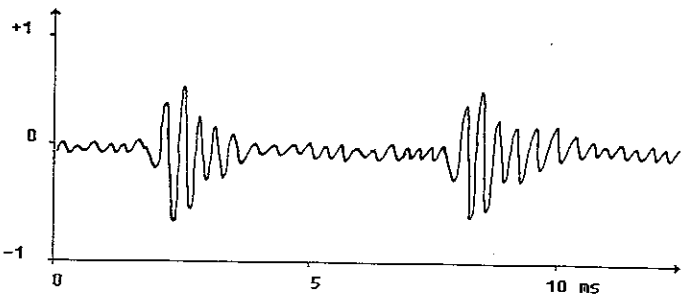

Fig. 7. — Signal issu d'un filtrage dans la région formantique.

[Speech signal filtered in formantic region.]

6. Jusqu'ici le système est rigoureusement additif, c'est-à-dire que l'on reconstitue le signal original en sommant les diverses FOF « naturelles » détectées. Nous

pouvons aussi reconstituer un signal synthétique grâce à une modélisation des FOF par une famille de fonctions mathématiques simples, ce qui autorise à nouveau la manipulation des caractéristiques acoustiques du signal. Des analyses-synthèses ont été menées pour diverses voix masculines et féminines. La partie basse du spectre (dans la région du premier formant, ou en dessous) présente des difficultés de modélisation et de détection qui altèrent dans certains cas la qualité de la synthèse (les deux ou trois premiers harmoniques semblent alors mal reconstitués). Néanmoins la qualité de synthèse est satisfaisante : des tests d'écoute informels montrent que le signal synthétique est perceptivement très proche de l'original.

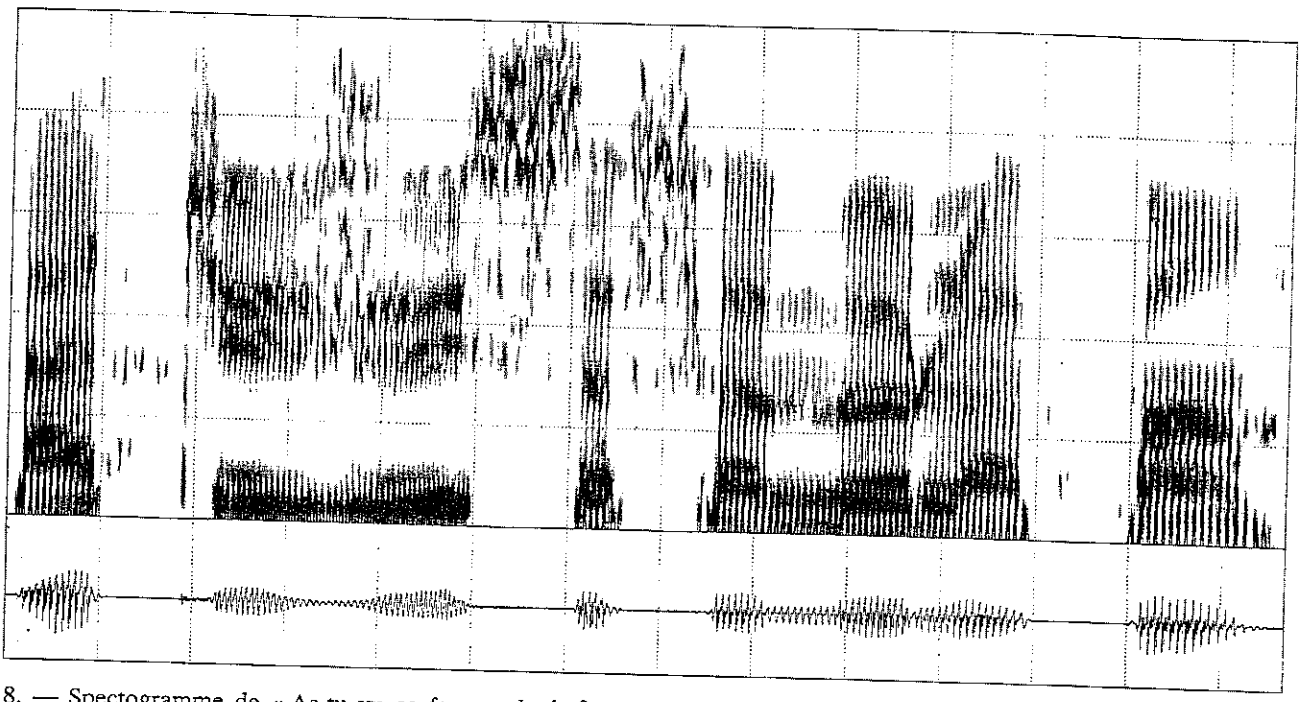

Fig. 8. — Spectogramme de « As-tu vu ce fameux lapin ? » parole naturelle.

[Spectrogram of the French sentence « As-tu vu ce fameux lapin ? » natural speech.]

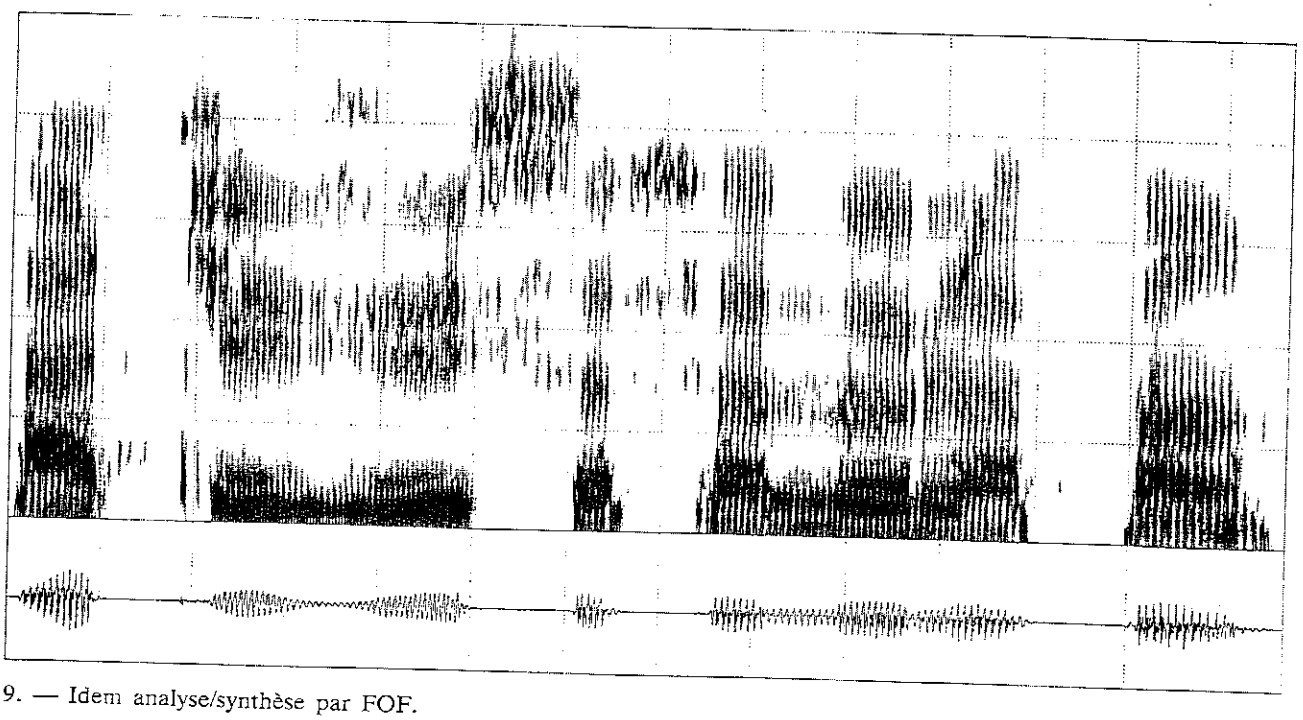

Fig. 9. — Idem analyse/synthèse par FOF.

[Idem Formant-Wave-Function analysis/synthesis.]

Conclusion.

Notre système de décomposition d'un signal de parole en fonctions d'ondes élémentaires utilise la connaissance *a priori* des régions de maximum spectral ; nous exploitons ainsi cette particularité du signal de parole, qui est de posséder un spectre « à formants ».

Un tel système permet de garder automatiquement une grande précision fréquentielle, due à la qualité de l'analyse par prédiction linéaire, et de ne pas altérer les évolutions temporelles rapides, par le filtrage et la détection qui suivent.

Notre méthode offre à la fois un mode d'analyse-synthèse automatique et un mode de synthèse à partir de paramètres acoustiques (formants, Fondamental...). Elle est potentiellement puissante pour la manipulation du signal vocal (tests psycho-acoustiques), son analyse en éléments de divers niveaux (« grain », formants, périodes de voisement, explosions...) et sa synthèse à partir de paramètres perceptivement et acoustiquement pertinents. Néanmoins il paraît nécessaire de l'améliorer, en particulier pour la modélisation de la partie grave du spectre, et de créer les outils qui permettent d'exploiter ses potentialités. Nous souhaitons pouvoir gérer des paramètres (parfois en grande quantité) à des niveaux d'observation différents. L'utilisation de techniques conceptuellement proches, comme l'analyse en ondelettes, s'inscrit également dans notre perspective.

Bibliographie

[1] RODET X., POTARD Y. et BARRIÈRE J. B., The Chant project : from the synthesis of the singing voice to synthesis in general, *Comput. Music J.* 8 (1980) N° 3.

[2] LIÉNARD J. S., Analyse à très court terme de la parole : un outil et quelques directions de recherche, *15es JEP Paris* (1985).

[3] KRONLAND-MARTINET R., GROSSMANN A. et MORLET J., Analysis of Sound Patterns Throught Wavelet Transforms, The International Journal of Pattern Recognition and Artificial Intelligence (special issue on expert systems and pattern analysis) 1987.

[4] RODET X., DELATRE J. L. et RAZZAM M., Construction du signal vocal dans le domaine temporel, 10es JEP Grenoble (1979).

[5] HARRIS F., On the use of windows for harmonic analysis with the discrete Fourier transform. *Proc. IEEE* **66** (1978) N° 1.

[6] DÉCHELLE F. et D'ALESSANDRO C., Rapports de DEA TAI. Université Paris 6 (1984).

[7] RODET X., Time Domain Formant-Wave-Function Synthesis. In Spoken Language Generation and Understanding, Ed. J. C. Simon (D. Reidel publishing Compagny, Dordrecht Holland) 1980.

[8] RODET X. et DEPALLE P., Synthesis by rule : LPC diphones and calculation of formant trajectories, *IEEE ICASSP* (1985).

[9] ATAL B. S. et REMDE J. R., A New Model of LPC Excitation for Producing Natural-Sounding Speech at Low Bitrates, *IEEE ICASSP* (1982).

[10] RODET X., DEPALLE P. et POIROT G., Analyse et synthèse de la voix parlée et chantée par modélisation de l'enveloppe spectrale et de l'excitation, 16es JEP Hammamet (1987).

[11] LIÉNARD J. S., Speech analysis and reconstruction using short-time, elementary waveforme *IEEE ICASSP* (1987).

[12] MAKHOUL J. et COSELL L., Adaptative Lattice Analysis of Speech. *IEEE Trans. Circuits Syst.* **28-6** (1981).

[13] CROCHIERE R. E., A Weighted Overlap-Add Method of Short-Time Fourier Analysis/Synthesis. *IEEE Trans. ASSP* **28-1** (1979).

[14] D'ALESSANDRO C. et LIÉNARD J. S., Decomposition of the Speech Signal into Short-Time Waveforms Using Spectral Segmentation. *IEEE ICASSP* (1988).