Représentation du signal de parole par une somme de fonctions élémentaires

Christophe d'Alessandro

THESE de DOCTORAT de L'UNIVERSITÉ PARIS 6

Spécialité:

INFORMATIQUE

présentée

par *Monsieur* **Christophe d'Alessandro**

pour obtenir le titre de DOCTEUR DE L'UNIVERSITÉ PARIS 6

Sujet de la thèse:

Représentation du signal de parole par une somme de fonctions élémentaires

soutenue le **17 Avril 1989**

devant le jury composé de:

Monsieur	*Xavier*	Rodet	*président*
Monsieur	*Jean-Sylvain*	Liénard	*rapporteur*
Monsieur	*Patrick*	Flandrin	*rapporteur*
Monsieur	*Jean-Claude*	Risset	*examinateur*
Monsieur	*Jean-Luc*	Schwartz	*examinateur*

Représentation du signal de parole par une somme de fonctions élémentaires

Thèse soutenue par C. d'Alessandro, le 17 Avril 1989.

Résumé

Parmi les méthodes de représentation du signal, pour le traitement automatique de la parole, la représentation par un ensemble discret d'événements spectro-temporels localisés ou *formes d'ondes élémentaires* offre des perspectives prometteuses.

Une approche triple préside à l'étude menée dans ce mémoire. D'abord, les conditions et les algorithmes pour une reconstruction exacte du signal depuis sa représentation se déduisent de l'interprétation de méthodes non-paramétriques classiques, transformée de Fourier à court terme et transformée en ondelettes, en terme de formes d'ondes élémentaires dans le domaine temporel. Les relations entre l'analyse du signal par des modèles fonctionnels du système auditif périphérique et l'analyse granulaire, mode particulier de représentation en formes d'ondes élémentaires sont ensuite étudiées. Enfin, une représentation en formes d'ondes élémentaires basée sur un modèle de production du signal de parole est détaillée, ainsi qu'un procédé d'analyse/synthèse automatique.

Des applications à l'analyse de la parole, qui s'appuient sur ce nouveau mode de représentation, sont envisagées. Pour la synthèse de parole, principale motivation de ce travail, une structure originale de synthèse et un procédé d'obtention automatique des paramètres sont proposés et discutés.

Mots clef: Parole, analyse, synthèse, représentations temps-fréquence, modèles auditifs

Abstract

A new method for speech signal representation, based on elementary waveforms, is proposed and its application in the field of automatic speech processing is discussed.

The decomposition of the speech signal into a set of well-localized time-frequency discrete elements was studied from three viewpoints. First, formal considerations and algorithms are derived by an interpretation of classical non-parametric methods (short-time Fourier and wavelet transform) as a time-domain elementary waveform representation. Second, the relationship between a particular mode of elementary waveform representation, the granular analysis, and an analysis of the acoustic signal produced by models of the peripheral auditory system is explored. Finally, a model-based elementary waveform speech representation, is presented and used in an automatic analysis-synthesis system.

Some applications using this new representation for speech analysis are proposed. Speech synthesis was the first aim of this work, and an original structure of synthesis, including an automatic system for parameters estimation, is described and discussed.

Keywords: Speech analysis, synthesis, time-frequency representation, auditory modelling

THESE de DOCTORAT de L'UNIVERSITÉ PARIS 6

Spécialité:

INFORMATIQUE

présentée

par *Monsieur* **Christophe d'Alessandro**

pour obtenir le titre de DOCTEUR DE L'UNIVERSITÉ PARIS 6

Sujet de la thèse:

Représentation du signal de parole par une somme de fonctions élémentaires

soutenue le **17 Avril 1989**

devant le jury composé de:

Monsieur	*Xavier*	Rodet	*président*
Monsieur	*Jean-Sylvain*	Liénard	*rapporteur*
Monsieur	*Patrick*	Flandrin	*rapporteur*
Monsieur	*Jean-Claude*	Risset	*examinateur*
Monsieur	*Jean-Luc*	Schwartz	*examinateur*

dédicacé: à mes parents, Gisèle et Jean.

Représentation du signal de parole par une somme de fonctions élémentaires

Thèse soutenue par C. d'Alessandro, le 17 Avril 1989.

Résumé

Parmi les méthodes de représentation du signal, pour le traitement automatique de la parole, la représentation par un ensemble discret d'événements spectro-temporels localisés ou *formes d'ondes élémentaires* offre des perspectives prometteuses.

Une approche triple préside à l'étude menée dans ce mémoire. D'abord, les conditions et les algorithmes pour une reconstruction exacte du signal depuis sa représentation se déduisent de l'interprétation de méthodes non-paramétriques classiques, transformée de Fourier à court terme et transformée en ondelettes, en terme de formes d'ondes élémentaires dans le domaine temporel. Les relations entre l'analyse du signal par des modèles fonctionnels du système auditif périphérique et l'analyse granulaire, mode particulier de représentation en formes d'ondes élémentaires sont ensuite étudiées. Enfin, une représentation en formes d'ondes élémentaires basée sur un modèle de production du signal de parole est détaillée, ainsi qu'un procédé d'analyse/synthèse automatique.

Des applications à l'analyse de la parole, qui s'appuient sur ce nouveau mode de représentation, sont envisagées. Pour la synthèse de parole, principale motivation de ce travail, une structure originale de synthèse et un procédé d'obtention automatique des paramètres sont proposés et discutés.

Mots clef: Parole, analyse, synthèse, représentations temps-fréquence, modèles auditifs

Abstract

A new method for speech signal representation, based on elementary waveforms, is proposed and its application in the field of automatic speech processing is discussed.

The decomposition of the speech signal into a set of well-localized time-frequency discrete elements was studied from three viewpoints. First, formal considerations and algorithms are derived by an interpretation of classical non-parametric methods (short-time Fourier and wavelet transform) as a time-domain elementary waveform representation. Second, the relationship between a particular mode of elementary waveform representation, the granular analysis, and an analysis of the acoustic signal produced by models of the peripheral auditory system is explored. Finally, a model-based elementary waveform speech representation, is presented and used in an automatic analysis-synthesis system.

Some applications using this new representation for speech analysis are proposed. Speech synthesis was the first aim of this work, and an original structure of synthesis, including an automatic system for parameters estimation, is described and discussed.

Keywords: Speech analysis, synthesis, time-frequency representation, auditory modelling

Table des matières

REMERCIEMENTS

Il m'est un agréable devoir de commencer ce mémoire en témoignant ma gratitude à ceux qui en ont rendu possible l'élaboration et l'achévement.

Je voudrais tout spécialement remercier mon jury:

Monsieur Xavier Rodet, mon directeur de thèse;

Monsieur Jean-Sylvain Liénard;

Monsieur Patrick Flandrin;

Monsieur Jean-Claude Risset;

Monsieur Jean-luc Schwartz;

J'aimerais exprimer ma reconnaissance amicale envers Mademoiselle Michèle Castellengo et le Laboratoire d'Acoustique Musicale de l'université Paris VI.

Je tiens également à remercier l'équipe parole du LAFORIA de l'université Paris VI, et le groupe Chant/Formes de l'IRCAM.

Je dois à Messieurs Jean-Sylvain Liénard et Joseph Mariani de m'avoir accueilli au sein de leur laboratoire: que tous les membres du LIMSI trouvent ici l'expression de ma reconnaissance.

Sans le soutien de ma femme, Nathalie d'Alessandro, ce travail n'aurait pu aboutir.

INTRODUCTION

L'étude et le traitement de la parole à l'aide d'une machine emprunte un passage obligé par la représentation du signal. Les possibilités ultérieures de traitement ou d'analyse de la réalité acoustique reposent sur les propriétés de la représentation initiale choisie.

De nombreuses méthodes, basées sur des analyses spectro-temporelles ou dérivant de modèles physiques ou mathématiques de production ou de perception de la parole, ont été décrites et utilisées, avec leur cortège d'avantages et de contraintes. Toutes ces méthodes doivent composer avec les mêmes types de problèmes, et chacune propose d'y répondre à sa façon: compromis entre résolution temporelle et résolution fréquentielle, relation avec la perception, adéquation à un modèle de production, existence et complexité des algorithmes.

Quelques unes de ces méthodes s'attachent explicitement à la représentation du signal de parole comme une somme de fonctions mathématiques assez simples, bien localisées dans le plan spectro-temporel, et pertinentes du point de vue de la production et/ou de la perception de la parole: elles forment l'objet de ce mémoire. De telles fonctions seront regroupées sous le terme générique de *formes d'ondes élémentaires.*

Afin d'illustrer ce propos, trois représentations dans le plan temps-fréquence d'un même signal de parole sont portées sur la figure .1. La première représentation est utilisée, de façon satisfaisante, comme paramétrisation acoustique dans des systèmes de reconnaissance automatique: pour chaque trame de 12.5 ms, le logarithme de l'énergie issue d'un banc de 16 filtres répartis sur une échelle de Bark, est figuré par la noirceur du tracé. Les événements spectro-temporels utilisés occupent approximativement un pavé de $12.5 \times 1.ms.Bark$ dans le plan, et sont au nombre de $16 \times 80 = 1280$ par seconde. Ces événements ne permettent pas de distinguer les phénomènes de production (voisement, barres d'explosions, formants par exemple). L'analogie avec la perception, très simplifiée, se limite à l'utilisation d'une échelle fréquencielle dérivée des études psychoacoutique pour l'analyse spectro-temporelle. Aucune possibilité de reconstruction du signal à partir de cette représentation ne résiste à son extrême simplicité. La seconde représentation, spectrogramme en bande large, a joué et joue encore un rôle fondamental en acoustique. Le tracé suit le même principe que précédement, avec des trames temporelles plus courtes, $1.5ms$, et un banc de 128 filtres régulièrement répartis, de 300 Hz de largeurs de bande. Les $128 \times 667 = 85376$ événements spectro-temporels par seconde occupent alors un pavé d'environ $1.5 \times 300ms.Hz$ dans le plan. Au prix de ce raffinement d'analyse, les phénomènes importants pour la production ou la perception sont bien apparents sur la représentation. Néanmoins, les événements spectro-temporels utilisés pour l'analyse ne sont pas directement pertinents de ce double point de vue:

chaque événement spectro-temporel de production (par exemple une période de voisement dans la région du second formant) se traduira par un grand nombre d'événements spectro-temporels sur la représentation (pour cet exemple de l'ordre de 100 ou 200), qui ne seront que d'une médiocre utilité pour estimer les paramètres utiles en analyse ou en synthèse (par exemple la fréquence centrale du formant, le calage temporel de la période etc.). La troisième représentation est le tracé dans le domaine spectro-temporel des formes d'ondes élémentaires obtenues par un procédé proposé plus loin, dans la dimension amplitude-temps. Ici la notion de trame et de banc de filtre disparaît au profit d'événements spectro-temporels localisés et adaptés au signal. Au nombre d'environ 1000 par seconde (comme pour la première représentation), ces événements rendent directement compte sur leurs paramètres des phénomènes de production: pour reprendre l'exemple précédent, la région du second formant pour une période de voisement sera représentée par une ou deux formes d'ondes, dont les fréquences centrales seront celles du second formant, qui représenteront aussi le comportement temporel de la période. De plus, comme avec la seconde représentation, on peut reconstruire le signal, et l'analyse visuelle est particulièrement instructive.

Ce troisième type de représentation constitue ainsi un outil puissant de génération, de manipulation et d'analyse du signal de parole, depuis l'analyse acoustique jusqu'à sa représentation comme *forme* à des niveaux de traitement supérieurs. L'enjeu de ce travail est la preuve, ou une esquisse de preuve, de cette assertion.

Le support physique utilisé naturellement pour la transmission de parole est le milieu aérien, où se propagent des variations de pression constituant le signal acoustique. Ce signal acoustique, ou plutôt son analogue, le signal électrique et sa représentation numérique, apparaît donc comme l'objet *mathématique* premier que l'on doit traiter dans tout acte de communication impliquant la machine.

Ce signal possède des caractères spécifiques liés à *l'appareil vocal* qui l'a d'abord produit.

De même il montre des propriétés d'importance et d'intérêt variables pour le *système auditif*, et par suite pour les systèmes chargés de le percevoir.

Il révèle ainsi une nature triple: objet mathématique, produit par un certain dispositif, et destiné à être perçu par un autre dispositif.

La notion de formes d'ondes élémentaires sera considérée ici dans ces trois contextes:

- L'analyse mathématique du signal, où le compromis entre résolution temporelle et résolution spectrale a suscité de nombreux travaux visant à une représentation par des fonctions élémentaires localisées en temps et en fréquence;
- Les modèles d'oreille, en particulier de l'oreille interne, où s'effectue une sorte de décomposition spectro-temporelle du signal acoustique;
- Les modèles de production du signal de parole: une décomposition en événements spectraux et temporels localisés peut se déduire du fonctionnement de l'appareil vocal;

Ces trois points supposent des signaux classés par ordre de généralité décroissante, les derniers étant exclusivement vocaux, les seconds exclusivement acoustiques et les premiers sans hypothèses particulières quant à leur provenance ou à leur nature.

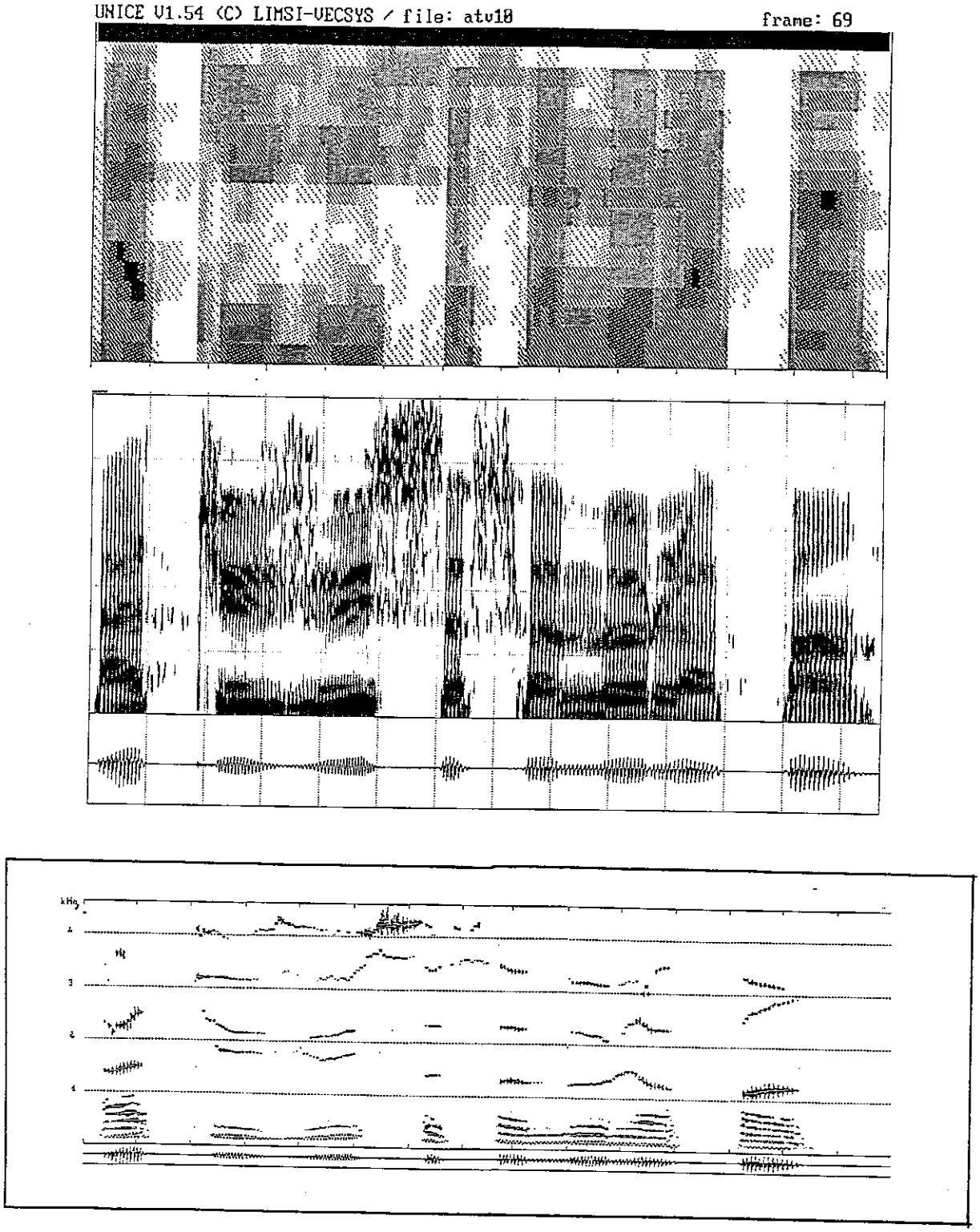

Figure .1: trois représentations spectro-temporelles

Le premier chapitre introduit des méthodes de théorie du signal pour la représentation en somme de formes d'ondes élémentaires, en s'appuyant sur le regain d'intérêt suscité par les méthodes non-paramétriques pour le traitement de la parole: développement récent de la transformée en ondelettes, compréhension et usages de plus en plus subtils de l'analyse de Fourier à court terme. Le point de vue adopté est l'interprétation de ces méthodes depuis *un ensemble discret de fonctions élémentaires d'analyse, dans le domaine temporel.* Les conditions sur les formes d'ondes élémentaires et l'échantillonnage spectro-temporel pour reconstruire le signal avec une précision donnée s'étudient naturellement dans ce cadre. Pour le traitement du signal de parole, hormis peut être le codage, cette interprétation ne présente tout son intérêt que lorsque un sens physique ou perceptif peut être attaché aux fonctions élémentaires.

La perception par une machine et l'étude du signal de parole peuvent profiter des connaissances sur le seul système réellement performant que l'on connaisse: l'oreille. En postulant une certaine indépendance entre traitements périphériques et traitements centraux, le second chapitre analyse quelques modèles fonctionnels du système auditif périphérique, afin de déduire des contraintes auditives pour une représentation spectro-temporelle du signal, et développe alors une réflexion originale sur les relations entre modèles auditifs et analyse granulaire. Ce type d'analyse, proposée par J.S. Liénard, se situe directement dans la perspective évoquée à la fin du paragraphe précédent: une représentation spectro-temporelle, sans utiliser de modèles *a priori*, permet d'estimer des paramètres de composantes physiquement et/ou perceptivement pertinentes tout en conservant la possibilité de reconstruire le signal dans le domaine temporel à partir des formes d'ondes élémentaires. Une justification de cette approche, qui éclaire les analogies avec le fonctionnement du système auditif périphérique à la lumière de modèles fonctionnels, est détaillée. La description d'un système d'analyse synthèse permet d'envisager des modes de traitements basés sur un ensemble de formes d'ondes: détection d'un ensemble d'événements acoustiques, articulatoires, regroupement suivant des critères de cohérence, adaptation à des systèmes de reconnaissance.

Le troisième chapitre est consacré à la synthèse de parole, qui constitue le domaine d'intérêt premier de ces recherches. Après le rappel de la théorie classique de production de la parole, les principaux modèles du signal interprétables en formes d'ondes élémentaires sont cités: représentations sinusoïdales, représentation à formants en parallèle, représentations par des sinusoïdes modulées. Plusieurs expérimentations sont alors présentées, prenant leur source dans la synthèse par Formes d'Ondes Formantiques définie par X.Rodet. Cette méthode est reconsidérée dans le cadre général d'une représentation du signal de parole en formes d'ondes élémentaires qui s'appuie sur un modèle de production. Des solutions sont proposées, qui permettent l'extension de la méthode pour le traitement des segments non-voisés, ou d'excitation mixte. De même, la région grave du spectre, où la forme d'onde de débit glottique prédomine, nécessite un raffinement d'analyse, traité ici par une paramétrisation sinusoïdale de la forme d'onde. Cette représentation en formes d'ondes élémentaires, basée sur un modèle de production peut s'interpréter comme une généralisation de la synthèse à formants en parallèle, dans le domaine temporel. Cependant des différences importantes existent: la décomposition source/filtre s'estompe au profit d'une plus grande homogénéité entre les divers modes de production, la région grave du spectre est traitée de façon plus fine. Le modèle complet de synthèse du signal de parole ainsi construit se situe dans le do-

maine acoustique, de façon purement externe: les objets premiers de la synthèse sont les formes d'ondes élémentaires. Ce parti pris s'éloigne des modèles de synthèse à formants classiques, ensemble de modules qui simulent chacun une part du fonctionnement acoustique de l'appareil vocal. Après la définition du modèle de synthèse vient le problème de l'estimation de ces paramètres. Le choix s'est porté sur l'implantation d'un système automatique d'extraction des paramètres de synthèse: la qualité de la resynthèse est une justification expérimentale de la qualité du modèle adopté. Un système automatique d'analyse-synthèse construit suivant ces principes est présenté, utilisant un ensemble varié de méthodes: modélisation et segmentation spectrale, filtrage, segmentation et modélisation temporelle des signaux filtrés, resynthèse. Plusieurs applications qui ont motivé la définition de ce système sont évoquées: modification spectro-temporellement localisées du signal de parole et synthèse en particulier.

Chapitre 1

DEVELOPPEMENT DU SIGNAL DE PAROLE EN SOMME DE FONCTIONS ELEMENTAIRES

1.1 représentation du signal dans un plan d'information

1.1.1 introduction

Le signal de parole se présente comme une variation de pression atmosphérique, rayonnant essentiellement depuis la bouche et les narines.

Le signal parvenant au récepteur (oreilles, microphones, ...) se trouve considérablement déformé dans sa dimension amplitude/temps en fonction de la localisation relative du locuteur et du capteur, mais aussi des conditions d'environnement et d'ambiance sonore. L'analyse temporelle des signaux de parole, qui offre néanmoins de précieuses méthodes, ne saurait suffire pour appréhender les invariants présents dans le signal ou pour comprendre les performances remarquables de l'analyse réalisée par le système auditif humain.

Les travaux de Fourier au siècle dernier, ont posé les bases de *l'analyse spectrale*, dans le domaine fréquentiel conjugué du domaine temporel (le produit d'un temps et d'une fréquence est sans dimension). Un traitement par des méthodes fréquentielles ne saurait pas plus qu'un traitement purement temporel apporter une réponse satisfaisante au problème de la représentation d'un signal non-stationnaire, comme la parole.

Depuis les célèbres travaux de Wigner, de Gabor, puis de Ville, s'est développé un ensemble de méthodes, les représentations temps-fréquence, qui permettent de transformer un signal monodimensionnel comme le signal de parole en un signal bidimensionnel, de la variable temps et d'une autre variable: fréquence, échelle par exemple. Une représentation dans un espace à trois dimensions, le temps, la fréquence et l'amplitude par exemple, permet l'étude des caractéristiques du signal [25] [26].

Parallèlement à ces résultats sur la représentation de signaux continus, le traitement des signaux et systèmes numériques a introduit l'utilisation d'un réseau discret de points, ou *pavage*, dans le domaine des variables du signal. Un exemple de ce type de réseau, qui

ne dépend que du temps, est donné par le théorème d'échantillonnage. De nombreuses fonctions peuvent servir pour représenter un signal comme une somme de translatées temporelles [4] [46].

L'objet de ce chapitre est de rappeler les bases de la représentation d'un signal sous la forme d'une somme de signaux élémentaires à court terme et à bande limitée, ou formes d'ondes élémentaires, fonctions du temps qui dépendent de l'instant et d'un autre paramètre, la fréquence ou l'échelle. La forme générale de cette sorte de représentation, ou ψ est une forme d'onde élémentaire, c un coefficient complexe, a et b deux variables (instant/fréquence, instant/échelle) et t le temps, est donnée par:

$$x(t) = \int_a \int_b c(a,b)\psi(a,b;t)dadb \tag{1.1}$$

sous certaines conditions l'ensemble des $\psi(a,b)$ peut être rendu discret, par utilisation d'un pavage du plan (a,b):

$$x(t) = \sum_n \sum_k c^{n,k}\psi^{n,k}(t) \tag{1.2}$$

en substituant au temps t l'échantillon m, la formulation discréte et échantillonnée de 1.1 devient:

$$x_m = \sum_n \sum_k c^{n,k}\psi_m^{n,k} \tag{1.3}$$

Toutes ces représentations sont définies et utiles dans des cadres très différents (théorie du signal, analyse fonctionnelle, physique quantique ...), mais l'exposé sera limité aux besoins du traitement du signal de parole.

Cet examen sera envisagé du point de vue des formes d'ondes élémentaires d'analyse utilisées, et ainsi un nombre important de représentations possédant des propriétés remarquables ne seront pas examinées: seules des représentations qui dérivent d'une décomposition explicite sur un ensemble de fonctions élémentaires seront considérées. De plus, des représentations qui auraient pu s'insérer dans ce cadre n'ont pas été retenues, représentation de Fourier-Bessel, décomposition sur une base de fonctions sphéroïdales aplaties [83] par exemple. Leur intérêt pratique pour le traitement de la parole reste à l'heure actuelle à démontrer.

L'étude de deux points de vue complémentaires, d'une utilisation constante sous de nombreuses formulations différentes, sera par contre développée: la transformation de Fourier à court terme et la transformation de Gabor qui s'y rapporte (représentation temps/fréquence), puis la transformation en ondelettes (représentation temps/échelle).

Les deux approches sont connues et utilisées, au moins sous certains aspects, depuis les années quarante. Un formalisme achevé et des algorithmes efficaces pour les signaux numériques, suivis d'une utilisation massive sont disponibles pour la transformation de Fourier à court terme depuis le milieu des années soixante-dix. L'utilisation de bancs de filtres passe-bande de largeur de bande constante, comme le vocodeur de phase et le spectrographe acoustique peuvent s'y rattacher.

Pour la transformation en ondelettes, si on la considère comme une forme particulière de filtrage linéaire, des algorithmes spécifiques existent également depuis assez longtemps. Néanmoins le formalisme explicite de la transformation en ondelettes, et

le développement d'algorithmes en tant que tels, ne datent que de quelques années et font toujours l'objet d'un effort considérable. Au plan temps-fréquence de l'analyse de Fourier il peut être alors commode de substituer d'autres *plans d'information.* Les quelques années (... mois) passés ont apporté une riche moisson de résultats fondamentaux et d'applications qui seront évoqués.

Une approche plus fondamentale des représentations temps-fréquence [79] consiste à définir au départ les propriétés souhaitables et à examiner le mérite des représentations proposées: la représentation de Wigner-Ville semble alors exhiber des propriétés remarquables parmi toutes les formes envisageables [33] [2] [34]. La reconstruction du signal avec cette représentation est problématique, et un relativement faible nombre d'applications à la parole existent pour l'instant [33] [17] [16]: elle ne sera pas examinée ici.

1.1.2 transformée de Fourier

Les travaux de Fourier [36] ont été à l'origine des résultats sur l'analyse et la synthèse spectrale. L'analyse fonctionnelle contemporaine permet de comprendre la série et l'intégrale de Fourier dans le cadre de la théorie des distributions de Schwartz [81] [57], généralisant la notion de mesure puis de fonction.

Il est utile de distinguer la nature des signaux que l'on traite, des définitions différentes s'appliquant à des signaux différents: signaux certains ou signaux aléatoires. Cependant, ces distinctions ne seront pas prises en compte dans ce chapitre, mais dans le chapitre traitant des modèles de production où le sens physique des signaux considérés amène naturellement à considérer ces deux classes.

Le succès de la transformée de Fourier est imputable à l'utilisation massive du filtrage linéaire, dont les fonctions propres sont les exponentielles complexes. Si la relation entre l'entrée x et la sortie y d'un système R est de la forme:

$$y(t) = \int R(\theta)x(t-\theta)d\theta \tag{1.4}$$

(le symbole $\int$ représente l'intégrale $\int_{-\infty}^{+\infty}$ sur $\mathcal{R}$ tout entier, sans précision supplémentaire, de même pour la somme $\sum \leftrightarrow \sum_{-\infty}^{+\infty}$, prise sur $\mathcal{Z}$).

RS est un filtre linéaire ou convolueur, ou système linéaire invariant dans le temps. La réponse aux exponentielles complexes $e^{pt}, p \in \mathcal{C}$ devient:

$$y(t) = \int R(\theta)e^{p(t-\theta)}d\theta \tag{1.5}$$

$$y(t) = H(p)e^{pt} \tag{1.6}$$

La décomposition d'une fonction complexe de la variable réelle en somme (continue ou discrète) de fonctions exponentielles complexes, naturelle pour les fonctions périodiques, conduit à la définition de la transformée de Fourier [72]:

$$\tilde{x}(\nu) = \int x(t)e^{-2i\pi\nu t}dt \tag{1.7}$$

et de la transformée de Fourier inverse:

$$x(t) = \int \tilde{x}(\nu) e^{2i\pi\nu t} d\nu \tag{1.8}$$

La convergence de l'intégrale n'est assurée qu'en posant des hypothèses sur la nature du signal: énergie finie, décroissance rapide, en se plaçant dans le cadre plus général de la théorie des distributions. La transformation de Fourier des signaux périodiques restreint l'intégrale de 1.7 à une période principale, et celle de 1.8 à une somme discrète.

1.1.3 signal analytique et fréquences

La transformation de Fourier introduit une première notion de fréquence. Connaître l'amplitude et la phase présente à une fréquence donnée implique dans ce cadre la donnée du signal sur une durée infinie. D'autres notions de fréquence ont été définies, en considérant des intervalles temporels différents d'observation du signal: fréquence instantanée pour un intervalle infiniment bref, fréquence instantanée pour une analyse de Fourier à court terme, fréquence moyenne sur une plage de temps donnée pour rechercher des composantes dominantes.

Un signal réel $x(t)$ répondant à la symétrie hermitienne,

$$\tilde{x}(\nu) = \tilde{x}^*(-\nu) \tag{1.9}$$

la valeur du spectre pour les fréquences négatives se déduit de la donnée des fréquences positives, et donc n'apporte rien, dans un certain sens, pour la connaissance de $x(t)$. Il est donc intéressant de considérer le signal complexe obtenu par suppression des fréquences négatives (qui sera seul représenté dans plusieurs représentations temps/fréquence) [85]. Cette notion de *signal analytique* a été introduite par Gabor et Ville.

$$x_a(t) = x(t) + i x_q(t) \tag{1.10}$$

où $x_q(t)$ représente le signal en quadrature de $x(t)$, obtenu par transformation de Hilbert:

$$x_q(t) = \frac{1}{\pi} \oint x(\tau) \frac{d\tau}{\tau - t} \tag{1.11}$$

$\oint$ signifiant intégrale en valeur principale au sens de Cauchy.

Une généralisation de la notion de *partie utile* d'un signal est proposée par Bertrand [10]. La partie utile est reliée au signal analytique par un facteur constant, comme classe d'équivalence de signaux invariants par le groupe des déphasages et des atténuations. On peut ne considérer qu'un représentant pour les signaux qui se déduisent les uns des autres par des transformations (réelles et causales) du type:

$$\tilde{x}(\nu) \longrightarrow \alpha e^{i\beta(sgn(\nu))} \tilde{x}(\nu) \tag{1.12}$$

où les constantes α et β rendent compte des atténuations pures et des déphasages purs et où $sgn(\nu)$ signifie signe de ν.

La notion de signal analytique permet de définir l'enveloppe e du signal et sa fréquence instantanée $\omega/2\pi$ [27]:

$$e(t) = (x^2(t) + x_q^2(t))^{\frac{1}{2}} \tag{1.13}$$

$$\omega(t) = \frac{d}{dt} arg(x_a) \tag{1.14}$$

L'enveloppe instantanée garde un sens physique d'enveloppe temporelle pour tous les types de signaux [9].

Par contre, il faut remarquer que la fréquence instantanée d'un signal n'a pas forcément un rapport simple avec le spectre fréquentiel tel qu'il est défini par la transformation de Fourier.

La fréquence instantanée est un puissant moyen de représentation pour les signaux à bande étroite $\Delta\nu/\nu_0 << 1$ ou signaux quasi-monochromatiques [60]: il devient possible dans le cas limite de signaux monochromatiques d'identifier fréquence instantanée et fréquence au sens de Fourier [80] [56]. La distribution d'énergie sur l'axe des fréquences instantanées offre des applications comme étages de paramétrisation acoustique pour un système de reconnaissance [26] [27].

Une autre notion de fréquence instantanée, différente à la fois de la fréquence au sens de Fourier et de la fréquence instantanée au sens de Gabor est définie comme dérivée temporelle de la phase d'un spectre de Fourier à court terme [35] [37]. Plusieurs études récentes prouvent l'intérêt de ce spectre pour estimer des caractéristiques du signal de parole (formants par exemple) [29] [87] [91] [52].

Enfin, pour les signaux à bande large (le qualificatif *bande large* est ici pris par référence au spectrographe acoustique, ou en considérant des largeurs de bande de quelques centaines de *Hertz*) une notion de fréquence dominante en moyenne peut être définie, par comptage des passages par zéro [56]:

$$\nu_{pz} = \pi \times N_{zc} \tag{1.15}$$

ou N_{zc} représente le nombre moyen de passages par zéro par unité de temps.

La fréquence estimée dépend bien sûr de l'intervalle d'observation, mais cette notion revêt dans certains cas un sens physique et pratique incontestable, malgré une définition mathématiquement imprécise. Des applications pour l'estimation de paramètres acoustiques seront considérées dans le cadre du chapitre 2.

L'utilisation des passages par zéro, du signal ou d'une version filtrée de ce signal, a connu un certain succès pour la reconnaissance de parole ou sa synthèse [30] [13].

Un troisième type de méthodes temporelles utilise la corrélation entre les signaux pour exhiber des périodicités qui s'y trouvent, en exploitant le pendant temporel de la notion de fréquence au sens de Fourier qui est la présence de périodicités. Le spectre fréquentiel de puissance est ainsi construit par un procédé temporel.

1.1.4 notion d'échelle

La représentation d'un signal par une méthode fréquentielle implique la signification physique de mouvement périodique, sous-jacente à la notion de fréquence. La recherche d'une résolution fréquentielle fixée conduit à choisir la durée des fonctions élémentaires d'analyse de façon invariante lorsque l'on passe d'une bande d'analyse à une autre:

leurs formes sont donc très différentes (nombre d'oscillations différent à des fréquences d'analyse différentes).

Une alternative à ce point de vue est de considérer des fonctions élémentaires de forme constante, et donc dépendantes d'un facteur d'échelle. La durée des fonctions élémentaires devient ainsi explicitement dépendante de la fréquence et au plan d'information se substitue le plan temps/échelle. Suivant la structure des signaux à représenter, l'un ou l'autre des choix peut s'avérer préférable: la notion d'échelle semble par exemple appropriée à l'étude des courbes fractales qui présentent des caractéristiques identiques à des échelles différentes. De même lorsque les signaux comportent de brusques discontinuités, l'étude à des échelles différentes permet de mettre à jour des évolutions rapides des coefficients. Les transitoires de la parole peuvent relever de cette problématique.

La complémentarité de l'examen d'un phénomène à des échelles différentes est reconnue et exploitée depuis les débuts de l'analyse de Fourier à court terme, il s'agit par exemple des différents filtres d'analyse offerts par les spectrographes analogiques (sonagraphes) des années quarante, des filtres en bande d'octave.

Cependant, partout ou la fréquence porte un sens physique sûr, pour des signaux quasi-stationnaires, l'intérêt d'une analyse en échelle peut s'amoindrir de par la résolution plus faible de l'aigu du spectre.

1.1.5 concentration en temps et en fréquence

Dans le cas limite de la transformée de Fourier le signal se développe en une somme infinie de fonctions orthogonales simples, de durée infinie. Le compromis entre précision fréquentielle et précision temporelle touche une ultime extrémité puisqu'il faut connaître le signal pendant une durée infinie pour calculer la valeur du spectre à une fréquence donnée, et puisqu'il faut connaître un spectre de bande passante infinie pour calculer le signal à un instant donné: dans le plan temps-fréquence les fonctions élémentaires d'analyse peuvent se représenter par des droites infinies parallèles à l'axe des temps.

A l'autre extrémité, [38] on peut définir une analyse purement temporelle en utilisant les mesures de Dirac à l'instant t, $\delta_t(\tau)$, comme fonctions élémentaires généralisées d'analyse.

$$x(t) = \int x(\tau)\delta_t(\tau)d\tau \tag{1.16}$$

avec

$$\delta_t(\tau) = \delta_0(t-\tau) \tag{1.17}$$

Dans le plan temps-fréquence, les fonctions élémentaires d'analyse peuvent se représenter par des droites infinies parallèles à l'axe des fréquences.

Les considérations précédentes formulées par Gabor dans les années quarante amènent à définir la relation entre la précision qu'il est possible d'obtenir conjointement en fréquence et en temps [70] en dehors de ces deux cas limites. L'enjeu est la représentation d'un signal comme somme de formes d'ondes élémentaires possédant une concentration spectro-temporelle bien définie.

Par analogie avec le principe d'incertitude gouvernant l'espace des phases (ou plan d'information) constitué par les variables conjuguées position et impulsion en mécanique quantique, on définit la valeur quadratique moyenne $\bar{t^2}_x$ de l'époque, ou instant d'occurrence d'un signal élémentaire:

$$\bar{t^2}_x = \frac{\int x(t)t^2x^*(t)dt}{\int x(t)x^*(t)dt} \tag{1.18}$$

où $x(t)$ représente un signal complexe (le signal analytique pour x réel).
la valeur quadratique moyenne de la fréquence $\bar{\nu^2}_x$ du signal élémentaire:

$$\bar{\nu^2}_x = \frac{\int \tilde{x}(\nu)\nu^2\tilde{x}^*(\nu)d\nu}{\int \tilde{x}(\nu)\tilde{x}^*(\nu)d\nu} \tag{1.19}$$

La durée effective de x (écart quadratique moyen), et sa largeur de bande effective vaut alors:

$$\Delta t_x = [2\pi\overline{(t-\bar{t}_x)^2_x}]^{\frac{1}{2}} \tag{1.20}$$

$$\Delta \nu_x = [2\pi\overline{(\nu-\bar{\nu}_x)^2_x}]^{\frac{1}{2}} \tag{1.21}$$

et l'on montre que:

$$\Delta t_x \Delta \nu_x \geq \frac{1}{2} \tag{1.22}$$

Nous verrons que la durée effective et la largeur de bande effective peuvent s'obtenir dans un sens un peu différent, mais de façon pratiquement plus simple en utilisant le théorème d'échantillonnage de Nyquist-Shannon.

1.22 va être utilisé pour définir une transformée utilisant des signaux élémentaires dans un certain sens optimaux.

1.2 représentation de Gabor

1.2.1 logons

D'après le principe de concentration spectro-temporelle, parfois abusivement dénommé principe d'incertitude par analogie avec la mécanique quantique, le calcul du signal qui minimise le produit des écarts quadratiques moyens:

$$\Delta t \Delta \nu = \frac{1}{2} \tag{1.23}$$

introduit le signal élémentaire de Gabor x_g:

$$x_g(t_0, \nu_0; t) = e^{-\alpha^2(t-t_0)^2} e^{2i\pi\nu_0 t + i\phi} \tag{1.24}$$

avec t_0, ν_0 et ϕ des constantes interprétables comme l'instant de maximum d'amplitude du signal élémentaire, la fréquence et la phase de l'oscillation porteuse.
La constante α règle Δt et $\Delta \nu$.

$$\Delta t = \frac{\sqrt{\pi}}{\sqrt{2}\alpha} \tag{1.25}$$

$$\Delta \nu = \frac{1}{\sqrt{2\pi}}\alpha \tag{1.26}$$

La transformation de Fourier du signal élémentaire de Gabor vaut:

$$\tilde{x}_g(t_0, \nu_0, \nu) = e^{-(\frac{\pi}{\alpha})^2(\nu-\nu_0)^2} e^{-2i\pi t_0(\nu-\nu_0)+i\phi} \tag{1.27}$$

Chaque *fonction élémentaire* permet donc de définir un rectangle dans le plan temps-fréquence de côtés Δt et $\Delta\nu$, centré au point (t_0, ν_0), dont la valeur est associée à la pesée c^{t_0,ν_0} du signal en ce point, et qui sera dénommé *logon.*

Un pavage du plan tout entier peut se réaliser à l'aide de logons, ce qui aboutit au développement:

$$x(t) = \sum_n \sum_k c^{nk} e^{-\pi \frac{(t-n\Delta t)^2}{2(\Delta t)^2}} e^{-2i\pi kt/\Delta t} \tag{1.28}$$

en choisissant des logons (non optimaux pour 1.22) de surface unité: $\Delta\nu\Delta t = 1$.

1.2.2 calcul des coefficients

Les coefficients c^{nk} dépendant du temps en n et de la fréquence en k, doivent être obtenus par approximations successives, car les signaux élémentaires ne sont pas orthogonaux (un certain recouvrement spectro-temporel existe entre des fonctions élémentaires adjacentes) [5].

Une forme intégrale de la représentation de Gabor existe [47], les indices n et k devenant les variables temporelles et fréquentielle t et ν:

$$x(\tau) = \int\int p(t,\nu) x_g(t,\nu;\tau) d\nu dt \tag{1.29}$$

avec calcul des coefficients par *pesée*, ou intercorrélation du signal x sur le signal $x_g^*(t,\nu)$:

$$p(t,\nu) = \int x(\tau) x_g^*(t,\nu;\tau) d\tau \tag{1.30}$$

On peut en déduire une expression des coefficients discrets c^{nk} [7]. Remarquons que le calcul de 1.29 et 1.30, par intercorrélation, possède une forme semblable à ce que l'on rencontrera dans la transformée en ondelettes.

Une extension de 1.29 1.30 a été proposée [66], en substituant au signal élémentaire de Gabor un *noyau de développement* ψ de la forme suivante:

$$\psi(t,\nu;\tau) = G(t+\tau) e^{2i\pi\nu(t-\beta\tau))} \tag{1.31}$$

où α est un paramètre réel et où l'unique condition sur la fonction réelle $G(t)$ est:

$$\int |G(t)|^2 dt = 1 \tag{1.32}$$

Les signaux élémentaires choisis par Gabor présentent l'attrait théorique de minimiser le produit $\Delta t \Delta \nu$, mais le développement des techniques d'analyse spectrale à court terme, qui peuvent être rapprochées des idées de Gabor, a amené plutôt l'utilisation d'autres types de signaux élémentaires pour le traitement de la parole. Néanmoins, des applications existent dans les domaines voisins, traitement d'image par exemple [84].

Dans un cas pratique, l'enveloppe gaussienne des signaux de Gabor doit subir une troncature, qui introduit des rebonds spectraux. Les propriétés spectrales de l'enveloppe des signaux élémentaires utilisés semblent prévaloir sur l'économie théorique proposée dans le cas gaussien. Le pavage régulier du plan temps-fréquence de la transformée de Gabor sera une des caractéristiques partagées par l'analyse de Fourier à court terme.

Toutes les signaux élémentaires d'analyse possèdent un Δt constant qui entraîne un effet de moyennage temporel constant pour chaque fréquence: un signal élémentaire comportera un nombre d'oscillations proportionnel à sa fréquence d'analyse.

Le développement de l'analyse de Fourier à court terme permet d'envisager la transformation de Gabor avec troncature temporelle comme un cas particulier, et propose des méthodes rapides de calcul des coefficients dans le cas de signaux discrets. La formulation continue 1.29 possède une interprétation directe dans ce cadre.

1.3 représentation de Fourier à court terme

1.3.1 trois interprétations de la transformée de Fourier à court terme

Dans la continuité des travaux de Gabor, une transformée de Fourier à court terme a été définie pour considérer le spectre d'un signal sur une plage temporelle localisée [74] [75].

A partir du signal à une dimension $x(\tau)$ on définit un signal à deux dimensions $x^w(t,\tau)$ qui représente à t fixé le signal x vu à travers la *fenêtre d'analyse* w centrée en t, dénommé *trame d'analyse*, et à τ fixé la contribution des différentes trames (centrées en t) à cet instant τ [75]. Une discussion sur le choix des fonctions w se trouve dans une section suivante.

$$x^w(t,\tau) = x(\tau)w(t-\tau) \tag{1.33}$$

On retrouve le signal x, en intégrant par rapport à t:

$$\begin{aligned} \int x(\tau)w(t-\tau)dt &= x(\tau)\int w(t-\tau)dt & (1.34)\\ &= x(\tau)\int w(t)dt & (1.35) \end{aligned}$$

La transformée de Fourier du signal x^w, bidimensionnel, peut se décomposer en deux transformées de Fourier partielles, par rapport à la première ou à la seconde variable, notées $\tilde{x}^w(t,\nu)$ et $\tilde{x}^w(\nu,\tau)$.

$$\tilde{x}^w(t,\nu) = \int x^w(t,\tau)e^{-2i\pi\nu\tau}d\tau \tag{1.36}$$

$$= \int w(t-\tau)x(\tau)e^{-2i\pi\nu\tau}d\tau \tag{1.37}$$

la transformée de Fourier inverse:

$$x^w(t,\tau) = w(t-\tau)x(\tau) \tag{1.38}$$
$$= \int \tilde{x}^w(t,\nu)e^{2i\pi\nu\tau}d\nu \tag{1.39}$$

donc:

$$x(t) = \int \tilde{x}(t,\nu)e^{2i\pi\nu t}d\nu \tag{1.40}$$

si la fenêtre est normalisée à $w(0) = 1$.

Remarquons que l'origine des temps est fixe dans cette formulation, alors que la fenêtre d'analyse est glissante. Ainsi l'exponentielle d'analyse est continue, et la cohérence des phases à ν fixé est assurée.

La représentation spectro-temporelle de $x(\tau)$ par $\tilde{x}^w(t,\nu)$ peut recevoir trois interprétations, dont deux sont désormais classiques.

Tout d'abord, en reconnaissant la forme d'une convolution [1] dans l'expression de $\tilde{x}^w(t,\nu)$, si on la considère comme une fonction de t, ce qui signifie que l'on fait glisser la fenêtre d'analyse, on peut réécrire l'analyse comme:

$$\tilde{x}^w(t,\nu) = w(t) * x(t)e^{-2i\pi\nu t} \tag{1.41}$$

ce qui permet d'interpréter la transformation de Fourier à court terme comme le filtrage linéaire (convolution temporelle par le filtre de réponse impulsionnelle w(t)) du signal $x_\nu(t)$ résultant de $x(t)$ modulé par $e^{-2i\pi\nu t}$: w(t) porte ainsi le nom de filtre d'analyse. Les fenêtres d'analyse spectrale peuvent être alors considérées comme réponses impulsionnelles de filtres passe-bas. La formule de synthèse correspond à la sommation continue (intégration) d'un ensemble d'oscillateurs commandés, dans le temps et pour chaque fréquence, par les sorties des filtres passe-bas précédents.

On peut également envisager ce processus comme la sommation des sorties f_ν d'un ensemble de filtres passe-bandes de réponses impulsionnelles $w(t)e^{2i\pi\nu t}$, obtenus lorsqu'on fixe la fréquence ν.

$$x(t) = \int f_\nu(t)d\nu \tag{1.42}$$

avec:

$$f_\nu(t) = \tilde{x}^w(\nu,t)e^{2i\pi\nu t} \tag{1.43}$$
$$= (w(t) * x(t)e^{-2i\pi\nu t})e^{2i\pi\nu t} \tag{1.44}$$
$$= w(t)e^{2i\pi\nu t} * x(t) \tag{1.45}$$

Une autre façon d'interpréter la transformation de Fourier à court terme est de considérer une analyse par blocs temporels. Chaque bloc s'obtient par fenêtrage, à t fixé et donne un spectre de Fourier à court terme: c'est un procédé symétrique du

précédent où les valeurs du spectre à un instant t sont obtenues par sommation sur τ. La sommation avec recouvrement (en t) des transformées inverses de chaque spectre à court terme redonne d'après 1.35 le signal initial à un instant τ:

$$x(\tau) = \frac{\int\int \tilde{x}^w(t,\nu)e^{2i\pi\nu t}d\nu dt}{\int w(t)dt} \tag{1.46}$$

On rencontre parfois le terme de synthèse par formes d'ondes pour l'interprétation par blocs de la transformation de Fourier à court terme. Chaque signal à court terme $x^w(t,\tau)$ représente une forme d'onde, qui n'est pas une forme d'onde élémentaire: la décomposition est temporelle mais non pas fréquentielle. L'application de cette interprétation au traitement de la parole peut être fructueuse, comme nous le verrons.

L'interprétation en terme de banc de filtres permet de mettre en lumière la largeur de bande due au filtre d'analyse: c'est une interprétation dans le domaine spectral. La seconde interprétation au contraire permet de cerner explicitement la durée d'analyse, et représente une décomposition temporelle. La figure 1.1 résume l'interprétation en banc de filtres et l'interprétation en blocs.

Une dernière interprétation qui se rapporte à la représentation de Gabor, consiste à envisager la représentation comme la somme dans le domaine temporel d'un ensemble de formes d'ondes élémentaires: c'est l'interprétation en formes d'ondes élémentaires. En croisant les deux interprétations précédentes, la décomposition temporelle de l'interprétation par blocs et la décomposition fréquentielle de l'interprétation en banc de filtres, on définit des formes d'ondes élémentaires $o_{t,\nu}$ centrée au point (t,ν) du plan temps-fréquence, de la forme:

$$o^w(t,\nu;\tau) = w(t-\tau)e^{2i\pi\nu(\tau-t)} \tag{1.47}$$

Ces formes d'ondes se déduisent par translation en temps et en fréquence du noyau de développement:

$$o^w(\tau) = w(\tau) \tag{1.48}$$

la forme d'onde élémentaire conjuguée est:

$$o^{w*}(t,\nu;\tau) = w(t-\tau)e^{-2i\pi\nu(\tau-t)} \tag{1.49}$$

et la pesée $p_{t,\nu}$ du signal x, ou intercorrélation par la forme d'onde élémentaire conjuguée vaut:

$$\begin{aligned} p_{t,\nu} &= \int x(\tau)o^{w*}(t,\nu;\tau)d\tau & (1.50)\\ &= \int x(\tau)w(t-\tau)e^{-2i\pi\nu(\tau-t)}d\tau & (1.51)\\ &= \tilde{x}^w(t,\nu)e^{2i\pi\nu t} & (1.52) \end{aligned}$$

On retrouve donc la transformée de Fourier à court terme, à un facteur de phase près qui rend compte d'une origine des temps fixe, alors que la forme d'onde élémentaire d'analyse induit une origine des temps glissant avec la fenêtre.

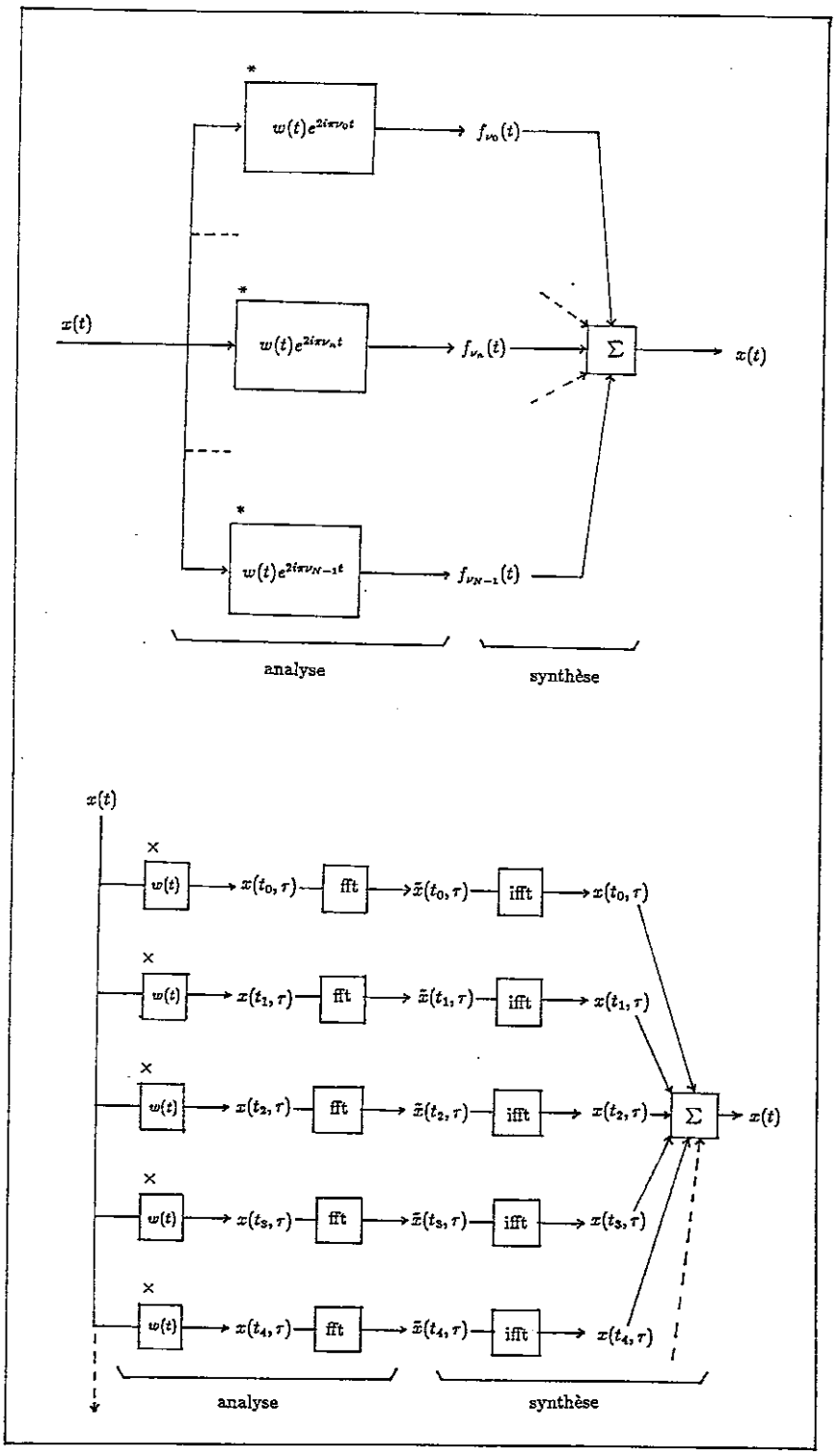

Figure 1.1: interprétation en banc de filtres puis par blocs de la transformée de Fourier à court terme.

La synthèse se déduit de la sommation de toutes les formes d'ondes élémentaires affectées de leur poids:

$$\int\int p_{t,\nu} o^w(t,\nu;\tau)d\nu dt = \int\int \tilde{x}^w(t,\nu)e^{2i\pi\nu t}w(t-\tau)e^{2i\pi\nu(\tau-t)}d\nu dt \quad (1.53)$$

$$= \int x(t,\tau)w(t-\tau)dt \quad (1.54)$$

$$= x(\tau)\int w^2(t-\tau)dt \quad (1.55)$$

$$= x(\tau)\int w^2(t)dt \quad (1.56)$$

par intégration par rapport à ν et 1.33. D'où la formule de synthèse:

$$x(\tau) = \frac{1}{\int w^2(t)dt}\int\int p_{t,\nu} o^w(t,\nu;\tau)dtd\nu \quad (1.57)$$

Cette interprétation généralise le résultat de Helström 1.29 1.30 [47] qui considère le cas d'une fenêtre d'analyse gaussienne afin de calculer la transformée de Gabor continue. La figure 1.3 présente l'interprétation en formes d'ondes élémentaires, qui apparaît sous une forme comparable dans le cas de la transformation en ondelettes.

1.3.2 échantillonnage

Le traitement numérique résulte d'une double discrétisation du signal analogique [55] [68] [69].

Un signal discret s'obtient par échantillonnage du signal analogique à des instants t_n. Les instants d'échantillonnage sont en général choisis périodiques:

$t_n = nT_e$ où T_e représente le pas d'échantillonnage et $1/T_e = F_e$ la fréquence d'échantillonnage.

Si le support du spectre de ce signal est borné, une fréquence d'échantillonnage égale au double de la plus haute fréquence présente dans le signal, permet de le reconstituer parfaitement à partir de sa représentation échantillonnée. L'échantillonnage du signal périodise son spectre à la période f_e, et dans l'intervalle fondamental $[-f_e/2, f_e/2[$ le spectre du signal échantillonné s'obtient par repliement du spectre du signal continu: pratiquement, le signal continu est traité de façon à supprimer l'effet indésirable de ce repliement.

L'échantillonnage temporel du signal est le premier exemple de représentation d'une fonction par un ensemble de fonctions élémentaires. A partir du cas idéal de l'analyse impulsionnelle, une discretisation du temps permet de reconstituer exactement un signal représenté par un ensemble de valeurs aux points d'échantillonnage. Plusieurs types de fonctions élémentaires ont été proposées: celles qui exigent le taux d'échantillonnage le plus bas sont les sinus cardinaux données par le théorème de Shannon. Ces fonctions ne sont pas de durée finie, et la représentation du signal par la somme des translatées d'un type plus pratique de fonctions (fonctions de Lagrange) a été étudiée [46].

Un signal numérique s'obtient à partir d'un signal discret par quantification. dont le choix est motivé par la dynamique que l'on souhaite obtenir.

transformation de Fourier discrète

Un nouveau type de transformation de Fourier a été défini pour les signaux numériques, en échantillonnant également le spectre du signal.

Soit une séquence de N nombres $x_m, 0 \leq m \leq N-1$ on définit la transformée de Fourier discrète par:

$$\tilde{x}_k = \sum_{m=0}^{N-1} x_m \Omega_N^{-km} \tag{1.58}$$

avec

$$\Omega_N = e^{\frac{2\pi i}{N}} \tag{1.59}$$

et la transformée de Fourier discrète inverse:

$$x_m = \frac{1}{N} \sum_{k=0}^{N-1} \tilde{x}_k \Omega_N^{mk} \tag{1.60}$$

On obtient ainsi une autre séquence de N nombres.

Il est intéressant de comprendre les relations entre le signal continu, le signal échantillonné, le spectre du signal continu, le spectre obtenu par transformation de Fourier discrète.

L' échantillonnage temporel donne un spectre continu périodique, dont la période vaut la fréquence d'échantillonnage. Ce spectre s'obtient sur une période fréquentielle fondamentale par repliement du spectre du signal continu: il est donc égal au spectre du signal continu si l'on a pris soin de le choisir avec une bande passante limitée à $f_e/2$.

L'échantillonnage fréquentiel va induire une double approximation: on obtient une version échantillonnée du spectre continu du signal continu replié, sur une période fondamentale fréquentielle.

De plus la transformée de Fourier discrète d'un signal étant un signal échantillonné périodique, sa transformée inverse redonne un signal échantillonné périodique, égal au signal échantillonné initial dans l'intervalle temporel fondamental [0,N[.

Les fonctions élémentaires d'analyse sont ici égales à des segments d'exponentielles complexes, racines $N^{ième}$ de l'unité (ou si l'on veut au produit temporel d'exponentielles complexes par une fenêtre carrée de N échantillons). Ces exponentielles complexes forment une base de fonctions orthogonales pour les signaux de N échantillons (donc qui durent NT_e secondes).

Le succès de la transformation de Fourier discrète a été favorisé par l'existence d'algorithmes de calcul efficaces, les transformations de Fourier rapides [19].

La transformée de Fourier discrète introduit une double discrétisation du plan spectro-temporel, ainsi que l'utilisation de fonctions élémentaires de durées et de bandes passantes finies. Avec les réserves et les précautions précédentes, la représentation de de Fourier à court terme est implémentée sur les machines numériques en utilisant la transformée de Fourier discrète.

1.3.3 fenêtres d'analyse spectrale

Une fenêtre temporelle est appliquée au signal à traiter, pour ne prendre en compte qu'un nombre fini d'échantillons. Cette limitation de la durée d'observation du signal temporel entraîne l'apparition du phénomène de Gibbs: oscillations autour des discontinuités de la transformée de Fourier.

Le choix de la fenêtre temporelle permet d'atténuer l'étalement spectral (phénomène de distribution), dû également à la manipulation de séquences de durée limitée.

Un signal échantillonné ou continu peut, par l'utilisation de fenêtres temporelles, n'être considéré que comme une suite de blocs de signal, placés à des instants donnés: on parlera alors de traitement par *trame*. Une trame d'analyse s'obtient par application d'une fenêtre sur le signal. L'espacement des trames conditionne leur recouvrement temporel, et les possibilités de reconstituer ou non le signal original à l'aide de ces trames. Les conditions sur l'espacement des trames sont dictées par les propriétés spectrales de la fenêtre: considérant sa largeur de bande effective (par exemple donnée par 1.21) le théorème de Shannon permet de prévoir l'échantillonnage temporel des trames.

De même, la durée de la fenêtre permet par une autre application du théorème de Shannon d'estimer le nombre de bandes de fréquences utiles pour reconstituer un spectre complet.

Ainsi l'étude des propriétés des différentes fenêtres temporelles [49] éclaire le traitement par trame comme processus de sous échantillonnage spectral et temporel, définissant leurs mérites relatifs à un certain nombre de critères.

Dans le domaine spectral, l'utilisation d'une fenêtre carrée $w(n) = 1$ si $-N/2 \leq n \leq N/2 - 1$ revient à la convolution par sa transformée de Fourier:

$$W_r(\nu) = \sum_{n=\frac{-N}{2}}^{\frac{N}{2}-1} e^{-2i\pi\nu n} = \frac{\sin \pi\nu N}{sin\pi\nu} \tag{1.61}$$

De nombreuses fenêtres ont été proposées, qui ramènent le signal temporel près de zéro aux extrémités afin de diminuer ces phénomènes indésirables: ainsi le signal périodisé du signal fenêtré n'a plus de discontinuité aux extrémités. On peut également choisir des fenêtres qui rendent continues les dérivées du signal périodisé jusqu'à un ordre élevé.

Des tests comparatifs existent [46] pour évaluer la qualité des fenêtres d'analyse spectrale en fonction de nombreux critères comme:

- amplitude du plus haut lobe secondaire par rapport au lobe principal dans le domaine spectral;
- degré de corrélation de séquences aléatoires pour un recouvrement donné des trames d'analyse;
- largeur de bande du bruit blanc équivalent;
- produit $\Delta t \times \Delta\nu$

Les fenêtres d'analyse spectrales peuvent être considérées comme des réponses impulsionnelles finies de filtres passe-bas. Le nombre N de points lors de leur échantillonnage est choisi pair (avec la même symétrie qu'une transformée de Fourier discrète, c'est à dire un décalage d'un échantillon vers les temps négatifs si la fenêtre est centrée en zéro).

Un compromis courant en traitement de la parole est d'utiliser la fenêtre de Hamming:

$$w(n) = \alpha + (1-\alpha)\cos[\frac{2\pi}{N}n] \tag{1.62}$$

pour une fenêtre de N points, ce qui donne dans le domaine spectral:

$$W(\nu) = \alpha D(\nu) + \frac{(1-\alpha)}{2}[D(\nu - \frac{1}{N}) + D(\nu + \frac{1}{N})] \tag{1.63}$$

avec:

$$D(\nu) = \frac{sin(\pi\nu N)}{sin(\pi\nu)} \tag{1.64}$$

α vaut 0.54 pour la fenêtre de Hamming.

Le premier lobe secondaire se trouve environ -42 dB sous le lobe principal, et un recouvrement de 4 (une trame tous les N/4 points) semble raisonnable [3], vu la largeur du lobe principal: la largeur de bande effective est prise ici comme la largeur du lobe principal, et non d'après 1.21. La durée effective est la durée (finie) de la fenêtre.

Les fenêtres de Blackman-Harris sont préférables pour l'atténuation des lobes secondaires (jusqu'à -92 dB), mais apportent une plus grande largeur du lobe principal, et nécessitent donc un recouvrement plus important des trames d'analyse (environ une trame tous les N/6 points).

$$w(n) = a_0 + a_1\cos[\frac{2\pi}{N}n] + a_2\cos[\frac{2\pi}{N}2n] + a_3\cos[\frac{2\pi}{N}3n] \tag{1.65}$$

pour une fenêtre de N points.

a_0 vaut 0.35875, a_1 0.48829, a_2 0.14128 et a_3 0.01168 pour les caractéristiques spectrales considérées.

Un troisième type de fenêtre, d'un intérêt particulier pour notre propos est la fenêtre gaussienne, enveloppe des signaux élémentaires de Gabor qui possède la propriété de minimiser le produit de la durée effective moyenne par la largeur de bande effective, pour une fenêtre de durée infinie. Cependant, une version tronquée de cette fenêtre présente des qualités inférieures à celles des deux fenêtres précédentes. Les propriétés de cette fenêtre induisent par ailleurs des relations particulières, sous forme d'équations aux dérivées partielles, entre le spectre d'amplitude et le spectre de phase des signaux analysés, qui permettent en ne connaissant que l'un des deux spectres de retrouver l'autre [77].

1.3.4 échantillonnage spectro-temporel

Toutes les formules présentées pour la transformation de Fourier à court terme possèdent une contrepartie en termes de signaux échantillonnés en temps, et en fréquence par utilisation de la transformation de Fourier discrète, avec les contraintes qui l'accompagnent.

Si x_m est maintenant un signal échantillonné, on peut définir de façon analogue à 1.33 le signal bidimensionnel $x^w_{n,m}$ par:

$$x^w_{n,m} = x_m w_{n-m} \tag{1.66}$$

où $w(n)$ représente une fenêtre échantillonnée.

La transformation de Fourier à court terme des signaux numériques résulte d'un double échantillonnage, en temps et en fréquence par utilisation de la transformation de Fourier discrète. Si la séquence temporelle possède N points, la transformée de Fourier à court terme s'écrit, d'après 1.58 et 1.37:

$$\tilde{x}^w_{n,k} = \sum_{m=0}^{N-1} x^w_{n,m} \Omega_N^{-km} \tag{1.67}$$

$$= \sum_{m=0}^{N-1} x_m w_{n-m} \Omega_N^{-km} \tag{1.68}$$

et la transformée de Fourier inverse, en utilisant 1.60 et 1.39:

$$x^w_{n,m} = w_{n-m} x_n = \frac{1}{N} \sum_{k=0}^{N-1} \tilde{x}^w_{n,k} \Omega_N^{km} \tag{1.69}$$

Le nombre de bandes d'analyse est fixé par la taille de la transformée de Fourier discrète, donc par le nombre de points de la séquence temporelle. Pratiquement la fenêtre d'analyse est moins longue que la séquence temporelle (pour préserver la résolution temporelle en augmentant la résolution spectrale) que l'on complète par des zéros.

De plus, les propriétés spectrales et temporelles du filtre d'analyse permettent de ne pas calculer les transformées directes et inverses pour chaque échantillon, mais d'utiliser des trames assez espacées. Un choix judicieux permet même de représenter un signal avec en moyenne un seul échantillon spectral par échantillon temporel [75].

Dans l'interprétation en banc de filtres, il s'agit d'échantillonner les sorties des filtres passe-bas d'analyse. Il peut alors être utile d'introduire un filtre interpolateur pour la resynthèse: le filtre de synthèse. Si la sortie des filtres doit être rééchantillonnée (interpolée) tous les R échantillons, et si f_n est la fenêtre ou filtre de synthèse [21]:

$$\tilde{x}^w_{n,k} = \sum_r f_{n-rR} \times \tilde{x}^w_{rR,k} \tag{1.70}$$

Dans l'interprétation en blocs, il suffit d'espacer les blocs d'analyse et l'interpolation de synthèse se fait par les recouvrements-additions.

Dans l'interprétation en formes d'ondes élémentaires, un espacement régulier entre les trames et entre les bandes d'analyse donne un pavage rectangulaire de l'espace temps-fréquence. Le théorème de Shannon permet de choisir ce pavage de façon à obtenir une reconstruction exacte du signal. La transformée de Gabor est un cas particulier de pavage rectangulaire, dont on peut obtenir une approximation en tronquant l'enveloppe des signaux élémentaires pour obtenir un filtre d'analyse à réponse impulsionnelle finie. La figure 1.2 montre le type de réseau utilisé, dans le plan temps-fréquence.

Ce pavage régulier du plan temps-fréquence entraîne un effet de moyennage des évolutions rapides du signal (beaucoup de périodes dans un seul signal élémentaire) qui dépend de la fréquence d'analyse et du filtre d'analyse. Il a donné des représentations extrêmement utiles du signal de parole, comme le spectrographe.

L'échantillonnage spectro-temporel est dans ce cas choisi pour faire apparaître des phénomènes à une échelle compatible avec la production et la perception du signal étudié: pas temporel de l'ordre de la milliseconde, largeur de bande des filtres de quelques dizaines de Hertz (bande étroite) ou de quelques centaines de Hertz (bande large). Il est certain que le spectrographe ne permet d'observer qu'une partie des évolutions spectro-temporelles du signal analysé: la forme exacte d'un cycle vocalique représente par exemple une évolution trop rapide par rapport aux réponses impulsionnelles des filtres d'analyse pour être correctement représentée. Un effet obtenu en considérant le signal de parole comme stationnaire, sur une tranche de temps ou il ne l'est pas, a été soigneusement étudié dans [82] [52], dans le cas d'un spectrographe numérique utilisant une fenêtre de Hamming: des évolutions formantiques rapides (en fréquence et en amplitude, lors de transitions comme des plosives) ne sont pas du tout apparentes ou même apparaissent en contresens.

Des études existent également pour justifier la résolution utile lorsque l'analyse à court terme est considérée comme un procédé de codage (vocodeurs) [71].

1.3.5 applications de la représentation de Fourier à court terme

L'utilisation de la transformation de Fourier à court terme offre de très nombreuses variantes quant à l'exploitation des spectres obtenus (estimation de paramètres physiques, compression d'information ...) [73] [76] [15] [3] [42] [63] [36].

C'est un puissant moyen de modification du signal vocal: modification des durées, de la mélodie, filtrage linéaire, filtrage adaptatif, soustraction spectrale, déréverbération.

L'utilisation des filtres de synthèse et d'analyse permet d'employer des cadences de décimation et d'interpolation différentes des spectres à court terme, ce qui se révèle très utile pour modifier la durée d'un signal, même de façon non uniforme.

Le traitement séparé (interpolation/décimation) des phases et des amplitudes permet de modifier la fréquence fondamentale du signal, lorsqu'il est quasi-périodique.

L'application d'une pondération aux échantillons spectraux à court terme réalise un filtrage linéaire, et l'on peut (sous certaines conditions quant aux fenêtres d'analyse et de synthèse) faire évoluer le gain du filtre pour réaliser un filtrage évoluant dans le temps. Cette utilisation est très répandue de par la (relative) rapidité des calculs et la commodité d'implantation de filtres de gains arbitraires, compte tenu des contraintes de la fenêtre d'analyse. Si $t_{n,m}$ représente la réponse impulsionnelle d'un filtre (évoluant dans le temps en n), l'utilisation d'un sous échantillonnage de facteur R et d'une TFD

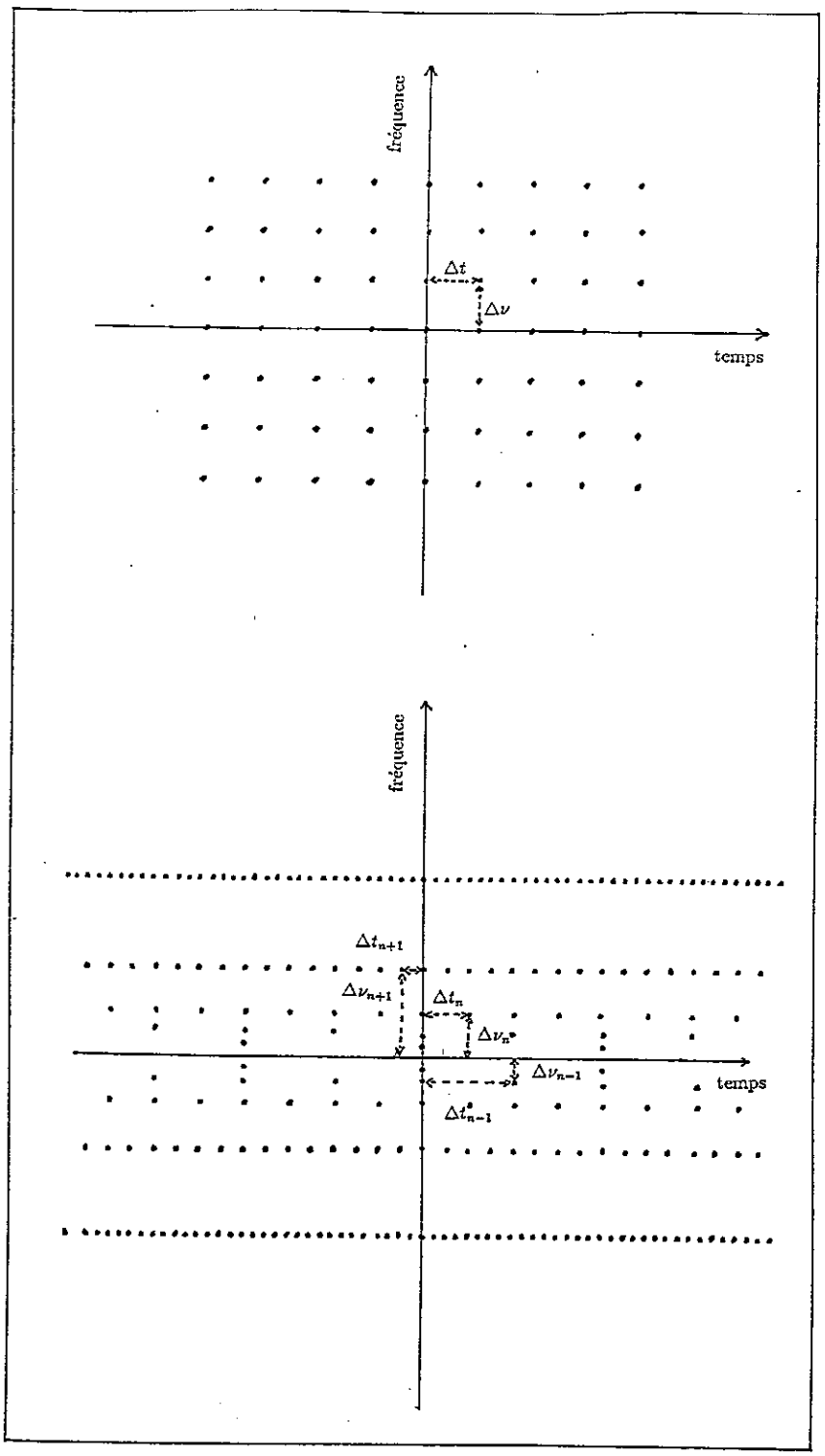

Figure 1.2: réseaux dans le plan temps-fréquence: transformée de Fourier à court terme et transformée en ondelettes.

sur N échantillons avec les filtres w_n d'analyse et f_n de synthèse, on montre que la réponse impulsionnelle véritable du système ainsi réalisé vaut:

$$\bar{t}_{n,m} = \sum_p \sum_s f_{n-sR} \times w_{sR-n+m} \times t_{sR,m-pN} \tag{1.71}$$

Si la réponse impulsionnelle n'évolue pas dans le temps $t_{n,m} = t_{n-m}$, la modification du signal de parole par multiplication par $\tilde{t}_{n,m}$ des échantillons spectraux réalise une convolution par $t_{n,m}$, donc un filtrage linéaire. On retrouve les méthodes de convolution rapide (ou fast convolution) basée sur la transformée de Fourier discrète en utilisant des fenêtres d'analyse et de synthèse rectangulaires.

1.3.6 algorithmes de calcul

Le succès de l'analyse-synthèse de Fourier à court terme est indissociable de la puissance de calcul offerte par les transformées rapides [19]. Le nombre de points par trame d'analyse, fixé par l'algorithme ne pose concrètement pas de problème et peut être relativement indépendant de la taille de la fenêtre d'analyse. Il décidera du nombre de bandes de fréquence disponibles. L'utilisation de la transformée de Fourier discrète implique une origine des temps glissant avec la fenêtre d'analyse. Un facteur de phase s'introduit alors.

Un autre algorithme de calcul rapide, l'analyse spectrale différentielle [50] conserve au contraire la cohérence des phases en utilisant une origine des temps fixe. Dans ce cas, l'utilisation d'une fenêtre carrée (avec les inconvénients spectraux qu'elle implique) conduit à un calcul rapide des coefficients, et permet de retrouver directement la fréquence instantanée. Le calcul est basé sur la remarque suivante; si on utilise une origine des temps fixe et une fenêtre rectangulaire de N points:

$$\tilde{x}_{n,k} = \sum_{m=0}^{N-1} x_m \Omega_N^{-km} \tag{1.72}$$

alors une différentiation montre que:

$$\tilde{x}_{n,k} - \tilde{x}_{n,k-1} = (x_k - x_{k-N})\Omega_N^{-nk} \tag{1.73}$$

ce qui donne une loi simple et rapide à calculer de réactualisation des échantillons spectraux pour une bande d'analyse donnée.

1.3.7 pavages adaptés

Différentes méthodes ont été proposées pour affiner la représentation de Fourier à court terme, afin de s'affranchir du pavage régulier dans le plan temps-fréquence.

La méthode MMWM (Modified Moving Window Method) [52], permet d'ajuster les noeuds du réseau en fonction du signal analysé, dans les limites d'un pavé. Le coefficient correspondant à un pavé spectro-temporel sera affecté au meilleur point possible de ce pavé, et non à son centre. En écrivant les coefficients de la transformée de Fourier à court terme, avec la fenêtre w, sous forme polaire:

$$\tilde{x}^w(t,\nu) = A(t,\nu)e^{i\phi(t,\nu)} \tag{1.74}$$

la formule de synthèse devient, en utilisant l'interprétation en formes d'ondes élémentaires 1.54:

$$x(\tau) = \int\int A(t,\nu)w(t-\tau)e^{2i\pi\nu(\tau-t)+i\phi(t,\nu)}d\nu dt \tag{1.75}$$

où $A(t,\nu)$ et $w(t-\nu)$ sont des quantités réelles variant lentement dans le temps, devant les variations de la phase. Ce sont donc les variations de la phase qui permettent de rechercher les points spectro-temporels qui ont une contribution maximale pour l'intégrale.

En annulant les dérivées partielles de l'argument de l'exponentielle par rapport au temps t et à la fréquence ν. on obtient le couple $(\bar{t},\bar{\nu})$ correspondant au retard de groupe de la pesée au point (t,ν):

$$\bar{t} = t - \frac{1}{2\pi}\frac{\partial\phi(t,\nu)}{\partial\nu} \tag{1.76}$$

et à sa fréquence instantanée:

$$\bar{\nu} = \frac{1}{2\pi}\frac{\partial\phi(t,\nu)}{\partial t} \tag{1.77}$$

Ainsi à partir d'un pavage régulier, une représentation adaptée au signal dans un certain sens, se déduit donc de l'information de phase locale. Cette adaptation est restreinte aux limites d'un pavé.

L'interprétation par blocs, ou formes d'ondes spectralement composées, de la transformée de Fourier à court terme offre de riches possibilités d'adaptation au signal de parole [45] [15]. Les blocs d'analyse sont centrés sur les instants importants du signal: les périodes de voisement. Ce traitement implique une détection de la fréquence de voisement, et une décision voisé/non voisé. Les périodes de voisement sont en moyenne d'un ordre de grandeur comparable à l'espacement des trames pratiquement utilisé, ce qui autorise la synchronisation des blocs de transformées de Fourier à court terme sur le fondamental. Une application directe à la modification des durées découle de ce synchronisme. Les autres modifications prosodiques (fondamental) exploitent le traitement spectral permis par la représentation de chaque bloc dans le domaine fréquentiel. Cette notion de forme d'onde, sans décomposition fréquentielle, est un puissant outil de manipulation du signal de parole, dans une certaine mesure complémentaire de l'approche développée dans le troisième chapitre: décomposition suivant les instants importants du signal complémentaire d'une décomposition suivant les composantes spectrales importantes du signal. Pour le traitement de la parole non voisée, les trames d'analyse sont distribuées sur l'axe des temps de façon irrégulière, pour éviter un bruit tonal lors des modifications, avec une densité suffisante pour une reconstruction de bonne qualité. Le pavage adapté du plan temps-fréquence est régulier dans la dimension fréquentielle, et irrégulier dans la dimension temporelle.

La liaison entre le regroupement des formes d'ondes par une décomposition temporelle adaptée, et le regroupement fréquentiel par une décomposition spectrale adaptée offre une perspective prometteuse.

1.4 représentation en ondelettes

1.4.1 transformée en ondelettes

Les insuffisances de l'analyse de Fourier, pour l'analyse de phénomènes physiques présentant des composantes d'échelles spectrales et temporelles très différentes, ont conduit à l'utilisation de systèmes dépendant non du temps et de la fréquence, mais du temps et d'un *facteur d'échelle*. Pour de tels systèmes, le rapport de la largeur de bande à la fréquence centrale d'analyse demeure constant [40] [90]:

$$\frac{\Delta\nu}{\nu} = constante \tag{1.78}$$

La durée des réponses impulsionnelles de tels systèmes est alors en relation inverse avec leurs fréquences centrales.

En terme de formes d'ondes élémentaires, les fonctions d'analyse se déduisent toutes d'une même fonction par translation dans le temps et contraction ou dilatation temporelle suivant le facteur d'échelle [61] [64] [65] [54]. Ce procédé se distingue de l'analyse de Fourier à court terme classique, dont les formes d'ondes élémentaires d'analyse se déduisent d'une même fenêtre (ou enveloppe) temporelle par translation dans le temps et les fréquences.

Les analogies entre le plan temps-fréquence et l'espace des phases de la physique quantique [6] apportent un autre éclairage au problème de la représentation par un ensemble discret d'ondelettes. La transformation de Fourier à court terme est liée au groupe de Weyl-Heisenberg des décalages en temps et en fréquence, qui place les noyaux d'analyse sur une grille rectangulaire dans le plan, alors que la transformation en ondelettes se rapporte au groupe affine des décalages et des dilatations [44] [43].

La transformation en ondelettes n'est pas le seul type de représentation à délaisser le groupe des translations en temps et en fréquence. Pour représenter les signaux à large bande, une classe de représentation temps-fréquence prend en compte le groupe affine des changements d'horloge [11] [12]:

$$(t,\nu) \longrightarrow (a(t+b), \frac{\nu}{a}) \tag{1.79}$$

pour $a > 0, b \in \mathcal{R}$, mais ne sera pas considéré ici.

Un autre point de vue familier au domaine du traitement de la parole conceptuellement proche de la transformation en ondelettes, est le codage en sous bandes [20] [31]. Des formulations arborescentes (arbres binaires) permettent d'obtenir une suite de filtres emboités, de largeurs de bandes décroissantes avec le nombre de pas de la décomposition.

Les coefficients de transformée en ondelettes d'un signal $x(t)$, ou pesée, s'écrivent de façon analogue à 1.52:

$$p_{b,a} = \frac{1}{\sqrt{a}} \int o^*(\frac{t-b}{a}) x(t) dt \tag{1.80}$$

$$= \sqrt{a} \int \tilde{o}(a\nu)\tilde{x}(\nu) e^{2i\pi\nu b} d\nu \tag{1.81}$$

ou $o((t-b)/a) = o_{b,a}(t)$ est l'ondelette d'analyse à l'intant b et à l'échelle a.

L'ondelette $o_{b,a}(t)$ doit répondre à certaines conditions:

$$\int o(t)dt = 0 \tag{1.82}$$

ce qui signifie que l'ondelette oscille de façon à présenter une moyenne nulle. La condition d'admissibilité:

$$c_o = \int |\tilde{o}(\nu)|^2 \frac{d\nu}{|\nu|} < \infty \tag{1.83}$$

implique qu'il n'y ait pas d'énergie au voisinage de la fréquence zéro en pratique (cette condition est dérivée du problème de la représentation du groupe des translations et dilatations).

Si le signal $x(t)$ est d'énergie finie, on peut utiliser la formule de synthèse:

$$x(t) = \frac{1}{c_o} \int \int p_{b,a} \frac{1}{\sqrt{a}} o(\frac{t-b}{a}) \frac{dbda}{a^2} \tag{1.84}$$

Pour utiliser la transformation comme représentation temps-fréquence, il peut être commode de demander en plus que la fonction $\tilde{o}(\nu)$ soit nulle pour les fréquences négatives (qu'elle soit un signal analytique): une façon simple de construire des ondelettes est alors d'adopter une démarche analogue à la transformation de Fourier à court terme en fenêtrant une exponentielle complexe.

La transformation peut recevoir deux interprétations, de même que la transformation de Fourier à court terme, dans le cas particulier d'ondelettes obtenues par fenêtrage.

D'une part on reconnaît dans 1.81, si l'on fixe a, la forme d'une convolution en b. On peut considérer la transformation en ondelettes comme un filtrage linéaire, par un banc de filtres linéaires possédant la propriété $\Delta\nu/\nu = C^{te}$, dont les caractéristiques sont fixées par le paramètre a, pour les fréquences centrales, et par la forme de l'ondelette pour la forme des filtres.

D'autre part, la transformée en ondelettes peut s'interpréter comme la pesée du signal sur une famille de fonctions se déduisant de $o(t)$. La figure 1.3 résume ces deux interprétations.

Par contre l'interprétation par blocs de la transformée de Fourier à court terme ne peut plus posséder d'équivalence pour la transformée en ondelettes, puisqu'elle considère toutes les bandes fréquentielles sur une durée fixée, à un instant fixé. Ici, la notion d'instant d'analyse persiste alors que la notion de durée d'analyse dépend de la fréquence. Il suffit de considérer la transformée d'une mesure de Dirac dans les deux cas: pour la transformation de Fourier à court terme on obtient dans l'espace temps-fréquence un rectangle de longueur la fenêtre d'analyse, et de hauteur la bande passante de la transformée (demi-fréquence d'échantillonnage par exemple), pour la transformation en ondelettes on obtient un triangle, puisqu'augmenter la fréquence revient à diminuer la durée.

Plusieurs travaux appliqués au traitement des signaux utilisent comme ondelette une exponentielle complexe enveloppée par une gaussienne [41] [54]:

$$o_m(t) = e^{-t^2/2} e^{2i\pi\nu_0 t} \tag{1.85}$$

dont la transformée de Fourier vaut:

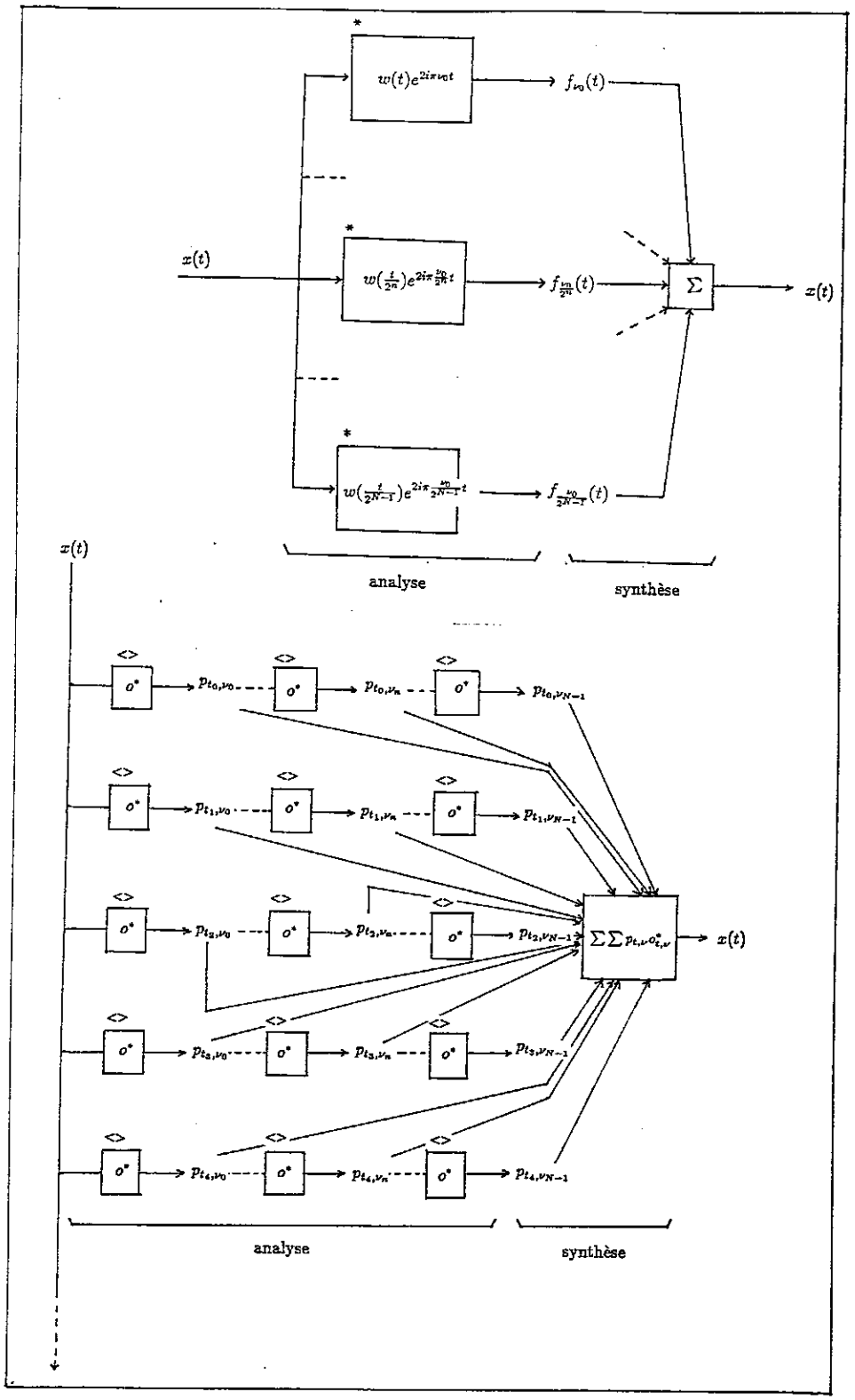

Figure 1.3: deux interprétations de la transformée en ondelettes: banc de filtres, et développement en somme de formes d'ondes élémentaires.

$$\tilde{o_m}(\nu) = e^{-(\nu-\nu_0)^2/2} \tag{1.86}$$

où ν_0 est la fréquence centrale à l'échelle choisie comme unité. Cette fonction n'est toutefois pas une ondelette admissible, $\tilde{o_m}(0) \neq 0$.

Il faut remarquer que parmi les ondelettes obtenues par fenêtrage, l'ondelette à enveloppe gaussienne tronquée n'est pas systématiquement préférable: l'étude des fenêtres d'analyse spectrale conduit plutôt à l'utilisation d'autres types d'enveloppes. La figure 1.4 montre, à trois échelles différentes l'enveloppe d'une ondelette basée sur une fenêtre de Hamming, dans la dimension temps/fréquence/amplitude.

Les formes intégrales de 1.81 comme pour la transformée de Fourier permettent par ailleurs d'accéder à la transformation en ondelettes des dérivées d'un signal en choisissant la dérivée de l'ondelette analysante comme ondelette analysante (par intégration par parties).

En se plaçant dans le domaine fréquentiel, on peut constater la situation en quelque sorte symétrique des deux approches, comme le montre la figure 1.5. Dans le premier cas, le nombre de cycles par ondelette varie avec la fréquence, mais la résolution fréquentielle est constante. Dans le second cas la résolution fréquencielle varie avec la fréquence, mais le nombre de cycles par ondelette est constant.

1.4.2 échantillonnage en temps et en fréquence

La transformation en ondelettes continue d'un signal présente une redondance intrinsèque (représentation d'une fonction d'une variable par une fonction de deux variables) qui impose des relations particulières aux coefficients de 1.81. La valeur de la transformée en un point (a_0, b_0) peut s'exprimer en fonction des valeurs dans le voisinage de ce point par:

$$p_{b_0,a_0} = \int\int K((b_0 - b)/a, a_0/a) p_{b,a} \frac{dadb}{a^2} \tag{1.87}$$

où $K(b, a)$, le noyau reproduisant peut s'exprimer sous la forme:

$$K(b, a) = \frac{1}{c_o} \int o^*((t - b)/a) o(t) dt \tag{1.88}$$

La région spectro-temporelle où la contribution de K est non négligeable fixe le domaine d'intégration dans le voisinage de (a_0, b_0) qui permet de reconstituer le coefficient en ce point.

Par un résultat analogue (bien qu'à deux dimensions) au théorème de reconstruction d'un signal échantillonné par des sommes de translatées, on peut définir un échantillonnage en échelle et en temps pour reconstruire le signal par des sommes de translatées/dilatées [24], à une précision donnée.

Les formules précédentes s'étendent au cas des signaux numériques. Le choix du filtrage à $\Delta\nu/\nu = C^{te}$ et l'application du principe de concentration fait apparaître un pavage du plan temps-fréquence assez différent de celui de la transformée de Fourier à court terme.

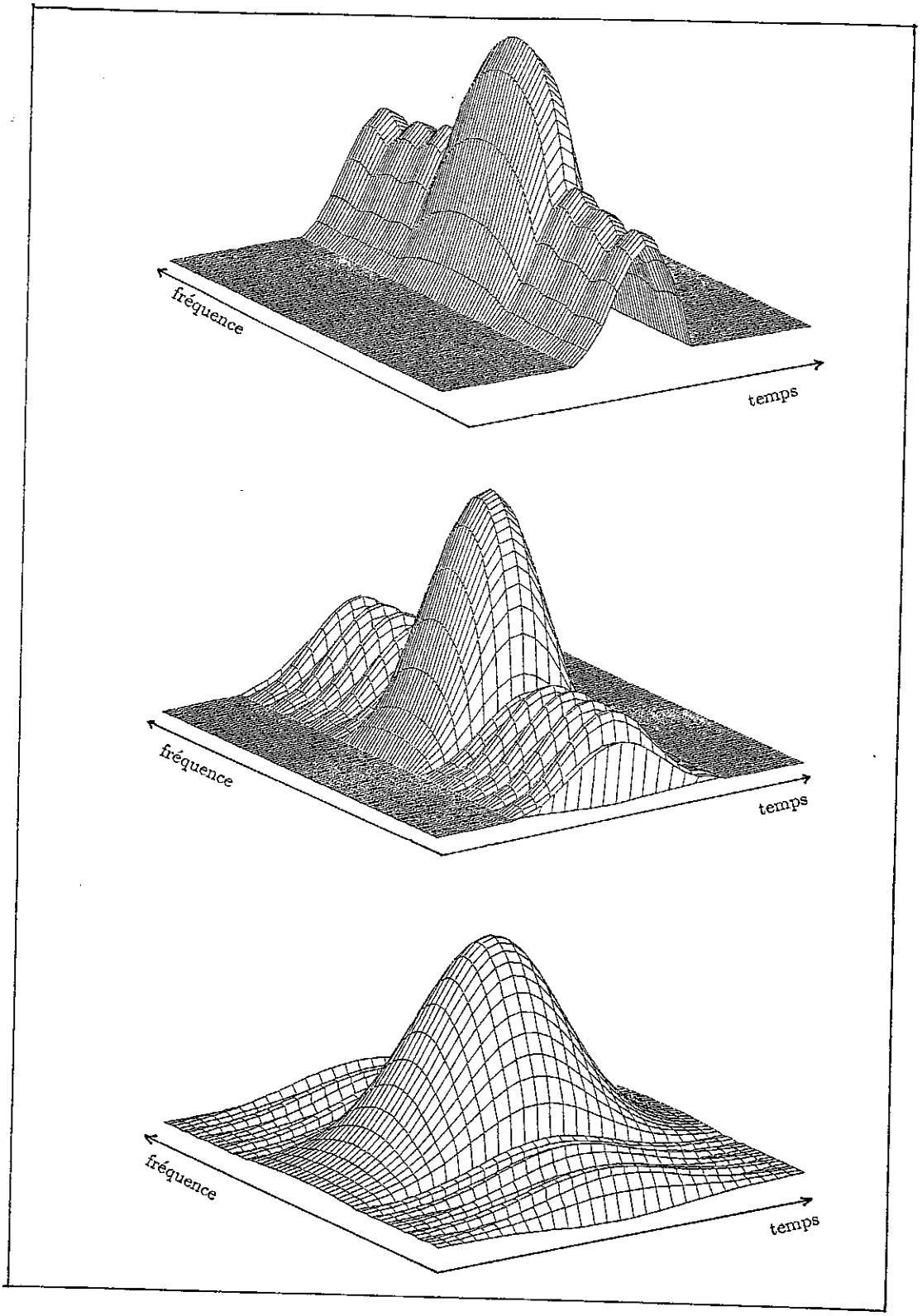

Figure 1.4: ondelette de Hamming, à trois échelles (octaves)

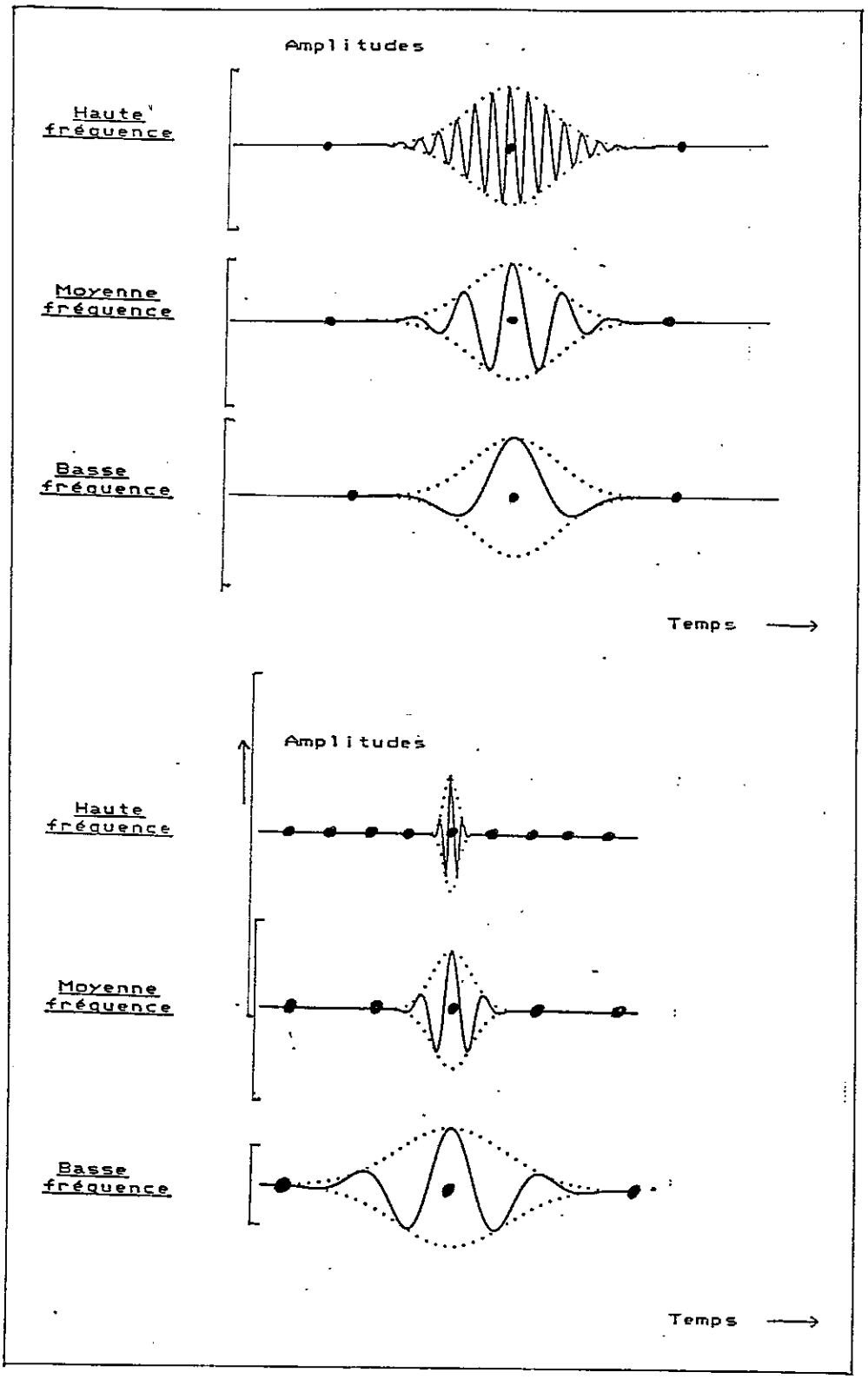

Figure 1.5: formes d'ondes élémentaires, transformée de Fourier à court terme, puis transformée en ondelettes (d'après Morlet)

L'étude des conditions sur les pavages acceptables peut se mener par des considérations classiques de traitement du signal, en fonction des largeurs de bande et des durées des ondelettes par application double du théorème de Shannon.

En effet, les ondelettes d'analyse ne se présentent plus toutes comme des rectangles identiques: leur durée s'allonge lorsque leur fréquence descend, et leur largeur de bande diminue.

Un pavage du plan par des fonctions de produit $\Delta t \times \Delta \nu$ constant n'est plus rectangulaire dans la dimension temps-fréquence. La figure 1.2 montre le type de réseau rencontré, en regard de celui de la transformée de Fourier à court terme. Cependant, si l'on change l'échelle linéaire de temps par une échelle en nombre de cycles, et l'échelle linéaire de fréquence par une échelle logarithmique (échelle d'octaves), le pavage redevient rectangulaire: c'est la représentation cycles-octaves. Le plan d'information de la transformation cycle-octave peut s'avérer plus pertinent, ainsi qu'un plan intermédiaire, le plan temps-octave (transformée en voix). C'est néanmoins le plan temps-échelle qui est le plus utilisé comme représentation pour les signaux acoustiques.

1.4.3 analyses graduées et ondelettes

Un point de vue formel sur les transformations dépendant d'un facteur d'échelle provient de l'analyse fonctionnelle.

Dans l'espace des signaux d'énergie finie, on définit une *analyse graduée* par une suite croissante de sous-espaces fermés $V_j, j \in \mathcal{Z}$ telle que [62] [23]:

- l'intersection de tous les V_j est réduite à la fonction identiquement nulle, et sa réunion est dense (donc tout signal d'énergie finie peut être considéré comme la limite d'une suite de signaux appartenant aux V_j);
- un signal $x(t)$ appartient à V_j si et seulement si $x(2t)$ appartient à V_{j+1};
- si un signal appartient à V_0, alors pour tout décalage temporel entier, $x(t_k), k \in \mathcal{Z}$ appartient à V_0;
- il existe (au moins) une fonction $g(t)$ de V_0 telle que ses translatées $g(t-k), k \in \mathcal{Z}$ forment une base inconditionnelle de V_0. Une base inconditionnelle est telle que les coefficients du signal sur cette base sont uniques et forment une famille absolument sommable (pour la norme usuelle);

Intuitivement cela revient à obtenir une approximation de plus en plus précise d'un signal par des signaux appartenant aux espaces emboités, où chaque espace définit une résolution. De plus, on passe d'une résolution à la résolution supérieure (deux fois plus fine) par un changement d'échelle (contraction de facteur 2). Une fonction unique g permet de décrire par ses translatées temporelles toutes les fonctions du plus petit espace V_0 [59].

Une telle structure se retrouve dans le codage en sous-bandes utilisant des filtres miroir en quadrature [67]. La notion de résolution est inversement proportionnelle à celle de largeur de bande, et la construction des espaces emboités se fera donc à l'envers partant du plus grand vers le plus petit (en suivant ainsi le déroulement des algorithmes

d'analyse). Si un signal possède une bande passante α limitée (par exemple la demi-fréquence d'échantillonnage), on peut décomposer le spectre en deux bandes par des *filtres miroir*: un filtre passe-haut et un filtre passe-bas. Les deux sous-bandes ont bien sûr une largeur de bande moitié de la bande totale, et le signal initial $x(t)$ (appartenant à l'espace des signaux à bande passante limitée par α) devient la somme de deux signaux $x_1(t)$ et $x_2(t)$ dont la largeur de bande est limitée par $\frac{\alpha}{2}$. En ne considérant que x_1, il appartient à l'espace des signaux à bande limitée par $\frac{\alpha}{2}$. On peut réitérer le procédé et obtenir une nouvelle représentation de $x(t)$ par quatre signaux dont les largeurs de bande sont limitées par $\frac{\alpha}{4}$. En ne considérant encore que le signal de bande passante $\frac{\alpha}{4}$ dont la bande est la plus grave, on obtient une approximation moins fine de $x(t)$. Il suffit de réitérer le procédé pour construire une analyse graduée jusqu'à la résolution la plus grossière choisie, qui correspond à V_0. Le codage en sous-bande ne cherche évidemment pas à dégrader la résolution, mais profite du fait que le codage de 2^n signaux de largeurs de bandes $\frac{\alpha}{2^N}$ est plus avantageux que le codage d'un signal de bande passante α.

Un exemple plus formel d'analyse graduée se rapporte à l'approximation par fonctions spline. V_0 est le sous-espace constitué par les signaux $m-1$ fois (m entier) continuement dérivables tels que leur restriction à chaque intervalle $[k, k+1[, k \in \mathcal{Z}$ soit un polynôme de degré $\leq m$. Le signal $g(t)$ est la fonction *basic spline* d'ordre m.

En ajoutant des hypothèses sur $g(t)$ (r fois continuement dérivable, avec toutes les dérivées à décroissance rapide) on montre qu'il existe des fonctions $\psi(t)$ telles que:

$$\sqrt{2^j}[\psi(2^j t - k)]_{(j,k)\in\mathcal{Z}^2} \tag{1.89}$$

forment une base orthogonale de l'espace des fonctions de carré sommable sur $\mathcal{R}$.

Sans détailler la construction de cette famille de fonctions (qui dérivent toutes de l'ondelette d'analyse $\psi(t)$) on peut sommairement la décrire de la façon suivante. A partir du signal $g(t)$ on définit un signal $\phi(t)$ tel que ses translatées forment une base orthonormée de V_0. Les propriétés de l'analyse graduée permettent de déduire de ϕ une base orthonormée de V_1. En considérant le complément orthogonal de V_0 dans V_1, le signal $\psi(t)$ peut être construit à partir de ϕ.

Donc les ondelettes d'analyse (translatées/dilatées de ψ) forment des bases orthonormées des compléments orthogonaux de l'analyse graduée.

Si l'on réexamine le codage en sous-bandes à la lumière de l'ondelette ψ, on retrouve l'interprétation en banc de filtres. Le complément orthogonal dans chaque sous-bande est constitué par les signaux dont la bande passante est la demi-bande supérieure de la sous-bande considérée. Ainsi, la construction des compléments orthogonaux est analogue à celle de l'analyse graduée, si à chaque étape on conserve la demi-bande supérieure de la bande la plus grave, obtenue par le filtre miroir dans la bande considérée, plutôt que la bande la plus grave à chaque étape. La construction donne finalement un banc de N filtres adjacents, la largeur de bande du $N - i^{ième}$ valant $\alpha/2^{i+1}$.

Le second exemple entraîne la construction d'ondelettes sous forme de fonctions affines par morceaux, ou polynomiales par morceaux.

Des bases orthonormées d'ondelettes à support compact ont été construites, [23] et les premières applications au traitement des signaux acoustiques apparaissent [28].

1.4.4 algorithmes de calcul

Contrairement à l'analyse de Fourier, il n'existe pas aujourd'hui un algorithme de calcul rapide universellement répandu pour l'analyse en ondelettes. Cependant certaines régularités dans le calcul peuvent être exploitées et des travaux récents proposent des algorithmes efficaces, qui sont passibles d'une rapide diffusion.

Deux points de vue sont possibles:

- en considérant la transformation en ondelettes comme un filtrage par des filtres à réponse impulsionnelle finie, on peut utiliser les méthodes classiques de convolution, en particulier la convolution rapide, une variante de la représentation de Fourier à court terme. Chaque bande d'analyse utilise une transformation de Fourier rapide de taille différente, en fonction de la durée de l'ondelette d'analyse. Suivant le nombre d'échantillons de la réponse impulsionnelle, une implémentation efficace est également possible par le calcul de la convolution dans le domaine temporel: de nombreuses comparaisons entre ces procédés existent [14].

- en considérant maintenant la transformation en ondelettes comme un procédé dépendant d'un facteur d'échelle, on remarque que pour passer d'une bande d'analyse à la précédente, en octave, les calculs ont une forme identique et très simple: il y a deux fois plus d'ondelettes à prendre en compte sur la grille, mais ces ondelettes ont une largeur de bande moitié ce qui entraine une quantité de calcul identique à chaque échelle.

Le second point de vue a tout d'abord été présenté pour le codage de parole, et repris récemment spécifiquement pour la transformation en ondelettes [58]. Un algorithme, *l'algorithme à trous* procède des mêmes principes et montre une structure similaire [48].

Pour les filtres miroirs en quadrature, l'algorithme est le suivant. A partir du signal échantillonné x_n de bande passante α, on définit une suite emboitée de filtres en réduisant la largeur de bande et le nombre d'échantillons temporels par deux à chaque étape. Pour une étape, la séparation en deux sous bandes égales de largeur de bande $\frac{\alpha}{2}$ se fait à l'aide de deux filtres possédant h_n^1 et h_n^2 comme réponse impulsionnelle, avec la relation *miroir*:

$$\tilde{h}^1(\nu) = \tilde{h}^2(\alpha - \nu) \tag{1.90}$$

Après filtrage par h_n^1 et h_n^2, les signaux obtenus x_n^1 et x_n^2 peuvent être sous-échantillonnés d'un facteur deux. On montre que la reconstruction du signal x_n à partir de x_n^1 et x_n^2 peut être parfaite, en rééchantillonnant et en utilisant deux filtres interpolateurs possédant k_n^1 et k_n^2 comme réponse impulsionnelle, sous les conditions suivantes:

- h_n^1 est un filtre passe-bas à réponse impulsionnelle finie, possédant un nombre pair d'échantillons (par exemple une fenêtre d'analyse spectrale);

- les fonctions de transfert H^1 et H^2 vérifient $H^1(z) = H^2(-z)$, c'est à dire que les filtres d'analyse sont en quadrature;

- les filtres d'analyse sont en miroir: $\tilde{h}^1(\nu) = \tilde{h}^2(\alpha - \nu)$;

- les relations entre filtres d'analyse et de synthèse sont: $H^1(z) = K^1(z)$ et $H^2(z) = -K^2(z)$;

En utilisant maintenant la demi-bande inférieure, échantillonnée à une fréquence moitié, il est possible de réitérer le procédé avec une quantité de calculs identiques, puisque la plus grande durée des filtres mis en jeu se compense par le plus faible nombre d'échantillons.

Une structure pyramidale apparaît, par bande d'octave. Pratiquement, il peut être utile d'affiner la résolution par rapport au minimum requis par la théorie en prenant plusieurs bandes d'analyse par octave, par exemple pour une représentation visuelle de type spectrographique.

Du codage en sous-bandes de la parole découlent ainsi des méthodes pour l'implémentation d'une analyse en ondelettes orthogonales (les sous-bandes doivent posséder la propriété $\Delta\nu/\nu = C^{te}$).

Les figures 1.6 à 1.11 donnent des exemples d'analyse multirésolution. Une analyse en ondelette orthogonale à été implémentée par Filtres Miroir en Quadrature à réponse impulsionnelle finie et à phase linéaire. La conception des filtre relève de la méthode classique de fenêtrage, avec optimisation des coefficients obtenus pour satisfaire au conditions précédentes.

Les figures 1.6, 1.8, 1.10 représentent l'analyse à cinq résolutions différente, respectivement d'un segment de voyelle /a/, de fricative /s/, et un /t/, avec les premier cycles de voisement. En bas est porté le signal original.

Les figures 1.7, 1.9, 1.11 représentent également l'analyse, respectivement d'un segment de voyelle /a/, de fricative /s/, et un /t/, avec les premier cycles de voisement. Le signal du bas est le signal original, le second est l'analyse à la résolution la plus faible et les cinq suivants les signaux de détails additionnels pour passer d'une résolution à une résolution plus fine [86].

1.5 conclusion

Le point de vue adopté pour la représentation en temps et en fréquence du signal de parole par des méthodes non-paramétriques s'est limité à une catégorie particulière de méthodes qui utilisent une décomposition, formulée de façon continue puis discrète, du signal sur un ensemble de fonctions élémentaires données *a priori*. Les conditions sur la reconstruction exacte du signal, les différents plans d'information et les caractéristiques du pavage spectro-temporel choisi ont été discutés, suivant deux axes principaux: la transformée de Fourier à court terme et la transformée en ondelettes. Un point de vue particulier, celui d'une représentation discrète envisagée depuis les formes d'ondes élémentaires d'analyse a présidé au choix des résultats présentés, parmi le vaste ensemble rencontré dans de nombreux travaux: il ne s'agit pas ici d'un exposé exhaustif mais d'une sélection des éléments qui éclairent un propos particulier.Ainsi, la transformée de Fourier à court terme n'est pas généralement envisagée comme un développement en somme de formes d'ondes, et la représentation en ondelettes déborde considérablement du cadre étroit qui a été considéré.

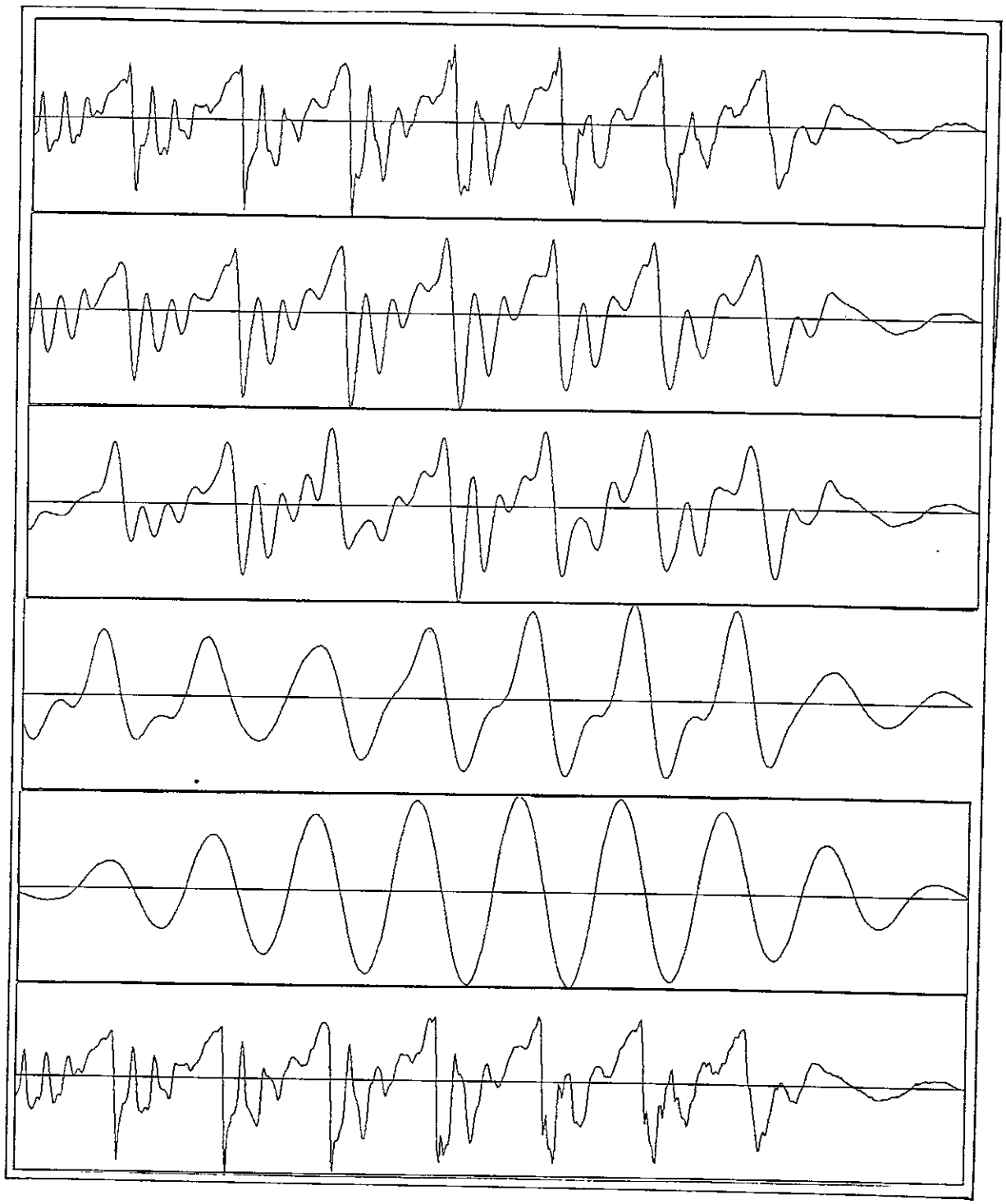

Figure 1.6: analyse multirésolution d'un /a/ par transformation en ondelettes orthogonales utilisant des filtres miroir en quadrature (voir texte).

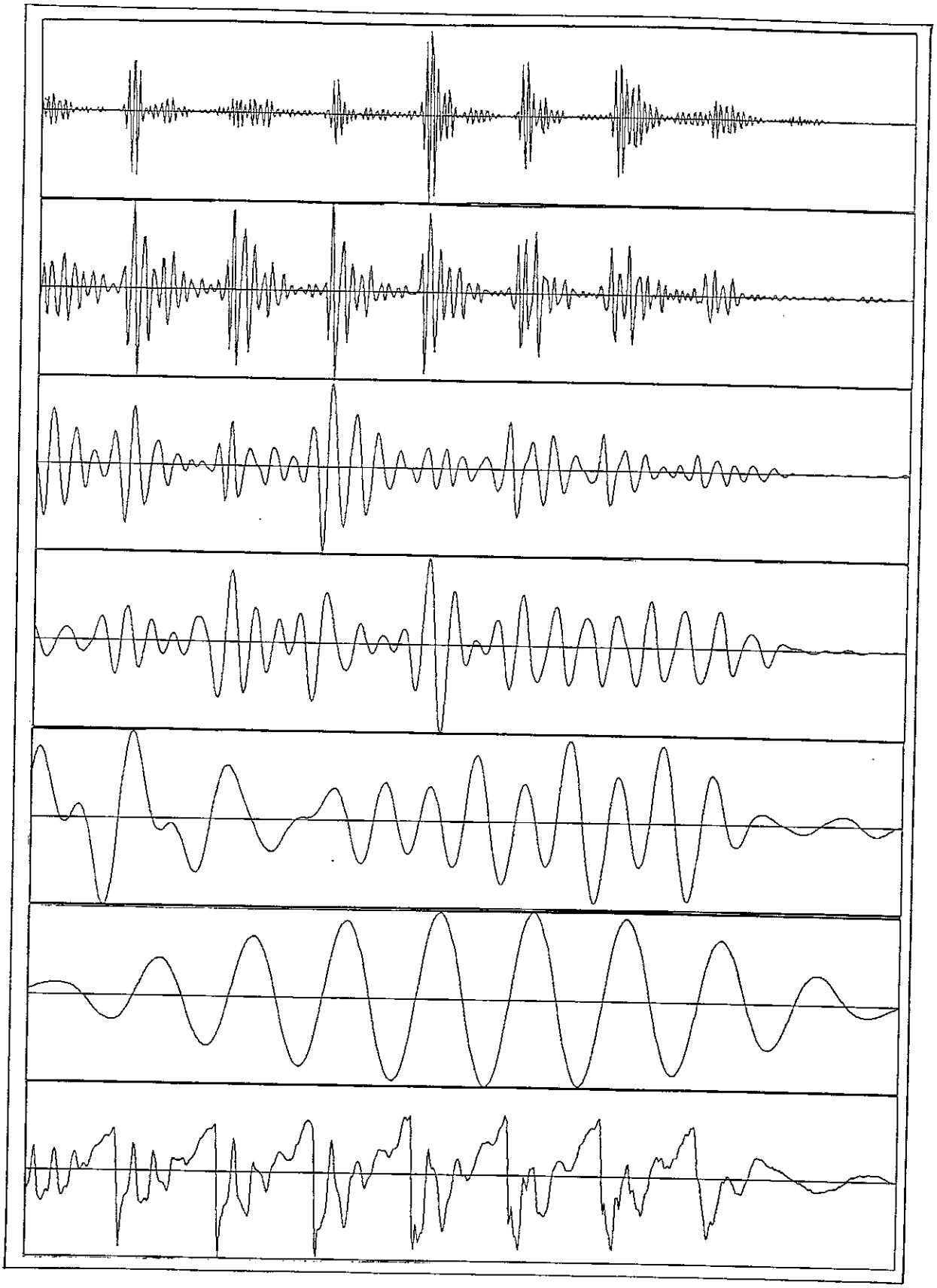

Figure 1.7: première résolution et détails additionnels d'un /a/ par transformation en ondelettes orthogonales utilisant des filtres miroir en quadrature (voir texte).

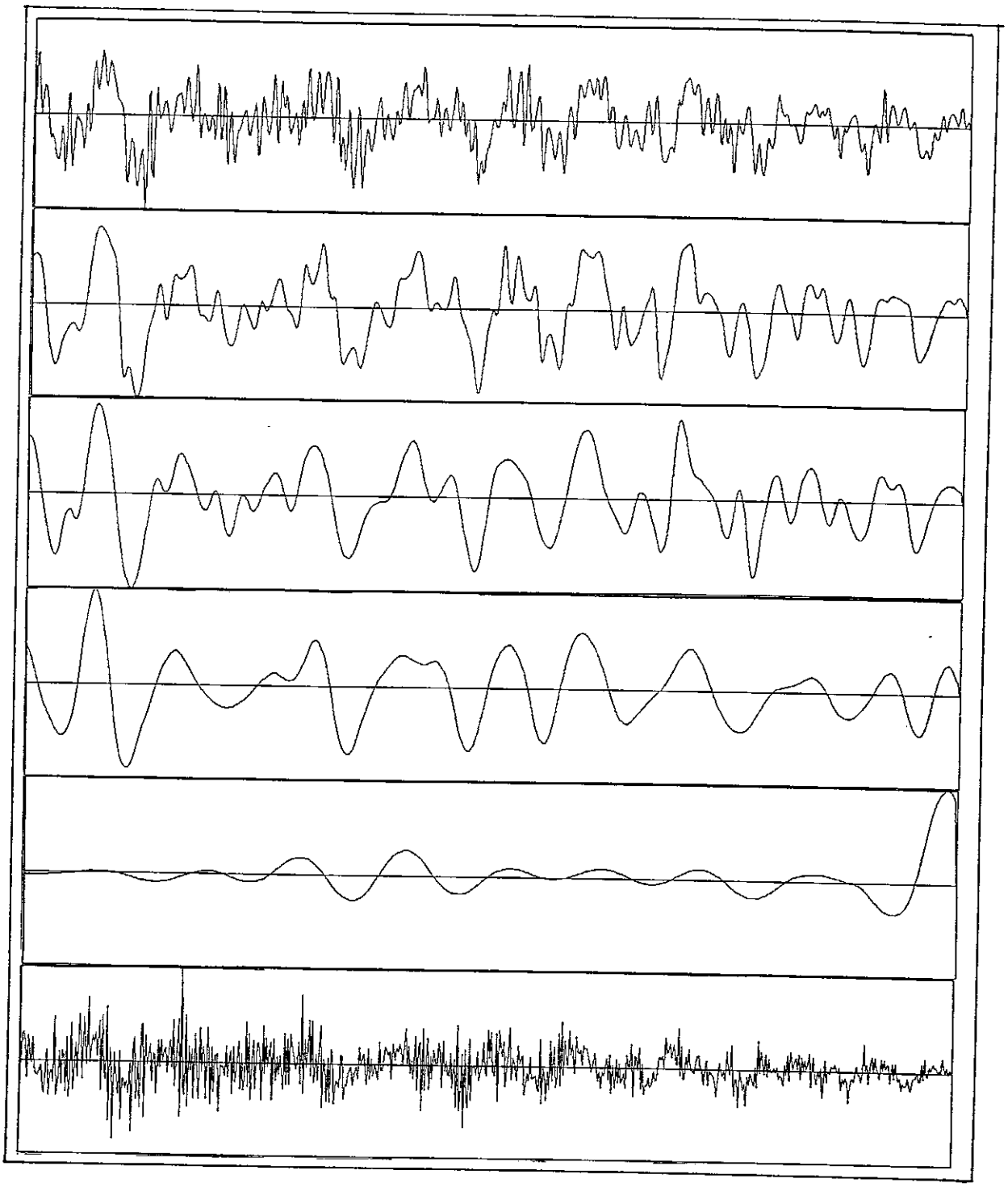

Figure 1.8: analyse multirésolution d'un /s/ par transformation en ondelettes orthogonales utilisant des filtres miroir en quadrature (voir texte).

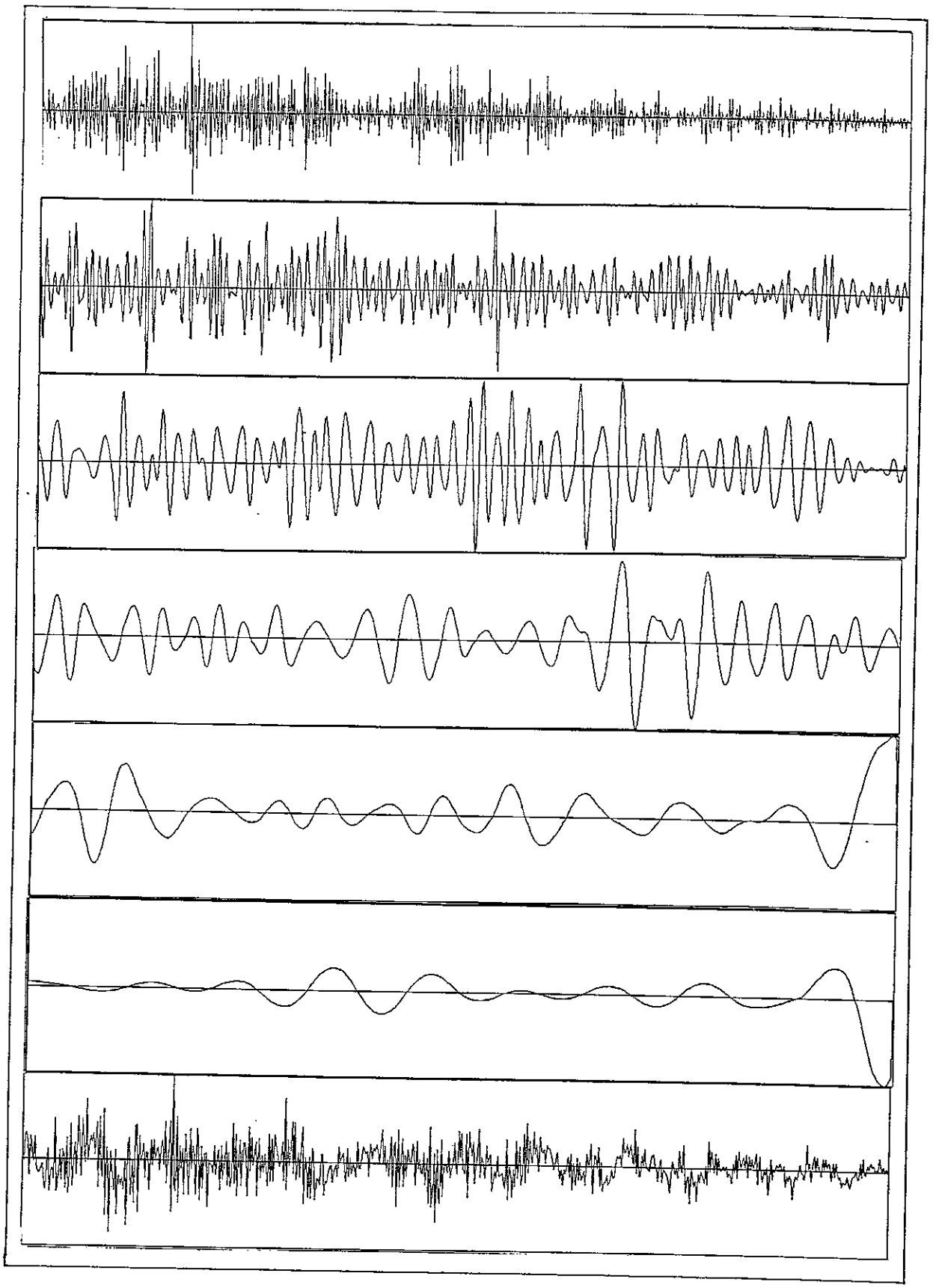

Figure 1.9: première résolution et détails additionnels d'un /s/ par transformation en ondelettes orthogonales utilisant des filtres miroir en quadrature (voir texte).

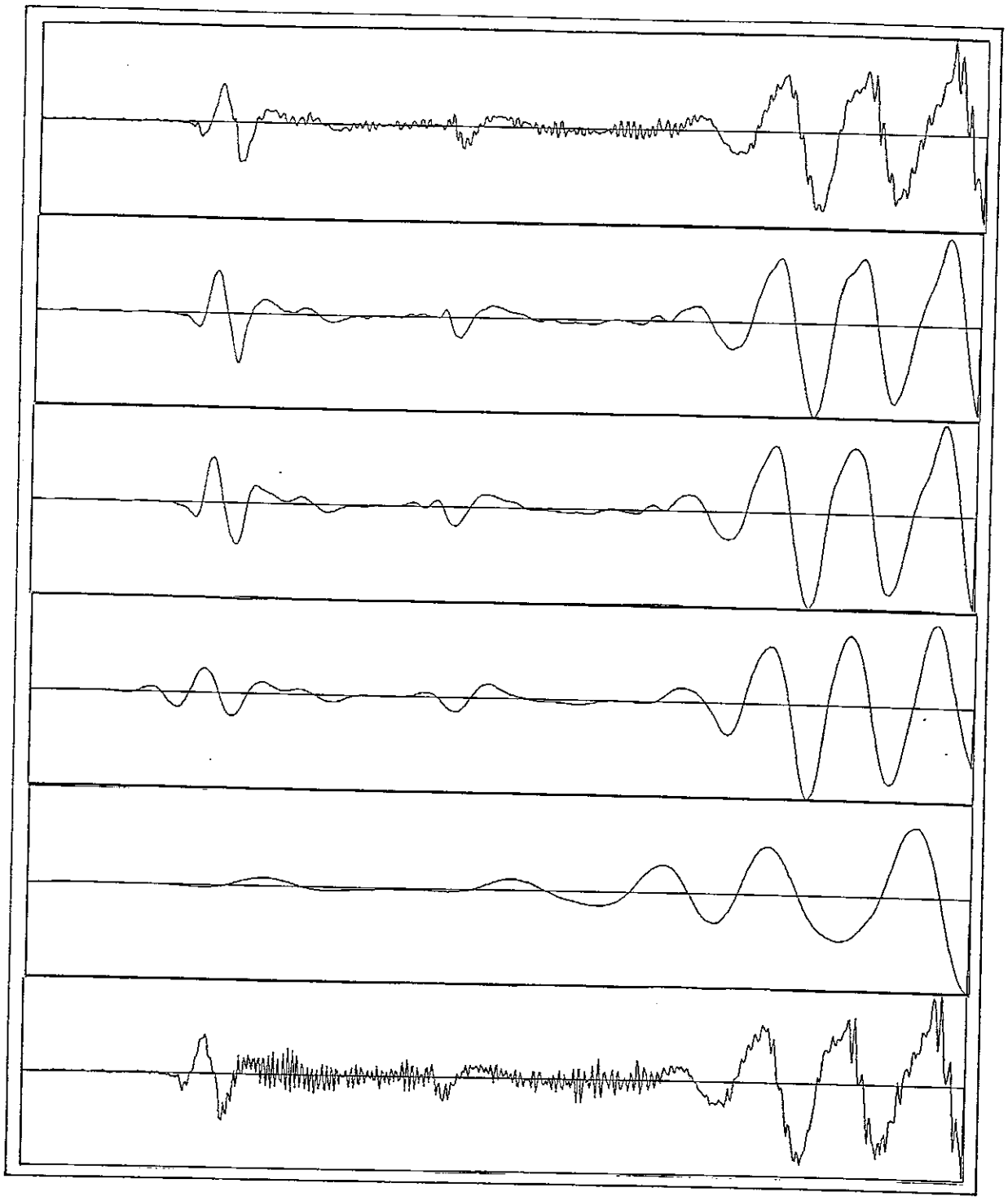

Figure 1.10: analyse multirésolution d'un /t/ par transformation en ondelettes orthogonales utilisant des filtres miroir en quadrature (voir texte).

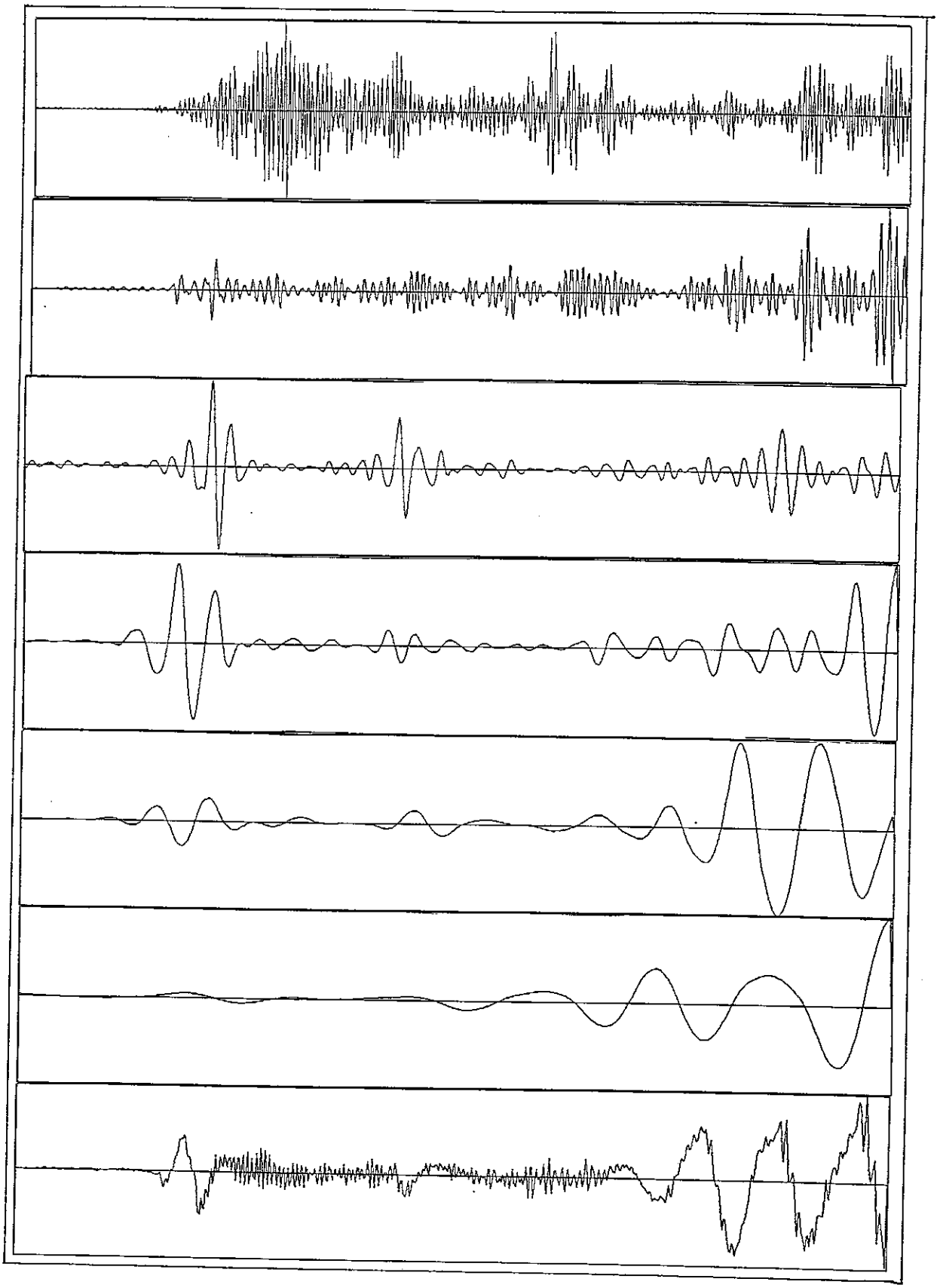

Figure 1.11: première résolution et détails additionnels d'un /t/ par transformation en ondelettes orthogonales utilisant des filtres miroir en quadrature (voir texte).

Ce type de méthodes se prête à trois interprétations complémentaires: l'interprétation en banc de filtres, à résolution constante ou à résolution relative constante, l'interprétation du traitement par blocs, et finalement l'interprétation en formes d'ondes élémentaires qui a été plus particulièrement considérée. Les trois interprétations offrent de multiples possibilités de traitement.

L'avantage offert par ces méthodes est de présenter un cadre formel puissant pour la représentation du signal: orthogonalité des formes d'ondes analysantes, réseaux discrets dans le plan d'information, reconstruction du signal.

Par contre, les coefficients obtenus ne possèdent pas directement de pertinence en tant que paramètres d'un modèle de production ou de perception de la parole. Il s'agit donc d'imposer des contraintes supplémentaires issues du domaine d'application: analyse, synthèse, transformation du signal par exemple.

Une évolution prometteuse est le choix de fonctions élémentaires et de réseaux qui ne sont plus fixés *a priori*, mais adaptés à la nature du signal analysé en tenant compte de contraintes perceptives ou de modèles de production. Les deux chapitres suivants exposent des réalisations dans cette direction.

Bibliographie Chapitre 1

Bibliographie Chapitre 1

[1] M. H. Ackroyd, 1970. *Short time spectra and time-frequency energy distribution* JASA, Vol. 50, No. 5, 1970.

[2] J. F. Allard, J. C. Valiere, R. Bourdier, 1988. *Broadband signal analysis with the smoothed pseudo-Wigner distribution* JASA, Vol. 83, No. 3, Mars 1988.

[3] J. B. Allen, 1977. *Short term spectral analysis, synthesis, and modification by discrete Fourier transform.* IEEE transaction on ASSP, Vol. ASSP-25, No. 3, Juin 1977.

[4] J. Arsac, J. C. Simon, 1960. *Représentation d'un phénomène physique par des sommes de translatées.* Annales de radio-électricité, tome XV, No. 61, Juillet 1960.

[5] R. Balian, 1981. *Un principe d'incertitude fort en théorie du signal ou en mécanique quantique.* Comptes-Rendus Académie des Sciences de Paris, t. 292, 1 Juin 1981.

[6] V. Bargmann, P. Butera, L. Girardello, J.R. Klauder, 1971. *On the completeness of the coherent states.* Reports on mathematical physics, Vol. 2, No. 4, 1971.

[7] M. J. Bastiaans, 1980. *Gabor's Expansion of a Signal into Gaussian Elementary Signals.* Proceedings of IEEE, Vol. 68, No. 4, Avril 1980.

[8] M. Baudry, B. Dupeyrat, 1975. *Analyse du signal vocal dans sa représentation amplitude/temps. Commande vocale d'un ordinateur en temps réel* 6ème Journées d'étude sur la Parole du GALF.

[9] C. Berthomier, 1976. *Représentation d'un signal dans le plan fréquence instantanée-temps.* Thèse d'état, Université Paris VI.

[10] P. Bertrand, 1983. *Représentation des signaux dans le plan temps-fréquence.* La recherche aérospatiale, Année 1983, No. 1(janvier-Février).

[11] J. Bertrand, P. Bertrand, 1988. *Time-frequency representation of broad-band signals.* Proceedings of IEEE-ICASSP-88.

[12] P. Bertrand, J. Bertrand, 1988. *Représentation temps fréquence des signaux à large bande.* La recherche aérospatiale, Année 1985, No. 5(Septembre-Octobre).

[13] L. Brillouin, 1959. *La science et la théorie de l'information* Masson, Paris, Réedité par les éditions Jacques Gabay, Paris, 1988.

[14] C. S. Burrus, T. W. Parks, 1985. *DFT/FFT and convolution algorithms, theory and implementation* John Wiley & sons, New York, 1985.

[15] F. Charpentier, 1988. *Traitement de la parole par analyse-synthèse de Fourier: applications à la synthèse par diphones.* Thèse E. N. S. T., Paris, 1 Juillet 1988.

[16] D. Chester, F. J. Taylor, 1984. *The Wigner distribution in speech processing applications* Journal of the Franklin institute, Pergamon Press, Vol. 318, No. 6, Décembre 1984.

[17] L. Cohen, C. A. Pickover, 1986. *A comparison of joint time-frequency distributions for speech signals* IEEE international symposium on circuits ans systems.

[18] J.M. Combes, A. Grossman et P. Tchamitchian (editeurs), 1989. *Wavelets, Time-frequency methods and phase space*, Springer-Verlag, Berlin, 1989.

[19] J. W. Cooley, J. W. Tukey, 1965. *A method for the machine calculation of complex Fourier series.* in *Digital signal processing.*, edité par L. R. Rabiner et C. M. Rader, IEEE press, New York 1972.

[20] R. E. Crochiere, S. A. Webber, J. L. Flanagan, 1976. *Digital coding of speech un sub-bands.* Proceedings of IEEE-ICASSP-76.

[21] R. E. Crochiere, 1980. *A weighted overlap-add method of short-time Fourier analysis/synthesis.* IEEE transaction on ASSP, Vol. ASSP-28, No. 1, Février 1980.

[22] I. Daubechies, 1987. *Orthonormal bases of compactly supported wavelets.* Rapport technique des *AT&T* Bell Laboratoties, Murray Hill, 1987.

[23] I. Daubechies, 1987. *Orthonormal bases of compactly supported wavelets.* Rapport technique des *AT&T* Bell Laboratoties, Murray Hill, 1987.

[24] I. Daubechies, 1987. *The wavelet transform, time-frequency localization and signal analysis.* Rapport technique des *AT&T* Bell Laboratoties, Murray Hill, 1987.

[25] C. Demars, 1987. *Représentations temps-fréquence, Eléments de bibliographie.* Notes et documents LIMSI, 87-2, Octobre 1987.

[26] C. Demars, 1988. *Représentations temps-fréquence et paramétrisation du signal de parole, Eléments de monographie.* Notes et documents LIMSI, 88-12, Juillet 1988.

[27] C. Demars, 1985. *Application des $\lambda^{2\cdot}$-distributions à la reconnaissance de la parole.* 14ème Journées d'Etude sur la Parole du G.A.L.F.

[28] C. Dorize, 1988. *Ondelettes discretes orthogonales et signaux acoustiques* Mémoire d'ingénieur en acoustique du C.N.A.M.

[29] G. Duncan, B. Yegnanarayana, H. A. Murthy, 1988. *Non-parametric formant estimation from the group delay function* Proceedings of speech '88, 7th FASE symposium, Edinbourg, Aout1988.

[30] B. Dupeyrat, 1975. *Reconnaissance de la parole. Méthode des passages par zéro du signal. Reconnaissance de voyelles isolées* Thèse de 3ème cycle, Université Paris VI.

[31] D. Esteban, C. Galand, 1977. *Application of quadrature mirror filters to split band voice coding schemes.* Proceedings of IEEE-ICASSP-77.

[32] J. L. Flanagan, 1972. *Speech analysis, synthesis and perception.* Springer-Verlag, Berlin.

[33] P. Flandrin, 1987. *Représentation temps-fréquence des signaux non-stationnaires..* Thèse d'état, Institut National Polytechnique de Grenoble.

[34] P. Flandrin, 1988. *Time frequency and time scale* IEEE Fourth Annual ASSP workshop on Spectrum estimation and Modeling, Minneapolis, Août 1988.

[35] D. H. Friedman, 1985. *Instantaneous-frequency distribution vs. time: an interpretation of the phase structure of speech.* Proceedings of IEEE-ICASSP-85.

[36] J. Fourier, 1822. *Théorie analytique de la chaleur* Firmin Didot, Paris, Réedité par les éditions Jacques Gabay, Paris, 1988.

[37] D. H. Friedman, 1987. *Formulation of a vector distance measure for the instantaeous-frequency distribution (IFD) of speech.* Proceedings of IEEE-ICASSP-87.

[38] D. Gabor, 1946. *Theory of communication.* Journal of the IEE No 93-1946, Londres.

[39] C. Galand, D. Esteban, 1983. *Design and application of parallel quadrature mirror filters (PQMF).* Proceedings of IEEE-ICASSP-83.

[40] G. Gambardella, 1971. *A contribution to theory of short-time spectral analysis with nonuniform bandwidth filters.* IEEE Transaction on circuit theory, Vol. ct-18, No 4, July 1971.

[41] P. Goupillaud, A. Grossmann, J. Morlet, 1984. *Cycle-octave and related transforms in seismic signal analysis.* Geoexploration, No. 23, Elsevier Sciences Publishers, Amsterdam.

[42] D. W. Griffin, J. S. Lim, 1984. *Signal estimation from modified short-time Fourier transform* IEEE transaction on ASSP, Vol. ASSP-32, No. 2, Avril 1984.

[43] A. Grossmann, J. Morlet, 1984. *Decomposition of functions into wavelets of constant shape, and related transforms.* Mathematics and physics, lectures on recent results, World Scientific Publishing Co., Singapour, 1985.

[44] A. Grossmann, J. Morlet, T. Paul 1985. *Transforms associated to square integrable group representation. Part I: General results.* J. Math. Phys. Vol. 26, No. 10, Octobre 1985.

[45] C. Hamon, 1988. *Synthèse de la parole par concaténation de formes d'ondes* 17éme Journées d'étude sur la parole de la Société Francaise d'Acoustique.

[46] F. J. Harris, 1978. *On the use of windows for harmonic analysis with the discrete Fourier transform.* Proceedings of the IEEE, Vol. 66 No 1., Janvier 1978.

[47] C. W. Helström, 1968. *An expansion of a signal into gaussian elementary signal.* IEEE transaction on information theory, Vol. 12, No. 1, Janvier 1968.

[48] M. Holschneider, R. Kronland-Martinet, J. Morlet, P. Tchamitchian, 1987. *A real-time algorithm for signal analysis with the help of the wavelet transform.* Proceedings 'Ondelettes, méthodes temps-fréquence et espace des phases', CIRM Luminy, 14-18 Décembre 1987, à paraître Springer-Verlag.

[49] J. C. Horvat, R. Houdas, 1970. *Représentation d'une fonction par une somme de translatées.* Thèse de troisième cycle, université Paris VI, 12 Juin 1970.

[50] A. J. E. M. Janssen, 1984. *Gabor representation and Wigner distribution of signals.* Proceedings of IEEE-ICASSP-84.

[51] P. Jardin, 1985. *Evaluation des performances d'une technique de Fourier glissant (analyse spectrale différentielle) au traitement des signaux de parole..* Thèse de 3ème cycle, Université Paris VI.

[52] K. Kodera, R. Gendrin, C. de Villedary 1978. *Analysis of time-varying signals with small BT values.* IEEE Trans. on ASSP, Vol ASSP-26, No. 1, Février 1978.

[53] R. Kronland-Martinet, J. Morlet, A. Grossmann, 1987. *Analysis of sound patterns through wavelet transforms.* International Journal of Pattern Recognition and Artificial Intelligence, vol. 1, 1987.

[54] R. Kronland-Martinet, 1988. *The use of the wavelet transform for the analysis, synthesis and processing of speech and music sounds.* Computer Music Journal, MIT Press, Vol. 12 No. 4, Décembre 1988.

[55] M. Kunt, 1981. *Traitement numérique des signaux.* Dunod, Paris.

[56] Madhu Sudan Gupta, 1975. *Definition of instantaneous frequency and frequency measurability.* American Journal of Physics Vol. 43, No. 12, Decembre 1975.

[57] P. Malliavin, 1982. *Intégration et probabilités, Analyse de Fourier et analyse spectrale.* Masson, Paris.

[58] S. G. Mallat, 1987. *A theory for multiresolution signal decomposition: the wavelet representation.* Rapport technique MS-CIS-87-22, University of pennsylvania, Philadelphie, Mai 1987.

[59] S. G. Mallat, 1987. *Multiresolution approximation and wavelets.* Rapport technique MS-CIS-87-87, University of pennsylvania, Philadelphie, Septembre 1987.

[60] L. Mandel, 1974. *Interptretation of instantaneous frequency.* American Journal of Physics Vol. 42, Octobre 1974.

[61] Y. Meyer, S. Jaffard, O. Rioul, 1987. *L'analyse par ondelettes.* Pour la science, Septembre 1987.

[62] Y. Meyer, 1987. *Les ondelettes.* Séminaire "Ondelettes, méthodes temps-fréquence et espace des phases", Marseille, 14-18 Décembre 1987.

[63] J. A. Moorer, 1977. *Signal processing aspects of computer music: a survey.* Proceedings of the IEEE, Vol. 65, No. 8, Aout 1977.

[64] J. Morlet, 1984. *Introduction à la représentation cycle octave.* Tract technique de l'O.R.I.C., Rueil-Malmaison.

[65] J. Morlet, 1983. *Sampling theory and wave propagation.* Issue in acoustic signal/image processing and recognition, NATO ASI series, Springer-Verlag, Berlin, 1983.

[66] L. K. Montgomery, I. S. Reed, 1967. *A generalisation of the Gabor-Helstrom transform.* IEEE transaction on Information Theory, Vol. IT-13, Avril 1967.

[67] H. J. Nussbaumer, 1983. *Complex quadrature mirror filters.* Proceedings of IEEE-ICASSP-83.

[68] A. V. Oppenheim, R. W. Schafer, 1975. *Digital signal processing.* Prentice-Hall, Englewood Cliffs, New Jersey.

[69] A. V. Oppenheim, 1978. *Applications of digital signal processing.* Prentice-Hall, Englewood Cliffs, New Jersey.

[70] A. Papoulis, 1977. *Signal analysis.* MacGraw-Hill, New York.

[71] C. R. Patisaul, J. C. Hammet Jr, 1975. *Time frequency resolution experiments in speech analysis and synthesis* JASA, Vol. 58, No. 6, Décembre 1975.

[72] B. Picinbono, 1977. *Eléments de théorie du signal..* Dunod, Paris.

[73] M. R. Portnoff, 1981. *Short-time Fourier analysis of sampled speech.* IEEE transaction on ASSP, Vol. ASSP-29, No. 3, Juin 1981.

[74] M. R. Portnoff, 1976. *Implementation of the digital phase vocoder using the fast Fourier transfom.* IEEE transaction on ASSP, Vol. ASSP-24, No. 3, Juin 1976.

[75] M. R. Portnoff, 1980. *Time-frequency representation of digital signals and systems based on short-time Fourier analysis.* IEEE transaction on ASSP, Vol. ASSP-28, No. 1, Février 1980.

[76] M. R. Portnoff, 1981. *Time-frequency modification of speech based on short-time Fourier analysis.* IEEE transaction on ASSP, Vol. ASSP-29, No. 3, Juin 1981.

[77] M. R. Portnoff, 1979. *Magnitude-phase relationships for short-time fourier transforms based on Gaussian analysis windows* Proceedings of IEEE-ICASSP-79.

[78] L. R. Rabiner, B. Gold, 1975. *Theory and application of digital signal processing* Prentice-Hall, Inc., Englewoods Cliffs, New Jersey.

[79] A. W. Rihaczek, 1968. *Signal energy distribution in time and frequency* IEEE transaction on Information Theory, Vol. IT-14, No. 3, Mai 1968.

[80] J. Shekel, 1953. *"Instantaneous" Frequency.* Proceedings of the I. R. E., Avril 1953.

[81] L. Schwartz, 1966. *Théorie des distributions.* Hermann, Paris.

[82] H. F. Silverman, Y. T. Lee, 1987. *On the spectrographic representation of rapidly varying speech* Computer speech and language, Vol. 2, No. 2, Juin 1987.

[83] D. Slepian, H. O. Pollak 1961. *Prolate sheroidal wave function, Fourier analysis and uncertainty* Part I, II, The Bell system technical journal ,Janvier 1961, Part III, The Bell system technical journal ,Juillet 1962.

[84] M. Turner, 1986. *Texture discrimination by Gabor functions* Biological Cybernetics, Vol. 55, 1986.

[85] J. Ville, 1948. *Théorie et applications de la notion de signal analytique.* Câbles et transmissions, Vol. 2, 1948.

[86] E. Valence, C. d'Alessandro, 1989. *Transformation en ondelettes orthogonales et filtres miroirs en quadrature.* Notes et documents LIMSI, à paraître.

[87] S. J. Walsh, P. M. Clarkson, 1987. *Speech enhancement using modelling and replacement of the instantaneous phase signal* Digital signal processing-87, V. Cappelliniand A. G. Constantinides (editeurs), Elsevier Science Publisher, Amsterdam, 1987.

[88] N. Wiener, 1930. *Generalized harmonic analysis* Acta Mathematica, vol. 55.

[89] W. Wokurek, F. Hlawatsch, G. Kubin, 1987. *Wigner distribution analysis of speech signals* Digital signal processing-87, V. Cappellini et A. G. Constantinides éditeurs, Elsevier Science Publishers.

[90] J. E. Youngberg, S. F. Boll, 1978. *Constant-Q signal analysis and synthesis* Proceedings of IEEE-ICASSP-78.

[91] B. Yegnanarayana, D. K. Saika, T. R. Krishnan, 1984. *Significance of group delay functions in signal reconstruction from spectral magnitude or phase* IEEE transaction on ASSP, Vol. ASSP-32, No. 3, Juin 1984.

Chapitre 2

SYSTEME AUDITIF ET FORMES D'ONDES ELEMENTAIRES

2.1 perception auditive et analyse acoustique

2.1.1 introduction

Référence obligée des systèmes d'analyse du son, l'oreille reste le récepteur privilégié des actes de parole et son fonctionnement porte une part de responsabilité, au même titre que l'appareil vocal, dans la conformation du signal de parole.

Les méthodes d'analyse du signal acoustique par la machine se situent principalement dans le cadre de la théorie du signal [3]. Néanmoins leur succès est lié d'une part à l'existence d'algorithmes de calcul efficaces, et d'autre part aux analogies que l'on peut relever avec le fonctionnement de l'appareil auditif humain. L'oreille reste l'étalon de tout système de perception artificielle.

Historiquement, le traitement automatique de la parole a importé un certain nombre de concepts issus d'études perceptives, en commençant par des contraintes psycho-acoustiques, puis plus récemment en incorporant des modèles sophistiqués du système auditif périphérique et de la perception de la parole[19] [42] [60]. Cette introduction apporte une amélioration indiscutable, pour des systèmes de reconnaissance automatique, dans les situations où la comparaison entre les performances humaines et celles des méthodes classiques de traitement du signal est le plus défavorable: en présence de bruit par exemple.

Ce chapitre débute par la description des résultats classiques de psycho-acoustique et de perception de la parole, puis de modèles du système auditif périphérique. Le propos est d'éclairer les relations entre ces données auditives et les formalismes mathématiques du chapitre 1, afin de soumettre un système de représentation en formes d'ondes élémentaires qui rende compte dans une certaine mesure de l'organisation perceptive, et qui forme un outil pour le traitement acoustique du signal de parole.

Cette décomposition prendra la forme générale d'une somme de fonctions élémentaires; une reconstruction aussi exacte que possible du signal à partir de sa description en formes d'ondes élémentaires est indispensable. On devra également retrouver de façon assez directe certains phénomènes observés en perception auditive dans les paramètres des formes d'ondes élémentaires, par des procédés comparables dans une certaine mesure

à ceux mis en évidence dans le système auditif périphérique.

2.1.2 psychologie et physiologie de la perception auditive

Les études sur l'audition s'appuient sur deux piliers: l'analyse du fonctionnement du système auditif, en particulier de *l'oreille* au sens de *système auditif périphérique*, et la psychologie de la perception auditive [31].

Le premier domaine collecte les données physiologiques et essaye éventuellement de construire des modèles reproduisant certains traits du fonctionnement de l'oreille. Le système auditif ne se limite pas à sa composante périphérique. Cependant la connaissance des processus du système nerveux central est beaucoup moins avancée, alors que les données disponibles sur les signaux issus du système périphérique, au niveau du nerf auditif, sont assez abondantes. Des méthodes expérimentales éprouvées existent depuis une quarantaine d'années. Toute l'information acoustique disponible pour les centres supérieurs transitant par le nerf auditif, celui-ci représente un lieu de choix pour comprendre l'analyse acoustique effectuée par le système auditif.

L'interaction réciproque entre les traitements centraux et les traitements périphériques se trouve fortement atténuée, ou supprimée par les conditions expérimentales habituelles (animaux anesthésiés). Les processus actifs d'adaptation restent inaccessibles pour ce type d'expérience, et seront négligés ici dans la définition d'un système initial assez simple. Le nerf auditif représente une des frontières sur lesquelles bute actuellement l'expérimentation physiologique. En ignorant les données des centres supérieurs, et en négligeant leur action sur le système périphérique, il s'agit de se situer ici au niveau de l'analyse plus que de la perception, même si l'on suppose en retrouver les propriétés. Cette position doit être comprise avec toutes les réserves précédentes: la vision cybernétique d'un centre d'analyse auditive puis d'un centre de traitement est sans doute très largement éloignée de la réalité, mais constitue encore une approximation utile pour notre propos.

La psychologie de la perception traite du signal à un niveau de représentation supérieur [25]. Il peut être commode de distinguer ici arbitrairement psycho-acoustique et perception de la parole proprement dite. Une façon de tracer une frontière entre les deux domaines se base sur les matériaux de test. Le premier domaine utilise des objets sonores élémentaires (clics, sinusoïdes, bruits, bouffées de sons purs ...) et déduit des lois sur la perception de paramètres acoustiques simples. Le second considère des objets complexes, segments phonétiques par exemple, dont la perception implique des mécanismes et des approches variées: association linguistique, jugement esthétique, ... L'interprétation des résultats d'expériences devient alors particulièrement délicate, et les modèles produits plus difficiles à situer et à justifier. Les méthodes de psychologie expérimentale forment l'essentiel de l'outil de recherche psycho-acoustique: les connaissances physiologiques sont insuffisantes pour rendre compte ou expliquer entièrement les processus mis en jeu à ce niveau.

Les deux sources de connaissances disponibles sur l'audition humaine fournissent donc des résultats de nature différente, obtenus par des procédés expérimentaux différents. Des relations existent entre les deux domaines, mais la compatibilité entre modèles physiologiques et psychologiques peut s'avérer problématique [81].

2.1.3 analyse auditive et analyse acoustique

La recherche des caractéristiques spectro-temporelles de l'analyse effectuée par le système auditif peut emprunter deux voies.

L'étude *externe*, ou psycho-acoustique, dégage des traits perceptivement pertinents dans le signal de parole, traits acoustiques pour chaque catégorie phonétique par exemple. Une évaluation des performances souhaitables, comme les bandes passantes du filtrage auditif ou l'échelle de hauteurs, ou bien une validation des résultats en analyse-synthèse relèvent de cette étude. Cependant, l'intervention des centres de traitement supérieurs impose aux modèles psycho-acoustiques une forme assez éloignée d'une représentation spectro-temporelle du signal. De plus, des sensations quantifiables par le sujet ne se rencontrent que pour des données dont l'ordre de grandeur est compris dans certaines limites. La perception globale de phénomènes et l'interaction entre plusieurs traitements rendent délicate l'étude à un niveau de détail fin de l'analyse acoustique humaine par des procédés externes.

L'étude *interne*, physiologique, est actuellement limitée au système périphérique, et ne se conçoit que chez l'animal anesthésié ou sur des cadavres [9]. La validation des résultats expérimentaux de l'étude interne s'effectue par corrélation avec ceux de l'étude externe. L'analyse cochléaire de signaux vocaliques par exemple essayera d'exhiber des formants. Parmi la masse de données disponibles, seules celles qui se rapportent à des faits perceptifs seront retenues. Ces données sur les processus de traitement périphérique peuvent faire directement l'objet de modèles informatiques du traitement acoustique auditif.

2.2 psycho-acoustique et perception de la parole

2.2.1 psycho-acoustique

sensation de hauteur

La hauteur perçue est reliée au taux de répétition de la forme d'onde d'un son. La fréquence fondamentale présente donc une contribution dominante, mais pas unique, puisqu'il faut aussi considérer l'intensité, le spectre et son évolution, la durée du son et son enveloppe, la présence d'autres sons.

La hauteur perçue d'un son, lorsqu'il est possible d'en percevoir une, comme par exemple lors d'une voyelle, répond approximativement à une échelle logarithmique en fonction de la fréquence de ce son. La notation musicale est ainsi la première échelle de notation mélodique qui rende compte de la perception. Cependant son rapport à l'échelle physique des Hz est complexe: il est impossible d'accorder correctement un instrument à clavier en doublant la fréquence en Hz pour passer d'une octave à l'autre.

L'échelle psycho-acoustique des *mels* représente la perception des hauteurs:

$$f_{mels} = 2595 \log_{10}(1 + \frac{f_{Hz}}{700}) \tag{2.1}$$

Cette échelle est construite pour doubler la sensation de hauteur lorsque l'on double la fréquence exprimée en *mels*.

La perception des hauteurs, d'une remarquable finesse, semble faire intervenir plusieurs types de traitements simultanés, et offre à l'investigation psycho-acoustique un champ d'expérimentation vaste, fécond et riche en paradoxes:

- fondamental absent, ou hauteur virtuelle: un son possédant plusieurs partiels, multiples d'une même fréquence f_0, possède une hauteur perçue à cette fréquence alors qu'aucune énergie ne s'y trouve;
- décalage de hauteur virtuelle: si l'on décale chaque composante du son précédent de $n\ Hz$, le son n'est plus exactement harmonique mais on perçoit tout de même une hauteur entre f_0 (la différence entre deux composantes) et $f_0 + n$;
- écoute analytique/écoute synthétique: la hauteur d'un son complexe semble évoluer différemment suivant le mode et l'intension d'écoute;
- hauteur de répétition: un son répété avec un décalage temporel Δt assez court présentera une hauteur de $\frac{1}{\Delta t}$, quelque soit sa composition spectrale (un bruit par exemple);
- sons paradoxaux: dont la hauteur semble croître et décroître simultanément [65];

L'explication psychologique et physiologique de la sensation de hauteur reste un problème fondamental de l'audition.

sensation de force

La sensation de force d'un son répond à une sensibilité différentielle relative (seuil différentiel d'intensité rapporté à l'intensité) constante, résultat souvent dénommé *loi de Weber*. Pour des zones de niveau et de fréquence moyens, la sensation varie approximativement comme le logarithme de l'excitation, en rendant égaux tous les échelons de sensation, résultat souvent dénommé *loi de Fechner*.

En fonction de leur fréquence, des sons d'intensité physique constante seront perçues avec des intensités différentes. Des courbes d'iso-sonie, le long desquelles la sensation d'intensité reste constante en fonction de la fréquence, décrivent ce phénomène depuis le seuil d'audition jusqu'au seuil de douleur. Ces courbes tracées grâce à des signaux élémentaires (sinusoïdes) ne restent pas forcément valides dans un contexte différent: écoute musicale, parole.

Plusieurs échelles logarithmiques permettent de mesurer l'intensité physique d'un son en *décibels* [50]:

- dB_{spl}: le niveau de référence est de $0.0002 dynes/cm^3$;
- dB_{sl}: le niveau de référence est le plus doux son pur audible;
- dBA: qui intègre les courbes d'iso-sonie;

Une échelle psycho-acoustique, l'échelle des *sones*, permet de mesurer la force perçue d'un son:

$$S_{sones} = C \times p^{0.6} \tag{2.2}$$

où p représente la pression et C une fonction de la fréquence.

filtrage psycho-acoustique et bandes critiques

L'appareil auditif procédant à une sorte d'analyse spectrale, il s'agit de déterminer les caractéristiques des filtres supposés. Les résultats physiologiques et psychologiques diffèrent assez nettement sur ce point. La psychologie introduit, en fonction de la fréquence, des largeurs de bande critiques. Suivant l'intervalle fréquentiel entre deux sons purs, ou la largeur de bande d'un bruit par exemple, la sensation d'intensité semble répondre à une addition des *intensités physiques*, au sein de la même bande critique, ou des *intensités perçues*, dans deux bandes critiques distinctes. La sensation de battement également ne semble possible que si les deux sons purs tombent dans la même bande critique.

Une échelle fréquentielle, l'échelle de Bark, liée à la notion de bande critique a été définie [84]. Un approximation [78] de cette échelle est:

$$f_{Hz} = 600sh(z_{Bark}/6) \tag{2.3}$$

Pour un banc de filtre en échelle de Bark, l'amplitude du gain G avec une largeur de bande de 1 Bark à -3 dB du sommet est donné par la formule:

$$10\log_{10} G(z_{Bark}) = 7.00 - 7.5(z_{Bark} - 0.215) - 17.5[0.196 + (z_{Bark} - 0.215)^2]^{\frac{1}{2}} \tag{2.4}$$

L'usage de l'échelle de Bark est très répandu dans les systèmes de traitement automatique de la parole. Dans des modèles du système auditif périphérique ce genre de filtres est parfois préféré aux filtres plus proches des données physiologiques.

effet de masque

Le seuil de perception d'un son, en présence d'un autre son, dépend de l'intensité et de la fréquence de ce dernier: il produit un effet de masque [36].

Des courbes résument, pour des sons purs, l'effet de masque produit sur des sons de fréquences diverses par d'autres sons d'amplitudes et d'intensités variables. On peut en tirer les conclusions suivantes:

- l'effet de masque est maximum lorsque les fréquences du son masqué et du son masquant sont voisines;
- l'effet de masque croît moins vite que le niveau du son masquant;
- l'effet de masque des fréquences graves sur les fréquences aigües est plus important que celui des fréquences aigües sur les fréquences graves;

sons subjectifs

L'oreille peut percevoir des sons alors qu'aucune énergie n'est présente dans la région spectrale correspondante du signal acoustique: ce sont les sons subjectifs.

L'oreille interne semble le lieu de formation des sons subjectifs. On peut les classer en deux catégories: harmoniques subjectifs et sons de combinaison. Leur production relève de mécanismes non-linéaires analogues à une distorsion.

Les premiers apparaissent lorsqu'un son pur est présenté avec une intensité assez importante: pour une fréquence de $1000Hz$ une intensité d'au moins $50dB_{sl}$ au dessus du seuil de perception est requise.

Les sons de combinaisons constituent deux classes: les sons différentiels et les sons de sommation. Les différentiels produits par deux sons purs f_1 et f_2 présentent la forme suivante:

$$f_1 - n(f_2 - f_1) \tag{2.5}$$

Les termes $f_1 - f_2$ (différentiel quadratique) et $2f_1 - f_2$ (différentiel cubique) dominent. Le son de sommation, $f_1 + f_2$ est plus difficilement perceptible.

2.2.2 perception de la parole

La perception de la parole ne se déduit pas seulement des résultats obtenus en psycho-acoustique. Le long apprentissage des sons de parole, dans une culture et à une époque donnée (accent, vocabulaire) , la fonction première de la parole comme transmission d'information linguistique (association linguistique), les relations au corps parlant (théorie motrice) se combinent et compliquent notablement les études perceptives [24].

L'étude perceptive permet d'évaluer l'importance des structures acoustiques présentes dans le signal de parole. [62]. Elle présente un point de passage obligé, même pour les modèles de la physiologie auditive: en effet on sera tenté de rechercher dans la masse d'observations disponibles uniquement celles qui semblent les plus significatives au niveau *macroscopique* de l'audition.

de la production à la perception

Comme la plupart des phénomènes sonores naturels, la parole se rapporte étroitement au système qui l'a produite: la fonction essentielle de la perception des sons est de permettre l'identification des processus qui les ont produits.

Les indices perceptifs qui permettent de discriminer des entités phonétiques se rattachent en grande partie à leur mode de production, tout comme les indices acoustiques.

La *théorie motrice* resserre encore les liens entre production et perception, en postulant que la perception d'une entité phonétique passe entièrement par l'identification (inconsciente) du geste articulatoire qui l'a produite.

Même sans adopter complètement la théorie motrice, la recherche des gestes articulatoires dans le signal acoustique semble une approche fertile pour la perception de parole. Un jeu d'événements articulatori-phonétiques [2] a été construit pour la segmentation temporelle du signal de parole.

perception des sons vocaliques

La perception des sons de type vocalique (voyelles orales et nasales, semi-voyelles, liquides, occlusives nasales) repose sur l'identification de la conformation et/ou de l'évolution du conduit vocal. Le conduit vocal est caractérisé par un ensemble de formants et d'anti-formants (le chapitre 3 détaille ces notions). Dès les années cinquante, des études en synthèse de la parole ont prouvé que seulement deux formants permettent

de constituer le système vocalique complet d'une langue (l'Anglais en l'occurence), en initiant ainsi les recherches sur les *premier et second formant effectifs* et l'intégration à large bande [73].

L'étude psycho-acoustique des paramètres formantiques reste un enjeu de tout premier plan auquel sont confrontées toutes les théories de perception de la parole.

Le fait dominant dans la perception des spectres vocaliques stationnaires (voyelles longues par exemple) est la sensibilité très importante aux formants et l'indifférence aux régions inter-formantiques. Les arguments en défaveur de l'importance des formants dans les représentations auditives existent cependant, en particulier dans le cas de voix murmurée, ou de voyelles nasales pour lesquels la région grave du spectre semble occulter les paramètres formantiques [16].

Dans la parole continue la stationnarité des sons vocaliques est illusoire dès que la vitesse de prononciation est élevée. Suivant la position des voyelles et leur contexte, les spectres à court terme que l'on peut mesurer ont une forte tendance à ne pas ressembler à ceux obtenus pour des sons stationnaires, mais sont interprétés de façon identique par l'auditeur. D'autres caractéristiques que les caractéristiques formantiques interviennent donc, en particulier la durée des segments vocaliques, et la forme des évolutions spectrales, liées aux gestes articulatoires.

D'autre part, la perception semblable de spectres issus de systèmes aussi différents quantitativement qu'un conduit vocal masculin, féminin ou enfantin fait apparaître une invariance perceptive de forme, ou de rapport entre les grandeurs formantiques, plutôt qu'une sensibilité à des valeurs absolues.

perception des sons fricatifs

La perception des sons fricatifs (voisés ou non) se rapporte d'une part aux caractéristiques statiques du bruit (son centre de gravité en particulier) et d'autre part aux caractéristiques dynamiques (l'évolution de part et d'autre du phonème, comme indication du lieu d'articulation).

Des formants peuvent ainsi être visibles et sensibles dans un segment fricatif, en fonction du contexte vocalique.

perception des sons plosifs

Les plosives sont perçues grâce aux caractéristiques de l'explosion qui suit l'occlusion, et à l'indication du lieu d'articulation que révèlent les évolutions formantiques de part et d'autre de la consonne. Le retard d'apparition de la vibration vocalique après l'explosion (délai d'établissement vocalique), la durée de l'occlusion, sont également des indications d'importance. L'appréciation de ces durées peut être extrêmement fine, de l'ordre de quelques millisecondes.

L'étude des plosives a conduit à la théorie de la *perception catégorielle*: lorsque l'on fait varier continuement certains paramètres de plosives synthétiques, l'auditeur ne perçoit systématiquement qu'un phonème, puis brutalement un autre, mais pas d'intermédiaire. Cette théorie fait toujours l'objet de controverses [12] [59].

scènes auditives

L'être humain présente de remarquables capacités de détection d'un signal de parole, même dans une ambiance acoustique très défavorable. Le célèbre effet de *cocktail party*, poursuite d'une conversation plongée dans un bruit de fond bien supérieur en intensité, démontre cette capacité. Par analogie avec l'analyse de scènes visuelles, les regroupements de *flux acoustiques* (acoustic streaming) sont étudiés en introduisant le concept de scènes auditives [13] [56] .

Ce regroupement des objets acoustiques s'effectue grâce à des critères de cohérence interne, dont les plus pertinents semblent être la hauteur perçue [58], le délai d'apparition et le taux de synchronisation. De riches problèmes sur l'organisation mentale de l'information acoustique et sur les corrélations possibles avec l'analyse issue de modèles physiologiques peuvent être envisagés par cette approche.

2.2.3 conclusions

La revue de quelques résultats de psycho-acoustique fournit un ensemble de contraintes pour une représentation spectro-temporelle.

L'analyse fréquentielle auditive se déduit de divers tests de masquage et se résume par une échelle définie en 2.3, précieux résultat pour tout système prétendant représenter la perception.

L'aspect d'intensité de cette analyse conduit également à une échelle, définie en 2.2, qui peut être utilisée avec profit pour le traitement de la parole.

Des phénomènes non-linéaires (sons subjectifs), ainsi que des phénomènes d'interaction spectro-temporelle locale (masquage), de regroupement (streaming), d'intégration large bande ne peuvent pas directement se traiter au niveau de la représentation spectro-temporelle.

Les traits perceptifs pertinents appellent des dimensions spectro-temporelles variées, depuis l'explosion des plosives et l'analyse de la mélodie jusqu'au centre de gravité d'un bruit ou aux formants.

Ces données de nature très diverse doivent être accessibles simplement sur les paramètres issus de la représentation, ou en étudiant les relations au sein d'un ensemble de ces paramètres.

2.3 système auditif périphérique

2.3.1 anatomie du système auditif

Le système auditif humain peut se décomposer en *système auditif périphérique*, et *système auditif central*. A ces deux composantes géographiquement séparées s'ajoute le nerf auditif, qui relie l'oreille interne au tronc cérébral [38] [64]. Cette décomposition ne représente que très grossièrement une décomposition fonctionnelle: chaque composante joue un rôle passif et actif dans la manipulation des signaux acoustiques, en relation avec les autres composantes.

Le système auditif périphérique est constitué d'une paire *d'oreilles*. L'oreille elle même se décompose en trois parties: l'oreille externe, l'oreille moyenne et l'oreille in-

terne. Le premier schéma de la figure 2.1 présente l'anatomie de l'oreille. L'échelle n'est pas respectée entre oreille externe, moyenne et interne: ces deux dernières composantes sont grossies par rapport à la réalité.

oreille externe et oreille moyenne

L'oreille externe se compose du pavillon, ensemble de cartilagineux de forme complexe et du conduit auditif externe, long d'environ 25 mm et d'un diamètre d'environ 8 mm: figure 2.1, deuxième schéma.

L'oreille moyenne se compose de la membrane du tympan, qui obture le conduit auditif externe, de la caisse du tympan, cavité remplie d'air (susceptible d'une liaison avec la bouche par les trompes d'Eustache pour égaliser la pression interne de la caisse et la pression atmosphérique), de la chaîne des osselets qui relie la membrane tympanique avec la fenêtre ovale. Le *tympan* est une membrane élastique quasi-circulaire d'environ 1 cm de diamètre, concave vers sa face externe. La chaîne articulée des osselets comprend le *marteau*, dont le manche est inclu dans l'épaisseur du tympan, *l'enclume* et *l'étrier*. La figure 2.1, troisième schéma représente l'oreille moyenne.

oreille interne

L'oreille interne, ou *labyrinthe* est la partie de l'oreille la plus complexe, la plus importante du point de vue de l'analyse auditive, et celle dont le fonctionnement reste le plus mystérieux. La *cochlée* ou limaçon, cavité osseuse présentant la forme d'un tube est enroulée deux fois et demie en forme de spirale hélicoïdale. Elle possède deux ouvertures couvertes par de fines membranes: la *fenêtre ovale*, qui ferme le vestibule, reliée au dernier osselet de l'oreille moyenne (étrier) , et la *fenêtre ronde*.

La lame spirale, osseuse, suivie de la membrane basilaire sépare l'intérieur de la cochlée en deux rampes, la rampe vestibulaire reliée à la fenêtre ovale et la rampe tympanique reliée à la fenêtre ronde, qui sont remplies d'une solution salée, la *périlymphe*. Les deux rampes communiquent à l'extrémité, ou *apex* de la cochlée. Le canal cochléaire, rempli *d'endolymphe* est une cavité comprise entre la *membrane de Reissner* et la *membrane basilaire*. La membrane basilaire, longue de 32 mm, possède une largeur croissante de la base (0.04 mm) à l'apex (0.36 mm) et porte *l'organe de Corti*. L'organe de Corti porte lui même deux ensembles de *cellules ciliées*, séparées par une arche (tunnel de Corti formé par les piliers de Corti). Les cellules les plus externes sont les cellules ciliées externes, au nombre d'environ 20000, et les cellules les plus internes sont les cellules ciliées internes, au nombre de 3500 environ. Les cils des cellules ciliées se meuvent dans le *canal cochléaire*, et sont recouverts par la *membrane tectoriale*. Des fibres nerveuses, environ 30000, sont en contact avec les cellules ciliées, 95 % des fibres afférentes innervent les cellules ciliées internes (les moins nombreuses), et les 5 % restant les cellules ciliées externes. Chaque cellule ciliée interne reçoit une vingtaine de fibres. Les fibres nerveuses efférentes sont au nombre de 1000 environ dans la région des cellules ciliées internes et 800 dans celle des cellules ciliées externes. Le corps des fibres nerveuses se trouve dans le *ganglion spiral*, d'où part le nerf auditif vers les *noyaux cochléaires*. Le quatrième schéma de la figure 2.1 montre un cochlée déroulée, puis une coupe transversale de la cochlée, et enfin le canal cochléaire, avec la membrane basilaire, les cellules

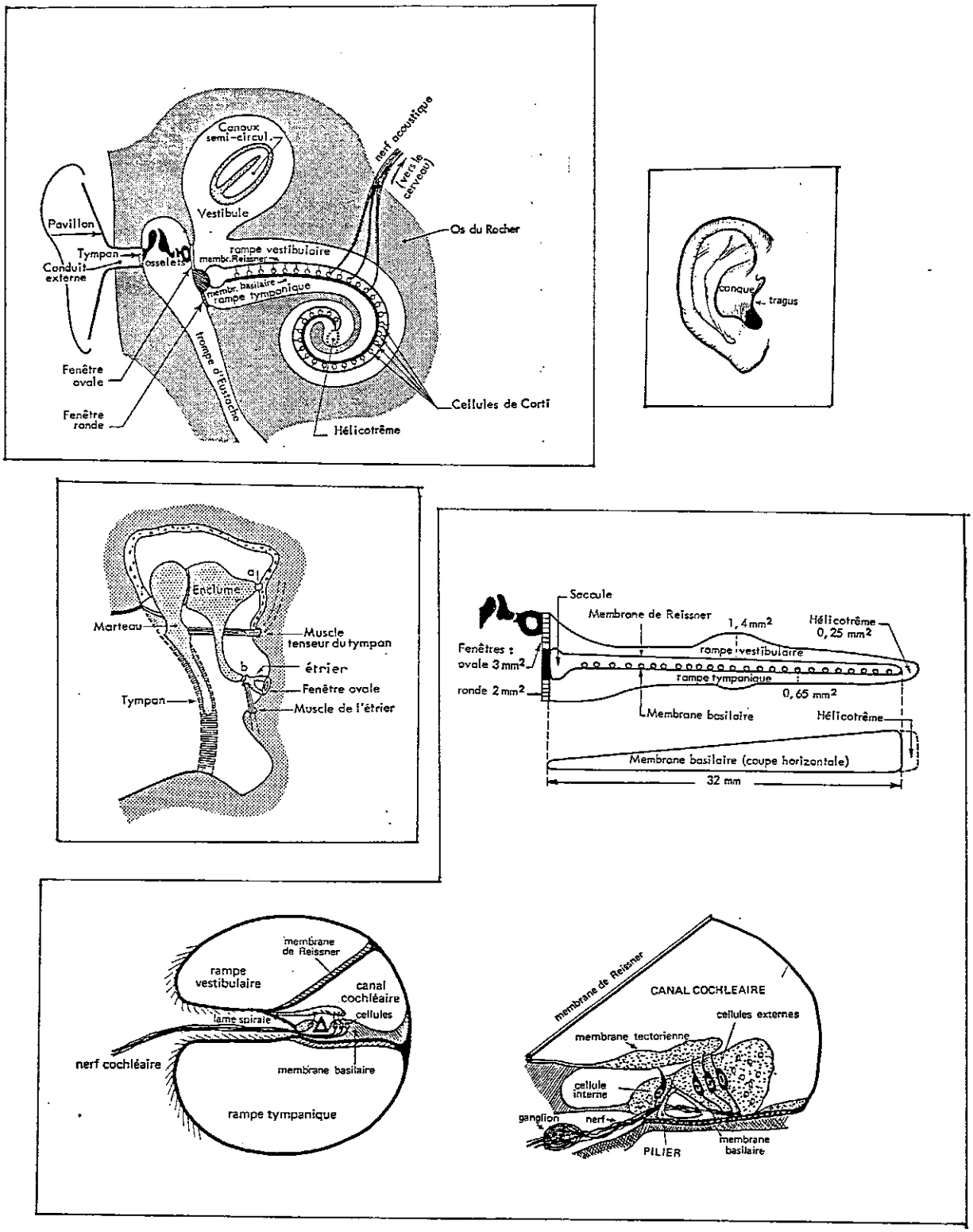

Figure 2.1: anatomie de l'oreille, d'après Leipp (voir texte).

ciliées et le départ du nerf auditif.

2.3.2 fonctionnement du système auditif périphérique

oreille externe

Beaucoup d'hypothèses sur les fonctions de l'oreille externe ont été émises: elle semble jouer un rôle dans la recherche de la direction du son et dans l'amplification d'une large zone de fréquences (le pavillon se comporte comme un cornet auditif et le conduit auditif comme un résonateur). Le rôle fondamental de l'écoute binaurale comme moyen d'orientation, d'extraction du relief acoustique, peut être négligé pour notre propos, et le développement qui suit ne concerne désormais qu'une écoute monaurale. L'amplification fréquentielle due à l'oreille externe possède un maximum vers 2.5 kHz.

oreille moyenne

L'oreille moyenne répond essentiellement à une fonction d'adaptation d'impédance acoustique et de protection contre les signaux d'intensité trop élevée. Elle assure la transmission mécanique de la pression d'air provenant de l'oreille externe au milieu liquide de la cochlée. Les variations de pression à la membrane du tympan sont multipliées lorsqu'elles parviennent à la fenêtre ovale à travers le système de levier constitué par les osselets, et les déplacements divisés d'autant. L'oreille interne n'est pas un système passif de par sa fonction de protection. Les sons de faible intensité provoquent un déplacement d'ensemble des osselets, alors que les sons d'intensité plus élevée provoquent un jeu des osselets les uns par rapport aux autres. La chaîne des osselets joue un rôle beaucoup plus important pour les sons graves, les sons aigus pouvant profiter de la transmission aérienne à travers l'oreille moyenne.

oreille interne

Dans la cochlée réside le mécanisme qui effectue la conversion du signal mécanique présent à la fenêtre ovale en signal électrique dans le nerf auditif. Cette transformation peut se décomposer en trois étapes:

- le mouvement de la fenêtre ovale provoque un mouvement du fluide dans la rampe vestibulaire et la rampe tympanique, ainsi qu'un mouvement de la membrane basilaire;
- le mouvement relatif de la membrane basilaire et de la membrane tectoriale provoque des contraintes sur les cellules ciliées;
- les cellules ciliées réagissent à ces contraintes et provoquent des potentiels d'actions dans les fibres nerveuses avec lesquelles elles sont en contact.

La surpression présentée par l'étrier à la fenêtre ovale se transmet presque instantanément au fluide à l'intérieur de la cochlée. Une onde progressive se crée le long de la membrane basilaire, si la fréquence du stimulus excède environ 50 Hz et une onde stationnaire en phase avec le signal d'excitation en dessous. Cette onde de déformation

grossit doucement, passe par un maximum, puis décroît brutalement. Ce maximum se localise près de la fenêtre ovale pour un son aigu, et d'autant plus éloigné qu'il est grave.

Le système *endo-cochléaire* (dans le canal cochléaire) possède un potentiel positif (de l'ordre de 80 mV). L'organe de Corti est le siège de plusieurs phénomènes électrophysiologiques. Deux phénomènes, dont le sens est toujours discuté, sont le *potentiel microphonique cochléaire* et *le potentiel de sommation.* Le second potentiel est une composante continue que l'on peut recueillir dans la cochlée à l'aide d'électrodes assez grosses ($50\mu m$ ou davantage). Le potentiel microphonique permet de retrouver une forme électriquement analogue au signal acoustique présenté. Ce signal électrique est à différencier de l'activité des fibres nerveuses, activité qui disparaît totalement lors de la mort du sujet, alors que le potentiel microphonique subsiste.

D'autres phénomènes électriques, plus intéressants ici, parcourent les fibres nerveuses. Une fibre au repos possède une polarisation entre sa membrane et l'intérieur. Une action (de nature électro-chimique) sur la fibre va provoquer une dépolarisation locale et induire une onde de dépolarisation: c'est le *potentiel d'action* de la fibre. Le troisième schéma de la figure 2.2 montre des potentiels d'action enregistrés dans une fibre du nerf auditif. Pour une fibre donnée, la valeur du potentiel d'action est toujours constante. Le potentiel d'action apparaît donc suivant une certaine statistique dans une fibre pour peu que son activation dépasse un certain seuil. Après la génération d'un potentiel d'action, la fibre ne répond plus pendant un court intervalle temporel: c'est la *période réfractaire* (environ $1ms$). En l'absence de stimulation, les fibres émettent spontanément entre 0 et environ 150 fois par seconde. On constate une sélectivité fréquentielle (courbe de réponse en fréquence, ou *courbe d'accord*) des fibres qui répondent préférentiellement, en fonction de leur emplacement le long de la membrane basiliaire, à une certaine fréquence (*fréquence caractéristique*). Le troisième schéma de la figure 2.3 montre des courbes d'accords, pour six fibres différentes, mesurées sur un chat anesthésié. Ce tracé, analogue à un spectre d'amplitude, a été obtenu en stimulant l'animal avec des sinusoïdes. La réponse impulsionnelle du filtre constitué par l'oreille interne jusqu'au niveau du nerf auditif peut être estimée par autocorrélogramme: un bruit blanc stimule une fibre donnée, et les potentiels d'actions sont enregistrés. les échantillons du bruit sont ensuite superposés et sommés de façon synchrone aux instants de déclenchement des potentiels d'action. Les schémas 1 et 2 de la figure 2.2 montrent les réponses impulsionnelles obtenues pour une même fibre, à des niveaux de stimulation différents, et les fréquences instantanées correspondantes. Cette propriété, la *tonotopie*, permet dans une certaine mesure de relier l'axe des fréquences avec l'axe spatial le long de la membrane basilaire. L'apex correspondant aux basses fréquences, et la partie proche de la fenêtre ovale correspondant aux hautes fréquences. Jusqu'à environ 3 ou 4 kHz, les réponses des fibres nerveuses sont synchronisées sur les instants forts du signal d'excitation. Les réponses coïncident avec un cycle, ou un nombre fixe de cycles de ce signal si elle ne peuvent se répéter aussi vite que la fréquence excitatrice ne le commanderait. Cet autre type de codage (*codage temporel*) des fréquences forme une explication complémentaire au codage tonotopique. L'histogramme de période permet de montrer cette réponse calée sur une phase particulière du signal. L'histogramme est obtenu en excitant l'oreille interne avec un signal périodique. Le nombre de potentiels d'action se produisant à une phase particulière de la période, sur un grand nombre de présentations, est porté en regard de cette période. Le quatrième schéma de la figure 2.2, pour un son excitateur

sinusoïdal de 1000 hz, et une fibre de fréquence caractéristique 1100 Hz, montre bien la réponse à l'alternance positive de la période, l'amplitude de l'histogramme suivant celle du signal, pour des intensités variées.

Au delà d'environ 4kHz, la réponse temporelle du nerf auditif semble beaucoup plus désorganisée et difficile à interpréter: les potentiels d'action ont une probabilité d'occurrence égale pour toute partie du cycle d'un son périodique.

Un signal plus intense provoque la décharge d'un nombre plus important de fibres, et d'un taux de décharge moyen plus important par fibre. Le premier schéma de la figure 2.3 montre la réponse de trois fibres à un signal de parole, avec et sans bruit de fond. Ces tracés représentent des histogrammes des temps post-stimulatoires: le stimulus est présenté de nombreuses fois, et le nombre de potentiels d'actions générés pendant chaque tranche de temps (l'unité est le nombre de potentiels d'action par seconde, spikes/s) qui suit la présentation est porté en regard de ce retard. Le troisième schéma de la figure 2.3 montre les histogrammes des temps post-stimulatoires en réponse de courts signaux de parole.

La quantité de potentiels d'action générés lors d'une stimulation avec un son stationnaire décroît dans 100 ou 200 ms qui suivent l'attaque de ce son: c'est l'adaptation à court terme. Le cinquième schéma de la figure 2.2 montre cet effet, l'excitation étant deux bouffées de son pur présentées avec des amplitudes variant de 13 à 49 dB_{spl}. L'adaptation due à la première bouffée modifie considérablement la réponse à la seconde. Cet effet est également visible sur le troisième schéma de la figure 2.3

L'oreille interne semble également le siège de phénomènes actifs qui ne seront pas détaillés ici [39], bien qu'ils semblent posséder une importance certaine.

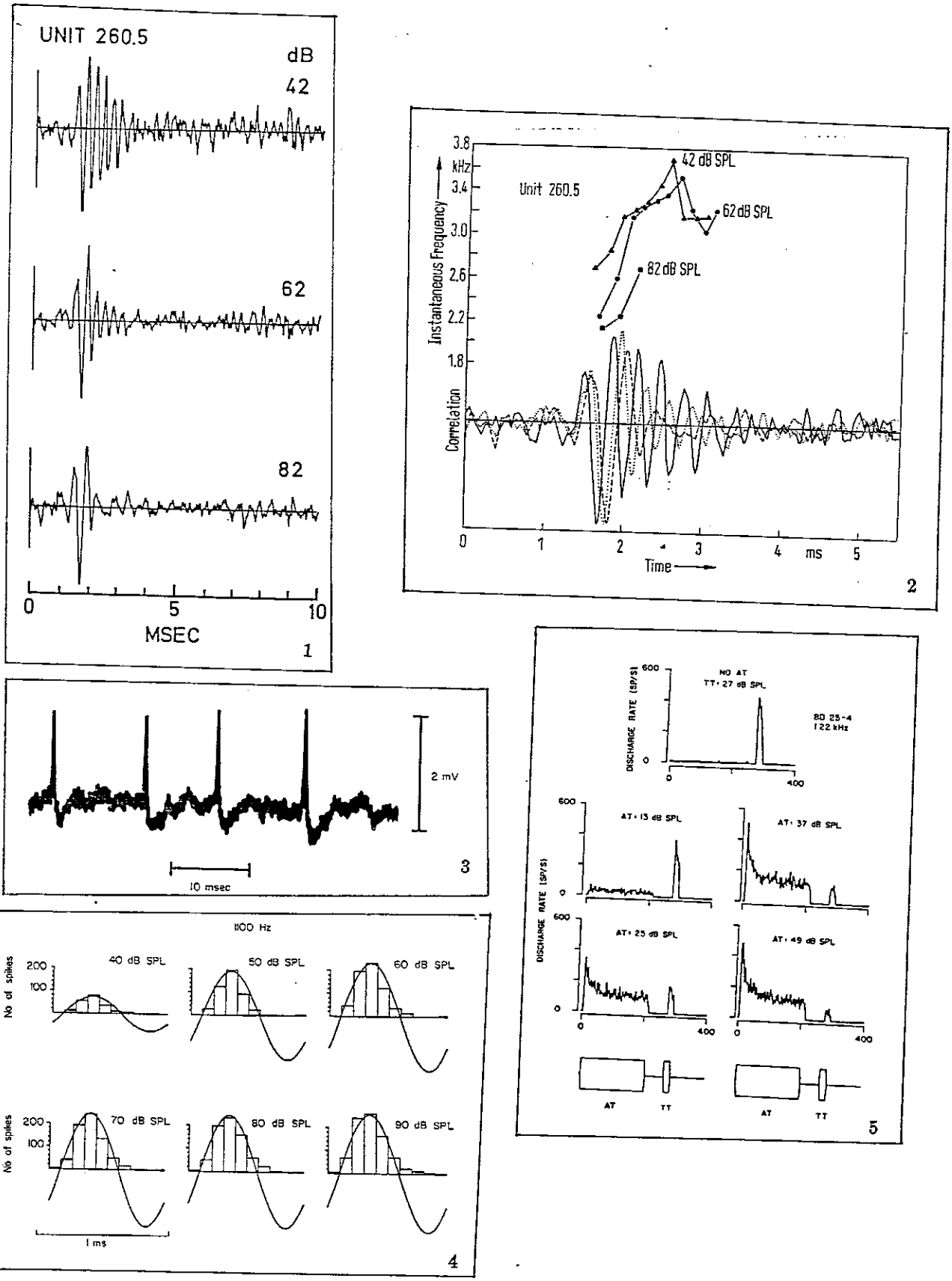

Figure 2.2: fonctionnement de l'oreille interne, d'après Moller et Nilsson, Moller, pickles, Rose, et Delgutte (voir texte).

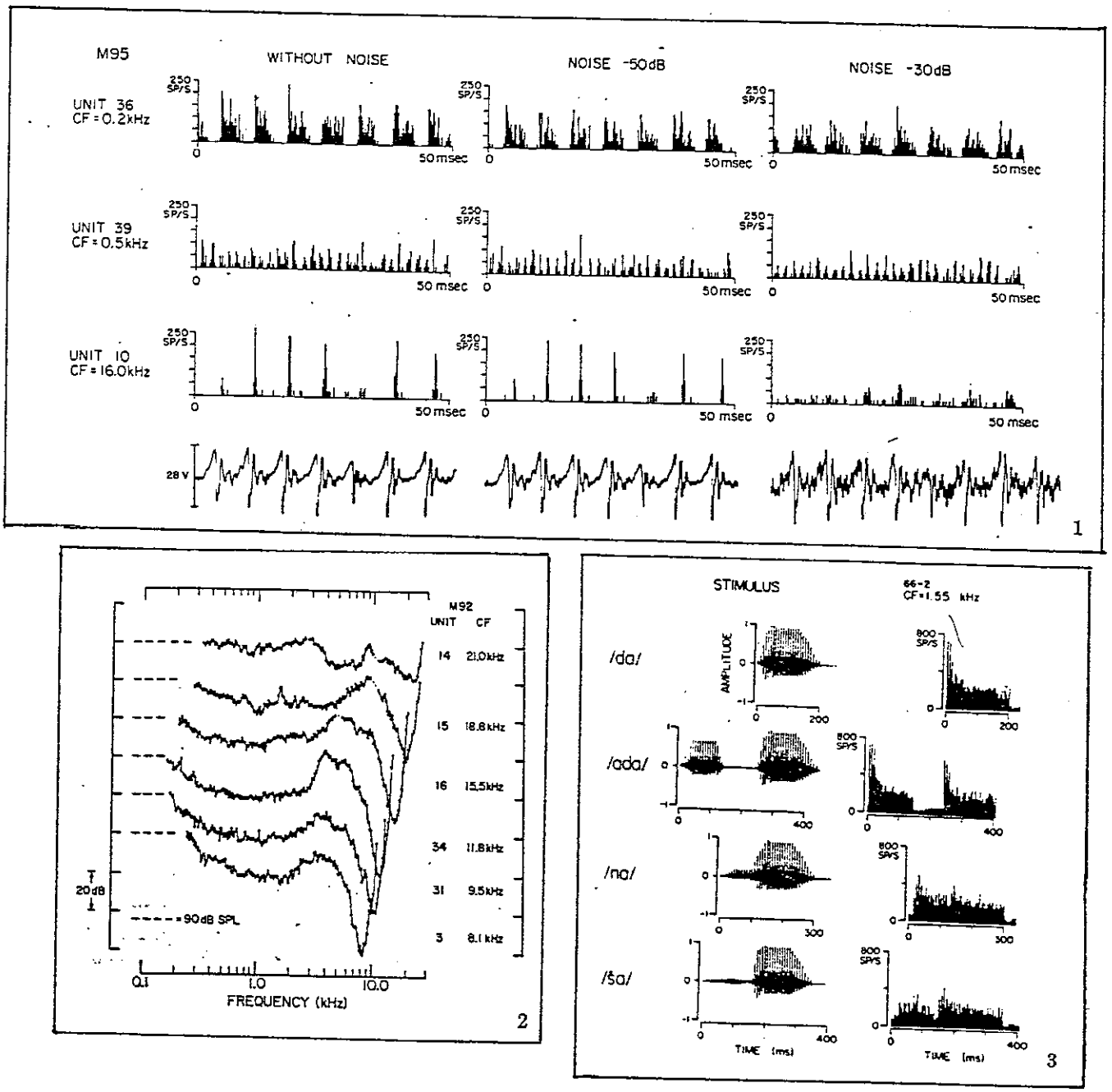

Figure 2.3: fonctionement de l'oreille interne, d'après Kiang et Moxon, Delgutte et Kiang (voir texte).

2.3.3 modèles du système auditif périphérique

modèles fonctionnels, modèles physiques

La modélisation de système auditif périphérique peut emprunter deux voies différentes.

Les modèles physiques s'intéressent au fonctionnement des différentes parties de l'oreille, acoustique de l'oreille externe, mécanique de l'oreille interne macro-mécanique cochléaire, micro-mécanique cochléaire, transduction dans l'organe de Corti [4] [5] [52]. Il s'agit de simuler le fonctionnement du système réel de façon interne.

Au contraire, les modèles fonctionnels tentent de représenter un fonctionnement d'ensemble, en se référant moins étroitement à la physiologie [17].

Notre propos, qui consiste à tirer de la modélisation auditive quelques traits saillants se rapportant à une décomposition en formes d'ondes élémentaires, se situe donc dans le cadre des modèles fonctionnels.

aspect fonctionnel de l'oreille externe et de l'oreille moyenne

La modélisation de l'oreille externe et de l'oreille moyenne s'effectue dans la plupart des modèles fonctionnels du système auditif de façon assez sommaire. Un simple filtre de préaccentuation, qui renforce le médium et l'aigu du spectre et évite ainsi d'accorder trop d'importance aux premiers harmoniques en accentuant les formants, est en général employé.

Si l'oreille externe peut être envisagée comme un système passif, en ne tenant pas compte de l'orientation adaptative de la tête, l'oreille moyenne dans sa fonction de protection et d'adaptation d'impédance est un système éminemment actif. Cette activité, modélisable par un filtre variable commandé par un processus d'adaptation et par sa propre sortie (processus reflexe) [14], n'est pas prise en compte dans de nombreux modèles fonctionnels. Les expérimentations sur des animaux anesthésiés déconnectent l'oreille externe et l'oreille moyenne de l'oreille interne afin d'éviter des résonances acoustiques qui risquent de biaiser les résultats.

aspect fonctionnel de l'oreille interne

L'essentiel des propriétés spectro-temporelles du système auditif périphérique se concentre donc dans l'oreille interne.

Constante dans la modélisation auditive, l'utilisation d'un nombre important d'opérateurs semblables en parallèle reflète la distribution des fibres nerveuses dans la dimension spatiale de la cochlée. Ces opérateurs réalisent toujours une sélection fréquentielle initiale qui reproduit les propriétés du filtrage cochléaire (rapporté souvent au filtrage psycho-acoustique). Ensuite, un ensemble de fonctions modélisent tout ou partie des effets d'adaptation à court terme, de redressement, de saturation, de suppression, de génération de sons subjectifs, de synchronisation ....

Une liste des propriétés souhaitables issues de l'observation peut être:

- tonotopie: une certaine fréquence a tendance à exciter préférentiellement une certaine zone le long de la membrane basiliaire, position répartie de l'aigu vers le grave (de la fenêtre ovale vers l'apex) [37], [40]. Cet ordonnancement géographique

se retrouve dans le nerf auditif [63]. Le filtrage ainsi réalisé répond à une échelle d'allure logarithmique, et il est courant de simuler cette analyse fréquentielle par un banc de filtre de largeur de bande 1 Bark (échelle issue de la psycho-acoustique). L'échelle acoustique fréquentielle est convertie en échelle spatiale le long de la membrane basilaire. La connectivité des fibres par rapport à la membrane basilaire met en correspondance cette échelle spatiale et les fréquences caractéristiques dans le nerf auditif. Le gain d'un filtre passe-bande (ou courbe d'accord d'une fibre) est assymétrique, avec une coupure franche de l'aigu et une coupure beaucoup plus douce des graves. Les largeurs de bande des filtres sont approximativement proportionnelles à leurs fréquences caractéristiques. Divers stimuli peuvent servir pour estimer les courbes d'accord, clics, bruit blanc, son pur. Les résultats sont stables pour tous les types d'excitation ;

- la courbe d'accord d'une fibre nerveuse dépend de l'intensité de l'excitation: elle est plus sélective pour de faibles intensités, et tend à s'élargir lorsque l'intensité s'accroît;

- le filtrage observé au niveau du nerf auditif n'est pas linéaire: divers modèles proposent un *second filtre*, et l'intervention d'opérateurs non-linéaires. Ces modèles permettent de rendre compte plus ou moins bien de la suppression à deux tons et des sons subjectifs. Un renforcement des contrastes spectraux peut en résulter;

- suppression à deux tons: phénomène d'atténuation de la réponse d'une fibre de fréquence caractéristique donnée à un son pur excitateur sous l'action d'un son suppresseur de fréquence voisine [1]. Ce phénomène est à distinguer de *l'inhibition latérale* due au système nerveux central;

- sons subjectifs: observés en psycho-acoustique ils sont déjà présents au niveau du nerf auditif. L'apparition des sons subjectifs est en partie imputable à l'existence de non-linéarités dans l'analyse cochléaire [70];

- adaptation à court terme: un effet d'adaptation à court terme intervient, provenant de la transduction des cellules cilliées aux fibres nerveuses [72]. Une fibre répondra plutôt aux changements d'intensité de l'excitation qu'à une excitation continue. Les parties transitoires du signal sont ainsi mises en valeur par rapport aux parties stables. Un renforcement des contrastes temporels s'ajoute donc au renforcement des contrastes fréquentiels dû au filtrage.

- codage non-linéaire de l'intensité: la sensation d'intensité est reliée à une échelle d'allure logarithmique. Le codage de l'intensité dans une fibre se traduit par une augmentation du taux de décharge moyen (depuis le taux de décharges spontanées), dès que l'excitation dépasse un certain seuil, et tant que l'excitation ne dépasse pas un autre seuil à partir duquel la réponse est saturée. Le codage de l'intensité dans un ensemble de fibres se traduit donc par une augmentation globale de l'activité, augmentation qui peut de plus présenter des propriétés de sélectivité fréquentielle par synchronisation des réponses temporelles sur une composante fréquentielle;

- redressement: les potentiels d'action générés, suivant une certaine statistique, présentent la forme d'impulsions, sensibles uniquement à une phase d'excitation. La forme de l'impulsion est constante pour une fibre donnée pour tous les niveaux d'excitation. Ce phénomène est parfois modélisé par un redressement simple alternance;

- codage temporel des fréquences: outre la tonotopie, une mesure délocalisée du spectre s'obtient en examinant les réponses temporelles d'une ou d'un ensemble de fibres. Ces réponses temporelles ont tendance à se synchroniser sur le signal en deçà d'environ 4 kHz. La répartition des décharges d'une fibre dans une période du signal excitateur suit approximativement la forme de cette période;

- distribution de l'énergie vibratoire retardée: de par la progression de l'onde de déformation sur la membrane basilaire, la réponse au grave du spectre est retardée par rapport à celle de l'aigu;

- la population des fibres nerveuses n'est pas homogène [44]: un petit sous-ensemble de fibres semble posséder un taux de décharges spontanées faible, et donc un seuil d'intensité élevé. De plus, il est probable que certaines fibres jouent un rôle particulier de marquage de certains événements acoustiques, phénomène observé par ailleurs en vision;

Nous allons passer en revue quelques modèles informatiques du système auditif périphérique, proposés dans le contexte du traitement automatique de la parole. Ces modèles semblent offrir un intérêt comme étage de paramétrisation acoustique de systèmes de reconnaissance, en particulier dans les cas assez mal traités par les méthodes classiques. Les modèles choisis sont plutôt fonctionnels qu'internes. La liste qui suit ne prétend pas à l'exhaustivité, mais présente des réalisations marquantes quant à l'interprétation en termes d'analyse spectro-temporelle.

spectrographe auditif

Le spectrographe acoustique, développé dans les années quarante, reste toujours un outil fondamental de recherche sur la parole. Une faible part des connaissances sur l'analyse auditive est incluse dans ce type de système. Nous reprenons les propositions de Carlson et Granström [16] pour améliorer la représentation spectrographique en se basant sur des principes d'analyse auditive.

L'espace temps-fréquence-amplitude utilise des échelles linéaires/linéaires/logarithmiques dans le spectrographe classique. Ici, la dimension temporelle reste inchangée (les effets de masquage temporel ne sont pas incorporés), la dimension fréquentielle utilise des filtres répartis sur une échelle psycho-acoustique (échelle de Bark, largeur de bande de 1 Bark), et l'amplitude est affichée sur une échelle dérivée des courbes d'isosonie (sones 2.2).

Outre l'analyse spectrale géographique qui résulte de la répartition d'énergie le long de la membrane basilaire, une analyse temporelle peut être ajoutée en recherchant les fréquences dominantes dans chaque filtre d'analyse. En affichant pour chaque fréquence d'analyse et à chaque instant l'amplitude en fonction du nombre de filtres dominés

par cette fréquence, des caractéristiques du signal de parole émergent mieux: premiers harmoniques, formants.

modèle de Ghitza

Dans ce modèle, un nombre important de filtres est utilisé pour la première étape d'analyse spectrale (85-100 filtres) [29]. Le gain des filtres suit fidèlement un modèle de courbe d'accord, et présente donc une assymétrie et une forme assez complexe. Les filtres montrent une coupure beaucoup plus franche du côté aigu que du côté grave et le rapport $\Delta f/f$ est approximativement constant. L'ensemble des filtres est uniformément distribué en suivant une échelle logarithmique.

Le traitement temporel des réponses de chacun de ces filtres intervient après l'analyse fréquentielle. La détection des fronts montants par un ensemble de niveaux, tous positifs et répartis sur une échelle logarithmique sur toute la dynamique du signal, modélise la suppression des parties négatives du signal et approche le taux moyen de décharge par niveau. Ainsi le taux de décharge moyen et une estimation logarithmique de l'intensité, en sommant les contributions des différents niveaux, sont disponibles pour une fibre. En répartissant l'inverse des intervalles temporels d'un ensemble d'intervalles récents (les 20 derniers) sur 100 échantillons fréquentiels, une fréquence est obtenue pour chaque niveau de chaque fibre. La somme pour une fibre de toutes les fréquences à tous les niveaux lui affecte une valeur spectrale dérivée de l'information temporelle. Le choix du nombre d'intervalles représentés induit une fenêtre d'analyse dont la durée dépend de la fréquence caractéristique: un facteur d'échelle s'introduit. Un spectre d'amplitude est dérivé de ces mesures en sommation des contributions de toutes les fibres: l'intensité à une fréquence mesure ainsi implicitement le nombre de régions dont le comportement est synchronisé. Ghitza propose donc une approche non tonotopique.

Ce modèle présente pour la reconnaissance en contexte bruité des performances supérieures aux méthodes classiques, en exploitant les propriétés spectro-temporelles locales du signal et le grand nombre de corrélations temporelles implicites qu'il contient [27].

Un test significatif du comportement du modèle a consisté à substituer aux filtres "cochléaires" d'entrée un banc de filtres par transformée de Fourier, avec fenêtrage de Hamming: les résultats semblent assez similaires. Le point déterminant reste l'utilisation d'information temporelle spectralement localisée (corrélation et synchronie) dans un grand nombre de bandes de fréquence plutôt que la forme exacte de l'analyse fréquentielle [28].

modèle de Lyon

La première étape de l'analyse est un filtrage: une centaine de filtres sont répartis sur une échelle de Bark. Le gain des filtres est réglé pour reproduire le comportement mécanique de la membrane basilaire. Plusieurs structures ont été proposées: cascade/parallèle avec des antirésonateurs et des résonateurs du second ordre, puis cascade de résonateurs du second ordre. Les filtres, fortement assymétriques, possèdent une résolution spectrale relativement bonne par leur pente assez raide au voisinage de la résonance, et une résolution temporelle relativement bonne par la pente douce vers les

graves qui induit une assez grande largeur de bande.

L'étape suivante est l'analyse temporelle. Les signaux issus de chaque bande d'analyse sont redressés (les parties négatives sont supprimées), et filtrés passe-bas (fréquence de coupure 1kHz).

Un mécanisme sophistiqué de contrôle automatique du gain, avec couplage entre les bandes d'analyse et une non-linéarité compressive, permet de restituer les phénomènes d'adaptation à court terme, d'adaptation dans l'oreille moyenne et de masquage.

L'exploitation du modèle cochléaire utilise essentiellement les données relatives aux décharges des fibres. Une étude des coïncidences entre ces décharges permet, au moins dans la visualisation, de retrouver des traits acoustiques du signal de parole: formants, fréquence de voisement. Les points saillants sont l'utilisation de modules complexes de contrôles automatiques de gains couplés, et l'expoitation des relations entre les réponses des diverses fibres. Le modèle peut délivrer en sortie une représentation graphique, ou *neurogramme*. La premier schéma de la figure 2.4 montre le tracé obtenu pour quelques périodes d'un signal voisé, les fréquences aigües sont en bas (base de la membrane basilaire), et les fréquences graves en haut.

Le modèle de Lyon, dans ses versions les plus récentes, incorpore des hypothèses sur des traitements possibles au delà du système périphérique [55] [51] [53]. Le calcul de l'autocoïncidence de la réponse du modèle (autocorrélation, fonction du retard) permet de compenser dans la dimension du retard, le déphasage entre les réponses de l'aigu et du grave introduit par la propagation de l'onde de déformation le long de la membrane basilaire. Le second schéma de la figure 2.4 montre l'autocoïncidence calculée sur le neurogramme précédent, et le dernier schéma explicite la relation entre les deux premiers.

modèle de Seneff

Le modèle de Seneff [80] [79] utilise un nombre plus faible de filtres (32-40) répartis sur une échelle de Bark (définie autrement que 2.3). La préaccentuation est réalisée par le banc de filtres, et chaque filtre est calculé pour présenter un gain d'amplitude assymétrique et un saut de phase à la résonance qui rendent compte de données physiologiques. Une structure cascade/parallèle implémente le banc de filtre. Différents types de gains ont été proposés, dont le détail diffère mais qui comportent tous les caractéristiques d'assymétrie et de répartition sur une échelle de Bark.

Une compression non-linéaire de l'enveloppe temporelle est réalisée ensuite en utilisant deux mécanismes de contrôle de gain automatique répondant à l'équation:

$$y(t) = \frac{x(t)}{(k + \int_{t-t_0}^{t} |x(\tau)| d\tau)} \tag{2.6}$$

où x représente l'entrée, y la sortie, k une constante et t_0 le temps d'intégration.

Le premier contrôleur de gain possède un temps d'intégration court ($\simeq 3ms$) et le second un temps plus long ($\simeq 40ms$). Un redressement simple alternance achève le traitement temporel. Une autre version du modèle se veut plus proche de la réalité physiologique, tant par les filtres que par la modélisation de la transduction entre cellules ciliées et fibres nerveuses. Elle diffère par l'ordonnancement du redressement et de la compression d'enveloppe et par le choix de cette compression, mais ne sera pas détaillée.

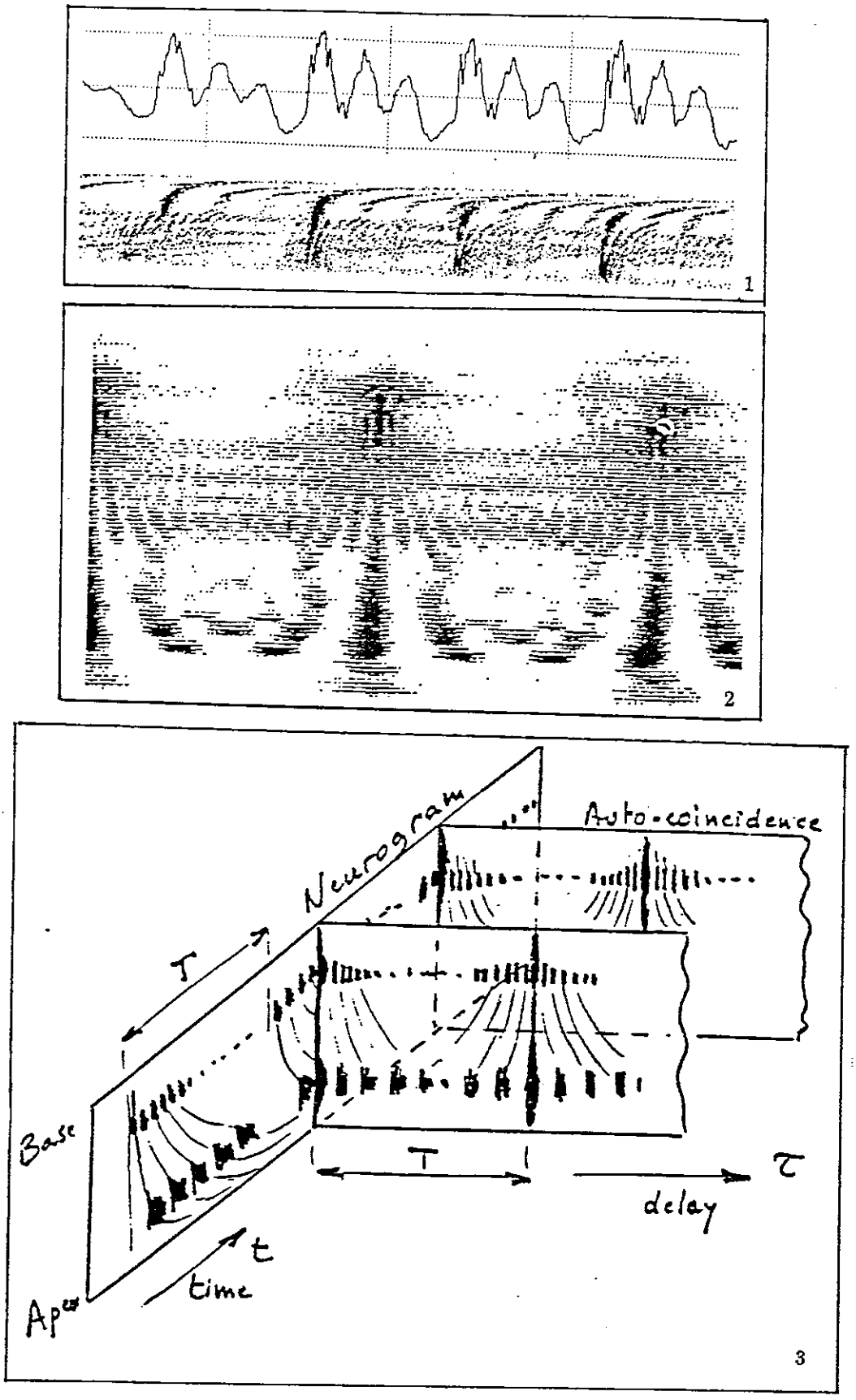

Figure 2.4: modèle de Lyon: neurogrammes et autocoïncidence, d'après Lyon et Liénard (voir texte).

Après cette étape initiale de traitement, l'estimation d'un spectre d'amplitude ou de la fréquence fondamentale du signal utilise un mécanisme de détection des synchronismes entre les voies d'analyse. Des *détecteurs de synchronie généralisés* effectuent le rapport du module de la somme et du module de la différence du signal issu de la bande analysée et de ce même signal retardé de τ.

Dans le cas d'une analyse spectrale le retard τ est égal à l'inverse de la fréquence caractéristique de la bande, et l'estimation du taux de synchronie fournit la dimension d'amplitude du spectre.

Dans le cas d'une recherche de fréquence fondamentale, la somme de tous les signaux issus du traitement initial passe dans un ensemble de détecteur de synchronie dont les retards sont réglés entre $2ms$ ($500Hz$) et $16ms$ ($62Hz$). Une variante de l'autocorrélation permet ainsi de calculer la fréquence fondamentale.

Le modèle considère un codage temporel dans des bandes fréquentiellement déterminées.

Une version modifiée de cette analyse, qui préserve le procédé en améliorant les détecteurs de synchronie, augmente le nombre de bandes d'analyse et affine les contrôles de gains pour les bandes aiguës. Ce système a fait l'objet de tests en tant que paramétrisation acoustique pour la reconnaissance [34] [35]. La comparaison avec une paramétrisation classique (18 coefficients cepstraux) fait apparaître un avantage lorsque la parole est dégradée par du bruit de synthèse ou dans des conditions réelles (à bord d'un hélicoptère).

modèle de Cooke

Un modèle informatique motivé par l'utilisation prometteuse de données perceptives dans le système de traitement automatique de la parole (la reconnaissance et l'analyse spectrographique en particulier) a été proposé par Cooke [18].

L'analyse met en oeuvre 61 canaux d'analyse, répartis sur une échelle de Bark, afin de couvrir la région fréquentielle 100-5000 Hz. Le modèle, canal par canal, possède trois principales étapes: filtrage, adaptation, extraction de caratéristiques temporelles.

La première étape est un filtrage passe-bande non linéaire, qui utilise deux filtres linéaires entre lesquels une non linéarité compressive (racine carrée anti-symétrique) est insérée. Le premier filtre présente un gain assymétrique (coupure plus raide des aigus), alors que le second est symétrique avec une bande passante moitié de la bande passante du premier filtre. Notons que ce second filtre ne semble plus être nécessaire en regard des théories actuelles, mais existe ici afin de rendre compte dans l'étape de filtrage initial d'effets de suppression à deux tons. Le temps de propagation de l'onde le long de la membrane basilaire est rendu par le déphasage induit par les filtres: l'aigu du spectre apparaît en avance par rapport au grave. Le premier schéma de la figure 2.5 montre la réponse du banc de filtres à un signal synthétique composé de trois formants.

L'étape suivante prend en compte les effets d'adaptation, en modélisant la transduction du mouvement mécanique le long de la membrane basilaire en potentiel d'action dans le nerf auditif. Cette transduction s'effectue en postulant la génération et la migration d'une population d'agents électro-chimiques (ou *neuro-transmetteurs*) dans les cellules ciliées. Le modèle propose un système à réservoirs multiples pour la génération et la disparition des neuro-transmetteurs [76]. L'entrée de ce système est l'enveloppe du sig-

nal disponible après le traitement initial, et la sortie représente une courbe d'enveloppe après adaptation par le mécanisme sophistiqué des réservoirs multiples.

Au codage de l'enveloppe adaptée des signaux filtrés viennent se superposer des informations fréquentielles obtenues en mesurant l'intervalle entre les passages par zéro du signal filtré dans chaque bande d'analyse: ce type simple d'analyse temporelle ressemble à celui mis en oeuvre dans la représentation spectrographique auditive.

Les sorties du modèle sont d'une part la mesure des passages par zéro dans chaque bande d'analyse, et d'autre part la mesure d'enveloppe adaptée. Le second schéma de la figure 2.5 montre les fréquences détectées, à partir du filtrage présenté sur le schéma précédent. La taille des cerles représente l'amplitude en sortie de l'étage d'adaptation.

modèle de Delgutte

Delgutte a proposé un modèle de traitement périphérique et a étudié sa réponse à des stimuli relevant des grandes catégories phonétiques. Un signal de parole en entrée produit en sortie un histogramme de décharge pour chacune des fibres [22].

Le premier stade de l'analyse est un filtrage linéaire, les gains des filtres étant dérivés de courbes d'accord. Le banc de filtres possède 12 canaux par octave dans la région 200-10000 Hz. Ici ce sont donc des données physiologiques qui décident du filtrage et non des données psycho-acoustiques.

Une détection d'enveloppe suit cette étape, puis des non-linéarités sans mémoire qui représentent le taux de décharge en fonction du niveau de l'excitation, pour rendre compte du seuil de décharge et de la saturation. Trois populations distinctes de fibres sont représentées par trois courbes différentes par leurs seuils et leurs dynamiques. Ainsi, des stimuli de niveaux aussi variés que ce que l'on peut rencontrer dans la réalité sont susceptibles d'être mieux considérés dans le modèle.

Un élément d'adaptation à court terme intervient ensuite, utilisant une constante de temps de $30ms$ et une constante de temps de quelques ms. Les phénomènes non-linéaires comme la suppression à deux tons, et les sons subjectifs ne sont pas pris en compte.

Outre ce calcul du taux moyen de décharge en fonction de la fréquence caractéristique, une analyse du comportement temporel avant la détection d'enveloppe permet d'obtenir une autre estimation spectrale.

Le modèle de Delgutte permet d'opérer, par le double codage temporel et tonotopique et par l'utilisation de familles de fibres possédant des comportements différents, une classification (qualitative) de segments phonétiques pour des niveaux d'excitation variés.

2.3.4 conclusions

L'analyse du fonctionnement du système auditif périphérique permet, de même que l'observation de résultats issus de la perception, de préciser des contraintes sur une analyse spectro-temporelle [74]. Les résultats peuvent provenir soit de l'expérimentation sur des animaux anesthésiés [69] [83] [20] [21] pour la perception d'éléments phonétiques de synthèse, ou de l'utilisation de modèles du système auditif périphérique. La distance

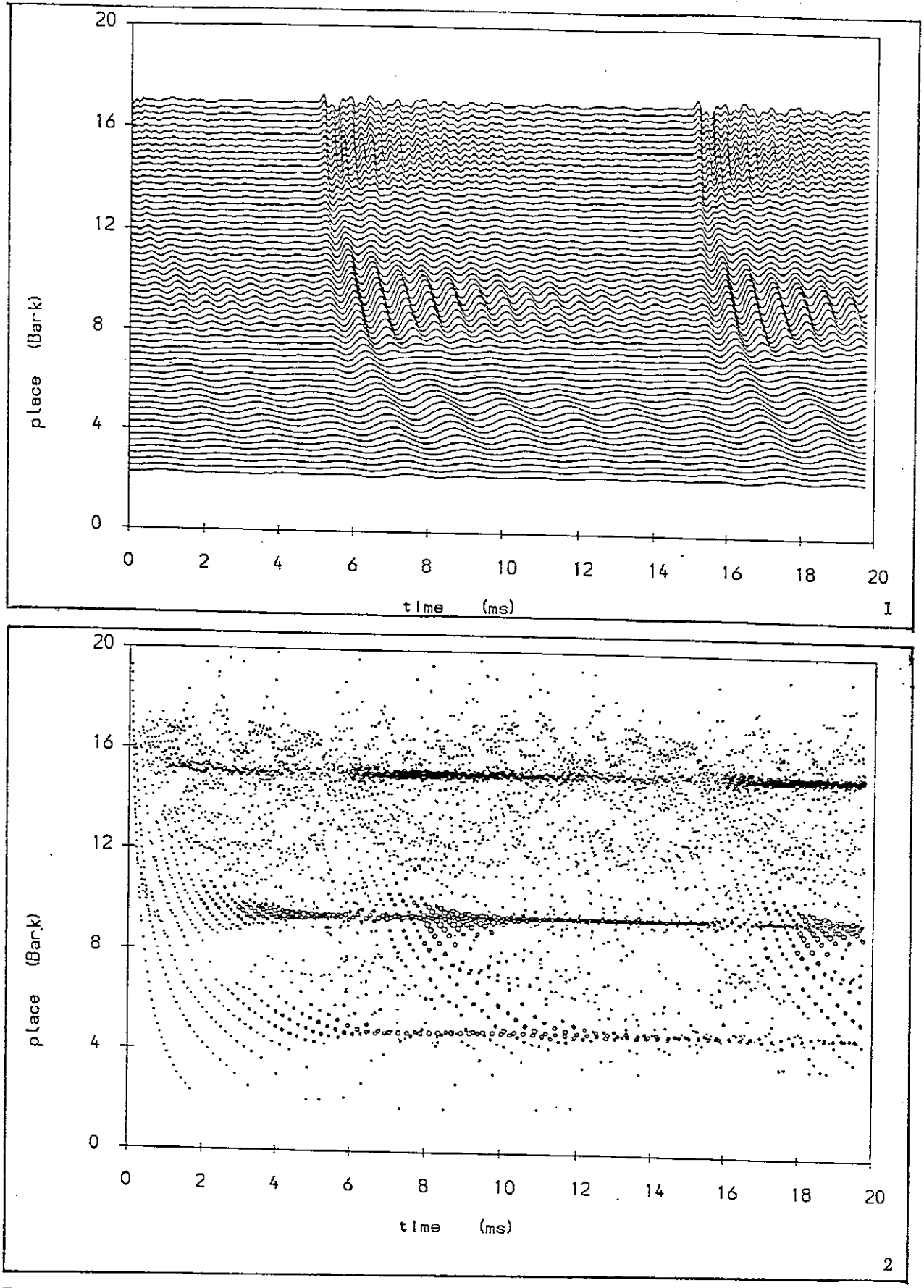

Figure 2.5: modèle de Cooke: banc de filtres et détection des fréquences, d'après Cooke, (voir texte).

accrue avec la réalité qu'impose la deuxième classe de résultats se compense par la commodité de réalisation et la possibilité d'effectuer des tests intensifs.

Les deux aspects du codage, géographique et temporel, contiennent des informations redondantes et complémentaires pour le codage des signaux de parole, et l'utilisation de nombreuses bandes d'analyse suivies d'un traitement plutôt spectral (tonotopique) ou plutôt temporel forment la base des simulations, ou des explications.

La première étape est un filtrage qui simule la tonotopie. La forme exacte des filtres ne semble pas présenter une importance critique: certains sont dérivés d'une répartition régulière sur l'échelle de Bark, d'autres combinent à la fois un effet passe-bande et passe bas. Ghitza semble montrer, pour une expérience donnée, que des filtres passe-bande simples conviennent aussi bien que des filtres déduits des études physiologiques. Par contre, des phénomènes comme la suppression à deux tons ou le renforcement des contrastes spectraux s'intègrent ou non dans cette étape.

L'interprétation des sorties de ce banc de filtres passe par une étape renforçant les contrastes temporels, par adaptation à court terme et saturation des réponses: les phénomènes transitoires émergent alors plus fortement que les phénomènes stationnaires.

Le codage tonotopique intervient comme mesure de la réponse le long de la cochlée, par comptage du taux de décharges nerveuses par fibre et par unité de temps. A cause de la saturation il faut considérer toutes les fibres d'une région spectrale pour estimer l'amplitude présente.

Les divers auteurs ont développé diverses méthodes pour traiter le codage temporel: il s'agit, sur une version redressée en simple alternance des signaux filtrés, (avant ou après adaptation) de détecter des périodicités. Une mesure de l'amplitude spectrale nécessite de regrouper les fibres dont la réponse est synchronisée: le taux de décharge pour une fibre ne forme plus dans ce cas une évaluation de l'amplitude spectrale.

2.4 application de contraintes perceptives pour la décomposition en signaux élémentaires

2.4.1 choix des contraintes

Les propriétés principales énoncées permettent d'entrevoir les relations possibles entre le traitement auditif périphérique et une analyse spectro-temporelle. Des notions de filtrage, de traitement temporel, de spectre, émergent plus ou moins nettement des études psychologiques et physiologiques.

L'intérêt pour la simulation informatique de modèles du système auditif périphérique en traitement automatique de la parole semble aujourd'hui en pleine croissance, et certaines caractéristiques issues de données physiologiques ou psychologiques sont incorporées dans tout système de traitement du signal acoustique.

propriétés psycho-acoustiques

Les propriétés psycho-acoustiques souhaitables concernent deux faits de nature différente: le filtrage psycho-acoustique et la représentation auditive du flux acoustique comme formes constituées d'un ensemble de formes d'ondes élémentaires.

Tout d'abord, une propriété spectrale, dont la contrepartie physiologique est discutée, concerne la répartition et la forme des filtres utilisés dans la première étape du traitement. L'échelle de Bark, issue des bandes critiques est un mélange entre une échelle des fréquences linéaires et une échelle logarithmique. Ce choix, adopté ici, offre les propriétés d'un filtrage à résolution constante dans le grave du spectre, et d'un filtrage à résolution relative constante au dessus d'environ 1000 Hz. L'aspect temporel du codage ne sera pas considéré en tant que contrainte psycho-acoustique, bien que des expériences permettent de le mettre en évidence par des tests de masquage: bouffées de sons purs haute fréquence ajoutées à un instant particulier de la période d'un son pur basse fréquence par exemple.

D'autre part, un choix plus fondamental et plus spéculatif est de tenter la justification de la décomposition du signal de parole en un ensemble discret de formes d'ondes élémentaires par des arguments sur le traitement des *formes* psycho-acoustiques. Entre la représentation du signal acoustique par une fonction du temps et de la fréquence et la représentation des objets mentaux qui s'y rapportent, la nécessité d'une représentation ou d'une suite de représentations emboîtées peut justifier notre démarche. Le choix opéré ici est en faveur d'une représentation très proche du signal dans la chaîne partant de l'acoustique pour aboutir au symbolique. Elles doit délivrer des objets tels que le maximum de notions perceptivement pertinentes puissent se déduire de façon aussi directe que possible d'un nombre réduit de ces objets, et de leurs relations. L'hypothèse principale est que la notion d'événement acoustique composé par un ensemble discret de formes d'ondes élémentaires se justifie par la possibilité d'isoler dans le flot perceptif des *formes* pertinentes. La recherche d'entités phonétiques [32], la notion classique de *gestalt*, celle plus récente de scènes auditives, les événements articulatori-phonétiques [82] sous-entendent des stratégies de découpage et de regroupement du continuum spectro-temporel présenté aux oreilles. La représentation en formes d'ondes élémentaires apporte une contribution originale, et adaptée à de tels traitements.

propriétés du système auditif périphérique

La modélisation du système auditif périphérique apporte un ensemble de contraintes. Il s'agit de s'inspirer du traitement *statistique* observé dans l'oreille interne pour proposer un traitement *déterministe* analogue par certains traits. Les propriétés retenues à ce stade sont:

- nombre assez important de canaux identiques et se recouvrant dérivés du filtrage physiologique: courbes d'accord du nerf auditif et répartition des filtres sur une échelle de Bark:
- renforcement des contrastes spectraux et temporels;
- compression d'amplitude;
- utilisation de l'amplitude présente dans chaque canal par analogie avec le codage tonotopique;
- détection des alternances du signal: redressement simple alternance, pour chaque bande:
- détection de périodicité dans les réponses temporelles: densité de passage par des seuils;
- examen des relations entre la réponse temporelle des différentes bandes, synchronie;

Le premier point peut se trouver en contradiction avec les données psycho-acoustiques, quant au détail de la forme des filtres. Ce détail a sans doute assez peu d'importance.

Le second point concerne les procédés de contraste, qui concentrent l'analyse sur les instants ou les fréquences dominants, permettant par exemple de retrouver des composantes physiques dans le signal. Inversement, le regroupement spectral ou temporel permet d'accentuer les contrastes en affectant à des points spectro-temporels privilégiés les contributions des points voisins. Le renforcement des contrastes temporels peut utiliser des échelles de temps différentes, plusieurs constantes de temps par exemple dans les contrôleurs automatiques de gain.

Les quatre derniers points concernent l'estimation des caractéristiques spectrales: un double traitement, tonotopique et temporel, semble exister au niveau du nerf auditif. Cette double approche trouve une équivalence en traitement du signal. Bien que l'utilisation du spectre d'amplitude soit largement majoritaire, des procédés récents pour estimer des paramètres acoustiques, ou pour l'analyse spectrale de parole noyée dans le bruit exploite des méthodes temporelles d'estimation spectrale: les passages par zéro d'un signal de parole fortement bruité semblent un indice acoustique plus sûr que l'enveloppe spectrale [26] [61].

A ces deux espèces de contraintes, il est souhaitable d'ajouter pour notre propos une reconstitution aussi parfaite que possible du signal de parole à toutes les étapes de sa décomposition.

2.4.2 analyse impulsionnelle, analyse granulaire

Des travaux développés depuis plusieurs années sous le terme *d'analyse impulsionnelle*, puis *d'analyse granulaire* forment la base de notre réflexion sur la représentation du signal de parole en signaux élémentaires par des arguments perceptifs [7].

analyse impulsionnelle

La première forme de ces travaux a été appelée analyse impulsionnelle [57] [45] . Le principe repose, en utilisant un nombre assez important de bandes d'analyse, sur la mesure des similarités entre les événements temporels rencontrés dans chaque bande. Une recherche des maxima d'enveloppe temporelle permet l'estimation des impulsions dans chaque bande. Les similarités entre bandes délivrent des impulsions globales à toutes les bandes, sans distinguer *a priori* entre les segments dont l'excitation est de nature différente: segments voisés, fricatifs, plosifs par exemple. Des essais menés en analyse synthèse utilisent la donnée de ces impulsions comme source d'excitation d'un vocodeur à canaux.

La première étape de l'analyse est un filtrage, par un banc de filtres non-déphaseurs, dans un nombre assez important de canaux (typiquement 32) qui couvrent la bande $(50 - 5000Hz)$ usuelle du traitement de la parole. Les filtres sont répartis sur une échelle de Bark, tous les $1Bark$ et avec un recouvrement de $0.5Bark$. L'implémentation choisie pour les filtres ($4eme$ ordre par deux cellules du second ordre) entraine des gains symétriques.

L'enveloppe des signaux présents dans chaque canal est ensuite estimée par redressement simple ou double alternance, et filtrage passe bas, ou des méthodes d'extraction d'enveloppe et de lissage similaires. Dans chaque canal des *impulsions* sont ainsi localisées: impulsion signifie ici un maximum local de l'enveloppe temporelle d'un signal. La détection d'enveloppe proposée est susceptible d'introduire des erreurs dans le grave du spectre.

A partir de ces informations spectro-temporellement localisées, l'enveloppe des signaux temporels dans plusieurs bandes de fréquence, une fonction de cohérence impulsionnelle est définie pour permettre de retrouver des impulsions globales dans le signal. La mesure de cette fonction utilise plusieurs critères sur la mesure des enveloppes temporelles ou de leurs dérivées temporelles. Le critère qui semble avoir obtenu les meilleurs résultats est un comptage du nombre de bandes où la pente de l'enveloppe est positive. L'échelle de temps dans la recherche des impulsions est de l'ordre de la milliseconde: les impulsions plus courtes ne sont pas considérées.

Le schéma global d'excitation ainsi calculé peut s'employer comme source d'excitation d'un vocodeur à canaux. Pour chaque impulsion globale il faut estimer l'enveloppe du spectre d'amplitude. Les résultats en analyse synthèse semblent satisfaisants. Le procédé peut se résumer par la formule:

$$s(t) = \sum_i \sum_{j=1}^{N} \alpha_i \gamma_j \delta(t - t_i) * h_j(t) \tag{2.7}$$

où $\delta(t-t_i)$ représente une impulsion à l'instant t_i, α_i son amplitude, $h_j(t)$ la réponse impulsionnelle et γ_j le gain (au sens de facteur d'amplitude) du jème filtre d'un banc

de N filtres. L'axe des fréquences est échantillonné par le nombre N de bandes.

Ce procédé offre des similitudes (en plus d'être exactement contemporain) avec le codage multi-impulsionnel par prédiction linéaire [8]. défini pour pallier les inconvénients d'une modélisation trop simpliste du signal d'excitation. Celui-ci est représenté par une suite d'impulsions dont on choisit l'emplacement et l'amplitude par minimisation récursive d'un critère d'erreur. Expérimentalement un nombre d'impulsions voisin de celui utilisé pour l'analyse impulsionnelle ($8-12$ toutes les $10ms$) permet de resynthétiser un signal de très bonne qualité, quasi-indiscernable de l'original. La recherche des impulsions est globale dans le cas multi-impulsionnel et local dans le cas de l'analyse impulsionnelle. La puissance et l'économie de la prédiction linéaire a assuré un très large succès à la méthode multi-impulsionnelle.

Une autre utilisation de la fonction de cohérence impulsionnelle est la recherche de la fréquence de voisement du signal de parole. Une analyse astucieuse de diverses situations possibles fait ressortir quelques cas ambigus dans les relations entre cohérence impulsionnelle et perception d'une hauteur [74]. En particulier le caractère global de la fonction la met en défaut lorsque plusieurs signaux localement cohérents sont présentés simultanément (trois trains périodiques d'impulsions présentés à trois filtres passe-bandes différents): plusieurs composantes autonomes sont perçues, chacune avec une hauteur propre parfaitement identifiable, alors que la fonction de cohérence impulsionnelle est faible.

Les glissandi rapides et périodiques de fréquence représentent une autre épreuve difficile pour la fonction. La cohérence verticale (temporelle) des impulsions dans chaque bande n'existe plus, bien qu'une hauteur soit perceptible sans ambiguïté.

analyse granulaire

Dans la continuité des travaux sur l'analyse impulsionnelle, l'analyse granulaire cherche à exploiter le caractère local des impulsions détectées, et à décomposer le signal de parole en une suite de fonctions élémentaires basées sur la connaissance du comportement temporel de chaque canal d'analyse [47] [48] [46]. Les impulsions sont d'abord recherchées en utilisant un banc de filtres. Des regroupements locaux d'impulsions entre les canaux adjacents peuvent diminuer leur quantité.

La forme expérimentale des *grains*, ou formes d'ondes élémentaires choisies, tant pour la modélisation dans chaque bande que pour la modélisation de plusieurs bandes regroupées, est celle d'une sinusoïde enveloppée par deux segments de cosinusoïde. Six paramètres caractérisent alors un grain: l'amplitude temporelle, l'instant du maximum (ou instant de référence), deux paramètres pour le temps de montée de la première partie de l'enveloppe et pour le temps de décroissance de la seconde partie, la fréquence porteuse et la phase relative à l'instant de référence. Spectralement ces fonctions élémentaires se présentent comme des réponses impulsionnelles de filtres passe-bas, dont le gain est lié aux paramètres d'enveloppe, et qui sont modulés à la fréquence de la porteuse (avec une certaine phase initiale).

Les premières étapes de l'analyse sont semblables à celles développées pour l'analyse impulsionnelle: filtrage, puis extraction d'enveloppe temporelle.

L'analyse granulaire va de plus représenter finement et modéliser les signaux dans chaque bande, plutôt que d'estimer une mesure globale de cohérence. Pour chaque im-

pulsion, la fréquence dominante est évaluée en mesurant les passages par zéro. Cette mesure est elle-même pondérée par l'emplacement du maximum de l'enveloppe temporelle dans la bande considérée. Les filtres étant régulièrement disposés sur une échelle de Bark, il est raisonnable de tenter l'approximation du signal temporel dans chaque canal en utilisant la connaissance des extrema de la courbe d'enveloppe temporelle et la mesure de la fréquence dominante pour chaque maximum. Ce procédé n'est pas mathématiquement exact, mais expérimentalement, l'erreur commise à ce stade n'est pas ou très peu perceptible.

A partir de ce premier résultat, la représentation du signal comme une somme de fonctions élémentaires dans chaque bande d'analyse, on peut regrouper des grains mis à jour. La représentation à ce niveau est très largement redondante, et l'extraction de paramètres acoustiques pertinents passe par celle de *structures* au sein de l'ensemble des grains.

Les grains qui, à un instant donné, présentent un comportement similaire sont regroupés. Ce regroupement doit représenter une même composante physique: par exemple la réponse à une impulsion glottique dans la région d'un maximum de la fonction de transfert du conduit vocal, ou bien l'explosion d'un plosive. Ce regroupement s'effectue par des procédés analogues à ceux développés dans les modèles du système auditif périphérique: mesure de la synchronisation ou de la simultanéité des réponses entre des bandes de fréquence adjacentes. Des paramètres acoustiques tels que les fréquences formantiques semblent émerger de façon robuste par le degré de ressemblance entre canaux voisins.

Le regroupement de canaux adjacents transforme donc la décomposition spectrale fixe initiale en une décomposition adaptée aux composantes, ou du moins aux régions spectrales privilégiées, du signal. Ce regroupement s'opère par un processus itératif, en ajoutant au canal qui répond le mieux dans une région du plan spectro-temporel une fraction du signal présent dans les canaux voisins, dont la réponse est plus faible mais proche.

Une analogie avec les modèles du système auditif périphérique envisagerait ces regroupements sous l'angle à la fois tonotopique (on regroupe sur la composante qui répond le plus fortement à une fréquence donnée), et temporel: (on regroupe les composantes synchronisées).

Le processus de regroupement induit une perte de qualité, entrainée par le cumul de l'imprécision dans chaque bande, qui reste néanmoins faible et n'entame pas l'intérêt de la méthode. Ce sont surtout les basses fréquences qui paraissent responsables de cette perte de qualité: les formes d'ondes élémentaires sont plus difficiles à isoler par l'enveloppe temporelle des signaux, puisque les composantes présentes dans cette région sont à bande étroite.

2.4.3 relation entre analyse en formes d'ondes élémentaires et modèles auditifs

contraintes retenues

D'après les propriétés déduites de l'analyse auditive, en s'inspirant des travaux menés en analyse impulsionnelle puis granulaire et des méthodes citées au premier

chapitre, il s'agit d'examiner et de discuter une décomposition en formes d'ondes élémentaires. La contrainte de reconstitution du signal à partir des résultats de l'analyse, à chaque étape de l'analyse, conduit à l'utilisation de méthodes particulières et à l'éviction de certains traitements, pourtant psychologiquement ou physiologiquement justifiés.

Les étapes de l'analyse proposée peuvent s'énoncer de la façon suivante, qui ne reflète pas systématiquement l'ordre d'enchainement des traitements:

- analyse du signal par un banc de filtres, avec un nombre assez important de bandes d'analyse;
- redressement du signal dans chaque bande;
- intégration temporelle dans chaque bande;
- décomposition temporelle, pour chaque bande, par le choix des instants de regroupement temporel;
- estimation spectrale par calcul de l'énergie présente dans chaque bande;
- estimation spectrale, pour chaque instant de regroupement temporel, par calcul temporel et local de la fréquence dominante;
- intégration spectrale, par recherche des fréquences de regroupement;

Les figures présentées dans cette section proviennent d'un logiciel d'analyse en formes d'ondes implémenté par P. Blanchet, en collaboration avec J.S. Liénard et l'auteur.

filtrage

La première étape produit un ensemble de signaux issus d'un banc de filtres [49]. Les filtres sont choisis en accord avec les données psycho-acoustiques sur une échelle de Bark. De plus, la contrainte de reconstruction impose des filtres à déphasage nul. Une échelle linéaire pour la répartition des filtres entraîne une résolution trop fine pour l'aigu du spectre, et une échelle logarithmique entraîne une résolution trop fine du grave, par rapport à la perception. L'échelle de Bark présente un compromis perceptivement justifié entre ces deux échelles, et l'analyse qui s'y rapporte un compromis entre transformée de Fourier à court terme et transformée en ondelettes. Comme ces deux types d'analyse s'interprètent dans des cadres communs (filtrage ou formes d'ondes élémentaires), la disparité introduite ne soulève pas de problèmes spécifiques. En particulier, si l'on s'intéresse à une représentation sous la forme d'une somme discrète, le pavage ne paraît régulier que dans un plan spectro-temporel en échelle linéaire/Bark. Par contre la décomposition exacte en formes d'ondes élémentaires reste possible dans ce cadre pour un choix adapté des ondelettes d'analyse, et de leur échantillonnage spectro-temporel.

A titre d'exemple, les figures 2.6, 2.7 montrent les réponses à une impulsion pour des bancs de 80 filtres distribués sur les trois types d'échelle. Seules les alternances positives sont portées, pour plus de clarté. Pour la figure 2.6 l'échelle est linéaire, la largeur de bande des filtres est calculée comme pour un spectrographe à bande large

(300 Hz) pour le schéma du haut, et comme pour un spectrographe à bande étroite (50 Hz) pour le schéma du bas. Dans le haut de la figure 2.7 l'échelle est logarithmique, comme pour une analyse par ondelettes, dans le bas il s'agit d'une échelle de Bark. La comparaison de ces deux figures éclaire donc les rapports entre ces types d'échelles. Les filtres présentés sont à phase nulle, pour pouvoir reconstruire le signal sans retard, et leurs réponses impulsionnelles et gains sont assez éloignés de ceux représentés sur les figures 2.2, 2.3.

Les canaux du banc de filtres se recouvrent spectralement et introduisent une forte redondance, utile pour la représentation graphique et pour des traitements ultérieurs.

Les réponses de ces trois types de bancs de filtres pour un même signal de parole sont portées sur les figures 2.8 et 2.9. Le premier schéma de la figure 2.8 se rapporte au premier schéma de 2.6: échelle linéaire avec 300 Hz de largeur de bande. Quatre formants sont visibles (il s'agit de quelques périodes d'un /a/), puisque la largeur de bande des filtres d'analyse est assez importante: le tracé représente le grossissement d'une portion de spectrogramme en bande large. Le second schéma de cette figure se rapporte à une échelle logarithmique, comme sur la figure 2.7. Le premier formant est décomposé en harmoniques, à cause d'un résolution très fine en basses fréquences, au contraire, les troisième et quatrième formants ne sont plus distincts, à cause de la résolution trop faible en haute fréquence. Le haut de la figure 2.9 présente le cas d'une échelle de Bark, comme sur la figure 2.7. Ce cas est clairement intermédiaire entre les deux précédents: dans le grave du spectre le premier formant est décomposé en harmoniques, mais la résolution dans le haut du spectre reste assez fine pour distinguer les deux derniers formants. Cette figure est à rapprocher de la figure 2.5: le déphasage entre grave et aigu du spectre présent dans le modèle basé sur la physiologie disparaît ici afin de pouvoir reconstruire le signal. Les bancs de filtres utilisent 80 bandes, et sont donc largement redondants. Les réponses dans le voisinage de composantes d'amplitude importante sont clairement synchronisées.

estimation spectrale des fréquences dominantes

Le filtrage représente en soi une façon d'estimer le spectre du signal: l'échantillonnage fréquentiel obtenu permet d'estimer les amplitudes spectrales par examen de l'énergie présente dans chaque bande. Un traitement de ce type se rencontre dans de nombreux systèmes de reconnaissance et présente une analogie avec le codage tonotopique. Cette estimation spectrale, d'une économie indiscutable, peut manquer de robustesse en présence de bruit.

redressement

Après l'étape de filtrage, un redressement des signaux dans chaque bande d'analyse va permettre de calculer leurs enveloppes temporelles, et d'estimer des fréquences dominantes. Il est remarquable, au moins aux intensités normales, que les réponses du nerf auditif soient sensibles seulement à une alternance du signal: le redressement possède ainsi une analogie physiologique. Le redressement simple alternance permet de détecter simplement les périodicités. Le second schéma de la figure 2.9 montre le signal du premier schéma après redressement. Ce tracé est à comparer avec celui du second

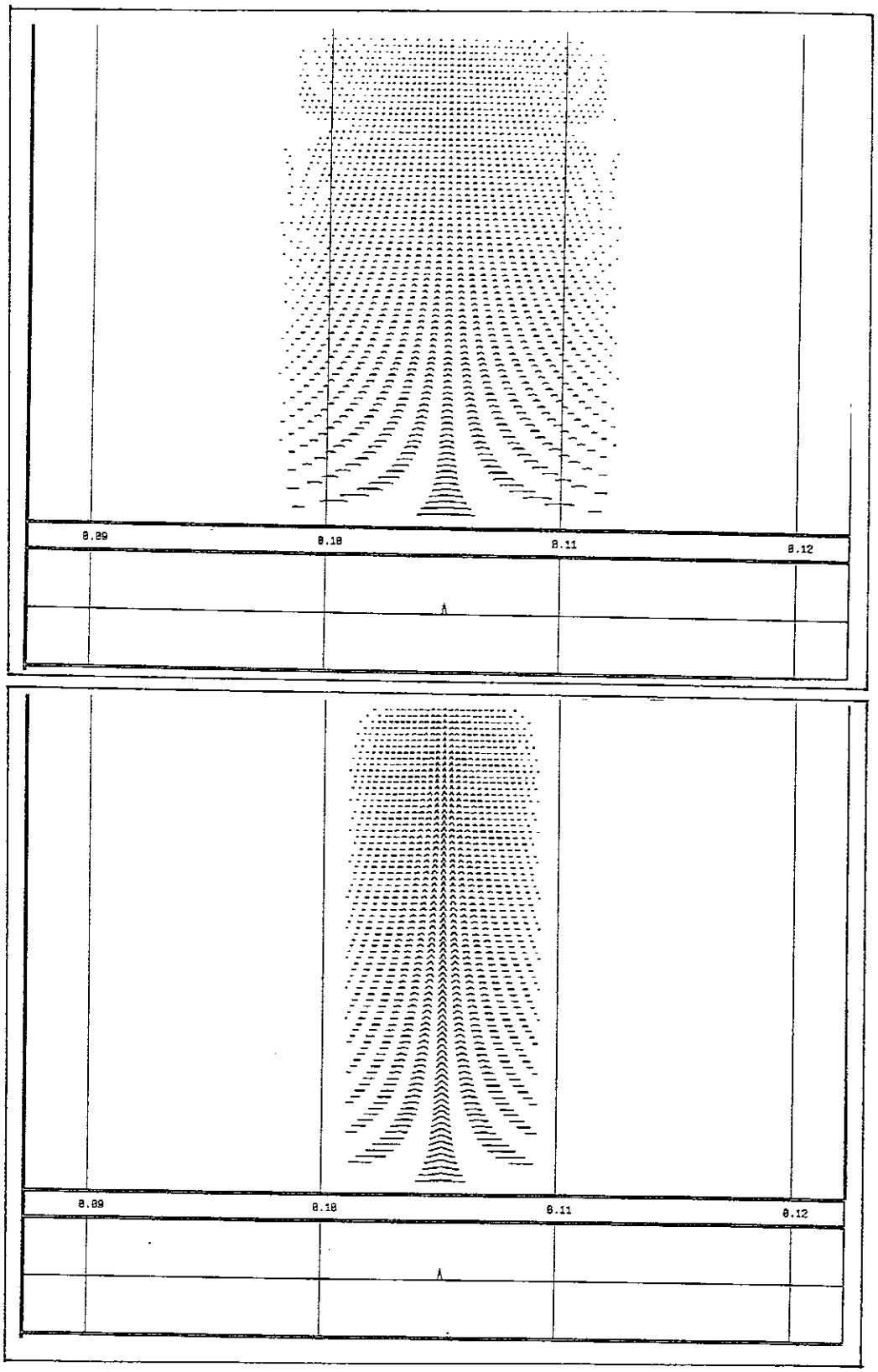

Figure 2.6: réponse impulsionnelle d'un banc de 80 filtres, de résolution constante 300 puis 50 Hz (voir texte)

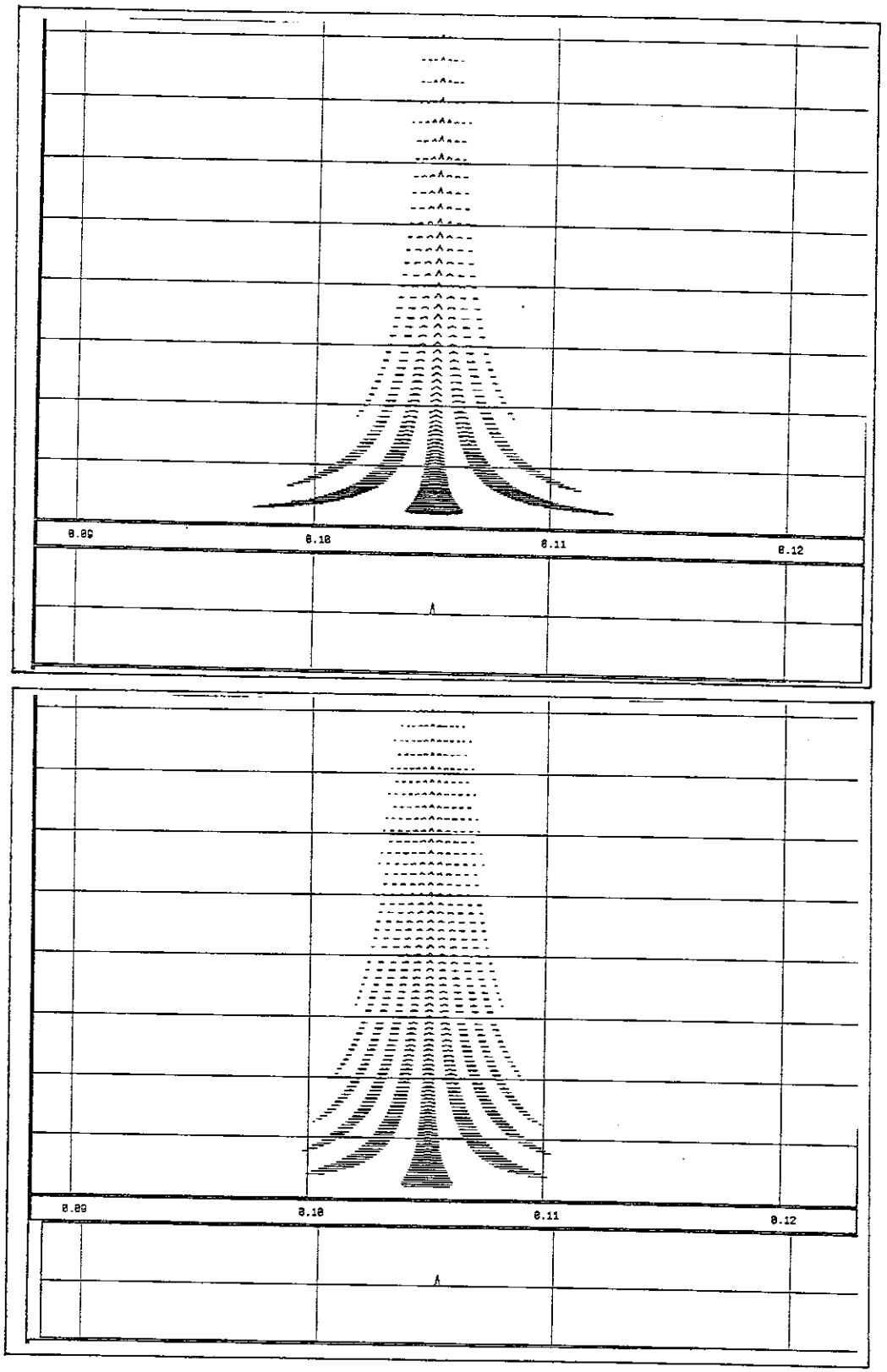

Figure 2.7: réponse impulsionnelle d'un banc de 80 filtres, de résolution relative constante 0.3, puis de résolution constante 1 Bark (voir texte)

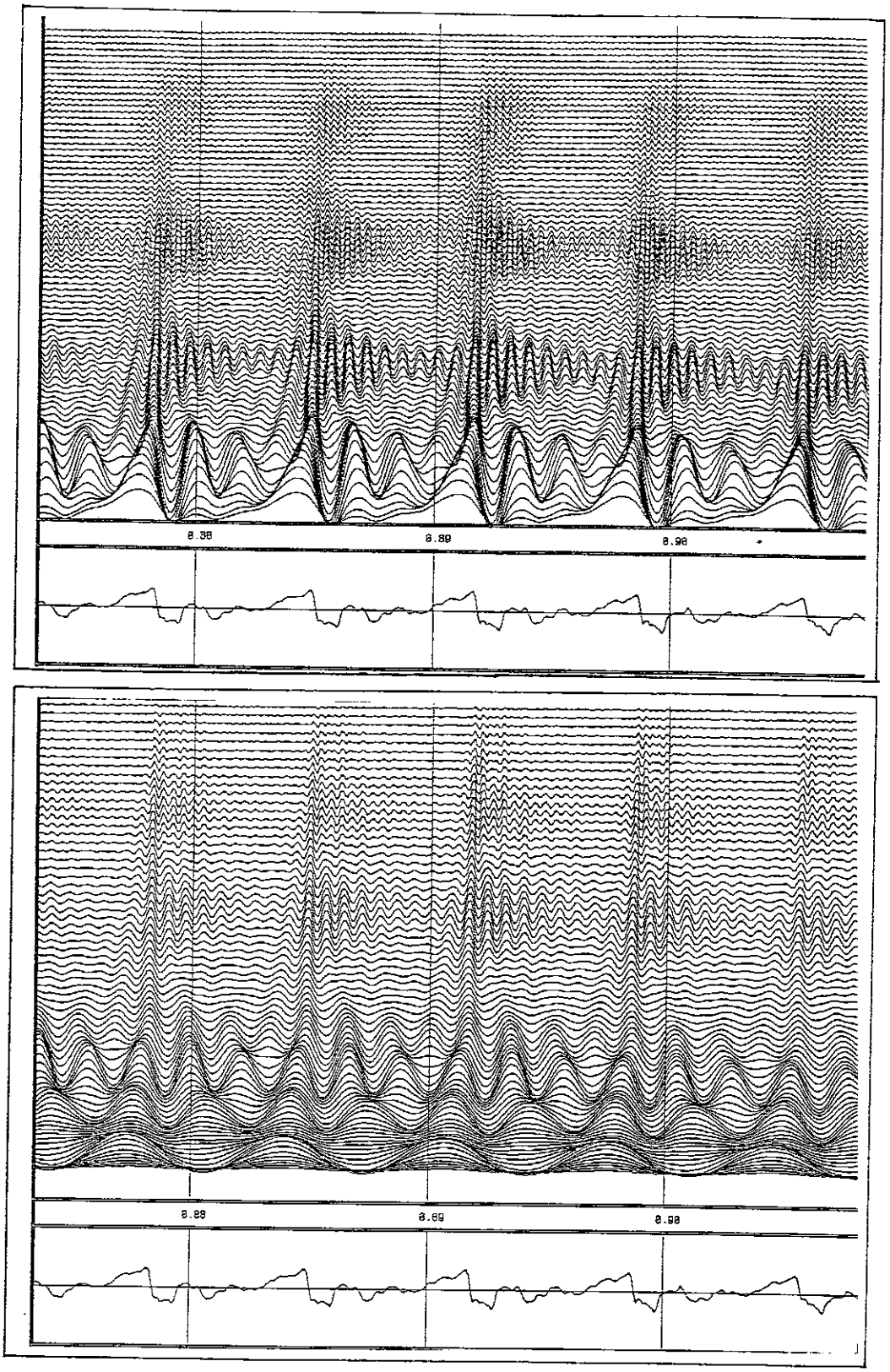

Figure 2.8: analyse d'un /a/ en 80 bandes, filtres de résolutions constantes 300 Hz, puis de résolutions relatives constantes 0.3 (voir texte)

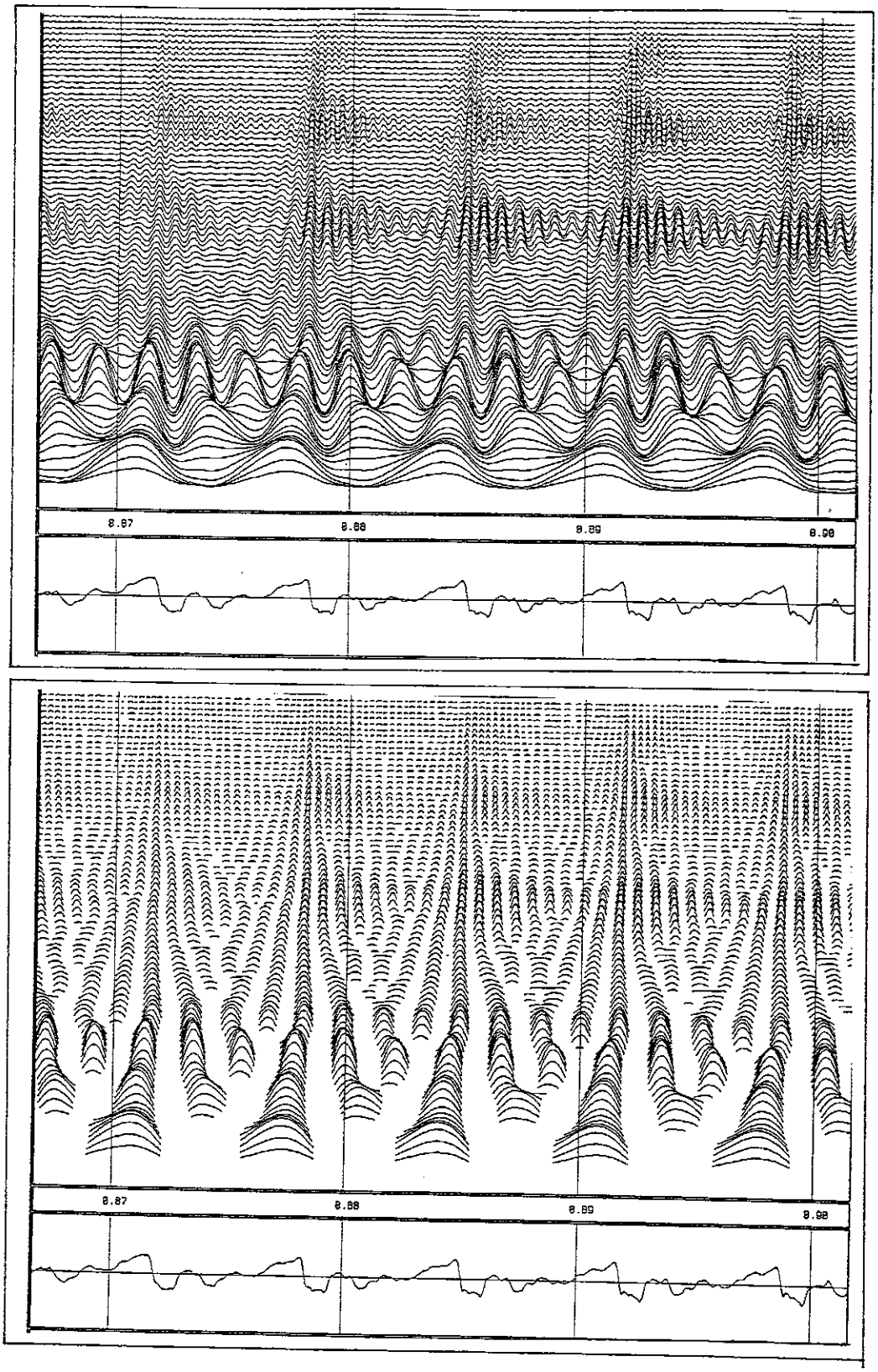

Figure 2.9: analyse d'un /a/ en 80 bandes, filtres de résolutions constantes 1 Bark. Signaux temporels, puis redressement simple alternance (voir texte)

schéma de la figure 2.4.

intégration temporelle et adaptation à court terme

Les signaux ainsi redressés préservent certaines caractéristiques temporelles liées à la fréquence de la bande à laquelle ils appartiennent. Une constante de temps commune à toutes les bandes intervient par l'intégration temporelle des signaux. Un filtrage passe-bas de fréquence de coupure d'environ $500Hz$, correspondant à une constante de temps de 2 ms, permet de définir une enveloppe à court terme.

Cette constante de temps est choisie de façon à observer dans le signal des phénomènes qui portent un sens du point de vue de la production, période fondamentale par exemple.

On peut rapprocher cette intégration temporelle d'une adaptation à court terme. En recherchant les extrema de l'enveloppe ainsi définie et en affectant au maximum un rôle particulier, comme dans l'analyse granulaire ou dans l'analyse impulsionnelle, on obtient un codage des variations du signal dans chaque bande. Un signal qui présente des caractéristiques constantes en regard de cette dimension temporelle sera représenté avec une seule forme d'onde, alors qu'un signal dont l'enveloppe évolue beaucoup sera représenté par plusieurs formes d'ondes.

La région spectrale qui entoure la fréquence de coupure de ce filtre d'intégration temporelle est bien sur le siège de phénomènes particuliers qui ne sont que des artefacts de calcul. Dans cette région les variations d'enveloppe et les composantes analysées sont d'un ordre de grandeur comparable. Ce problème peut provoquer une redondance parasite dans la décomposition temporelle qui suivra: les formes d'ondes obtenues dans un signal de parole voisée peuvent par exemple se répéter suivant un multiple du fondamental plutôt que suivant le fondamental lui-même.

Les regroupements temporels sont donc susceptibles de se justifier par la constante de temps de l'adaptation à court terme, bien qu'il ne s'agisse encore que d'une analogie. L'examen des dimensions temporelles du processus de production conduit également à une constante de temps de l'ordre de la milliseconde. Cette résolution propose un premier ordre de grandeur temporel pour le traitement de la parole. Les niveaux suivants, du point de vue des dimensions temporelles, seront par exemple le niveau des phonèmes ou niveau *segmental*, de l'ordre de la dizaine de grains, puis le niveau des prosodèmes, ou niveau *supra segmental*, de l'ordre de la centaine de grains.

décomposition temporelle

La connaissance des courbes d'enveloppe dans chaque bande d'analyse permet de décomposer le signal dans les régions des maxima. Ainsi, un procédé adapté au signal permet de le représenter par un ensemble de formes d'ondes élémentaires, dont la dimension fréquentielle est donnée par la largeur de bande d'analyse, et dont la dimension temporelle est donnée par l'intégration temporelle de l'étape précédente.

Ce procédé diffère d'une représentation temps-fréquence pour laquelle les fonctions élémentaires d'analyse sont réparties sur une grille régulière dans un système d'échelle donné. On peut par contre le rapprocher des analyses synchrones au fondamental que l'on rencontre couramment en traitement de la parole. Ici cependant, la recherche est

localisée fréquentiellement, et les formes d'ondes des différentes bandes de fréquence sont *a priori* indépendantes. La mesure de cette indépendance représente en soi une information sur le signal.

La forme exacte des fonctions d'enveloppe utilisées pour délimiter la région temporelle d'une forme d'onde ne semble pas revêtir une importance extrême: une reconstitution présentant de faibles ondulations d'enveloppe dues à ces fonctions ne sera perceptivement pas différente de l'original. La façon la plus simple de procéder sans introduire de points anguleux dans l'enveloppe est d'utiliser des segments de sinusoïdes, comme dans l'analyse granulaire. Le seul rôle de ces fonctions d'enveloppe temporelle étant de décomposer le signal de façon aussi lisse que possible, leurs caractéristiques spectrales n'ont pas beaucoup d'importance. En décomposant le signal par des segments de sinusoïde, la somme des fonctions d'enveloppe peut être égale à l'unité.

Le bas de la figure 2.10 présente la décomposition temporelle d'un /u/, en utilisant 20 bandes d'analyse, et le haut de la figure 2.11 celle du morceau de phrase / as-tu vu ce .../. Grâce à l'intégration temporelle choisie, les périodes de voisement sont bien visibles. Chaque forme d'onde est représentée à la fréquence centrale du filtre d'analyse, par un losange dont la longueur figure la durée, la hauteur l'amplitude, et le sommet l'instant de maximum de l'enveloppe temporelle.

estimation temporelle des fréquences dominantes

La décomposition temporelle précédente permet de connaître la localisation temporelle des maxima d'énergie du signal dans chaque bande. Il s'agit de détecter la fréquence dominante relative à un maximum temporel. Cette étape, dans l'analyse granulaire, utilise une mesure pondérée des passages par zéro. Dans le modèle de Ghitza, c'est le passage par un ensemble de seuils d'amplitude répartis sur une échelle logarithmique. Une approche complémentaire basée sur l'auto-corrélation est développée par Seneff par l'utilisation de détecteurs généralisés de synchronies. Une estimation spectrale par des méthodes temporelles résulte de ces calculs. Cette estimation spectrale est temporellement mieux localisée, à une échelle d'intégration temporelle qui résulte du filtrage passe bas précédent. Cette échelle est beaucoup plus faible, d'un rapport 10 environ, que celle résultant d'une analyse habituelle par trames. Une amélioration des performances semble résulter, en présence de bruit, [28] [35] de ce type d'estimation spectrale .

Le bas de la figure 2.11 et le haut de la figure 2.10 montre l'affichage des formes d'ondes aux fréquences détectées. Il s'agit des mêmes formes d'ondes que sur l'autre partie de ces figures. La détection des fréquences dominantes permet de faire apparaître la structure fréquentielle des signaux analysés: les harmoniques en basse fréquence, puis les formants sont visibles. La visualisation dans le plan temps/fréquence est obtenue par des méthodes très différentes d'un spectrographe classique, bien que des structures comparables apparaissent dans le tracé. La figure 2.12 montre la phrase /as-tu vu ce fameux lapin ?/, en utilisant 20 bandes d'analyse. La figure 2.13 montre la même phrase avec 80 bandes d'analyse: le tracé est renforcé près des événements spectro-temporels importants du signal, mais les deux figures sont assez comparables.

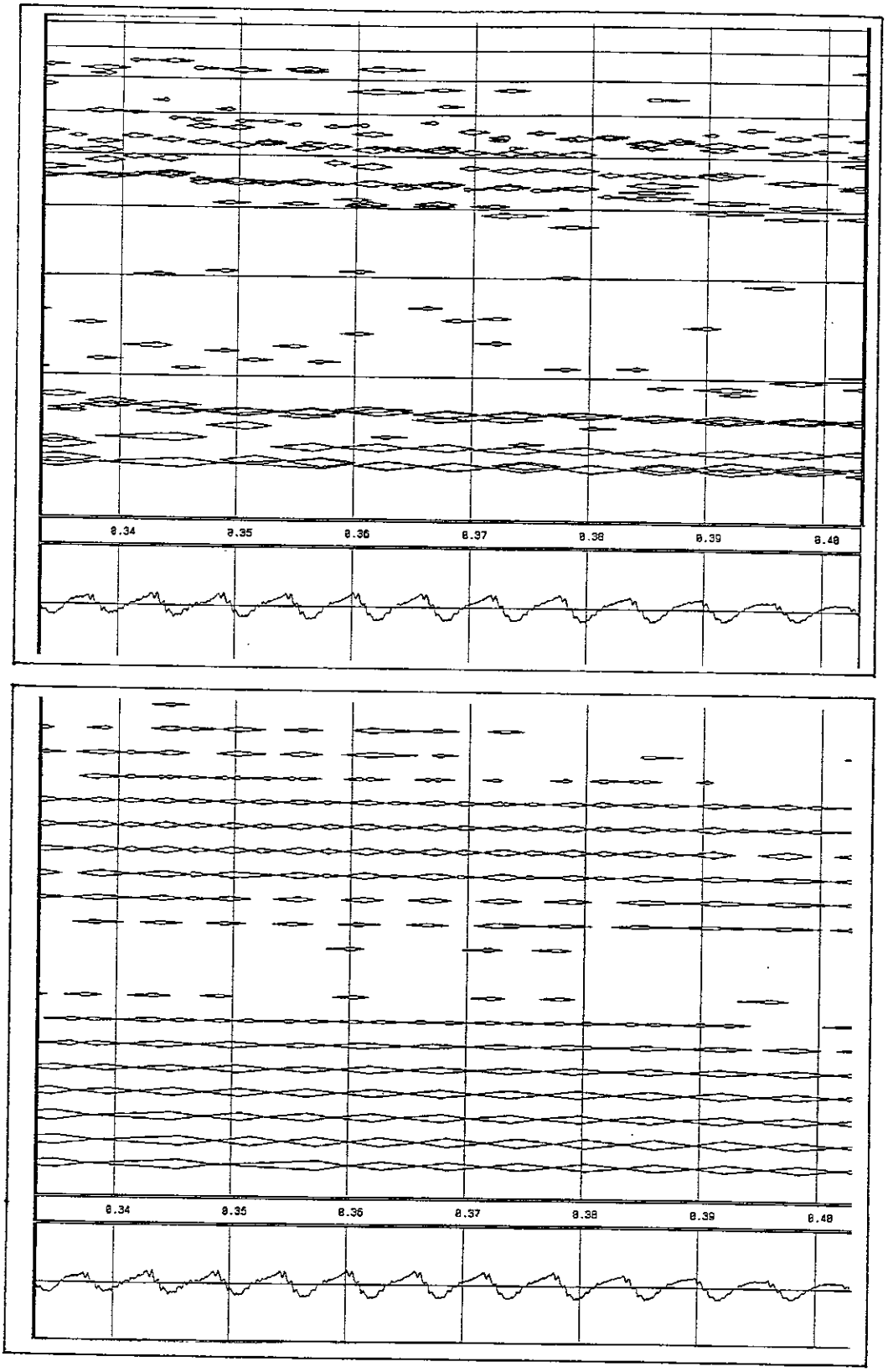

Figure 2.10: analyse d'un /u/ en 20 bandes, de résolution constante 1 Bark: détection des formes d'ondes dans chaque bande, recherche des fréquences dominantes (voir texte)

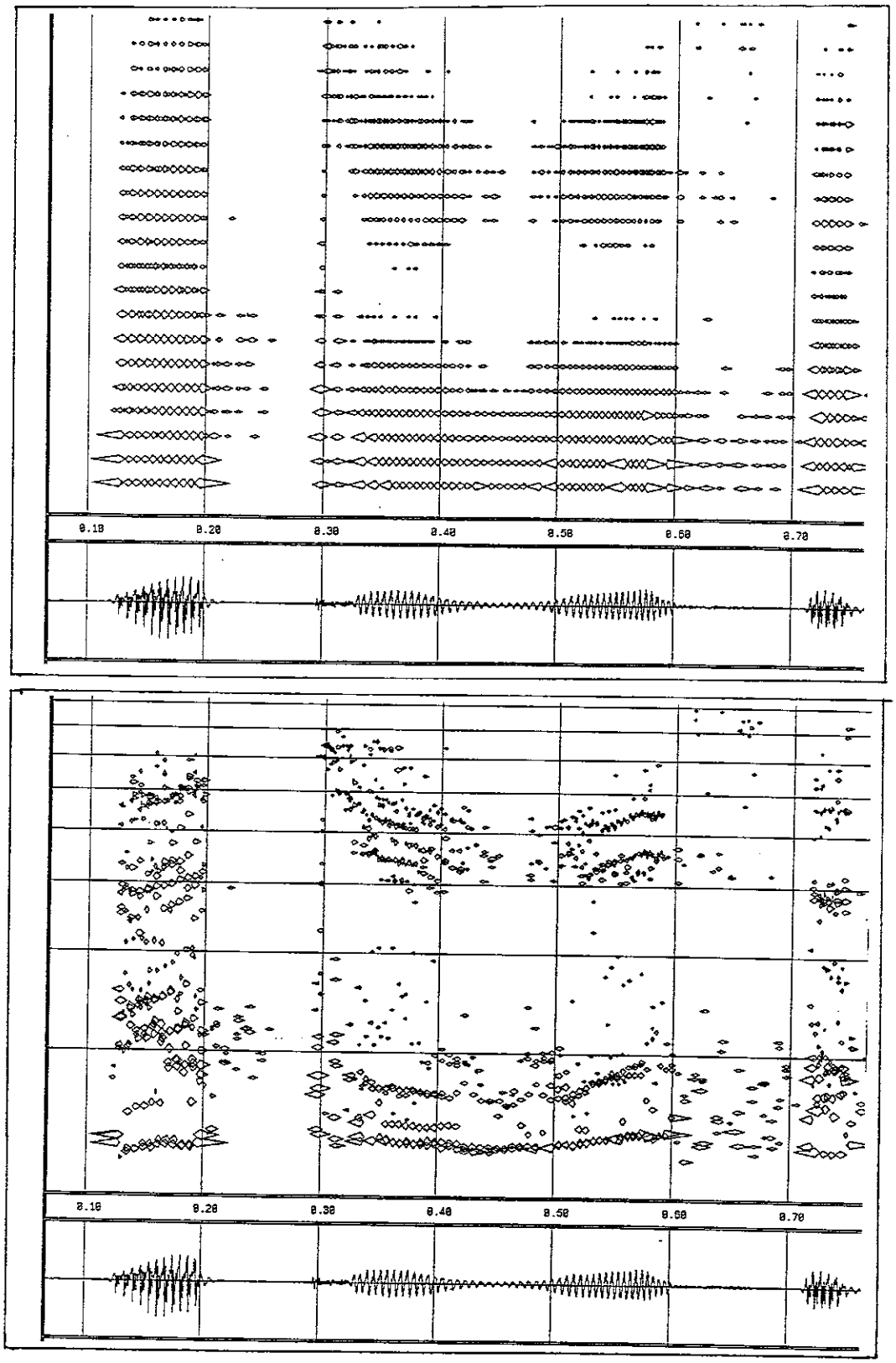

Figure 2.11: analyse de /as-tu vu ce ...en 20 bandes, de résolution constante 1 Bark: détection des formes d'ondes dans chaque bande, recherche des fréquences dominantes (voir texte)

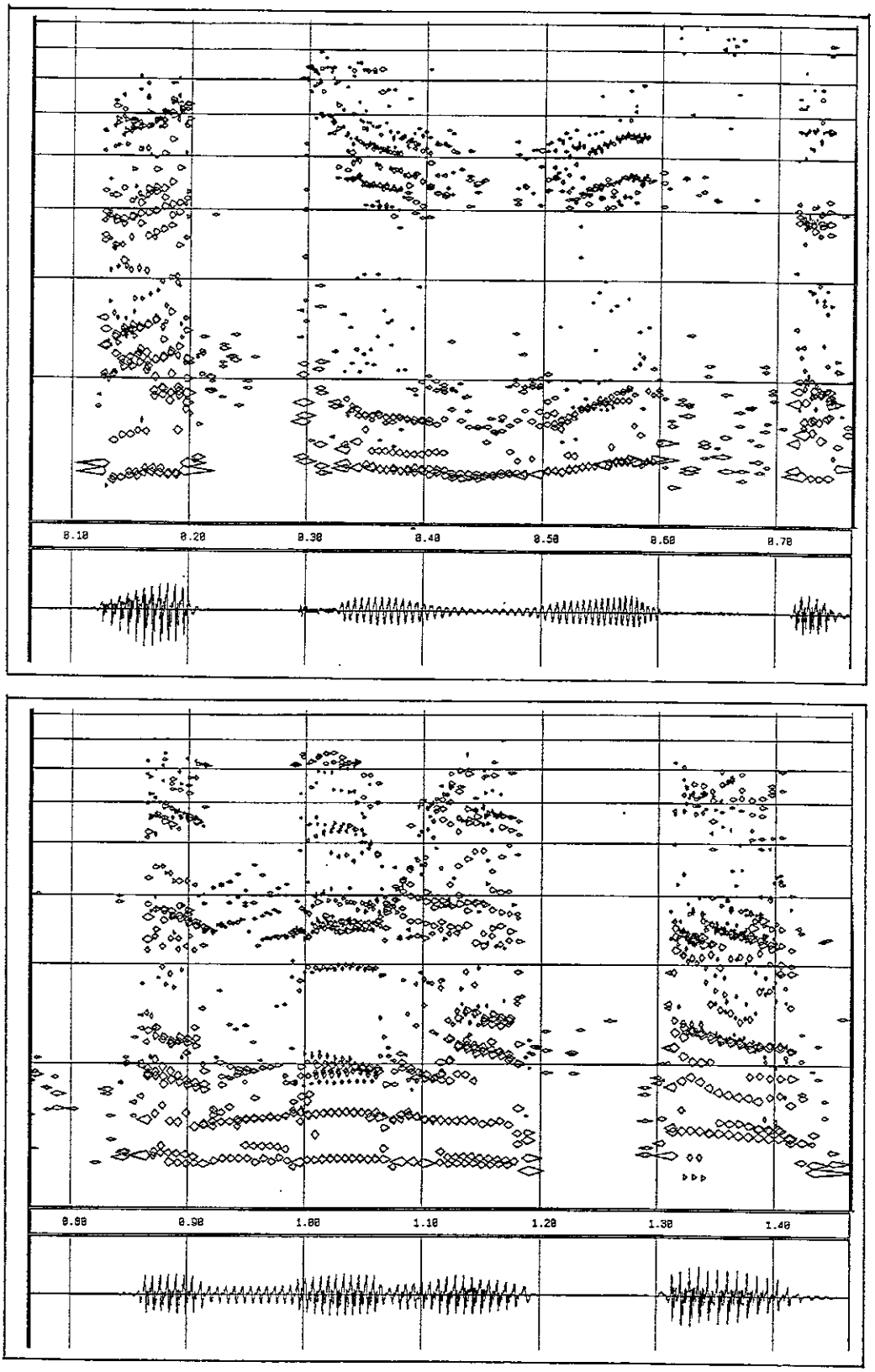

Figure 2.12: analyse de la phrases /as-tu vu ce fameux lapin ?./ en 20 bandes, de résolution constante 1 Bark, répartition régulière sur une échelle de Bark: détection des formes d'ondes dans chaque bande et détection de la fréquence dominante (voir texte).

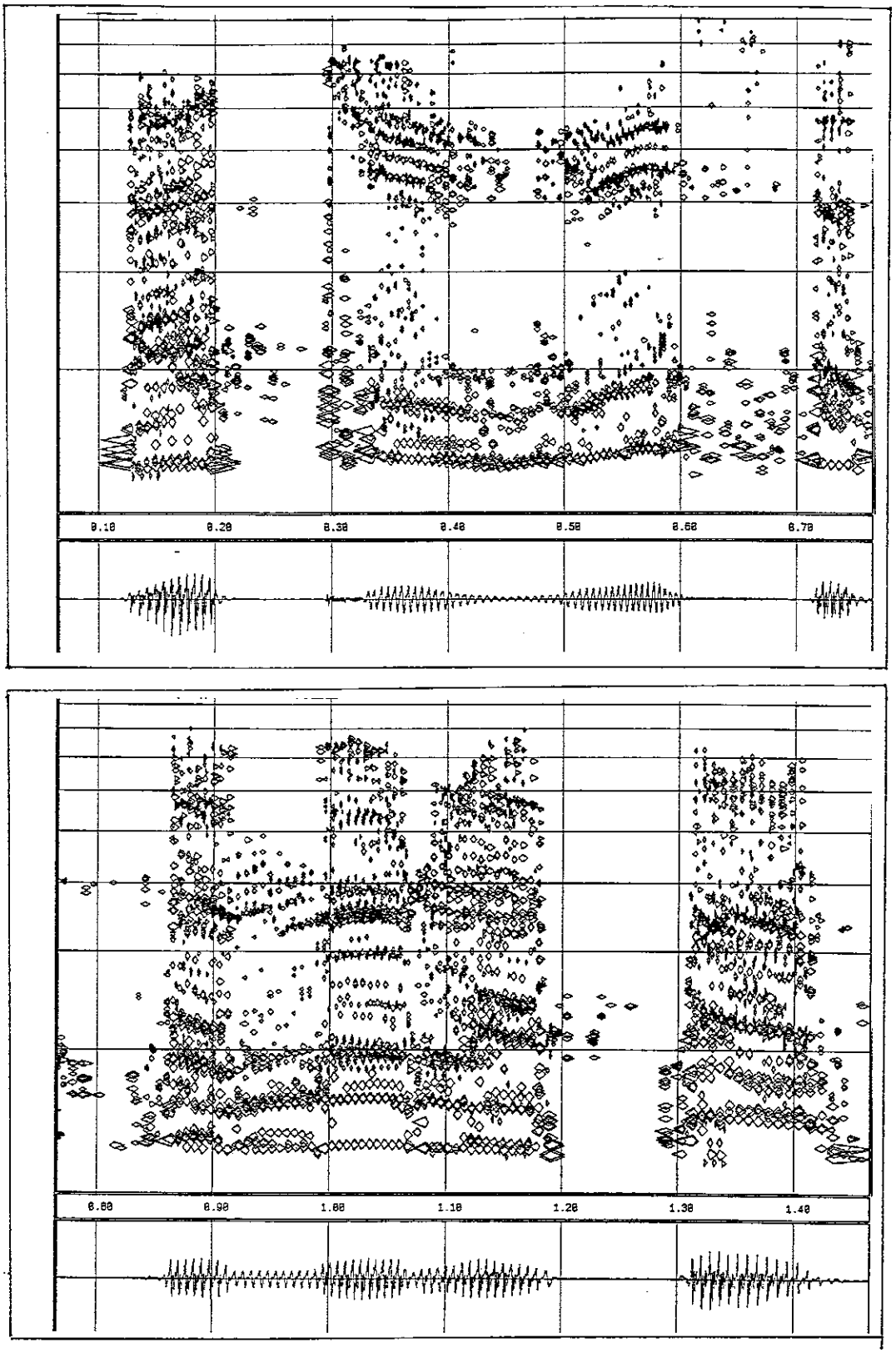

Figure 2.13: analyse de la phrases /as-tu vu ce fameux lapin ?./ en 80 bandes, de résolution constante 1 Bark, répartition régulière sur une échelle de Bark: détection des formes d'ondes dans chaque bande et détection de la fréquence dominante (voir texte).

décomposition spectrale

La redondance fréquentielle introduite par le large recouvrement des filtres d'analyse est nécessaire pour effectuer des corrélations entre canaux. Le regroupement des canaux comportant des composantes communes permet par contre de réduire cette redondance, une fois que les caractéristiques utiles du signal ont été extraites. La méthode de regroupement associe les canaux qui présentent des composantes fréquentielles communes en sommant les contributions de plusieurs canaux, dans l'analyse granulaire. Les grains sont ensuite seulement recherchés et modélisés. Ce procédé entraîne une dégradation de qualité à la resynthèse, car les signaux sont moins simples à cause de leur plus grande largeur de bande et des décalages temporels entre les bandes.

De même que pour les regroupements temporels, les regroupements fréquentiels entraînent une localisation fréquentielle de l'information en quelque sorte diffuse contenue dans les diverses bandes voisines. Cette concentration peut s'envisager comme une analogie avec l'augmentation du contraste fréquentiel dans l'oreille interne. L'analogie n'est pas exacte, et des phénomènes comme la suppression à deux tons par exemple ne sont pas explicitement pris en compte à ce niveau.

reconstructions exactes

Si l'on considère l'analyse en formes d'ondes élémentaires basée sur des critères perceptifs comme un procédé d'analyse/synthèse, les conditions pour une reconstruction parfaite, à chaque étape, doivent être explicitées.

La notion même de *reconstruction exacte* mérite cependant une attention particulière. La reconstruction du signal par les méthodes non-paramétriques citées aux chapitre 1 est mathématiquement parfaite, aux imprécisions de calcul près, et si l'on respecte un certain nombre de contraintes. De nombreuses méthodes de représentation du signal de parole, basées sur un modèle de ce signal, proposent un procédé de reconstruction qui présente une minimisation de l'erreur commise par rapport aux paramètres du modèle, et non une reconstruction mathématiquement exacte. Le jugement de l'oreille, en dernier ressort est seul capable de décider en pratique de la qualité d'une reconstruction, par des tests d'évaluations.

Les problèmes de reconstruction résultant de l'analyse granulaire vont être examinés à chaque étape.

La première étape, filtrage par N canaux de gain G_i $i \in [1, N]$ répartis sur une certaine échelle, conduit à une reconstruction parfaite en sommant la sortie de tous les canaux si:

$$G(\nu) = \sum_{i=1}^{N} G_i(\nu) = 1(\nu) \tag{2.8}$$

En pratique une telle condition est réalisée par des filtres se recouvrant fréquentiellement, et l'ondulation résultante du gain total $G(\nu)$ peut être rendue très faible.

L'étape suivante, qui risque d'introduire des erreurs de reconstruction, est la décomposition temporelle. On suppose que la courbe d'enveloppe, avec les restrictions évoquées pour le grave du spectre, est segmentée en extrema. Ici encore, canal par canal, une

reconstruction parfaite implique que les K_i fonctions de décomposition temporelle $d_{k_i}(t)$ indexée par k_i dans le ième canal, soient de somme égale à l'unité:

$$d_i(t) = \sum_{k_i=1}^{K_i} d_{k_i}(t) = 1(t) \tag{2.9}$$

Dans l'analyse granulaire, pour satisfaire cette condition les fonctions d'enveloppe sont formées de deux segments de sinusoïdes.

La dernière condition pour une reconstruction parfaite dans chaque bande, et donc globalement, est de calculer exactement le signal qui oscille à l'intérieur de chaque forme d'onde. Ce n'est possible de façon simple que si l'on suppose ce signal de fréquence fixe durant toute la forme d'onde. Le signal n'a *a priori* aucune raison de satisfaire cette contrainte: la largeur de bande des filtres en haute fréquence par exemple est bien trop importante. En regard de ces largeurs de bandes, le théorème de Shannon permet de prédire la fenêtre temporelle correspondante, mais l'intégration temporelle, qui a permis la décomposition temporelle, est en contradiction avec ce critère raisonnable. Cependant, si les formes d'ondes se placent aux points dominants du signal, dans le domaine spectro-temporel, les entorses au théorème d'échantillonnage semble perceptivement négligeables.

Si de plus on cherche à regrouper plusieurs canaux, le problème de la modélisation du signal oscillant dans chaque bande se complique d'autant, du moins si on cherche à conserver tous les signaux présents dans le banc de filtres.

L'étude en cours, sur la perception des formes d'ondes élémentaires utilisées, devra permettre de juger de la pertinence de leurs paramètres en tant qu'événements acoustiques isolés, et dans le contexte de la parole. L'évaluation de l'importance de ces paramètres doit permettre de préciser des critères de reconstruction auditivement valides.

La figure 2.14 résume les analogies entre des faits tirés de l'analyse auditive et des méthodes mises en oeuvre dans l'analyse granulaire.

2.4.4 perspectives d'applications

spectrographe

Les représentations spectrographiques basées sur des données auditives sont l'objet d'un intérêt croissant pour l'examen visuel du signal. Cette étude de la parole visible reste un outil indispensable pour l'appréhension de la parole ou des systèmes de traitement automatique.

L'analyse en formes d'ondes élémentaires offre différents types de visualisation. L'affichage des formes d'ondes élémentaires dans la dimension amplitude/temps aux fréquences correspondantes permet de visualiser des composantes du signal (harmoniques, formants, explosions, bruit fricatif par exemple). L'affichage des paramètres de ces formes d'ondes de façon symbolique dans le plan temps-fréquence offre des renseignements complémentaires.

Figure 2.14: résumé des analogies entre analyse granulaire et analyse auditive.

estimation des paramètres acoustiques

La représentation en formes d'ondes élémentaires de ce chapitre permet de réaliser plusieurs types d'estimation spectrale, et donc de rendre compte de paramètres physiques que l'on analyse par des méthodes spectrales: formants, harmoniques, centre de gravité d'un bruit fricatif. La précision temporelle de la méthode permet également de traiter des paramètres acoustiques que la phonétique considère, mais qui sont difficilement analysables par les méthodes classiques, à cause du fenêtrage et de l'utilisation de trames: structure temporelle du bruit fricatif, explosion de plosives par exemple. Certains de ces paramètres peuvent se relier, comme on l'a vu, à ceux utilisés dans des modèles informatiques du système auditif périphérique [77].

processus de regroupement et de traitement des formes acoustiques

La représentation a été conçue pour permettre l'étude du signal de parole à des niveaux d'observation différents: le regroupement de formes d'ondes élémentaires permet de constituer et de manipuler des *objets* acoustiques [32], puis phonétiques par exemple. Une description du signal comme structure composée d'objets de plus en plus simples, et qui présentent un caractère pertinent à un niveau de traitement donné forme donc une motivation importante de cette représentation: une voyelle par exemple peut se décomposer comme un ensemble d'impulsions complexes (périodes de voisement) ou comme un ensemble de formants, ces périodes ou formants étant eux-mêmes des ensembles de formes d'ondes élémentaires. La manipulation d'objets complexes pose cependant des problèmes informatiques difficiles, dont l'utilisation de la programmation orientée objet conçue pour des tâches dans un certain sens similaires [67] est une solution possible.

Le traitement de scènes auditives, la caractérisation d'événements articulatoires ou phonétiques s'inscrivent dans cette perspective.

reconnaissance par des méthodes connexionnistes

La représentation granulaire peut s'adapter à un système de reconnaissance basé sur l'utilisation d'un réseau de cellules de traitement simples [10] [43]. Un système de reconnaissance de parole, dont dérive un système de reconnaissance de caractères écrits, utilisant la propagation guidée de signaux dans un réseau d'automates à seuils présente actuellement un étage de paramétrisation acoustique dont les résultats sont voisins de ceux de l'analyse en formes d'ondes élémentaires: un ensemble d'éléments spectro-temporels localisés. Ce prétraitement est inspiré de données psycho-acoustiques, et l'application de l'analyse granulaire semble une évolution naturelle, qui permet un contrôle précis de la dégradation ou de la pertinence de l'analyse acoustique préliminaire.

2.5 conclusion

L'analyse de la perception des phénomènes acoustiques, en particulier de la parole, par le système auditif apporte deux types complémentaires de renseignements: des données psychoacoustiques qui sont issues d'une approche plutôt externe de la

perception sonore, et des données électro-physiologiques qui proviennent d'une approche plutôt interne. Les deux aspects de l'analyse auditive laissent apparaître une décomposition spectro-temporelle du signal acoustique. La modélisation du système auditif périphérique offre de plus des représentations explicites pour comprendre cette décomposition.

Par rapport aux méthodes classiques de paramétrisation du signal de parole, ces contraintes auditives soulignent le rôle de l'analyse temporelle, tant au niveau de l'estimation spectrale locale qu'au niveau de la constitution d'objets perceptifs plus globaux: période de voisement, explosions, structure du bruit de frication par exemple. Des dimensions spectrales et temporelles particulières se déduisent de l'analyse auditive, qui diffèrent des dimensions empiriques utilisées ordinairement.

L'analyse granulaire, analyse en formes d'ondes élémentaires inspirée de la modélisation auditive, offre un procédé automatique d'analyse/synthèse. Les paramètres des formes d'ondes élémentaires peuvent recevoir, dans certaines limites qui ont été évoquées, une interprétation non dénuée de sens du point de vue de l'analogie avec la perception.

Par rapport aux méthodes exactes du premier chapitre, une perte de précision paraît par contre inévitable. L'erreur d'estimation des paramètres de chaque forme d'onde élémentaire résulte de l'incompatibilité des choix de résolution et d'échantillonnage spectro-temporels issus de la perception.

Des applications variées au traitement automatique de la parole se présentent pour une telle analyse: spectrographie, analyse de paramètres acoustiques, traitement de formes acoustiques, reconnaissances par un réseau connexionniste par exemple.

La validation de la méthode en tant que représentation du signal d'un point de vue perceptif doit passer par l'étude psycho-acoustique des formes d'ondes élémentaires utilisées, en particulier dans le contexte du signal de parole.

Bibliographie Chapitre 2

Bibliographie Chapitre 2

[1] P. J. Abbas, M. B. Sachs, 1976. *Two-tone suppression in the auditory-nerve fibers: extension of a stimulus response relationship.* JASA, Vol. 59, No. 1, Janvier 1976.

[2] C. Abry, C. Benoit, L. J. Boë, R. Sock, 1985. *Un choix d'événements pour l'organisation temporelle du signal de parole* 15èmes Journées d'Etude sur la Parole du GALF.

[3] W.A. Ainsworth, 1988. *Speech recognition by machine.* Peter Peregrinus Ltd, Londres.

[4] J. B. Allen, 1980. *Cochlear modelling- 1980* Proceedings of IEEE-ICASSP-80.

[5] J. B. Allen, 1985. *Cochlear modelling* IEEE ASSP magazine, Janvier 1985.

[6] J. M. Aran, A. Dancer, J. M. Dolmazon, R. Pujol, P. Tran Ban Huy, 1988. *Physiologie de la cochlée* Edition INSERM, Paris.

[7] V. Asta, J. S. Liénard, F. Manceron, 1979. *L'icophone logiciel: un synthétiseur par formes d'ondes* 10èmes Journées d'Etude sur la Parole du GALF.

[8] B. S. Atal, J. R. Remde, 1982. *A new model of LPC excitation for producing natural sounding speech at low bit rates* Proceedings of IEEE-ICASSP-82.

[9] G. Von Békézy, 1960. *Experiments in hearing* MacGraw-Hill, New York.

[10] D. Béroule, 1985. *Un modèle de mémoire adaptative, dynamique et associative pour le traitement de la parole* Thèse de 3ème cycle, Université Paris XI, Mai 1985.

[11] D. Béroule, J. L. Schwartz, 1986. *Essai de formalisation de faits et hypothèses de physiologie concernant le traitement de l'information pour la reconnaissance automatique de la parole* 15èmes Journées d'Etude sur la Parole du GALF.

[12] C. Biétry, 1986. *Synthèse de haute qualité de la parole: étude multi-voix des caractéristiques individuelles de la parole.* Thèse de troisième cycle.

[13] A. S. Bregman, 1984. *Auditory scene analysis* IEEE seventh international conference on pattern recognition, Montréal, Juillet-Aout 1984.

[14] J. Caelen, 1979. *Un modèle d'oreille, analyse de la parole continue, reconnaissance phonémique* Thèse d'état, Toulouse, Juin 1979.

[15] J. C. Caerou, J. M. Dolmazon, V. S. Shupljakov, 1986. *Modélisation active de l'ensemble cochléaire: une nouvelle approche des non linéarités de fonctionnement* 15èmes Journées d'Etude sur la Parole du GALF.

[16] R. Carlson, B. Granström, 1982. *The representation of speech in the peripheral auditory system* Elsevier Biomedical Press, Amsterdam.

[17] I. A. Chistovich, M. P. Granstrem, V. A. Kozhevnikov, L. W. Lesogor, V. S. Shupljakov, P. A. Taljasin, W. A. Tjulkov, 1974. *A functional model of signal processing in the peripheral auditory system* Acustica, Vol. 31.

[18] M. P. Cooke, 1986. *A computer model of peripheral auditory processing uncorporating phase-locking, suppression and adaptation effects* Speech communication, Vol. 5, No. 3-4, Décembre 1986.

[19] R. A. Cole (éditeur), 1980. *Perception and production of fluent speech* Lawrence Erlbaum Associates, Hillsdale, New Jersey.

[20] B. Delgutte, 1980. *Representation of speech-like sounds in the discharge pattern of auditory-nerve fibers* JASA, Vol. 68, No. 3, Septembre 1980.

[21] B. Delgutte, N.S. Kiang, 1984. *Speech coding in the auditory nerve:*

I. Vowel like sound,

II. Processing schemes for vowel-like sounds,

III. Voiceless fricative consonants,

IV. Sounds with consonant-like dynamic characteristics,

V. Vowel in background noise. JASA, Vol. 75, No. 3, Mars 1984.

[22] B. Delgutte, 1984. *Codage de la parole dans le nerf auditif* Thèse d'état, Université Paris VI, Juin 1984.

[23] J. M. Dolmazon, L. Bastet, V. S. Shupljakov, 1977. *A functional model of the peripheral auditory system in speech processing* Proceedings of IEEE-ICASSP-77.

[24] J. L. Flanagan, 1972. *Speech analysis, synthesis and perception* Springer-Verlag, Berlin.

[25] P. Fraisse, 1966. *La psychologie expérimentale* Presses Universitaires de France, Paris.

[26] D. H. Friedman, 1979. *Multichannel zero-crossing intervals pitch estimation* Proceedings of IEEE-ICASSP-79.

[27] O. Ghitza, 1986. *Auditory nerve representation as a front-end for speech recognition in a noisy environment* Computer speech & Language, Vol. 1, No 1, Academic press, Décembre 1986.

[28] O. Ghitza, 1987. *Robustness against noise: the role of timing-synchrony measurement* Proceedings of IEEE-ICASSP-87.

[29] O. Ghitza, 1985. *A measure of in-synchrony regions in the auditory nerve firing patterns as a basis for speech vocoding* Proceedings of IEEE-ICASSP-85.

[30] R. Goldhor, 1983. *A speech signal processing system based on peropheral auditory model* Proceedings of IEEE-ICASSP-83.

[31] A. Gribenski, 1951. *L'audition* Presses Universitaires de France, Paris.

[32] S. Grovel, 1987. *Un outil de description et de manipulation des entités du signal de parole* Rapport de D.E.A. d'informatique, Université Paris XI, Septembre 1987.

[33] W. M. Hartmann, 1978. *The effect of amplitude envelope on the pitch of sine wave tones* JASA, Vol. 63, No. 4, Avril 1978.

[34] M. J. Hunt, C. Lefèbvre, 1986. *Speech recognition using a cochlear model* Proceedings of IEEE-ICASSP-86.

[35] M. J. Hunt, C. Lefèbvre, 1987. *Speech recognition using an auditory model with pitch-synchronous analysis* Proceedings of IEEE-ICASSP-87.

[36] A. J. M. Houtsma, T. D. Rossing, W. M. Wagenaars, 1987. *Auditory demonstration* Acoustical Society of America, CD 1126-061, Philips, Septembre 1987.

[37] B. M. Johnstone G. K. Yates, 1974. *Basilar membrane tuning curves in the guinea pig.* JASA, Vol. 55, No. 3, Mars 1974.

[38] W. D. Keitel, W. D. Neff (éditeurs), 1975. *Handbook of sensory physiology, Volume V, Auditory system* Springer-Verlag, Berlin.

[39] D. T. Kemp, 1978. *Stimulated acoustic emissions from within the human auditory system* JASA, Vol. 64, No. 5, Novembre 1978.

[40] N. Y. S. Kiang, E. C. Moxon, 1974. *Tail of tuning curves of auditory-nerve fibers.* JASA, Vol. 55, No. 3, Mars 1974.

[41] N. Y. S. Kiang, 1980. *Processing of speech by the auditory nervous system* JASA, Vol. 68, No. 3, Septembre 1980.

[42] K. F. Kraiss, J. Moral (éditeurs), 1976. *An introduction to human engineering* Verlag TÜV Rheinland GmbH, Cologne.

[43] J. Leboeuf, 1988. *Un sytème connexionniste appliqué au traitement automatique de la parole* Thèse de 3ème cycle, Université Paris XI, Octobre 1988.

[44] M.C. Liberman, 1978. *Auditory-nerve response from cats raised in a low-noise chamber* JASA, Vol. 63, No. 2, Fevrier 1978.

[45] J. S. Liénard, F. Manceron, 1981. *Analyse impulsionnelle de la parole: expériences préliminaires* 12èmes Journées d'Etude sur la Parole du GALF.

[46] J. S. Liénard, 1985. *Analyse à très court terme de la parole: un outil et quelques directions de recherche* 15èmes Journées d'Etude sur la Parole du GALF.

[47] J. S. Liénard, 1985. *Very short-time analysis of speech: a tool, some preliminary notions, and new illustration AT&T* Bell Laboratories report, non publié, Juin 1985.

[48] J. S. Liénard, 1987. *Speech analysis and reconstruction using short-time, elementary waveforms* Proceedings of IEEE-ICASSP-87.

[49] J. S. Liénard, C. d'Alessandro, 1989. *Wavelet transform and granular analysis of speech* in *Wavelets, Time-frequency methods and phase space*, J.M. Combes, A. Grossman et P. Tchamitchian editeurs, Springer-Verlag, Berlin, 1989.

[50] P. H. Lindsay, D. A. Norman, 1977. *Human information processing* Academic Press, Orlando, Floride.

[51] R. F. Lyon, 1982. *A computational model of filtering, detection, and compression in the cochlea* Proceedings of IEEE-ICASSP-82.

[52] R. F. Lyon, C. A. Mead, 1988. *Cochlear hydrodynamics demystified* Rapport technique du California institute of technologie, Caltech-CS-SR-88-4, Février 1988.

[53] R. F. Lyon, 1986. *Experiments with a computational model of the cochlea* Proceedings of IEEE-ICASSP-86.

[54] R. F. Lyon, N. Lauritzen, 1985. *Processing speech with the multi-serial signal processor* Proceedings of IEEE-ICASSP-85.

[55] R. F. Lyon, 1984. *Computational models of neural auditory processing* Proceedings of IEEE-ICASSP-84.

[56] S. MacAdams, 1984. *The auditory image: a metaphor for musical and psychological research on auditory organisation* in "Cognitive process in the perception of art", W. R. Crozier et A. J. Chapman (éditeurs), Elsevier Science Publishers.

[57] F. Manceron, 1982. *Contribution à l'analyse spectro-temporelle du signal de parole considéré comme une suite d'impulsions acoustiques.* Thèse de docteur-ingénieur, Université Paris XI, Janvier 1982.

[58] C. Marin, 1987. *Rôle de l'enveloppe spectrale dans la perception des sources sonores* Mémoire de DEA de phonétique, Université Paris III, Septembre 1987.

[59] D. W. Massaro, M. M. Cohen, 1983. *Categorical or continuous speech perception: a new test* Speech Communication, Vol. 2, No. 1, Mai 1983.

[60] T. Myers, J. Laver, J. Anderson (éditeurs), 1981. *The cognitive representation of speech* North-Holland publishing company, Amsterdam.

[61] R. N. Niederjohn, Meir Lahat, 1985. *A zero-crossing consistency method for formant tracking of voiced speech in high noise level.* IEEE transaction on ASSP, Vol. ASSP-33, No. 2 , Avril 1985.

[62] C. R. Patisaul, 1976. *Time-frequency resolution experiment in speech analysis and synthesis* JASA, Vol. 58, No. 6, Décembre 1976.

[63] R. R. Pfeiffer, D. O. Kim, 1975. *Cochlear nerve fiber responses: Distribution along the cochlear partition.* JASA, Vol. 58, No. 4, Octobre 1975.

[64] J. O. Pickles, 1982. *An introduction to the physiology of hearing* Academic Press, Londres.

[65] J. C. Risset, 1978. *Hauteur et timbre des sons* Bulletin d'audiophonologie, Vol. 8, No. 3, 1978.

[66] J. C. Risset, 1986. *Son musical et perception auditive* Pour la Science, Novembre 1986.

[67] X. Rodet, P. Cointe, 1984. *Formes: composition and scheduling of processes* Computer Music Journal, Vol. 8, No 3, Septembre 1984.

[68] T. D. Rossing, 1986. *Effects of signal envelope on the pitch of short sinusoidal tones* JASA, Vol. 79, No. 6, Juin 1986.

[69] M. B. Sachs, E. D. Young, 1979. *Encoding of steady-state vowels in the auditory nerve: representation in term of discharge rate.* JASA, Vol. 66, No. 2, Aout 1979.

[70] M. B. Sachs, E. D. Young, 1980. *Effect of nonlinearities on speech encoding in the auditory nerve* JASA, Vol. 68, No. 3, Septembre 1980.

[71] D. Schofield, 1985. *Visualisation of speech based on a model of the peripheral auditory system* Rapport du National Physical Laboratory, NPL DITC 62/85, Londre, Juillet 1985.

[72] M. R. Schroeder, J. L. Hall, 1974. *Model for mechanical to neural tranduction in the auditory receptor* JASA, Vol. 55, No 5, Mai 1974.

[73] J. L. Schwartz, 1987. *Représentation auditive de spectres vocaliques* Thèse d'état, Grenoble, Juillet 1987.

[74] J. L. Schwartz, 1982. *Essai de bilan sur l'organisation de l'information nerveuse à la sortie du système auditif périphérique* Rapport interne du LIMSI, Octobre 1982.

[75] J. L. Schwartz, P. Escudier, 1986. *Le système auditif humain comprend-il un mécanisme d'intégration spectrale à large bande?* 15èmes Journées d'Etude sur la Parole du GALF.

[76] H. A. Schwid, C. D. Geisler, 1982. *Multiple reservoir model of neurotransmitter release by a cochlear inner hair cell* JASA, Vol. 75, No 5, Novembre 1982.

[77] C. L. Searle, J. Zachary Jacobson, S. G. Rayment, 1979. *Stop consonant discrimination based on human audition* JASA, Vol. 65, No 3, Mars 1979.

[78] A. Sekey, B. A. Hanson, 1984. *Improved 1-Bark bandwidth auditory filter* JASA, Vol. 75, No. 6, Juillet 1984.

[79] S. Seneff, 1986. *A computational model for the peripheral auditory system: application to speech recognition research* Proceedings of IEEE-ICASSP-86.

[80] S. Seneff, 1984. *Pitch and spectral estimation of speech based on auditory synchrony model* Proceedings of IEEE-ICASSP-84.

[81] P. Slurovicz, J. L. Goldstein, 1983. *A central spectrum model: a synthesis of auditory-nerve timing and place cues in monaural communication of frequency spectrum* JASA, Vol. 73, No. 4, Avril 1983.

[82] Z. L. Wu, P. Escudier, J. L. Schwartz, R. Sock, 1988. *Caractérisation d'événements articulatori-acoustiques sur un modèle du système auditif périphérique: rôle de l'adaptation nerveuse et de l'inhibition latérale* 17èmes Journées d'Etude sur la Parole de la SFA.

[83] E. D. Young, M. B. Sachs, 1979. *Representation of steady-state vowels in the temporal aspects of the discharge pattern of population of auditory nerve fibers.* JASA, Vol. 66, No. 5, Novembre 1979.

[84] E. Zwicker, R. Feldtkeller, 1967. *L'oreille récepteur d'information.* Traduction Françaisepar C. Sorin, Masson, Paris.

Chapitre 3

DECOMPOSITION EN FORMES D'ONDES ELEMENTAIRES BASEE SUR UN MODELE DU SIGNAL DE PAROLE

3.1 modèles du signal de parole

3.1.1 modèles physiques et modèles mathématiques

utilisation de modèles en traitement de la parole

Le signal de parole possède des propriétés remarquables dues au système particulier qui l'a produit, l'appareil phonatoire, et aux contraintes phonologiques particulières d'une langue donnée, qui limitent le vaste ensemble d'objets sonores susceptibles d'être produits par l'homme.

Les modèles du signal (au sens large de modèles physiques ou modèles mathématiques) présentent de l'intérêt dans les trois principaux domaines du traitement automatique de la parole:

- en codage, les connaissances *a priori* sur le signal à coder permettent de diminuer la quantité de coefficients (qui deviennent des paramètres d'un modèle) à considérer: les cas extrêmes, vocodeur phonétique (50 bits par seconde, cas idéal) ou vocodeurs segmentaux (quelques centaines de bits par seconde) utilisent un modèle phonologique.

- en reconnaissance, l'identification des unités à reconnaître passe par celle des paramètres acoustiques pertinents, donc par un modèle sous-jacent.

- en synthèse, la nécessaire simplicité de commandes de tout dispositif de synthèse oblige à condenser la réalité acoustique en utilisant un nombre réduit de paramètres pertinents, tant au niveau segmental que supra-segmental.

L'utilisation de modèles du signal de parole apparaît dès le commencement du traitement artificiel (par autre chose que l'appareil auditif et l'appareil phonatoire) de la

parole, et se poursuit lors de son étude scientifique. La première représentation de la parole sous une forme différente du signal acoustique, l'écriture, débute par un changement du signifiant, représentation symbolique du signal sous forme pictographique. Alors que l'utilisation des idéogrammes cesse de se développer, l'écriture phonétique linéaire, puis alphabétique, s'accompagne de théories et de modèles phonétiques puis phonologiques du signal de parole [59].

L'avènement de l'ère scientifique issue de la renaissance occidentale s'accompagne des premiers modèles physiques pour la production du signal de parole. F. Bacon, au début du dix-septième siècle, affirme dans *New Atlantis*: *nous produisons encore à volonté des sons articulés et toutes les lettres de l'alphabet, soit les consonnes, soit les voyelles , que nous imitons ainsi que les différentes espèces de voix et de chants des animaux terrestres et des oiseaux.* M. Mersenne, vers 1630, propose dans l' *Harmonie Universelle* un projet de dispositif parlant proche de l'orgue à tuyaux: les voyelles proviennent de tuyaux à embouchure de flûte (le terme de *voyelle* persiste en facture d'orgue pour caractériser la sonorité de ce type de tuyaux), et divers dispositifs sont décrits pour imiter les consonnes. La célèbre machine de Von Kempelen (1791) présente la première forme achevée de synthétiseur de parole, basé également sur un modèle physique, analogue mécanique de l'appareil phonatoire.

Les technologies électriques ont favorisé au début les modèles physiques fonctionnels (vocoder et voder, modèle linéaire de production classique) puis les modèles physiques internes (analogue électrique de l'appareil phonatoire), et les modèles mathématiques du signal de parole.

validation d'un modèle

Une justification rigoureuse du point de vue mathématique ou physiologique des modèles fonctionnels n'est en général pas proposée, ni même sans doute possible.

De même, l'explication des phénomènes physiques sous-jacents pour un modèle mathématique reste dans la plupart des cas sommaire, indépendamment de l'efficacité de ce modèle.

Ainsi, la synthèse a toujours été et reste encore pour beaucoup de modèles la seule jusitification et le seul contrôle de la qualité des choix opérés: quelle que soit l'élégance mathématique du modèle ou sa finesse d'interprétation physique, c'est finalement ce critère perceptif qui est appliqué.

Notons qu'une évolution de ce critère a accompagné l'évolution des techniques et des connaissances, et que la parole *de bonne qualité* d'hier n'est plus considérée comme telle d'aujourd'hui. Des protocoles de test de la qualité pour la parole synthétique ont été proposés, et le problème d'obtention d'un jugement quantitatif reste aigu.

Le modèle proposé doit être compris avec cette restriction: la contrainte retenue est une représentation du signal aussi pertinente que possible du point de vue de la perception. Les paramètres du modèle doivent rendre compte des phénomènes de production perceptivement importants. La modélisation explicite de phénomènes de production complexes est au delà de ce propos, mais une explication de la production en termes intermédiaires, par exemple phonétiques, doit être possible.

3.1.2 production de la parole

L'appareil vocal va être rapidement décrit [39] afin d'introduire le modèle linéaire de production, et toutes les méthodes qui s'y réfèrent. Les principales catégories acoustiques de sons utilisées lors de la phonation seront ensuite citées.

L'appareil vocal humain, est susceptible de produire une grande variété de sons. La parole proprement dite n'exploite qu'un sous ensemble de cette riche palette. Les objets sonores utilisés pour parler dépendent en partie de la langue: des bruits de bouche (clics), par exemple, possèdent dans certaines langues une valeur phonétique qui est absente dans les langues indo-européennes.

les organes de la phonation

On peut décomposer l'appareil phonatoire en plusieurs sous-ensembles fonctionnels:

- les poumons et la trachée-artère;
- le larynx;
- le conduit vocal;

La source d'énergie nécessaire pour produire de la parole réside dans les muscles abdominaux et thoraciques. Les poumons comprimés sous l'action de cette musculature, qui, agissant ainsi de façon analogue à un soufflet, fournissent une pression d'air transformée en son à travers le larynx et le conduit vocal. La trachée-artère, conduit quasi-cylindrique, part des bronches pour aboutir au larynx.

Le larynx, ensemble de muscles, de cartilages articulés, de ligaments et de muqueuses, est compris entre la trachée-artère d'une part et la cavitée pharyngée de l'autre. Les cartilages et ligaments du larynx permettent d'actionner une paire de muscles, les cordes vocales (qui ne sont pas des cordes mais des bandes musculaires) dont l'ouverture porte le nom de glotte. Lorsque les cordes vocales sont séparées, l'air circule librement dans la glotte: un son peut néanmoins être produit, par exemple en voix chuchotée. Le rapprochement des deux membranes et leur accolement produit une fermeture de la glotte, puis, sous l'action de la pression d'air sub-glottique délivrée par les poumons, une impulsion acoustique. L'onde de débit glottique se répète d'après le cycle suivant (théorie myo-élastique):

- les cordes vocales sont accolées et la glotte fermée;
- sous l'action de la pression sub-glottique, les deux bandes de muscle tendent à s'amincir et à s'effacer à l'endroit de la fermeture;
- les cordes s'écartent et laissent passer un jet d'air;
- cette bouffée d'air entraîne une dépression sous la glotte et la force de Bernouilli jointe à l'élasticité des muscles tendent à refermer la glotte;
- les cordes vocales sont accolées et la glotte fermée ...;

C'est la vibration des cordes vocales qui fixe la fréquence fondamentale du signal vocal. La contraction des muscles du larynx contrôle très finement la périodicité du processus décrit plus haut. Les cordes vocales possèdent de plus plusieurs modes de vibration différents, suivant le degré et la forme de leur accolement: de ces modes résultent les différents *registres*: registre de poitrine, registre de tête ou de fausset. D'ordinaire dans la voix parlée un même registre est systématiquement employé, généralement le registre de poitrine chez le locuteur masculin et le registre de tête ou de poitrine chez le locuteur féminin.

Le conduit vocal se décompose en plusieurs parties: pharynx, cavité orale, cavités nasales. Le mouvement des organes articulateurs, lèvres, langue, luette, mâchoires, déterminent la conformation et les évolutions du conduit vocal.

les sons de la parole

L'ensemble d'organes rapidement décrits est susceptible de fonctionnements acoustiques variés. La parole n'en retient qu'un nombre réduit (en supposant une élocution "normale", sans cris, sans pleurs, sans hurlements, gémissements, chuchotements, rires et sans pathologie phonatoire particulière ...). Sans entrer dans le vaste champ de la phonétique articulatoire, nous allons en terme de source d'excitation et de fonction de transfert, grossièrement dresser une typologie des objets sonores rencontrés lors de la phonation [71].

On peut sommairement distinguer trois types de sources sonores (qui peuvent se combiner ou intervenir séparément):

- la vibration des cordes vocales, source quasi-périodique délivrant un signal (dit *voisé*) de durée indéfinie, dans les limites d'une expiration;
- un bruit de frication, signal aléatoire produit par un écoulement d'air turbulent dû à une constriction dans le conduit vocal, également de durée indéfinie;
- une rapide occlusion dans le conduit vocal (lèvres, dôme de la langue/palais, apex de la langue/incisives supérieures ...) dont le relâchement génère une impulsion acoustique. Ici par contre la durée échappe dans une grande mesure au contrôle du locuteur;

Les signaux issus de ces trois types de source sonore élémentaire sont transformés par le conduit vocal, et se propagent ainsi modelés à l'extérieur du système phonatoire par les ouvertures que représentent le nez et la bouche.

Le conduit vocal assure quant à lui la fonction de *filtrage* des signaux de source, le filtre mis en jeu pouvant évoluer avec une certaine rapidité. La cavité formée par le conduit vocal agit, comme toute cavité, en filtre acoustique que l'on peut supposer linéaire en première approximation. Cette cavité possède des résonances, et des anti-résonances (en particulier lors de l'utilisation de la dérivation nasale) qui *forment* ainsi le spectre du signal résultant.

Les organes articulateurs (luette, langue, lèvres, mâchoires) peuvent donner naissance à des évolutions très rapides de la conformation du conduit vocal, et donc de ses propriétés acoustiques(par exemple en parole voisée, d'une période à l'autre les résonances peuvent avoir considérablement évolué).

On peut ainsi distinguer plusieurs types de sons regroupés au sein de classes phonétiques.

les phonèmes du français

La première subdivision phonétique se rapporte au mode d'excitation du conduit vocal, et à la stabilité de ce dernier: c'est la séparation voyelles/consonne. Les voyelles correspondent d'une part à une excitation quasi-périodique (donc assez longue) délivrée par les cordes vocales, et d'autre part à une conformation assez stable du conduit vocal, du moins lorsque la voyelle est prononcée isolément [72] [14].

Dans la parole continue, l'effet de coarticulation dû au mouvement ininterrompu des muscles de l'appareil vocal altère la stabilité du conduit vocal même lors de l'émission des voyelles. Suivant l'ouverture de la dérivation nasale (par abaissement du velum), les voyelles seront nasales ou seront orales.

On distingue les différentes voyelles grâce à l'emplacement des premiers formants. En particulier les deux premiers formants permettent une discrimination grossière. Un schéma classique établit la répartition des voyelles dans le plan F1, F2 (premier formant, deuxième formant) et fait apparaître le *triangle vocalique* (limites de F1 et F2 en fréquence centrale pour les voyelles).

Lorsque l'excitation glottique coexiste avec une évolution rapide du conduit vocal, on assiste à la naissance de semi-voyelle. Suivant la vitesse d'élocution cette évolution sera perçue comme une entité indépendante (semi-voyelle) ou comme une suite de phonèmes distincts (diphtongue).

les consonnes se répartissent en quatre classes principales:

- les fricatives. Pour une fricative sourde, l'excitation est un bruit de friction. Le conduit vocal est ainsi perturbé en amont et en aval de la constriction. La contribution de la partie aval domine la "coloration" du signal de source. Par ailleurs, les cordes vocales peuvent entrer en vibration en même temps que le bruit de friction. Cette modulation du bruit crée une fricative voisée. Les divers lieux de constriction (lieux d'articulation) déterminent les diverses fricatives: fricatives dentales (articulées entre les incisives supérieures et la lèvre inférieure), alvéolaires (aux alvéoles derrière les incisives) et post-alvéolaire (en arrière des alvéoles).

- les plosives. Les plosives forment un sous ensemble des occlusives. Elles résultent du brusque relâchement d'une constriction dans le conduit vocal. Joint à une vibration des cordes vocales, on obtient une plosive voisée, sinon une plosive sourde. L'articulation peut être labiale, dentale ou alvéolaire, vélaire. Un silence précède l'explosion, durant le temps d'occlusion, et, comme pour les fricatives, un petit délai est nécessaire pour l'établissement du voisement.

- les nasales, ou occlusives nasales. Elles sont assez proches des plosives voisées, mais l'abaissement du voile du palais entraîne l'apparition d'anti-résonances dues à la dérivation nasale. L'occlusion de la cavité orale étant complète, le son rayonne depuis les narines.

- les liquides. Assez semblables aux semi-voyelles, elle résultent d'une excitation voisée et de rapides mouvements des articulateurs.

3.1.3 modèle linéaire de production du signal de parole

Les grandes catégories phonétiques décrites peuvent se dériver d'un modèle linéaire fonctionnel pour la production du signal de parole. Ce modèle *classique* de production comporte une source d'excitation e, un filtre f et un terme de rayonnement $r(t)$. On suppose tous les systèmes linéaires et découplés, ce qui n'est réaliste qu'en première approximation [26].

source d'excitation

La source d'excitation e peut adopter différentes formes, en fonction des catégories phonétiques précédentes. La source de voisement est idéalement un train quasi-périodique d'impulsions p passant dans un filtre passe bas qui restitue une réponse impulsionnelle u_g proche de l'onde de débit glottique [76].

$$e(t) = p(t) * u_g(t) = \sum_i \delta(t - ti) * u_g(t) \tag{3.1}$$

De nombreux modèles d'onde de débit glottique ont été proposés [30]. On peut les classer commodément en modèles définis dans le domaine spectral ou modèles définis dans le domaine temporel.

L'aspect temporel de l'onde de débit glottique [107] (ou plutôt de sa dérivée après l'action du terme de rayonnement hors de la tête) détermine souvent la forme grossière du signal temporel pour les voyelles, et les modèles temporels présentent ainsi un intérêt indiscutable. Le spectre de l'onde de débit glottique, qui est extrêmement variable dans le détail [62], , présente une pente d'environ $-12dB/octave$, ce qui explique l'importance perceptuelle de la source de voisement dans le grave du spectre.

Trois modèles temporels classiques pour la forme de l'onde de débit glottique sont:

- en triangle;
- modèle de Rosenberg [99];
- modèle de Fant [28];

Modèle triangulaire, d'amplitude α:

$$u_g(t) = \alpha(t/t1)$$

pour $0 \leq t < t1$

$$u_g(t) = \alpha(1 - ((t - t1)/t2)))$$

pour $t1 \leq t < t1 + t2$

$$u_g(t) = 0$$

pour $t1 + t2 \leq t < t0$
Modèle de Rosenberg, d'amplitude α:

$$u_g(t) = \alpha/2(1 - cos(t\pi/t1))$$

pour $0 \leq t < t1$

$$u_g(t) = \alpha cos(\pi/2(t - t1)/t2)$$

pour $t1 \leq t < t1 + t2$

$$u_g(t) = 0$$

pour $t1 + t2 \leq t < t0$

Modèle de Fant, d'amplitude α et de facteur de décroissance β:

$$u_g(t) = \alpha/2(1 - cos(t\pi/t1))$$

pour $0 \leq t < t1$

$$u_g(t) = (\alpha(\beta(cos(\pi(t - t1)/t2) - 1) + 1))_+$$

pour $t1 \leq t < t1 + t2$

$$u_g(t) = 0$$

pour $t1 + t2 \leq t < t0$

le symbole $()_+$ signifiant que l'on ne garde que la partie positive de la fonction. La figure 3.1 illustre ces trois modèles de forme d'onde de débit glottique dans les domaines temporel et fréquentiel.

Un modèle spectral proposé par Fant, possède quatre pôles $\hat{z}_i$. U_{g0} est une constante de gain [26]:

$$U_g(z) = \frac{U_{g0}}{\prod_{i=1}^{4}(1 - z/\hat{z}_i)} \tag{3.2}$$

Un modèle spectral plus simple, cité par Markel et Gray, est un simple filtre passe-bas du second ordre, avec la constante K [74]:

$$U_g(z) = \frac{1}{(1 - Kz^{-1})^2} \tag{3.3}$$

Le même modèle reste en général utilisable pour les explosions de plosive, la source d'excitation étant une impulsion isolée pour les sourdes, et un train d'impulsions dont l'une domine pour les sonores.

Pour le bruit de frication, la source d'excitation est un bruit blanc n. La source impulsionnelle et la source de bruit doivent pouvoir se mélanger, pour les fricatives voisées par exemple. Ce mélange est constant chez certains locuteurs (voix peu timbrée, souffle), et la dichotomie voisé/non voisé porte une grande part de responsabilité dans la qualité artificielle des systèmes de synthèse qui l'utilisent.

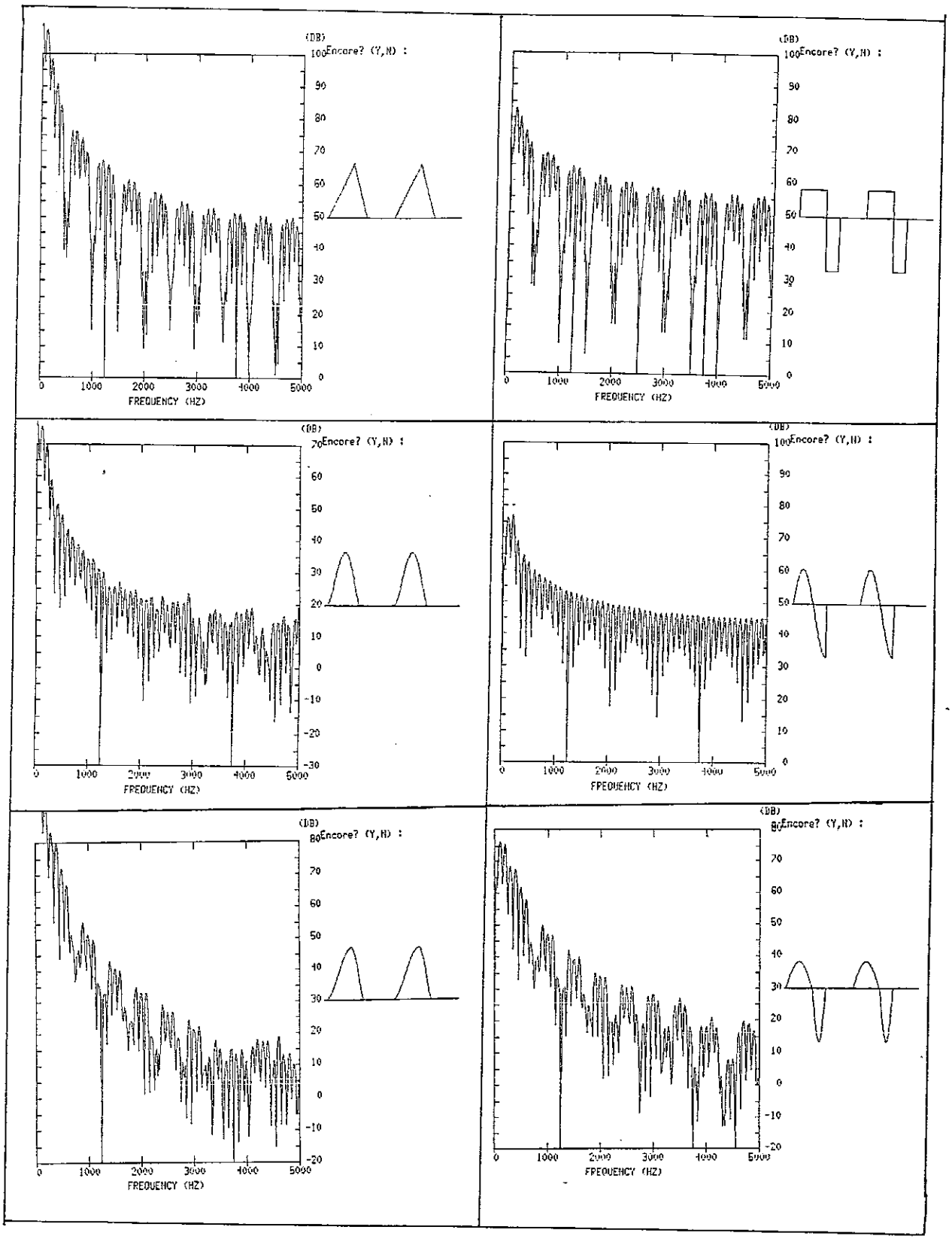

Figure 3.1: trois modèles de forme d'onde de débit glottique: en triangle, de Rosenberg, de Fant.

conduit vocal

Le filtre linéaire associé au conduit vocal possède des zéros et des pôles, afin de rendre compte des formants et des anti-formants. Lors de la production des voyelles orales, le conduit vocal est assimilable à un tube acoustique, de forme complexe qui ne comporte que des résonances, ou formants. Sa fonction de transfert (rapport de l'onde de débit aux lèvres à l'onde de débit glottique), comporte une infinité de pôles. Un terme correcteur $K_{p\infty}(z)$ doit donc être considéré si l'on ne conserve qu'un nombre fini N de pôles conjugués $\hat{z}_i$, $\hat{z}_i^*$. La constante K_p est reliée aux pertes par vibration des parois, (qui restent faibles $K_p \simeq 1$), et la glotte est supposée fermée:

$$V(z) = \frac{U_l}{U_g}(z) = \frac{K_p K_{p\infty}(z)}{\prod_{i=1}^{N}(1 - z/\hat{z}_i)(1 - z/\hat{z}_i^*)} \tag{3.4}$$

Pour prendre en compte le couplage avec les cavités nasales (voyelles nasales, consonnes nasales), ou avec les cavités en arrière de la source d'excitation, (cavités subglottique, pendant la partie du cycle vocalique ou la glotte est ouverte, cavités en arrière d'une constriction: occlusives, fricatives), des zéros $\bar{z}_j$, $\bar{z}_j^*$ doivent être insérés dans la fonction de transfert (avec les termes correcteur K_{zer} et $K_{zer\infty}$ correspondants pour un nombre M fini de zéros):

$$V(z) = \frac{U_l}{U_g}(z) = \frac{K_p K_{zer} K_{p\infty}(z) K_{zer\infty}(z) z \prod_{j=1}^{M}(1 - z/\bar{z}_j)(1 - z/\bar{z}_j^*)}{\prod_{i=1}^{N}(1 - z/\hat{z}_i)(1 - z/\hat{z}_i^*)} \tag{3.5}$$

rayonnement

Le dernier terme représente la conversion de l'onde de débit aux lèvres en onde de pression à une distance d de la tête. Spectralement, un pente d'environ $+6dB/octave$ en résulte, induisant une pente globale (avec celle de l'onde de débit glottique) d'environ $-6dB/octave$ pour le signal de parole. En considérant le rayonnement d'un petit baffle sphérique, à la surface d'une sphère, l'action de cette conversion peut être en première approximation assimilée à un filtrage de préaccentuation (filtrage passe-haut du premier ordre), ou dérivation. La forme d'onde pour une période vocalique est ainsi fortement influencée par la dérivée de l'onde de débit glottique. Ainsi la pression P à distance d de la tête est donnée par:

$$L(z) = \frac{P}{U_l}(z) = 1 - K_d z^{-1} \tag{3.6}$$

Avec $K_d \simeq 1$.

modèle source/filtre simplifié

Le spectre d'énergie d'une impulsion et la densité spectrale de puissance d'un bruit blanc étant identiques (l'autocorrélation d'un bruit blanc est une impulsion), dans le domaine spectral il est commode de simplifier encore ce modèle linéaire en choisissant une source de spectre plat, et en réunissant dans un même filtre les contributions de la glotte, du filtre associé au conduit vocal et du terme de rayonnement. Pour la parole voisée:

$$S(z) = P(z)U_g(z)V(z)L(z) \quad (3.7)$$
$$= P(z)H(z) \quad (3.8)$$

et pour la parole non voisée:

$$S(z) = N(z)V(z)L(z) \quad (3.9)$$
$$= N(z)H(z) \quad (3.10)$$

donc en regroupant les deux cas dans la source E:

$$S(z) = E(z)H(z) \quad (3.11)$$

Cette écriture commode est équivalente au modèle linéaire classique. l'aspect "temporel" (périodicité ou caractère aléatoire) de l'excitation se retrouve dans la source simplifiée E, alors que l'aspect spectral de cette excitation, jointe à l'action du conduit vocal et du rayonnement se retrouve dans le filtre H. Ce filtre comporte des pôles et des zéros, et l'on peut de plus supposer qu'il varie dans le temps, avec une réponse impulsionnelle $h(\tau, t)$.

Une simplification supplémentaire amène à considérer le filtre H comme un filtre tout-pôle. Si l'on choisit la forme de 3.6 pour le terme de rayonnement, et la forme de 3.3 pour l'onde de débit glottique, avec K proche de 1, le numérateur se simplifie et H devient tout-pôle.

$$S(z) = \frac{E(z)}{A(z)} \quad (3.12)$$

avec

$$A(z) = \frac{1}{H(z)} \quad (3.13)$$

et

$$A(z) = \sum_{i=0}^{M} a_i z^{-i} \quad (3.14)$$

Puisque A est tout-zéro.

3.1.4 modélisation spectrale par prédiction linéaire

prédiction linéaire

La modélisation du filtre présent dans le modèle linéaire simplifié est un problème important pour de nombreux traitements. Les méthodes de prédiction linéaire proposent une solution élégante à ce problème [74] [33].

La prédiction linéaire du signal de parole s'appuie sur la corrélation existante entre les échantillons adjacents du signal de parole. La connaissance d'un certain nombre p

de ces échantillons jusqu'à un instant donné $n-1$ permet de prédire approximativement l'échantillon suivant, noté $\hat{s}_n$:

$$s_n \simeq \hat{s}_n = \beta_1 s_{n-1} + \beta_2 s_{n-2} + \ldots + \beta_p s_{n-p} \tag{3.15}$$

donc l'erreur de prédiction ε_n entre signal prédit et signal véritable vaut:

$$\varepsilon_n = s_n - \hat{s}_n = s_n - (\sum_{i=1}^{p} \beta_i s_{n-i}) \tag{3.16}$$

l'erreur quadratique ε_n^2, en notant $\alpha_0 = 1$ et $\alpha_i = -\beta_i$ vaut:

$$\varepsilon_n^2 = (\sum_{i=0}^{p} \alpha_i s_{n-i})^2 \tag{3.17}$$

et l'erreur quadratique totale $\bar{\varepsilon}$:

$$\bar{\varepsilon} = \sum_{n=n0}^{n_1} \varepsilon_n^2 \tag{3.18}$$

en définissant:

$$c_{ij} = \sum_{n=n0}^{n_1} s_{n-i} s_{n-j} \tag{3.19}$$

on obtient pour l'erreur quadratique totale:

$$\bar{\varepsilon} = \sum_{i=0}^{p} \sum_{j=0}^{p} \alpha_i c_{ij} \alpha j \tag{3.20}$$

Les coefficients α_i sont recherchés en choisissant une fenêtre temporelle pour donner un sens à l'erreur totale, et en minimisant l'erreur par rapport aux coefficients. Les minima sont atteints en annulant les dérivées partielles de $\bar{\varepsilon}$ par rapport aux α_i.

La prédiction linéaire suppose que le filtre H est un filtre tout-pôle: un nombre de pôles suffisant permet néanmoins d'analyser correctement les segments qui exigeraient un modèle pôles-zéros (nasales, occlusives).

prédiction linéaire et modèle linéaire de production

La prédiction linéaire du signal de parole entretient des rapports étroits avec le modèle linéaire de production. En effet, en prenant des hypothèses réalistes sur les composantes du modèle linéaire (U_g est un filtre passe-bas du second ordre, L un filtre passe-haut du premier ordre, V un filtre tout pôle), il est possible d'identifier le modèle linéaire de production et le modèle auto-régressif obtenu par prédiction linéaire. L'erreur résiduelle ε_n reçoit alors une interprétation en terme de source d'excitation e, et le filtre A est associé au filtre prédicteur. L'identification du filtre A suppose un résiduel à spectre plat, ce qui en terme d'excitation se traduit par un bruit blanc ou une seule impulsion. La modélisation de la source d'excitation en prédiction linéaire peut donc simplement se réaliser par un générateur d'impulsion et un générateur de bruit blanc, avec une décision voisé/non voisé.

algorithmes

N points de signal étant disponibles en fenêtrant le signal de parole (trame d'analyse), le filtre de prédiction est évalué à des instants régulièrement espacés (généralement toutes les $10ms$ environ).

Différentes méthodes pour le calcul des coefficients de prédiction ont été proposées, unifiées par deux méthodes classiques, la méthode d'autocorrélation et la méthode de covariance. Ces deux méthodes diffèrent par la fenêtre $[n_0, n_1]$ pour calculer l'erreur quadratique totale $\bar{\varepsilon}$. La méthode d'autocorrélation fixe $n_0 = -\infty$ et $n_1 = +\infty$, ce qui entraîne que les coefficients c_{ij} prennent la forme de coefficients de corrélation à court terme. La méthode de covariance fixe $n_0 = p$ et $n_1 = N - 1$.

Ces méthodes évaluent le filtre prédictif à des instants séparés par un intervalle temporel fixe.

Une variante de la méthode d'autocorrélation utilise un algorithme récursif de calculs des coefficients de prédiction: c'est la méthode PARCOR. Les *coefficients de réflexion* ainsi obtenus peuvent s'identifier aux coefficients de réflexion d'un modèle physique multi-tubes du conduit vocal, et des algorithmes efficaces existent ??.

Des méthodes de prédiction linéaire adaptative utilisent une réalisation du filtre 3.14 sous forme de filtre en treillis [68] [69] [74] : ici un jeu de coefficients de prédiction est calculé pour chaque échantillon de signal. Ces méthodes adaptatives interdisent l'usage de fenêtre de durée finie, mais pour accorder plus de poids aux échantillons récents, utilisent une fenêtre infinie décroissante vers les échantillons négatifs (par exemple une exponentielle décroissante). Un tel algorithme sera utilisé par la suite.

La déconvolution entre source d'excitation et filtre par prédiction linéaire présente donc l'avantage de l'efficacité des calculs, et les paramètres formantiques peuvent se déduire de façon assez simple de cette analyse.

détection de formants par prédiction linéaire

L'extraction des formants par prédiction linéaire peut adopter deux chemins: la recherche des racines du filtre d'analyse a, ou bien la recherche des maxima du gain de ce filtre, obtenu par transformée de Fourier des coefficients de prédiction [70] [25] [24] . La première méthode est plus complexe (car il faut regrouper les racines obtenues pour trouver un maximum effectif), mais est exhaustive, en particulier dans le cas de maxima très proches. Les données brutes obtenues par cette recherche des maxima se rapportent aux formants dans $80 - 90\%$ des cas pour la parole voisée.

Néanmoins, l'obtention des formants oblige à des décisions supplémentaires: les courbes de trajectoires formantiques doivent être continues, et relativement lisses, en accord avec le processus physiologique qui les a produites. Si l'on numérote les formants, comme pour une description phonétique, les trajectoires des fréquences centrales peuvent se croiser, ce qui apporte des difficultés suppémentaires. Il s'agit de raccorder d'une trame sur l'autre les maxima obtenus de façon à conserver des courbes acceptables: la détection des maxima est donc un problème simple alors que celle des formants est un problème complexe.

excitation

Les défauts de qualité de l'analyse prédictive sont de deux natures: défaut de modélisation du système (modèle tout-pôle), et surtout défaut de modélisation de l'excitation, par excès de simplification.

L'analyse par prédiction linéaire classique prend une décision voisé/non-voisé pour chaque trame d'analyse. Dans le cas d'un signal voisé, un train périodique d'impulsions séparées par une période fondamentale constitue l'excitation et pour un signal non-voisé, un générateur de bruit. Le mélange entre ces deux sources n'est pas possible, et les paramètres d'excitation sont ainsi extrêmement simples.

Plusieurs méthodes ont été proposées pour pallier le manque de qualité dû à l'excitation:

- le codage séparé de la bande de base (Voice Excited Linear prediction) [40];
- le codage du résiduel (Residual Excited Linear Prediction);
- l'utilisation d'un dictionnaire pour coder le résiduel (Code Excited Linear Prediction) [103];
- le codage de l'excitation par un ensemble d'impulsions adapté au signal (Multi-Pulse Excited Linear prediction) [7] [57];

La prédiction linéaire multi-impulsionnelle apporte une solution au problème de l'excitation en considérant la source d'excitation comme un ensemble d'impulsions. La relation avec le modèle de production est ainsi perdue, ou du moins difficile à établir, en échange d'une qualité transparente. Le choix de l'emplacement et de l'amplitude des impulsions s'appuie sur une procédure récursive de minimisation de l'erreur (erreur quadratique moyenne entre le signal et le signal synthétisé avec n impulsions d'excitation). La validation de cette méthode est expérimentale (test d'écoute), et une valeur moyenne du nombre d'impulsions optimal par trame ($10ms$) varie entre 8 et 12. Il n'y a évidemment plus de distinction entre excitation voisée et non voisée.

3.1.5 autres méthodes de déconvolution source/filtre

méthodes homomorphiques

Un autre procédé efficace de déconvolution entre source d'excitation et conduit vocal est la déconvolution homomorphique [80]. Alors que le filtrage linéaire permet de séparer des composantes combinées linéairement, dans le cas de composantes combinées de façon non-linéaire (multiplication ou convolution), les méthodes homomorphiques permettent de se ramener au cas linéaire.

Le procédé le plus courant de déconvolution homomorphique pour la parole est le *cepstre*. Le cepstre est défini comme la transformée de Fourier inverse du logarithme du module de la transformée de Fourier d'un signal. Donc, en écrivant le signal de parole sous la forme de 3.11, on obtient la relation:

$$log(|\tilde{s}(\nu)|) = log(|\tilde{e}(\nu)\tilde{h}(\nu)|) \quad (3.21)$$
$$= log(|\tilde{e}(\nu)|) + log(|\tilde{h}(\nu)|) \quad (3.22)$$

et en notant $\breve{s}(\kappa)$ le cepstre de $s(t)$, qui dépend de la variable κ ou *quéfrence*, homogène à un temps:

$$\breve{s}(\kappa) = \breve{e}(\kappa) + \breve{h}(\kappa) \quad (3.23)$$

Si les contributions relevant du conduit vocal et les contributions relevant de la source d'excitation évoluent avec des rapidités différentes dans le temps, il devient possible de les séparer par application d'une simple fenêtre temporelle (filtre passe-bas pour le conduit vocal et filtre passe-haut pour la source d'excitation dans le domaine quéfrentiel).

Le conduit vocal possède une contribution fréquentiellement assez lisse, qui aboutit à un cepstre basse quéfrence: les échantillons cepstraux pour le conduit vocal sont près de l'origine.

Réciproquement, le spectre correspondant à la source d'excitation varie rapidement en fonction de la fréquence, tant pour une excitation voisée que pour une excitation bruitée: le cepstre correspondant sera haute quéfrence. La séparation des deux composantes peut donc s'effectuer dans de bonnes conditions, si le premier formant est assez éloigné du fondamental.

La déconvolution homomorphique, comme la prédiction linéaire, permet d'obtenir l'enveloppe spectrale du signal, que l'on assimile au gain d'amplitude du filtre H. Notons que la réponse en phase n'est pas considérée, et sera donc traitée en prenant des hypothèses supplémentaires: systèmes à phase minimum par exemple.

Ce procédé peut être employé pour détecter les formants et la fréquence fondamentale. Pour la détection des formants, le problème se pose de façon analogue au cas de la prédiction linéaire. Le filtrage basse quéfrence du cepstre permet d'obtenir l'enveloppe spectrale, qui comporte plusieurs maxima. La recherche de ces maxima fournit des données brutes, desquelles il s'agira d'extraire les formants, en appliquant des décisions sur la continuité et le lissage des trajectoires formantiques.

retard de groupe

Une autre approche non-paramétrique pour rechercher la localisation des formants utilise le retard de groupe [115] [80].

A partir du signal de parole est formé un signal à phase minimum par transformation de Fourier inverse du spectre d'amplitude. Un fenétrage temporel est appliqué à ce signal, de durée inférieure à celle d'une période de voisement. Une transformée de Fourier du signal obtenu permet de calculer son spectre de phase, et par dérivation par rapport à la fréquence, le retard de groupe. On prouve que le retard de groupe d'une cascade de résonateurs est la somme des retards de groupe de chaque résonateur, et qu'une haute résolution est obtenue dans ce cas.

Une simple recherche des maxima locaux du retard de groupe donne la position des formants, avec une bonne résolution même pour des voix aigues ou des formants très proches.

3.1.6 bande de base et interaction source/filtre

La *bande de base* du signal de parole peut être définie comme la région spectrale en dessous d'environ $800Hz$, ou comme la région moyenne du premier formant. Cette région possède des caractéristiques particulières, dues à la forte interaction entre la contribution de l'onde de débit glottique et la fonction de transfert du conduit vocal [27].

Dans le domaine temporel par exemple, Fant a montré que l'amortissement du premier formant n'est pas uniforme pendant un cycle vocalique, mais est beaucoup plus fort pendant la période d'ouverture de la glotte.

Dans le domaine spectral, la contribution de l'onde de débit glottique et du premier formant sont particulièrement difficiles à séparer: on constate ainsi parfois l'apparition d'un *formant glottique*, dû au spectre de l'onde de débit glottique.

Les modèles temporels simples pour les régions formantiques qui seront décrits plus loin sont donc insuffisants pour la bande de base. Une représentation sinusoïdale (comme il a été proposé en prédiction linéaire [40]) ou une représentation temporelle plus complexe semblent ici nécessaire.

3.2 représentations sinusoïdales

3.2.1 synthèse additive

La représentation de Fourier pour les signaux périodiques entraîne naturellement l'utilisation de sinusoïdes comme fonctions élémentaires pour représenter les signaux acoustiques. La synthèse musicale a très tôt utilisé de telles fonctions pour l'étude d'instruments de musique ou pour créer des sons inouïs [90] dans une méthode qui utilise un ensemble de sinusoïdes en rapport quasi-harmonique: c'est la *synthèse additive* [77]:

$$s(t) = \sum_{k=1}^{M} A_k(t) sin(t(2\pi k\nu_0 + 2\pi\phi_k(t))) \tag{3.24}$$

L'amplitude A_k et la déviation en fréquence ϕ_k varient lentement dans le temps, ν_0 est la fréquence fondamentale et k le rang de l'harmonique.

La production des sons non périodiques reste bien sûr possible. La synthèse additive n'a suscité un intérêt en parole que depuis l'existence de méthodes d'extraction automatique des paramètres.

Une méthode originale d'estimation des coefficients de chaque sinusoïde s'appuie sur la méthode de Prony [53] [52]. Le signal est considéré comme la somme de N sinusoïdes amorties, d'amplitudes A_i, de facteurs d'amortissement α_i, de fréquences fc_i et de phases ϕ_i:

$$x_n = \sum_{i=1}^{N} A_i e^{-\alpha_i n} cos(2\pi f c_i n + \phi_i) \tag{3.25}$$

La méthode de Prony consiste à former une matrice particulière M à partir des échantillons du signal temporel, chaque ligne est une séquence du signal d'entrée décalée d'un échantillon par rapport à la précédente, puis à relier simplement les coefficients de la somme précédente aux racines du modèle auto-régressif obtenu pour les vecteurs du noyau de M.

En négligeant les amortissements on obtient les coefficients des composantes sinusoïdales, et il reste à apparier d'une trame sur l'autre les sinusoïdes estimées.

3.2.2 vocodeur de phase

Le vocodeur de phase apparaît comme une interprétation de l'analyse/synthèse de Fourier à court terme, comme il a été rappelé au premier chapitre. Chaque bande d'analyse délivre une amplitude et une phase, qui serviront à piloter un ensemble de sinusoïdes en synthèse. Si le nombre d'échantillons spectraux est assez important, ce qui diminue en proportion la largeur de bande de chaque filtre, la définition spectrale peut être rendue suffisamment fine pour séparer chaque raie spectrale pour la parole voisée (pour un fondamental à 100 Hz et une bande passante de 5 kHz, l'utilisation de 128 points spectraux suffit pour remplir cette condition). Une représentation de type sinusoïdale, largement redondante, s'obtient donc à partir de l'interprétation en banc de filtres de la transformée de Fourier à court terme.

Cette interprétation sera utilisée avec profit par la suite pour les modèles sinusoïdaux du signal de parole, en remarquant que pour le modèle de production, seules les bandes spectrales localement dominantes méritent d'être conservées (sans hypothèse sur l'harmonicité du signal), pour chaque trame d'analyse.

3.2.3 codage par tons de la bande de base

En dehors de la synthèse additive, une des premières applications en analyse/synthèse automatique de la parole de sinusoïdes comme fonctions de base d'un modèle du signal est le codage de la bande de base [40] .

La bande de base est ici définie comme la région spectrale inférieure à environ 800 Hz. De par la pente de l'enveloppe spectrale du signal de parole, que l'on attribue au terme de rayonnement aux lèvres et à la source de voisement, la contribution du ou des premiers formants dans cette région est difficile à séparer de celle de l'onde de débit glottique.

Le codage par prédiction linéaire s'est rapidement tourné vers le codage de la bande de base comme source d'excitation du filtre de synthèse, à cause de la dégradation importante qui résulte du choix binaire voisé/non voisé et de la détection du fondamental. Une façon élégante de coder la bande de base consiste à la considérer comme la somme d'un certain nombre de sinusoïdes pures, ou tons:

$$bb_k = \sum_{i=1}^{n} a_k^i sin(\psi_k^i) \tag{3.26}$$

La phase instantanée ψ_k^i est l'intégrale de la fréquence instantanée ϕ_k^i:

$$\psi_k^i - \psi_{k-1}^i = 2\pi\phi_k^i \tag{3.27}$$

Le nombre n de tons mis en oeuvre est ici fixé (8 semble suffisant), mais aucune relation n'est exigée entre eux: en particulier aucune relation d'harmonicité.

Pour extraire les paramètres de chaque sinusoïde, l'auteur rejette l'utilisation de méthodes de détection du fondamental, jugées peu fiables, et propose une méthode prédictive basée sur le filtrage de Kalman. Les paramètres sont ainsi estimés pour chaque échantillon puis sous-échantillonnés et codés (un estimateur par ton).

L'algorithme exact n'est pas décrit, mais la méthode semble permettre la réalisation de vocodeurs de bonne qualité et robustes à un débit moyen.

3.2.4 représentation harmonique

harmoniques généralisés

Le *codage harmonique* du signal de parole a été défini pour la parole voisée, dont la structure quasi-périodique appelle une extension de la représentation des sons périodiques par un spectre de raies [6] [5].

La parole voisée peut se représenter comme le filtrage d'un train quasi-périodique d'impulsions:

$$s(t) = \int h(t, t-\tau)e(\tau)d\tau \tag{3.28}$$

où $h(t,\tau)$ représente la valeur à l'instant t de la réponse impulsionnelle du conduit vocal à une excitation à l'intant τ, et $e(t)$ un train quasi-périodique d'impulsions.

En généralisant de façon élégante le comportement spectral d'un train périodique d'impulsions à un train seulement quasi-périodique d'impulsions, on peut définir des *harmoniques généralisés* comme composantes spectrales de $e(t)$, indexées par k:

$$e(t) = \sum_k \Omega(t)e^{ik\Phi(t)} \tag{3.29}$$

Les exponentielles possèdent des amplitudes et des fréquences qui varient (lentement) dans le temps. Pour chaque harmonique généralisé de l'excitation, on peut obtenir un harmonique généralisé du signal voisé:

$$o_k(t) = \int h(t, t-\tau)\Omega(\tau)e^{ik\Phi(\tau)}d\tau \tag{3.30}$$

L'étude directe de ces harmoniques généralisés est difficile car leurs fréquences évoluent dans le temps, de même que le filtre associé au conduit vocal. Une déformation de l'axe temporel permet de transformer le train quasi-périodique d'impulsions en train périodique d'impulsions, et ramène à un cas plus simple, car on montre que cette opération se traduit avec une approximation suffisante par une opération simple sur h.

L'utilisation de la transformée de Fourier à court terme avec origine des temps glissante permet de passer des harmoniques généralisés à un spectre de raies généralisées, modèle non stationnaire de la parole voisée. L'estimation des paramètres passe par celle de la fréquence fondamentale à un instant donné.

utilisation en codage

La dépendance du codage harmonique envers la détection du fondamental reste une de ses faiblesses: cette détection qui peut présenter des difficultés importantes et induire des erreurs ne semble pas en toute généralité fiable. Un autre inconvénient résulte du traitement différent des segments voisés et non voisés, cette distinction n'étant pas toujours pertinente (fricatives voisées par exemple). Une économie et une qualité certaines résultent par contre de la simplicité de description des segments voisés: l'application essentielle de la méthode est le codage [112] [111] [75].

Le codeur complet utilise d'une part le codage harmonique généralisé, mais doit de plus transmettre un résiduel pour permettre le traitement des parties non voisées, et la correction des erreurs pour les parties voisées. Le codage harmonique exploite donc la propriété spectrale relative à une excitation quasi-périodique pour réduire le débit d'information. Les contraintes déduites sur les composantes spectrales proposent une analyse temporelle commune pour toutes les fréquences analysées, et imposent des relations sur les composantes fréquentielles (raies spectrales).

3.2.5 représentation sinusoïdale composite

Un autre type de représentation sinusoïdale est donné par la représentation sinusoïdale composite [100]. Le signal de parole est représenté par:

$$s(t) = \sum_{i=1}^{n} a_i sin(\omega_i(t) + \Phi_i) \tag{3.31}$$

L'analyse consiste à extraire les $2n$ coefficient a_i et ω_i en égalant l'autocorrélation du modèle et du signal de parole correspondant.

Une formule de synthèse équivalente a été proposée, par une méthode de prédiction linéaire modifiée [12]. La contrainte de normalisation des coefficients α_i de prédiction linéaire devient:

$$\sum_{i=0}^{p} \alpha_i^2 = 1 \tag{3.32}$$

On montre que dans ces conditions les pôles du modèle résident sur le cercle unité (donc il n'y a pas de terme d'amortissement dans la formule de synthèse). Le calcul des pôles permet d'estimer les coefficients de 3.31.

3.2.6 représentation sinusoïdale

modèle sinusoïdal

Une forme de représentation sinusoïdale qui ne recherche pas des composantes harmoniques, et dont les coefficients sont estimés de façon non-paramétrique, s'applique à un modèle sinusoïdal de la parole [63] [64] .

Cette approche se base sur la transformation de Fourier à court terme, d'une part pour l'estimation des sinusoïdes, d'autre part pour la déconvolution homomorphique des composantes du conduit vocal $h(t,\tau)$ et de l'excitation $e(t)$ [84] [66] .

Le modèle de production se présente de façon classique 3.28 comme la réponse d'un système linéaire évoluant dans le temps à une excitation exprimée sous forme sinusoïdale. Contrairement à la représentation harmonique, la généralisation de l'excitation par rapport au cas périodique ne s'appuie pas sur une généralisation de l'harmonicité, ce qui affranchit le système d'une détection très exacte du fondamental:

$$e(t) = \sum_{l=1}^{L(t)} a_l(t) e^{i(\int_{t_l}^{t} \omega_l(\sigma) d\sigma + \phi_l)} \tag{3.33}$$

Le nombre de sinusoïdes $L(t)$ dépend du temps, ainsi que les amplitudes $a_l(t)$, les fréquences ω_l. Les phases initiales ϕ_l dépendent de l'instant d'apparition de la sinusoïde t_l.

Si l'action du conduit vocal est représentée par sa fonction de transfert sous forme polaire:

$$H(t,\omega) = M(t,\omega) e^{i\Phi(t,\omega)} \tag{3.34}$$

le modèle complet s'écrit:

$$s(t) = \sum_{l=1}^{L(t)} A_l(t) e^{i\Psi_l(t)} \tag{3.35}$$

avec:

$$A_l(t) = a_l(t) M(t, \omega_l(t)) \tag{3.36}$$

et:

$$\Psi_l(t) = \int_{t_l}^{t} \omega_l(\sigma) d\sigma + \phi_l) + \Phi(t, \omega_l(t)) \tag{3.37}$$

estimation des paramètres

L'estimation de ces amplitudes et de ces phases, pour chaque trajectoire est un problème complexe. On montre que la minimisation d'un critère d'erreur (au sens des moindres carrés, et si le signal est échantillonné) pour une excitation exactement périodique, introduit un estimateur simplement relié à la transformation de Fourier discrète.

L'extension au cas non périodique propose de rechercher les fréquences ω_l comme les maxima du spectre d'amplitude à court terme obtenu.

Pour la parole voisée, l'analyse utilise une estimation moyenne de la fréquence de voisement (pour adapter la durée de la fenêtre d'analyse), qui peut ne pas être très précise. Pour les segments non-voisés, une représentation sinusoïdale reste valide si le spectre à court terme reste proche de la densité spectrale de puissance du bruit. Un échantillonnage spectral suffisamment fin permet donc de représenter ce type de signal, avec la même méthode: dans ce cas, c'est la finesse de l'échantillonnage désiré qui définit la durée fixe de la fenêtre d'analyse spectrale.

L'analyse proposée se déroule trame par trame. Un double problème apparaît donc aux frontières des trames: l'interpolation des coefficients et le choix des trajectoires de chaque sinusoïde.

Le choix des trajectoires est opéré en raccordant d'une trame à l'autre les composantes sinusoïdales. Le raccordement utilise un seuil sur la fréquence au delà duquel une composante non appariée disparaît (meurt) ou apparaît (naît).

Lorsque le signal se trouve décrit par un ensemble de trajectoires, naissant et disparaissant à des trames données, apparaît le problème du raccordement des paramètres estimés d'une trame sur l'autre le long de chaque trajectoire. Deux options, de débit différent pour le codage, sont possibles: l'interpolation "aveugle", analogue à celle de la méthode de recouvrement/addition, ou bien l'interpolation directe des valeurs pour les trois paramètres de chaque composante sinusoïdale, le long d'une trajectoire donnée.

L'interpolation par recouvrement/addition entraîne un débit élevé: pour une fenêtre triangulaire, une qualité de parole quasi-transparente implique une durée des trames temporelles assez courte ($\simeq 10ms$). Une durée double semble entraîner une nette dégradation (qualité "rauque" de la parole synthétique).

L'interpolation des amplitudes ne pose pas de problèmes particuliers dès que les trajectoires sont définies: une simple interpolation linéaire suffit. L'interpolation des fréquences et des phases au contraire peut exiger un déroulement de la phase, puisque

le spectre de phase n'en fournit que la valeur principale, modulo 2π. Une procédure est mise en oeuvre pour déterminer le meilleur entier K, au sens du lissage de la phase, diviseur de la phase modulo 2π. Les quatre coefficients d'une interpolation cubique de la phase et de la fréquence s'en déduisent en utilisant les quatre valeurs aux bornes de la trame. En effet, on peut ici assimiler la fréquence de la composante sinusoïdale obtenue comme maximum local du spectre de Fourier à court terme et la fréquence instantanée, dérivée de la phase.

La procédure d'analyse/synthèse est donc relativement simple. Les inconvénients liés à l'utilisation de trajectoires restent limités, puisque ces trajectoires ne concernent que des objets fréquentiels localisés et à bande étroite. L'utilisation de trajectoires qui soulève de nombreux problèmes et rencontre de nombreuses objections pour une analyse en formants, présente ici au contraire de remarquables qualités de robustesse. D'une part les problèmes de numérotation et de croisement des trajectoires ne se posent pas, et d'autre part la localisation spectro-temporelle des sinusoïdes augmente la robustesse, par rapport aux méthodes utilisant des objets de largeur de bande plus élevée.

modification du signal de parole

Le modèle sinusoïdal autorise également la transformation du signal de parole [85] . Il s'agit de retrouver dans les paramètres sinusoïdaux la contribution de la source, et celle du système, comme le prédisent les équations 3.35 3.36 3.37. Système signifie ici à la fois le filtre que constitue le conduit vocal, le terme de rayonnement aux lèvres et aux narines, et le terme d'onde de débit glottique. Cette décomposition source/filtre utilise la déconvolution homomorphique, qui permet d'estimer l'action du système sur les composantes sinusoïdales. Cependant l'évaluation de la phase du système ne se fait de façon simple qu'en prenant l'hypothèse d'un système à phase minimum (la phase s'obtient alors par la connaissance de l'amplitude): cette condition n'a aucune raison d'être vérifiée, et l'estimation de la phase de l'excitation est faussée par cette hypothèse. Le signal temporel synthétique est différent du signal naturel. De même, le signal d'excitation obtenu par un système à phase minimum, qui permet une excellente synthèse, semble peu réaliste comme source d'un modèle de production: même pour de la parole voisée il présente une allure très bruitée.

Une estimation plus réaliste de la phase de l'excitation introduit une notion de cohérence de l'excitation des différentes composantes. Le signal de source présentera une impulsion de voisement lorsque l'excitation présentera un ensemble cohérent de sinusoïdes (qui répondent en phase à l'instant t_0) [73] :

$$e(t) = \sum_{l=1}^{L(t)} a_l(t) e^{i(t-t_0)\omega_l} \tag{3.38}$$

t_0 peut être estimé au sens des moindres carrés. Ce procédé se rapproche de méthodes décrites dans le chapitre deux pour regrouper les réponses de différents canaux d'analyse en fonction d'un critère de cohérence des phases. Un sens physique apparaît pour la source d'excitation, et les signaux temporels naturels et synthétiques deviennent tout à fait semblables.

3.3 représentation par des sinusoïdes modulées

3.3.1 modèle à formants en parallèle statique

Alors que les modèles sinusoïdaux s'introduisent *naturellement* par l'analyse de Fourier des sons périodiques, puisque les exponentielles complexes sont les fonctions propres des filtres linéaires, le modèle acoustique de production entraîne l'utilisation de sinusoïdes amorties. Pour un système numérique, à bande limitée par la fréquence d'échantillonnage, les termes $K_{p\infty}$ et $K_{zer\infty}$ qui représentent la contribution en basse fréquence des pôles d'ordre élevé peuvent être considérés constants [99]. Une constante K résume les constantes de 3.5 et les modules des paires conjuguées de pôles et de zéros.

Si l'on ne considère que la fonction de transfert du conduit vocal, sous sa forme pôle-zéro la plus générale:

$$V(z) = Kz\frac{\prod_{i=1}^{M}(z-\hat{z}_i)(z-\hat{z}_i^*)}{\prod_{j=1}^{N}(z-\hat{z}_j)(z-\hat{z}_j^*)} \tag{3.39}$$

Les pôles de V sont simples, et le nombre de zéros est inférieur au nombre de pôles, ce qui permet de décomposer 3.39 en éléments simples, avec les constantes A_i:

$$V(z) = \sum_{i=1}^{L}\frac{A_i}{(z-\hat{z}_i)} + \frac{A_i^*}{(z-\hat{z}_i^*)} \tag{3.40}$$

Ce qui s'écrit dans le domaine temporel:

$$v(t) = \sum_{i=1}^{L} 2|A_i|e^{-\alpha_i t}cos(\omega_i t + \phi_i) \tag{3.41}$$

avec $\hat{z}_i = -\alpha_i + i\omega_i$ et $A_i = |A_i|e^{i\phi_i}$.

L'équation 3.41 pose le principe de la synthèse à formants en parallèle, qui utilise de simples résonateurs du second ordre comme sections parallèles. Pour une section, la réponse impulsionnelle s'écrit:

$$s(t) = Ae^{-\alpha t}cos(\omega t + \phi) \tag{3.42}$$

Le coefficient A_i est une fonction de tous les pôles et de tous les zéros, obtenu comme résidus de la fonction pour le $i^{ème}$ pôle:

$$A_i = lim_{z\to z_i}(z-z_i)V(z) \tag{3.43}$$

Les paramètres de ce modèle n'évoluent pas dans le temps, au moins d'une période fondamentale à l'autre. Cette supposition est irréaliste, pour une voix de fondamental assez bas ou pour une transition rapide (plosive/voyelle par exemple).

L'estimation des paramètres pour ce modèle n'est pas simple: cette difficulté reste un des désavantages de la synthèse à formants. De nombreuses méthodes de détection des formants ont été et sont toujours proposées, mais les ambiguïtés (croisement des formants, classement et numérotation des formants, formants proches) sont très délicates à résoudre.

Une variante de ce type de synthèse, la synthèse par Formes d'Ondes Formantiques (ou FOF), a été proposée pour calculer l'onde sonore directement dans la dimension temporelle. L'étude de cette représentation, qui constitue une grosse partie de ce travail sera détaillée plus loin.

3.3.2 modèle à formant en parallèle dynamique

Un modèle de représentation du signal de parole par des formants évoluant dans le temps a été proposé, extension *dynamique* de la représentation formantique en parallèle *statique* [15] .

Il est possible de définir des amplitudes, fréquences centrales et largeurs de bande qui évoluent dans le temps $A_n(t)$, $F_n(t)$, et $B_n(t)$:

$$F_n(t) = \frac{\omega_n(t)}{2\pi} \tag{3.44}$$

$$B_n(t) = \frac{-\alpha_n(t)}{\pi} \tag{3.45}$$

donc l'équation 3.42 devient, en choisissant les phases initiales nulles, sous forme complexe et pour un formant:

$$s(t) = A(t)e^{((\sigma(t)+i\omega(t))t)} \tag{3.46}$$

Une relation élégante permet de calculer les coefficients de 3.46. Si l'on recherche les coefficients variables A, σ, ω en les approximant par les développements suivants:

$$A(t) = A_0 e^{\sum_{k=1}^{N_A} A_k t^k} \tag{3.47}$$

$$\sigma(t) = \sum_{k=0}^{N_\sigma} \sigma_k t^k \tag{3.48}$$

$$\omega(t) = \sum_{k=0}^{N_\omega} \omega_k t^k \tag{3.49}$$

En utilisant une transformation de Fourier à court terme centrée exactement sur une période de voisement:

$$\tilde{s}(t,\nu) = \int_{\frac{-T_0}{2}}^{\frac{T_0}{2}} w(\tau)s(t+\tau)e^{-2i\pi\nu(t+\tau)}d\tau \tag{3.50}$$

ce qui suppose un amortissement tel que la contribution des périodes précédentes soit négligeable, les coefficients des développements 3.47, 3.48, 3.49 satisfont l'équation 3.51, si l'on choisit $N = N_A - 1 = N_\omega = N_\sigma$:

$$\frac{\partial\tilde{s}(t,\nu)}{\partial t} = \sum_{k=0}^{N}(-i)^k(k-1)(A_{k+1}+\sigma_k+i\omega_k)\frac{\partial^k\tilde{s}(t,\nu)}{\partial\nu^k} - i\nu\tilde{s}(t,\nu) \tag{3.51}$$

Cette propriété de dérivation spectro-temporelle permet d'obtenir les coefficients [32], en calculant les dérivées temporelles et fréquentielles de $\tilde{s}(t,\nu)$.

L'ordre N des développements doit pratiquement être assez petit, et l'application de cette méthode séduisante se heurte à plusieurs problèmes pratiques: analyse synchrone au fondamental, amortissement fort pour limiter l'influence d'une période sur l'autre. Les formants doivent être isolés avant d'appliquer la méthode et l'estimation des paramètres initiaux (qui doivent être connus *a priori*) reste délicate dans les transitions rapides. Pour le moment aucune application pratique en traitement de la parole n'a été réalisée avec ce modèle.

3.3.3 modèles à sinusoïdes modulées

La représentation du signal de parole par des sinusoïdes modulées a connu une période d'intérêt au début de la décennie précédente. Analyses alternatives aux analyses en fréquence par transformation de Fourier à court terme, deux méthodes basées sur des sinusoïdes modulées par des gaussiennes, ou par une fenêtre de Hanning ont été définies et mises en œuvre, puis appliquées à la reconnaissance de parole.

sinusoïdes modulées par une gaussienne

Le modèle à sinusoïdes modulées par des gaussiennes (GMC) a été obtenu comme solution d'une équation différentielle [73]. La relation de ce modèle avec la production n'est pas évidente *a priori*, car l'équation qui justifie le choix de ce modèle est déterminée par des critères plutôt abstraits.

Cette formulation s'appuie en effet sur la remarque que le signal de parole, filtré dans une bande de fréquence (ici des bandes centrées sur les maxima spectraux), pourrait être représenté correctement comme une somme de dérivées de la fonction gaussienne.

Il s'agit donc d'une représentation sur une famille de fonctions (comme celles du chapitre 1), en adoptant deux contraintes: le signal est préalablement segmenté spectralement en bandes, le signal est représenté à l'aide de fonctions qui satisfont une certaine équation différentielle.

Le choix de la seconde contrainte n'est pas basé sur un modèle de production, et une grande partie de l'effort consacré à cette représentation consiste à la justifier de ce point de vue. C'est plutôt l'aspect formellement attrayant (solution d'une équation différentielle) qui semble guider cette approche originale.

La fonction gaussienne $g = e^{-t^2/2}$ satisfait aux équations différentielles de la forme:

$$\ddot{g}(t) + \alpha t \dot{g}(t) + \beta g(t) = 0 \tag{3.52}$$

avec les conditions initiales $g(0) = C_1$ et $\dot{g}(0) = C_2$.

Pour analyser la parole, une généralisation de 3.52, autorise les décalages temporels, de phase, la compression et l'expansion de l'échelle temporelle des solutions. Markel a construit le modèle GMC, en utilisant comme fonctions élémentaires pour représenter la parole dans une bande de fréquence les solutions d'une variante de 3.52:

$$\ddot{y}(t) + (\frac{2\pi}{s})^2(t-C)\dot{y}(t) + (\frac{2\pi}{s})^2[N^2 - \frac{1}{2} + (\frac{2\pi}{s})^2\frac{(t-C)^2}{4}]y(t) = 0 \tag{3.53}$$

avec les conditions initiales

$$y(C) = Acos(\phi) \tag{3.54}$$

$$\dot{y}(C) = 2\pi\frac{A}{S}(N^2-1)^{\frac{1}{2}}sin(\phi) \tag{3.55}$$

La solution vaut:

$$y(t) = Ae^{\frac{-a(t-C)^2}{4}}cos(\omega(t-C)-\phi) \tag{3.56}$$

avec:

$$a = (\frac{2\pi}{S})^2 \tag{3.57}$$

et:

$$\omega = \sqrt{\frac{(2\pi)^2(N^2-1)}{S^2}} \tag{3.58}$$

La signification des cinq paramètres $ASC\phi N$ apparaît ainsi: A est l'amplitude du pic, S la durée autour du pic qui renferme presque toute l'énergie, C le centre de l'enveloppe temporelle, ϕ la phase rapportée à C, et N grossièrement la moitié du nombre de cycles sous l'enveloppe. Ce dernier paramètre peut être remplacé par un paramètre F, fréquence de la cosinusoïde (paramètres $ASC\phi F$).

Cette formulation donne un sens mathématique à l'idée intuitive de représenter la parole par une somme de fonctions élémentaires de ce type.

L'analyse parcourt cinq étapes pour calculer les paramètres GMC:

- filtrage du signal pour obtenir plusieurs bandes d'analyse: les filtres peuvent être adaptés au signal à analyser par détection des régions de maximum spectral (de façon manuelle). Le filtrage est effectué par transformation de Fourier avec fenêtre carrée, sans recouvrement/addition: des discontinuités sont créées aux frontières des trames.

- détection des extrema, pour rechercher l'emplacement d'une forme d'onde.

- calcul d'une enveloppe gaussienne qui s'ajuste aux extrema, pour une forme d'onde, suivant un critère d'erreur prédéfini. Un processus adaptatif récursif utilisant la ligne de plus grande pente du gradient est employé.

- génération de l'amplitude et de la phase en fonction du maximum de cette enveloppe.

- calcul du signal résiduel par soustraction de la représentation obtenue au signal naturel.

Le processus est mené indépendamment dans les différentes bandes d'analyse.

Un inconvénient de la méthode, dû à l'abstraction du modèle choisi (solution d'une équation différentielle, sans modèle physique sous-jacent), est sa tendance à générer beaucoup de formes d'ondes, 10 ou 20 pour une seule période de voisement. Ces formes

d'ondes ont alors un sens difficile à établir du point de vue de la production, même dans le cas simple des voyelles orales. De plus, le nombre de formes d'ondes générées s'accroît avec la fréquence centrale de la bande de fréquence d'analyse: pour une représentation en trois bandes on obtient respectivement 4, 8, et 15 formes d'ondes sur un segment d'un peu plus d'une période de voisement (18.25 ms).

Les paramètres formantiques peuvent être estimés grâce à la représentation GMC. Cette estimation pour un corpus de 10 voyelles synthétisées par un synthétiseur à formant série semble convenable. Cependant, les relations entre les paramètres $ASC\phi F$ et les paramètres formantiques restent plutôt complexes.

sinusoïdes modulées par une fenêtre de Hanning

Un ensemble similaire de paramètres $ASC\phi F$ est utilisé dans une autre forme de représentation du signal de parole par des sinusoïdes modulées: le modèle HMC (Hanning Modulated Cosine). L'équation 3.56 devient ici [13]:

$$y(t) = \frac{A}{2}(1 + cos(\frac{2\pi}{S}(t - C)))cos(2\pi F(t - C) - \phi) \tag{3.59}$$

Ici, le modèle n'est plus dérivé comme solution d'une équation différentielle, ni d'un modèle de production, mais se justifie expérimentalement par sa capacité à représenter de la parole filtrée.

Une autre différence d'importance avec le modèle GMC est le filtrage initial fixe (en bandes d'octave) pour le modèle HMC: les inconvénients d'une détection de formants sont alors évités.

Cette méthode a conduit à un système d'analyse/synthèse de qualité quasi-transparente, qui justifie expérimentalement le modèle, tant pour la parole voisée que pour la parole non voisée. La méthode utilise les étapes suivantes:

- filtrage en bande d'octave.
- détection des maxima sur les signaux filtrés.
- calcul des paramètres $ASC\phi F$ en fonction des extrema.

Un cas particulier peut être considéré pour les voix possédant un fondamental élevé, où le signal dans certaines bandes de filtre peut apparaître comme une sinusoïde: chaque cycle de la sinusoïde est alors représenté par une forme d'onde.

Une application des modèles GMC et HMC au décodage acoustico-phonétique a été menée, et semble avoir donné des résultats acceptables [81] [82]. Néanmoins ces deux méthodes n'ont pas connu de continuation au delà des travaux initiaux. La généralisation d'analyses bien plus efficaces (prédiction linéaire), le manque de lien avec un modèle de production, et l'absence de perspective sur l'utilisation de ce type de méthodes à l'époque, ont conduit l'abandon de cet axe de recherche. Le travail présenté ici est donc conceptuellement proche de ceux effectués, puis délaissés, il y a presque vingt ans: ce mémoire s'attache cependant à prouver que de riches perspectives et des applications existent pour ce type de méthodes en synthèse de parole et en analyse basée sur des analogies avec l'audition.

3.4 représentation en formes d'ondes et modèle de production

3.4.1 représentation formantique

Formes d'Ondes Formantiques

La décomposition parallèle de la fonction de transfert du conduit vocal 3.42 conduit dans le domaine temporel à des sinusoïdes (en changeant le cos de 3.42 en sin) amorties par des exponentielles (figure 3.2).

$$s(t) = Ae^{-\alpha t} sin(2\pi f_c t + \Phi) \tag{3.60}$$

Chaque forme d'onde peut s'interpréter comme la réponse impulsionnelle d'une section de la décomposition en parallèle, et ainsi être associée à un formant. L'expression temporelle se calcule simplement à l'aide des trois paramètres d'un formant. La méthode de synthèse par Formes d'Ondes Formantiques, ou FOF s'appuie sur cette remarque pour reconstituer un signal vocalique directement dans la dimension amplitude/temps [95].

Une FOF est une fonction temporelle, définie ici comme la réponse impulsionnelle d'une section de la décomposition en parallèle de la fonction de transfert du conduit vocal. Les fonctions du type de 3.60 constituent donc les premiers exemples de FOF.

La première utilisation des FOF proposait de les extraire directement de segments de parole naturelle [9], mais les avantages d'une expression analytique (plus coûteuse en calcul que la première alternative) ont conduit à choisir des FOF dont la forme se rapproche de 3.60 [92] [93].

La FOF le plus généralement utilisée s'écrit:

$$fof(t) = \Lambda(t) sin(\omega_c t + \phi) \tag{3.61}$$

avec l'enveloppe temporelle $\Lambda(t)$:

$$\Lambda(t) = 0 \tag{3.62}$$

pour $t \leq 0$

$$\Lambda(t) = \frac{A}{2}(1 - cos(\beta t))e^{-\alpha t} \tag{3.63}$$

pour $0 < t \leq \frac{\pi}{\beta}$

$$\Lambda(t) = Ae^{-\alpha t} \tag{3.64}$$

pour $t > \frac{\pi}{\beta}$.

Les paramètres formantiques sont la fréquence centrale $\frac{\omega_c}{2\pi}$, la largeur de bande $\frac{\alpha}{\pi}$, et l'amplitude temporelle.

Cette FOF diffère donc de 3.60 par un paramètre $\frac{\pi}{\beta}$ de temps de montée, ou temps d'excitation, de l'enveloppe temporelle. Dans l'interprétation source/filtre, l'introduction

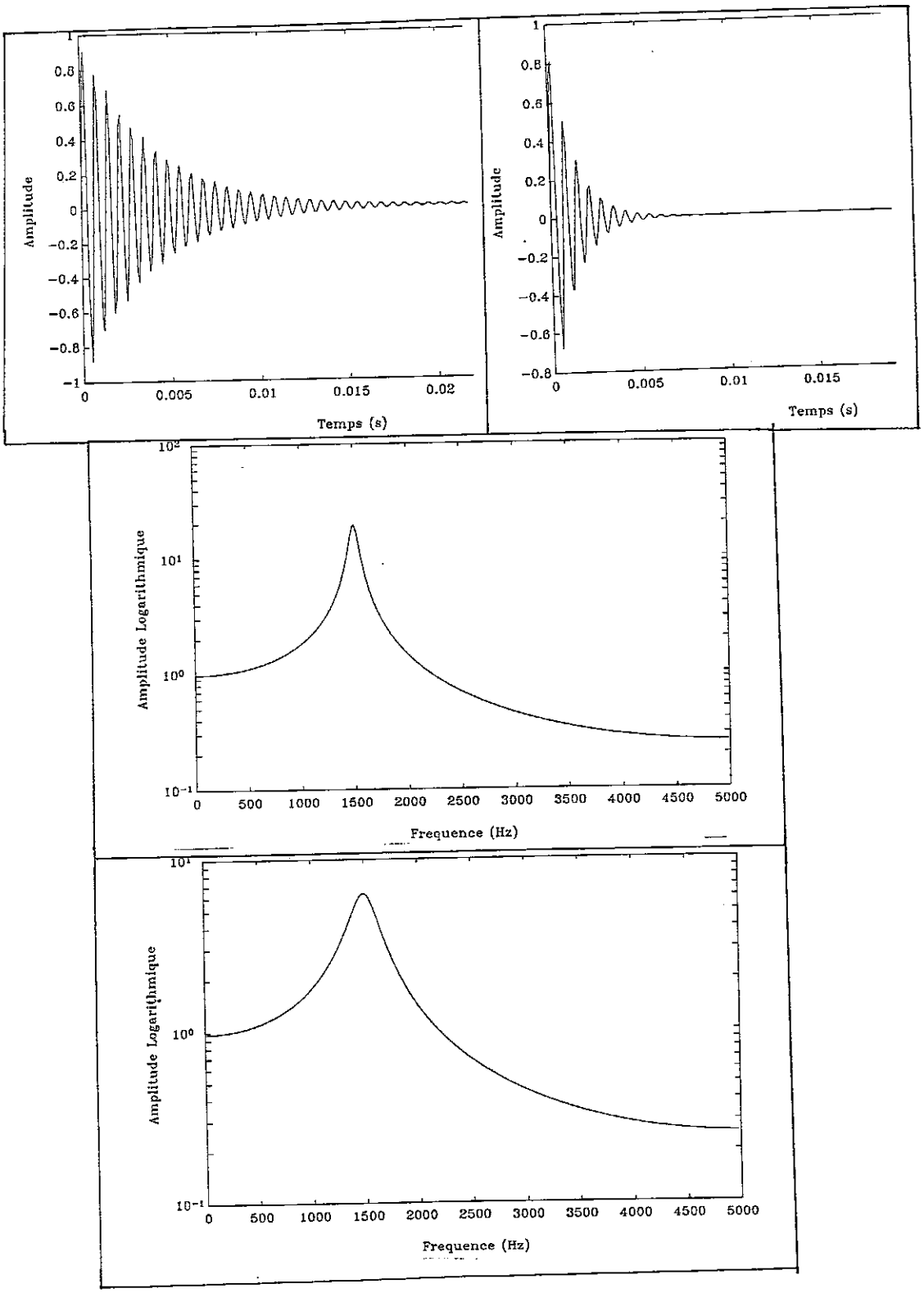

Figure 3.2: réponse impulsionnelle et spectre d'amplitude logarithmique de deux résonateurs: fréquence centrale 1500 Hz, largeur de bande 80 Hz et 250 Hz

de ce paramètre correspond à l'excitation d'un résonateur du second ordre par une impulsion qui n'est pas infiniment brève, ou bien à la réponse impulsionnelle d'un filtre qui n'est pas exactement un résonateur du second ordre, comme il sera discuté plus loin. Un contrôle plus fin de l'enveloppe spectrale, et de l'enveloppe temporelle est offert par le paramètre β, ce qui a justifié son introduction (figure 3.3).

La discussion qui suit sur l'allure spectrale des FOF basée sur [92]. L'enveloppe spectrale s'obtient par transformation de Fourier de l'enveloppe temporelle, puisque la modulation temporelle par une sinusoïde devient spectralement la convolution par deux impulsions de Dirac.

Le calcul du gain pour l'enveloppe de cette FOF donne:

$$\tilde{\Lambda}(\nu) = \frac{A\beta^2(e^{(-(\alpha+2i\pi\nu)\frac{\pi}{\beta})}+1)}{2(\alpha+2i\pi\nu)((\alpha+2i\pi\nu)^2+\beta^2)} \tag{3.65}$$

et son module:

$$|\tilde{\Lambda}(\nu)| = \frac{A\beta^2\sqrt{e^{-(2\alpha\frac{\pi}{\beta})}+2e^{-(\alpha\frac{\pi}{\beta})}cos(2\pi\nu\frac{\pi}{\beta})+1}}{2\sqrt{(\alpha^2+4\pi^2\nu^2)((\alpha^2+4\pi^2\nu^2)^2+\beta^2(\beta^2+2\alpha^2-8\pi^2\nu^2))}} \tag{3.66}$$

En prenant des valeurs plausibles en synthèse de parole:

$$\alpha \simeq 10^2\pi Hz \tag{3.67}$$

$$\beta \simeq 10^3\pi Hz \tag{3.68}$$

on peut considérer que $\beta^2 \gg \alpha^2$. De même en plaçant ν au voisinage de l'origine (c'est à dire au voisinage de la fréquence centrale pour la FOF), on peut considérer que $\beta^2 \gg \nu^2$ et que $cos(2\pi\nu\frac{\pi}{\beta}) \sim 1$. Au voisinage de l'origine, le module 3.66 peut donc être approché par:

$$|\tilde{\Lambda(\nu)}| \simeq \frac{A\beta^2\sqrt{e^{-2\alpha\frac{\pi}{\beta}}+2e^{-\alpha\frac{\pi}{\beta}}+1}}{2\sqrt{(\alpha^2+8\pi^2\nu^2)\beta^4}} = \frac{A(e^{-\alpha\frac{\pi}{\beta}}+1)}{2\sqrt{\alpha^2+8\pi^2\nu^2}} \tag{3.69}$$

le paramètre α règle donc, de façon indépendante de β en première approximation, la largeur de bande à $-6dB$ du sommet et le comportement spectral de la FOF au voisinage de la fréquence centrale. Ce comportement est similaire à celui du simple résonateur du second ordre.

Expérimentalement il apparaît que β règle de façon assez indépendante de α la forme de l'enveloppe spectrale en dessous de $-6db$ (ou "jupes" de la FOF).

Utile en synthèse, le paramètre β peut être introduit en analyse sous une forme simplifiée similaire lors de l'observation du signal filtré dans une région formantique.

L'amplitude du maximum spectral est donnée par 3.66, ce qui permet d'ajuster A pour obtenir la valeur voulue.

Une FOF est caractérisée par 6 paramètres:

- $f_c = \frac{\omega_c}{2\pi}$ sa fréquence centrale;

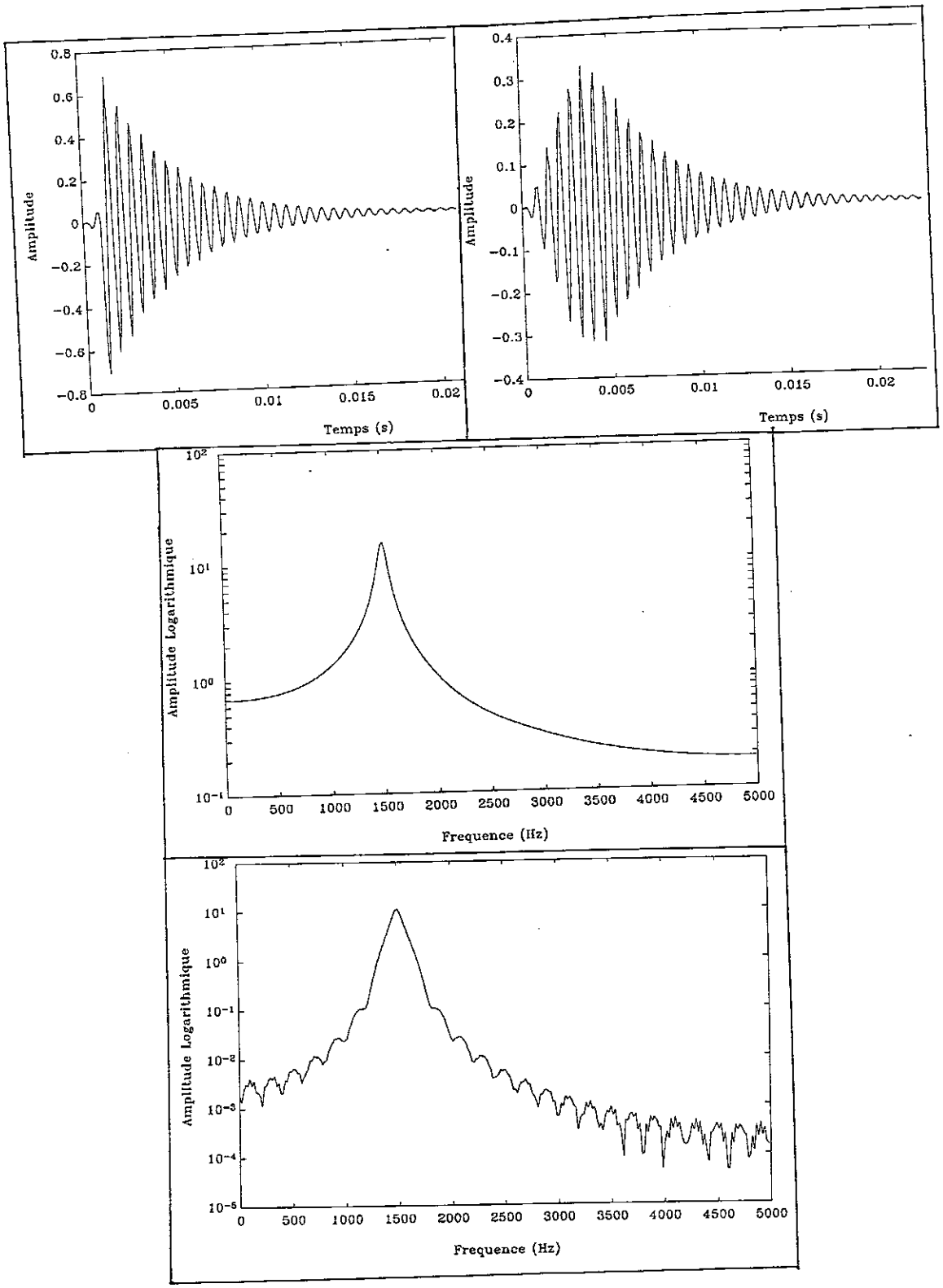

Figure 3.3: FOF et son spectre d'amplitude logarithmique: fréquence centrale 1500 Hz, largeur de bande 80 Hz, temps d'excitation 1 ms, puis 5 ms.

- $l_b = \frac{\alpha}{\pi}$ sa largeur de bande;
- A son amplitude temporelle;
- t_r son instant de référence, compté au début de la FOF ou au sommet de l'enveloppe temporelle;
- $t_a = \frac{\pi}{\beta}$ sa durée d'attaque, ou temps de montée de l'enveloppe temporelle;
- ϕ sa phase initiale;

Dans les réalisations pratiques, il peut s'avérer utile de tronquer les FOF, dont la durée n'est pas *a priori* finie. Le calcul digital des FOF entraîne nécessairement des FOF de durée finie, dès que leur amplitude passe sous le seuil de quantification. De plus, à cause de l'amortissement exponentiel, l'influence d'une FOF au bout d'une durée de quelques dizaines de millisecondes semble perceptuellement négligeable. Deux paramètres supplémentaires s'introduisent:

- d_{batt} instant de début d'amortissement forcé;
- a_{tten} durée de l'amortissement forcé;

et l'enveloppe temporelle avec amortissement forcé $\Lambda_{af}(t)$ devient:

$$\Lambda_{af}(t) = \Lambda(t) \tag{3.70}$$

pour $t \leq d_{batt}$

$$\Lambda_{af}(t) = \Lambda(t)\frac{1}{2}(1 + cos(\frac{\pi}{a_{tten}}(t - d_{batt})) \tag{3.71}$$

pour $d_{batt} < t \leq d_{batt} + a_{tten}$.

synthèse de segments vocaliques

En se référant au modèle source/filtre simplifié, les FOF permettent de synthétiser des segments vocaliques par une méthode apparentée à la synthèse à formants en parallèle, dans le domaine temporel.

En utilisant une excitation (virtuelle) quasi-périodique, tous les segments purement voisés se représentent simplement par un ensemble de FOF. Le réglage de l'enveloppe spectrale permet de générer aussi bien des voyelles orales, que des voyelles nasales (en simulant des zéros dans la fonction de transfert grâce aux largeurs de bande et aux facteurs β), des liquides, des occlusives nasales, et même des plosives. Des voix chantées d'excellente qualité ont également été synthétisées par cette méthode: la simplicité du traitement de la fréquence fondamentale permet facilement de simuler le vibrato du fondamental, ses variations aléatoires de hauteur par exemple. La formule de synthèse FOF en parallèle s'écrit, pour N formants, et M impulsions d'excitation:

$$s(t) = \sum_{j=1}^{M}\sum_{i=1}^{N} \delta_0(t - t_j) * fof_i(t) \tag{3.72}$$

Contrairement aux modèles parallèles, la synthèse par FOF utilise la version la plus simple du modèle de production source/filtre. Toutes les caractéristiques spectrales du segment sont en effet concentrées sur les paramètres des FOF: l'excitation virtuelle est réduite à un simple train d'impulsions. Les termes d'onde de débit glottique, de rayonnement, de conduit vocal (avec ses pôles et ses zéros) sont décrits directement par le comportement spectral de la somme de FOF.

La synthèse par FOF possède de plus l'avantage de l'économique. Les algorithmes sont particulièrement bien adaptés aux calculs en virgule fixe, ce qui a permis assez tôt des implémentations de ce type de synthétiseurs en temps réel sur des micro-processeurs de traitement du signal [91] [18].

Une estimation automatique des paramètres de ce modèle pour la parole voisée peut se réaliser avec une analyse LPC et une détection du fondamental. La figure 3.4 montre les résultats d'une telle analyse et la figure 3.5 les sonagrammes des signaux naturel et synthétique obtenus. Pour chaque période de voisement un ensemble de FOF correspondant aux maxima de l'enveloppe spectrale est généré. La qualité est similaire à celle obtenue avec une bonne analyse LPC classique.

discussion de la synthèse par FOF en parole

Les limites de la synthèse par FOF peuvent se rapporter à trois principales catégories: limitations liées à l'enveloppe spectrale, limitations liées à la source d'excitation, défauts de modélisation pour certains segments de parole.

La première sorte de défauts est partagée par tous les modèles de synthèse en parallèle. Bien que théoriquement équivalents, dans les applications pratiques le gros avantage de ce type de synthèse par rapport à la synthèse en série est la possibilité de régler indépendamment l'amplitude de chaque formant. Par contre, la sommation de composantes indépendantes entraîne dans certains cas des interférences entre composantes voisines: si les relations de phase entre les harmoniques communs de deux formants différents sont opposées, une forte atténuation se produit. Des creux additionnels apparaissent ainsi dans l'enveloppe spectrale, comme s'ils résultaient de zéros de la fonction de transfert, qui sont difficiles à prévoir et à contrôler. La correction des phases de chaque FOF pour remédier à ce problème fait intervenir l'ensemble de ces fonctions, et ne semble pas offrir de solution simple.

La source d'excitation virtuelle, réduite simplement à un train périodique d'impulsions, ne semble pas réaliste pour synthétiser de la parole. Ce type d'excitation est responsable de la qualité métallique des sons de synthèse produits par divers dispositifs: prédiction linéaire classique, vocodeurs à formants. Même dans la parole voisée, par exemple les voyelles orales, le signal naturel ne présente pas un tel degré de synchronisation entre les diverses composantes. De plus, il est courant d'observer plusieurs maxima de l'enveloppe temporelle, que l'on peut rapporter à plusieurs impulsions d'excitations du conduit vocal dans le même cycle vocalique. Ces impulsions peuvent par exemple s'expliquer par la fermeture des cordes vocales (impulsion en général dominante) et par leur ouverture. Une petite quantité de bruit est observable même dans les segments voisés les plus réguliers, en haute fréquence. Pour rendre compte de cette inharmonicité, une excitation virtuelle plus élaborée qu'un simple train d'impulsions commun pour toutes les FOF doit être définie.

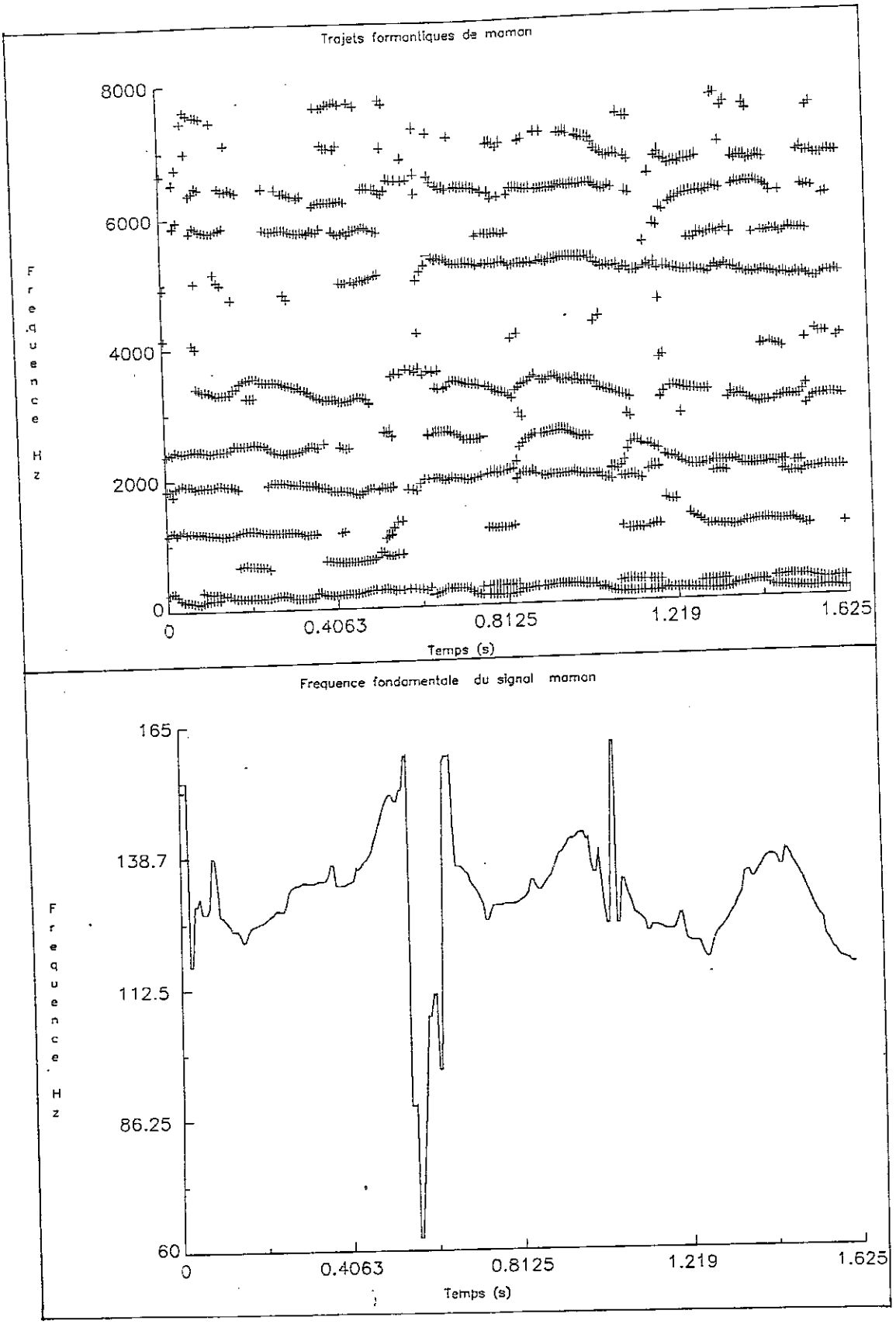

Figure 3.4: Détection des maxima spectraux par LPC et du fondamental pour la phrase (entiérement voisée) "Maman aimait une momie".

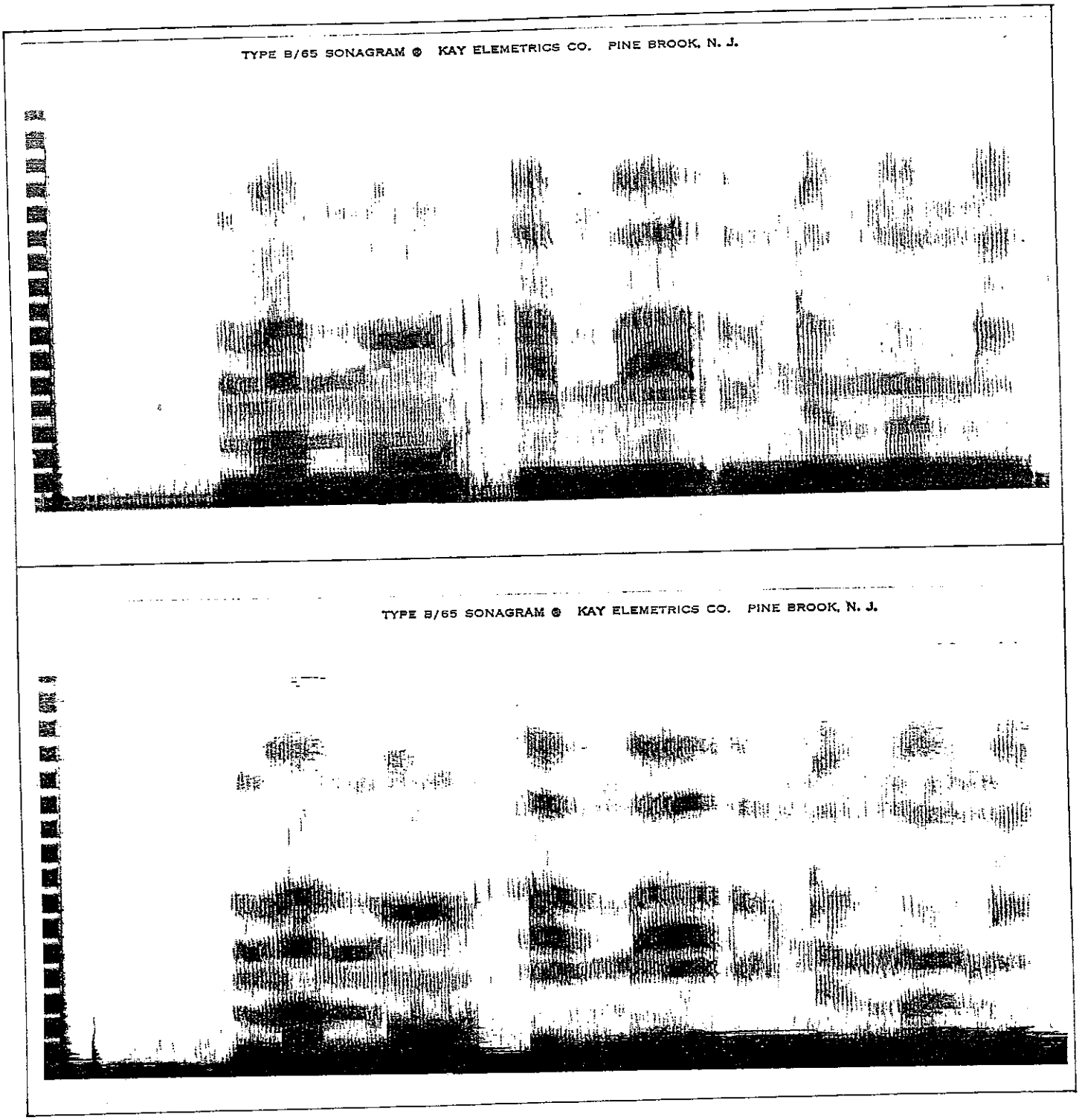

Figure 3.5: Signal naturel (haut) et signal synthétique (bas) issu de l'analyse précédente.

Certains phénomènes de parole souffrent plus particulièrement d'une source d'excitation simplifiée: les segments non périodiques (voisés et non voisés), les segments quasi-sinusoïdaux. Dans le cas des fricatives, une source impulsionnelle globale est clairement défectueuse. D'une façon générale le grave du spectre, dont l'importance perceptuelle est très grande, doit faire l'objet d'une attention particulière, assez difficile à réaliser avec un petit nombre de fonctions spectralement très simples.

Il faut de plus supposer que l'amortissement des FOF est suffisant pour que son effet soit négligeable au bout d'un faible nombre de périodes. De même, les paramètres formantiques sont supposés invariants pendant un cycle vocalique.

3.4.2 extension à la parole non-voisée

Une extension de la source d'excitation virtuelle de la synthèse par la FOF paraît indispensable pour représenter de nombreuses classes de sons. Cette extension consiste à permettre une excitation indépendante des différentes bandes formantiques. 3.72 devient alors, toujours pour N formants:

$$s(t) = \sum_{i=1}^{N} \sum_{j_i=1}^{M_i} \delta_0(t - t_{j_i}) * fof_{i,j_i}(t) \tag{3.73}$$

Cette extension généralise bien sûr le cas précédent. Elle généralise également, en le localisant, le cas d'une excitation multi-impulsionnelle. Ici, les impulsions multiples employées comme excitation concernent indépendamment chaque bande formantique: l'excitation est multi-impulsionnelle par bande de fréquence. Un sens physique reste attaché à l'excitation multi-impulsionnelle grâce à cette localisation. Il est possible de choisir les impulsions afin de produire un signal périodique ou bien un signal aléatoire dans chaque bande formantique. Si une impulsion d'excitation apparaît exactement au même instant dans toutes les bandes, on retrouve le cas d'une impulsion unique excitant le filtre global constitué des N sections parallèles. L'hypothèse, démontrée en analyse prédictive multi-impulsionnelle, que tous les types d'excitation peuvent se réaliser par une séquence d'impulsions est donc adoptée. Les séquences d'impulsions sont par contre ici localisées spectralement, ce qui permet de les interpréter de façon plus simple physiquement.

Un fait expérimentalement remarquable est le relativement petit nombre d'impulsions utiles pour générer un signal fricatif: comme en prédiction linéaire multi-impulsionnelle, en moyenne une dizaine d'impulsions toutes les 10 ms. La densité spectrale de puissance du bruit se règle, comme dans le cas d'un signal périodique par les caractéristiques spectrales des FOF utilisées. Il est ainsi remarquable, d'un point de vue pratique, que les mêmes fonctions puissent servir pour tous les types de signaux (au moins dans les régions formantiques), et que les instants d'apparition de ces fonctions permettent de régler le degré d'harmonicité de la région formantique considérée, quasi-indépendamment des régions voisines.

La source classique d'excitation pour la parole non-voisée est un bruit blanc Gaussien de moyenne nulle, et de variance σ^2. La fonction d'autocorrélation d'un tel bruit est une impulsion de Dirac puisque les échantillons sont décorrélés, et sa densité spectrale de puissance est la constante σ^2.

Dans le domaine temporel, ce type d'excitation revient à générer une nouvelle FOF à chaque échantillon, ce qui n'est pas très économique. Si l'on remplace cette excitation par un ensemble plus restreint d'impulsions, on peut néanmoins obtenir un signal aléatoire qui possède des propriétés fréquentielles assez équivalentes. Les figures 3.6 et 3.7 montrent un bruit blanc filtré par un filtre du second ordre en regard d'un bruit généré par FOF, avec les mêmes fréquences centrales amplitudes et largeurs de bande: le résultat est donc expérimentalement trés voisin.

Les formes d'ondes dominantes (par exemple sur l'enveloppe temporelle du signal) portent expérimentalement suffisamment d'information pour représenter le signal aléatoire correctement. Une justification théorique exacte utilise un pavage du plan temps-fréquence par un ensemble complet de fonctions élémentaires orthogonales, comme ceux présentés au premier chapitre: un signal aléatoire est représenté exactement par un ensemble de formes d'ondes dans ce cas. Pour une représentation basée sur un modèle, l'orthogonalité et la complétude ne sont plus des conditions *a priori*, et l'on s'éloigne d'un pavage théoriquement satisfaisant.

Des expériences ont été menées pour synthétiser des fricatives à l'aide de FOF. Les paramètres des FOF sont fixés, et seule l'excitation a un caractère aléatoire: des impulsions sont générées, uniformément réparties autour d'une fréquence moyenne donnée (par exemple 1000 Hz), dans un intervalle donné (par exemple plus ou moins 300 Hz). La figure 3.8 montre la syllabe /fy/ naturelle puis synthétisée par FOF: le bruit de frication avec une génération aléatoire des FOF et la partie voisée avec une génération synchrone au fondamental et une détection manuelle (visuelle) des paramètres.

Ce traitement impulsionnel de la source d'excitation autorise toutes les graduations entre une excitation strictement périodique et une excitation aléatoire (voir figure 3.9). La représentation de certains segments, comme les fricatives voisées, les plosives voisées, certaines voyelles même, exige une excitation très différente dans des bandes de fréquence différentes: excitation quasi-périodique en basse fréquence et aléatoire en haute fréquence par exemple.

3.4.3 représentation de la bande de base

L'utilisation de FOF présente des inconvénients pour la représentation de la partie la plus grave du spectre de parole. Alors que l'oreille présente une sensibilité importante dans cette région (l'échelle des Bark montre un comportement linéaire en dessous d'environ 800 Hz), la forme spectrale des FOF paraît trop simpliste pour une représentation fidèle. Plusieurs arguments penchent en faveur d'une représentation plus fine de la partie grave du spectre:

- l'examen de la réponse d'un modèle fonctionnel du système auditif périphérique montre que les premiers harmoniques sont résolus indépendamment.
- la fréquence centrale du premier formant est proche dans certains cas de celle du fondamental, et ces ordres de grandeur comparables empêchent l'utilisation de fonctions temporelles comme 3.61.
- dans les modèles à formants parallèles, la forme d'onde de débit glottique est en général plus complexe que de simples impulsions: la configuration du spec-

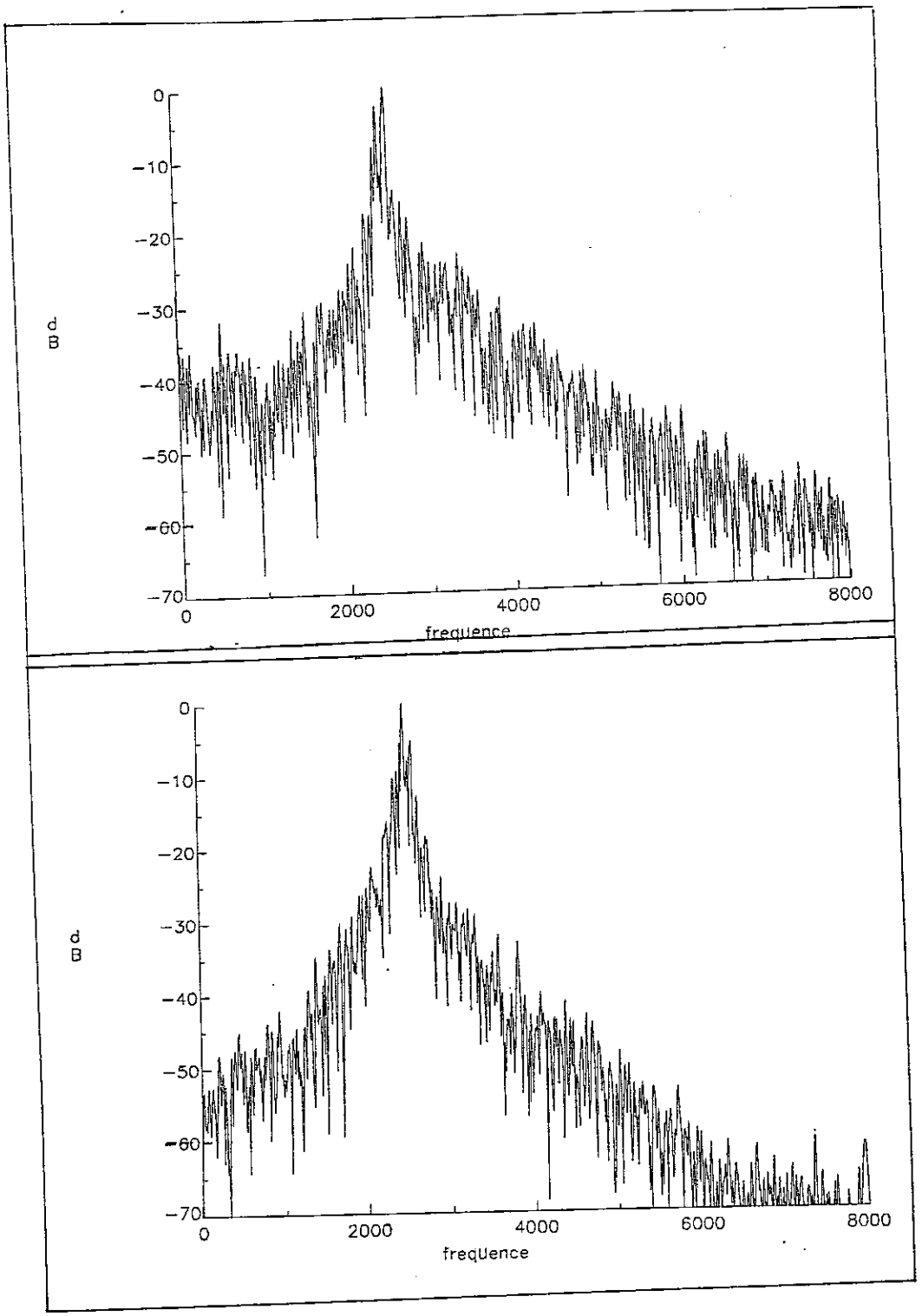

Figure 3.6: Spectres d'amplitude d'un bruit blanc filtré par un filtre du second ordre (haut), et d'un bruit généré par FOF (bas) (fréquence centrale 2500 Hz, largeur de bande 50 Hz).

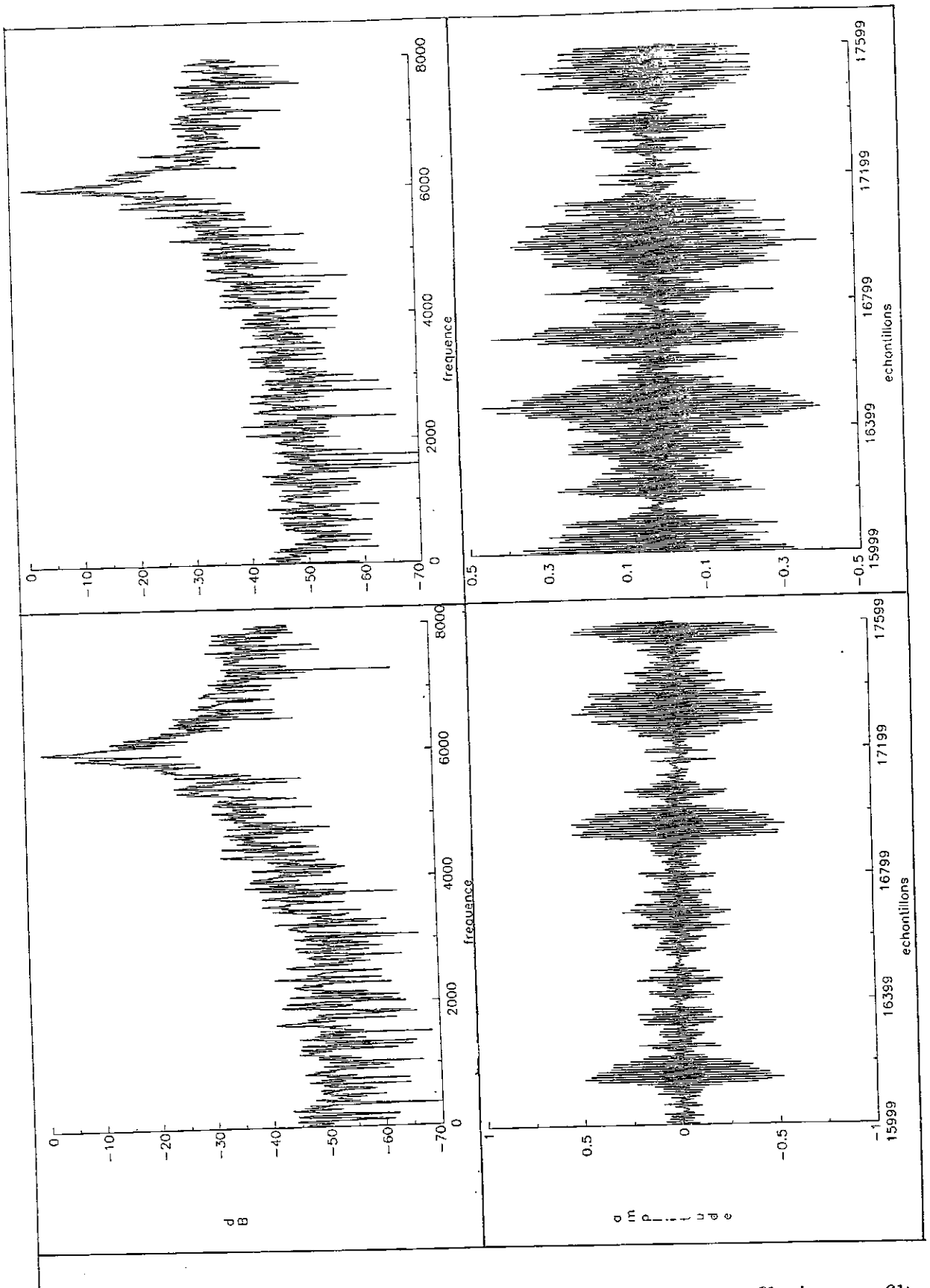

Figure 3.7: Spectres d'amplitude et signaux temporels d'un bruit blanc filtré par un filtre du second ordre (gauche), et d'un bruit généré par FOF (droite) (fréquence centrale 6000 Hz, largeur de bande 80 Hz).

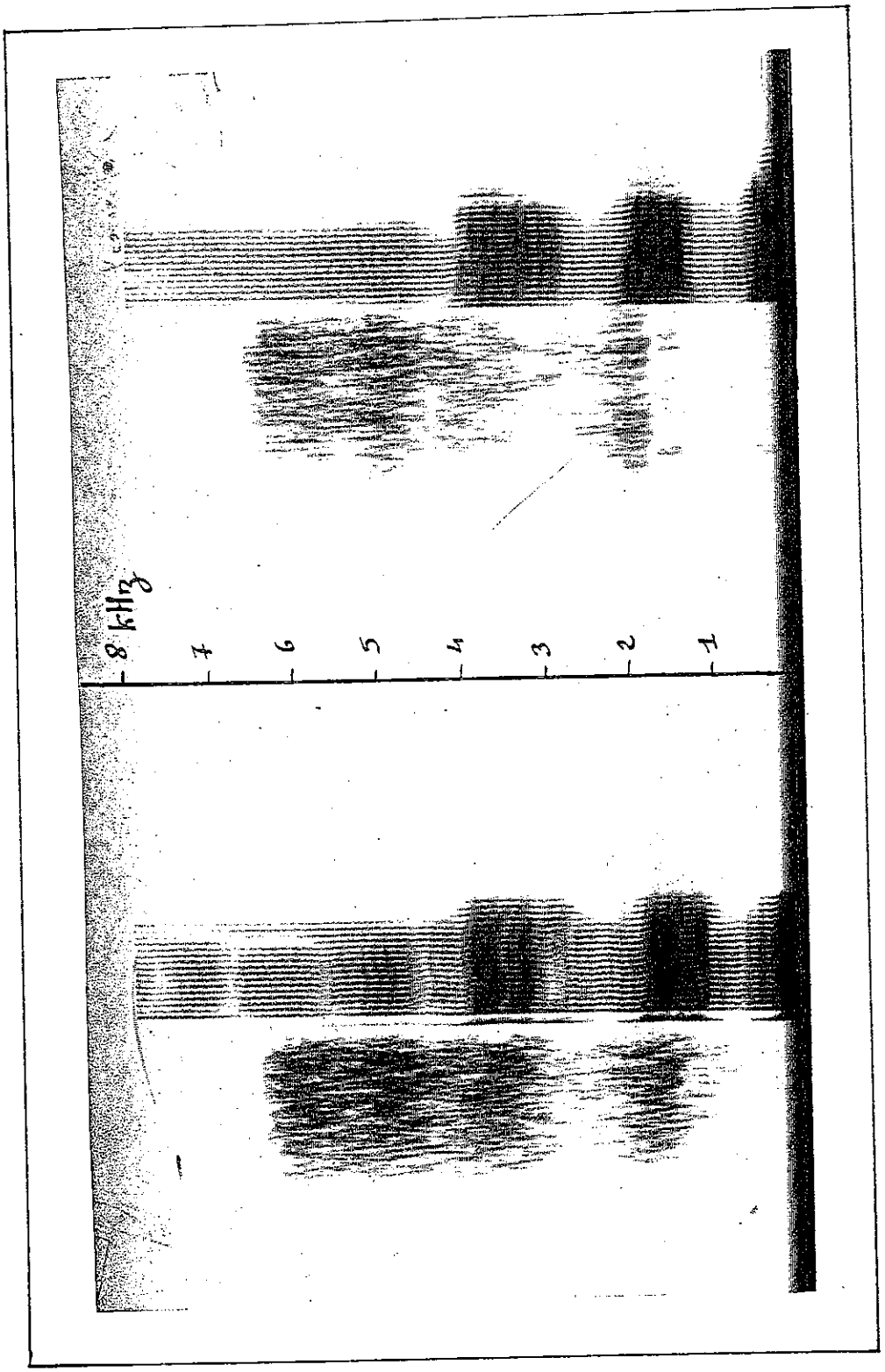

Figure 3.8: /fy/ naturel (droite) et synthétisé par FOF (gauche), avec estimation manuelle des paramètres.

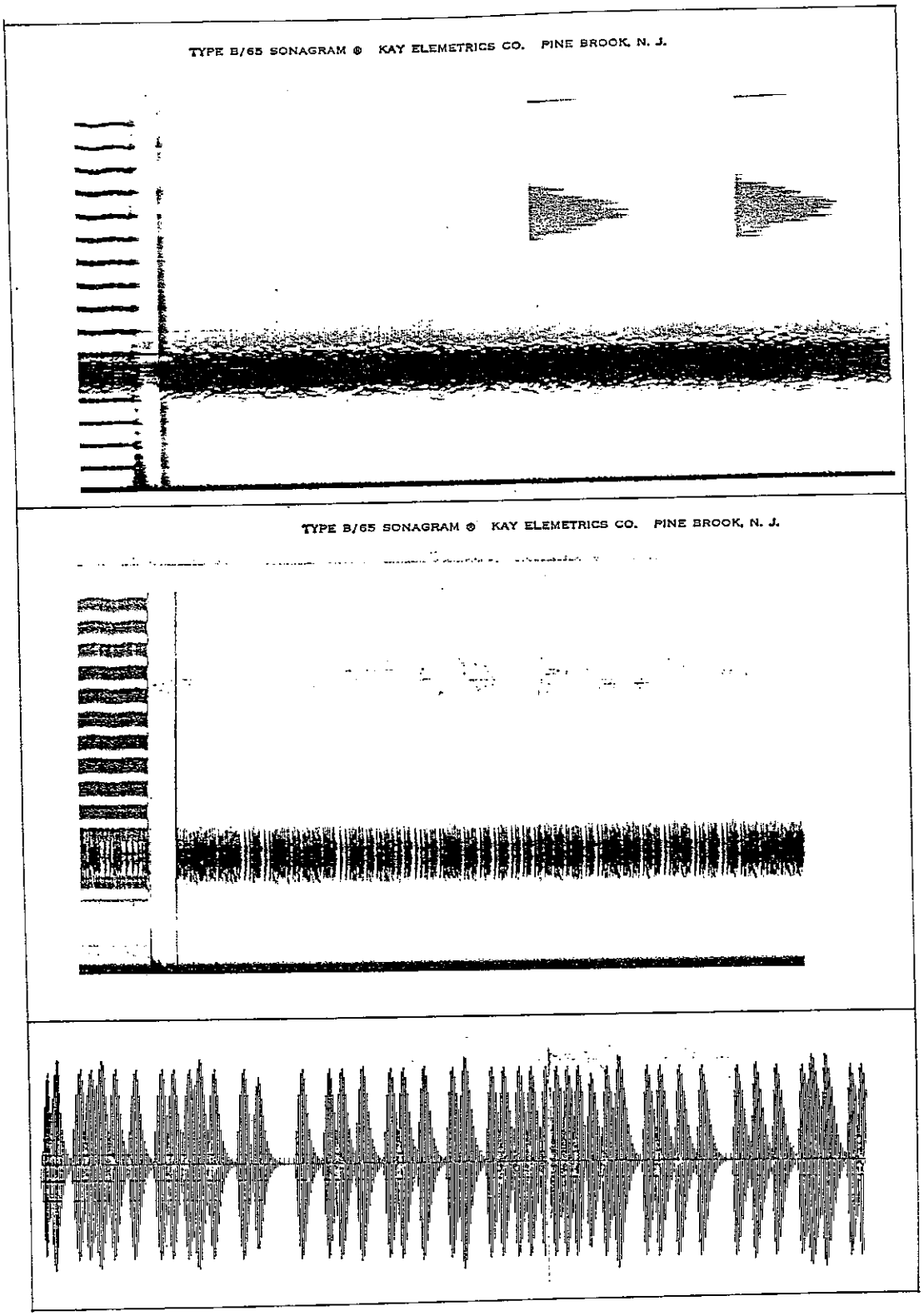

Figure 3.9: sonagramme à étroite puis à bande large, et signal d'un bruit à bande étroite généré par FOF. Ce bruit montre le principe de génération, mais présente une sonorité "granuleuse".

tre en basse fréquence est largement due à cette excitation. Dans le cas d'une représentation par formes d'ondes, où l'excitation virtuelle est particulièrement simple, les formes d'ondes elles-mêmes doivent assumer la complexité du spectre dans cette région.

Plusieurs méthodes sont envisageables pour traiter le raffinement impliqué par la *bande de base*. Le terme bande de base signifie ici la région spectrale autour du premier maximum du spectre d'amplitude.

Le choix adopté consiste à représenter par des fonctions temporelles le comportement du signal dans cette région, ces fonctions temporelles doivent de plus avoir un sens du point de vue du modèle de production. Les fonctions temporelles les plus simples et les plus naturelles que l'on puisse employer sont les sinusoïdes. Alors que dans les régions formantiques plusieurs harmoniques sont regroupés en une seule FOF, sans que la perte relative de précision sur la forme exacte de l'enveloppe spectrale ne représente un inconvénient trop grave, dans la bande de base chaque composante doit être traitée avec une bonne résolution. Le signal temporel de cette bande de base possède un comportement très varié, qui entraîne un nombre assez important de paramètres pour le représenter. On représentera donc ce signal par les sinusoïdes qui le constituent.

Pour un modèle sinusoïdal, le problème de l'interpolation des trois paramètres peut se résoudre simplement en considérant des segments de sinusoïdes à court terme: le recouvrement/addition de ces formes d'ondes permet une interpolation automatique sans se soucier de l'appariement des segments d'une trame à l'autre par exemple. La formule de synthèse devient:

$$s(t) = \sum_{i=1}^{K} \sum_{j_i=1}^{L_i} \delta_0(t - t_{j_i}) * foh_{i,j_i}(t) \tag{3.74}$$

avec les formes d'ondes harmoniques foh de la forme:

$$foh(t) = \Lambda(t) sin(\omega_c t + \phi) \tag{3.75}$$

et l'enveloppe temporelle $\Lambda(t)$:

$$\Lambda(t) = 0 \tag{3.76}$$

pour $t \leq 0$

$$\Lambda(t) = \frac{A}{2}(1 - cos(\beta_1 t)) \tag{3.77}$$

pour $0 < t \leq \frac{\pi}{\beta_1}$

$$\Lambda(t) = \frac{A}{2}(1 + cos(\beta_2(t - \frac{\pi}{\beta_1})) \tag{3.78}$$

pour $\frac{\pi}{\beta_1} < t \leq \frac{\pi}{\beta_1} + \frac{\pi}{\beta_2}$

Dans cette région spectrale, et de par leur choix, le critère de segmentation de ces formes d'ondes n'a plus la simplicité de celui utilisé pour les FOF. A priori un harmonique est une fonction dont l'enveloppe possède une évolution pour une durée de l'ordre de grandeur d'un segment plutôt que pour une durée de l'ordre de grandeur

d'une période de voisement par exemple. Ce n'est donc plus l'enveloppe temporelle qui permet de choisir la durée d'une forme d'onde. Deux choix sont alors possibles: soit adopter un critère arbitraire d'intégration temporelle, soit profiter de la périodicité des formes d'ondes pour définir leur durée.

Le premier cas fixe une valeur arbitraire pour β_1 et β_2, ce qui ramène à un traitement par trame. Cependant, la durée des formes d'ondes peut être adaptée au signal, par exemple comme durée d'un nombre fixe de période de la composante la plus grave. Les formes d'ondes ne sont alors plus indépendantes entre elles, mais un sens physique reste attaché à cette durée: pour les segments voisés elle représente une période de voisement.

Le second cas, s'appuie sur la périodicité locale des formes d'ondes pour choisir un β_1 et un β_2 adapté. Par exemple un nombre fixe de périodes par forme d'onde représente un choix qui rend indépendante chaque bande de fréquence harmonique, et différents chaque β_1 et β_2. L'inconvénient est que le sens physique de chaque forme d'onde ne s'obtient que par rapport aux formes d'ondes voisines.

Comme il a été mentionné pour la représentation sinusoïdale, aussi bien les segments voisés que les segments non-voisés peuvent se représenter par une somme de sinusoïdes. Aucune relation *a priori* n'est imposée alors sur ces formes d'ondes (harmonicité ...). Dans le cas de la parole voisée, les composantes seront bien sûr associées aux harmoniques, et dans le cas de parole non voisée elles permettent de reconstituer le comportement du signal en basse fréquence.

3.4.4 modèle complet

Le modèle complet regroupe les deux types de formes d'ondes, et la formule de synthèse s'écrit:

$$s(t) = \sum_{k=1}^{K} \sum_{l_k=1}^{L_k} \delta_0(t - t_{l_k}) * foh_{k,l_k}(t) + \sum_{i=1}^{N} \sum_{j_i=1}^{M_i} \delta_0(t - t_{j_i}) * fof_{i,j_i}(t) \tag{3.79}$$

Ce modèle représente la décomposition en parallèle d'un filtre qui regroupe les trois termes associés à la forme d'onde de débit glottique, au conduit vocal, au rayonnement aux lèvres et aux narines. Cette décomposition met en évidence des régions formantiques, suivant les pôles de ce filtre. Chaque région formantique est représentée par son comportement temporel, modélisé par des formes d'ondes formantiques, variantes de la réponse impulsionnelle d'un résonateur du second ordre. Une région formantique particulière, depuis la fréquence fondamentale jusqu'à la région du premier formant exige un traitement particulier. Une nouvelle décomposition en parallèle est effectuée, et le comportement temporel du signal dans cette région est paramétrisé grâce aux composantes sinusoïdales présentes. Ces composantes peuvent être identifiées aux harmoniques pour la parole voisée. La décomposition de la bande de base utilise des formes d'ondes différentes des FOF, qui sont des sinusoïdes enveloppées par des segments de cosinusoïdes. Ce raffinement de l'analyse en basse fréquence semble indispensable, à cause de la forme trop simple des FOF pour représenter finement une région où la fréquence fondamentale est d'un ordre de grandeur comparable avec celui des fréquences centrales formantiques.

segments vocaliques

Les paramètres des formes d'ondes, dans le cas de segments vocaliques (voyelles, liquides, nasales) reçoivent une interprétation simple. Dans le cas idéal, une seule excitation virtuelle apparaît par période de voisement. Les caractéristiques du filtre sont fixées par les FOF dans les régions formantiques, et par les formes d'ondes sinusoïdales dans la bande de base. Les fréquences des formes d'ondes sinusoïdales sont dans des rapports harmoniques, et leurs amplitudes fixent l'enveloppe spectrale dans le bas du spectre.

Lorsque la voix est très faible (fin de groupe prosodique, manque d'articulation), le signal prend une allure quasi-sinusoïdale. Les FOF sont donc quasiment absentes et seules les formes d'ondes sinusoïdales sont utiles dans ce cas.

fricatives

L'excitation virtuelle devient ici un train aléatoire d'impulsions, différent dans chaque bande de fréquence. La notion de formants doit être étendue pour une excitation non périodique: les maxima de l'enveloppe de la densité spectrale de puissance forment les candidats naturels pour cette extension. Ces maxima existent dans les fricatives, même si ils sont moins nombreux en général que les formants de la parole voisée [116]. Il est fréquent qu'un ou plusieurs formants des segments adjacents à une fricative soient déjà ou encore présents dans le bruit fricatif. Les paramètres formantiques règlent donc l'enveloppe de la densité spectrale de puissance, puisque le filtrage linéaire garde ce sens dans le cas de signaux aléatoires. L'excitation virtuelle possède un caractère aléatoire, et expérimentalement une impulsion toutes les millisecondes en moyenne suffit pour obtenir des périodogrammes plausibles pour des fricatives. Les fricatives voisées s'obtiennent en mélangeant une excitation quasi-périodique en basse fréquence avec une excitation aléatoire.

La voix chuchotée correspond à l'excitation du conduit vocal par un jet d'air turbulent créé par l'ouverture de la glotte. La structure formantique est bien apparente, et une décomposition en parallèle reste tout à fait valide. Dans chaque bande de fréquence l'excitation est un train aléatoire d'impulsions, comme pour les fricatives sourdes.

plosives

L'excitation virtuelle d'une plosive est une impulsion isolée. Il s'agit donc d'une troisième sorte de signaux, des transitoires très brefs. La production prédit que les gestes articulatoires les plus rapides sont ceux associés aux plosives, occlusion dans le conduit vocal suivie d'un brusque relâchement. L'aspect fréquentiel de cette explosion est décomposé en FOF et en formes d'ondes sinusoïdales. En général pour les plosives un ou plusieurs maxima sont également apparents dans l'enveloppe spectrale, ce qui jusitifie la décomposition en parallèle. L'explosion est accompagnée de bruit ou de quelques périodes de voisement pour une plosive sonore.

hypothèses et discussion

Ce modèle suppose que les paramètres de chaque forme d'onde soient invariants pendant la durée d'une forme d'onde. Cette hypothèse peut se rattacher à la stationnarité du modèle de production. Cependant l'indépendance de l'excitation entraînée par la décomposition en parallèle rend difficile la définition de l'intervalle de temps pendant lequel le signal est supposé stationnaire, puisque cet intervalle est lié à la durée des formes d'ondes, et à la fréquence de leur excitation virtuelle: en effet le recouvrement de deux formes d'ondes entraîne une interpolation des paramètres formantiques ou harmoniques.

Un modèle variant dans le temps peut se définir en autorisant les paramètres des formes d'ondes à devenir des fonctions du temps. Ceci prend un sens surtout pour les voix graves, pendant les évolutions rapides du conduit vocal (coarticulation plosive/voyelle par exemple). De même le comportement temporel des premiers formants peut subir une forte influence de la part de l'onde de débit glottique. La définition de ce modèle variant dans le temps semble une extension naturelle, même si les problèmes d'estimation des paramètres, évoqués dans la section suivante, deviennent alors plus complexes.

La représentation de la courbe d'enveloppe du spectre d'amplitude par des FOF profite de la relative indifférence de l'oreille aux creux de cette courbe. En particulier la création de vallées additionnelles à cause des interférences destructives inhérentes à la structure parallèle semble inévitable.

La forme du gain des FOF employées (qui est symétrique) n'autorise également qu'une approximation de l'enveloppe spectrale réelle. Mais la forme exacte de l'enveloppe spectrale tend à perdre de l'importance dans les régions formantiques: les propriétés de l'oreille montrent une résolution relative constante en haute fréquence. L'approximation par des FOF reste donc valide, même si la forme exacte de l'enveloppe spectrale n'est pas exactement respectée.

La forme de l'enveloppe temporelle des FOF montre une décroissance exponentielle. Cette décroissance est effectivement observable tant que la période fondamentale est suffisamment grande. Dans le cas de voix aiguës, la décroissance des composantes d'une période sur l'autre peut être assez faible et contredire cette hypothèse.

3.5 estimation des paramètres

3.5.1 méthode

choix d'une méthode

Un système d'analyse/synthèse a été bâti pour obtenir de façon automatique une décomposition du signal de parole sous la forme de 3.79.

Comme pour les décompositions en formes d'ondes élémentaires basées sur des données auditives, un ensemble varié de méthodes a été employé. Une méthode unifiée d'estimation des paramètres à l'aide d'un modèle simple du signal ne semble pas possible dans notre cas.

Une alternative aux procédés simples présentés ici, qui permettent à toutes les

étapes une vérification directe de la décomposition, serait peut-être à rechercher parmi la riche palette des méthodes paramétriques. La méthode de Prony en particulier, variante de la prédiction linéaire, permet d'ajuster un ensemble d'exponentielles complexes à une série temporelle. Le problème de l'excitation d'un tel système, en particulier l'emplacement des impulsions indépendamment pour chaque maximum spectral (pôle ou ensemble de pôles), ne semble pas être considéré actuellement. Le mélange entre modélisation spectrale et modélisation de l'excitation entraîne de difficiles problèmes actuellement en dehors du champ d'investigation des méthodes paramétriques.

La synthèse en accord avec la formule 3.72 est largement répandue dans les centres d'acoustique musicale. Des versions de cette synthèse en temps réel existent sur de petites machines depuis quelques années, prouvant la simplicité de mise en oeuvre et la puissance de la méthode. Par contre aucun système automatique d'estimation des paramètres, de même que l'extension pour la parole non-voisée, n'existaient.

Le premier système d'estimation semi-automatique des paramètres a été mis au point [20], basé sur les trajectoires formantiques. Ce système a évolué par la suite avec des options assez différentes [97] [98]. La synthèse automatique de parole voisée, avec détection du fondamental, synchronisation des FOF sur le fondamental, et estimation des paramètres par modélisation spectrale a été une application directe de ce système. La qualité était comparable à celle obtenue par une bonne prédiction linéaire classique pour les segments vocaliques, ou plosifs. Aucune possibilité de produire des fricatives n'était offerte avec les options choisies. Par contre une analyse automatique d'un sous ensemble de la parole, avec synthèse par formes d'ondes a ainsi été démontrée. Les principales faiblesses de cette analyse restaient d'une part la modélisation trop simple de l'excitation, comme en prédiction linéaire classique, et d'autre part la différence de qualité dans le grave du spectre. L'étape suivante a donc consisté à autoriser une indépendance de l'excitation dans les différentes régions formantiques, et le traitement plus raffiné de la bande de base.

processus d'analyse/synthèse

Le processus d'analyse synthèse s'appuie sur une modélisation spectrale *a priori* du signal de parole, ce qui évite le problème de regroupement des formes d'ondes élémentaires rencontré au second chapître.

Le processus d'analyse/synthèse parcourt les étapes suivantes (voir figures 3.10, 3.11), qui vont être détaillées:

- modélisation de l'enveloppe spectrale par prédiction linéaire.
- détection des maxima de l'enveloppe spectrale
- segmentation spectrale à l'aide de ces maxima: les régions formantiques et la bande de base sont définies à cette étape.
- filtrage dans les régions formantiques: des signaux temporels sont obtenus.
- calcul du module de la transformée de Fourier de la bande de base.
- segmentation spectrale de la bande de base.

- filtrage dans les régions harmoniques: des signaux temporels sont obtenus.
- détection dans les bandes formantiques et dans les bandes harmoniques des formes d'ondes, par calcul de l'enveloppe temporelle.
- estimation des paramètres de chaque forme d'onde.
- optimisation des paramètres de chaque forme d'onde.
- synthèse par 3.79 d'un signal à partir de sa représentation en forme d'onde.

Le traitement des bandes formantiques et des bandes harmoniques est en grande partie homogène. Le processus d'estimation des paramètres peut être optimisé récursivement, afin de réduire l'erreur au sens des moindres carrés entre le signal synthétique et le signal naturel.

3.5.2 modélisation spectrale

Plusieurs types de méthodes permettent d'obtenir une estimation du filtre f du modèle linéaire simplifié. Parmi ces méthodes, la prédiction linéaire offre l'avantage de la simplicité de mise en oeuvre et de la qualité du modèle. La prédiction linéaire montre une tendance, un nombre de coefficients de prédiction étant fixé, à produire une quantité assez constante de maxima spectraux: l'ordre du modèle doit donc être soigneusement choisi.

Afin d'accentuer les maxima spectraux d'ordre élevé, une préaccentuation est appliquée au signal avant la prédiction linéaire. Ce pré-traitement est un filtrage passe-haut du premier ordre, répondant à l'équation suivante:

$$y_n = x_n - \alpha x_{n-1} \tag{3.80}$$

pour la valeur de α choisie ($\simeq 0.95$), la préaccentuation est d'environ $+6dB/octave$.

Parmi les différents algorithmes de prédiction linéaire, l'analyse adaptative par un filtre en treillis, avec fenêtre d'oubli exponentielle, proposée dans [96] offre d'excellentes performances.

Le but de la modélisation spectrale est d'obtenir une estimation des régions spectrales proéminentes. L'ambition est donc moindre que celle de la détection de formants, qui implique un suivi des maxima de l'enveloppe spectrale. L'amplitude du gain du filtre est obtenue de façon classique par l'évaluation sur le cercle unité du polynôme formé par les coefficients de prédiction.

L'algorithme du filtre en treillis adaptatif délivre pour chaque échantillon un jeu de coefficients. Ces coefficients sont pris en compte à des intervalles réguliers, ou trames d'analyse. Ces trames d'analyse ne concernent ici que les premiers stades de la méthode: la décomposition ultérieure du signal temporel dans chaque bande filtrée permet de s'affranchir du découpage temporel fixe en trame d'analyse, comme nous allons le voir.

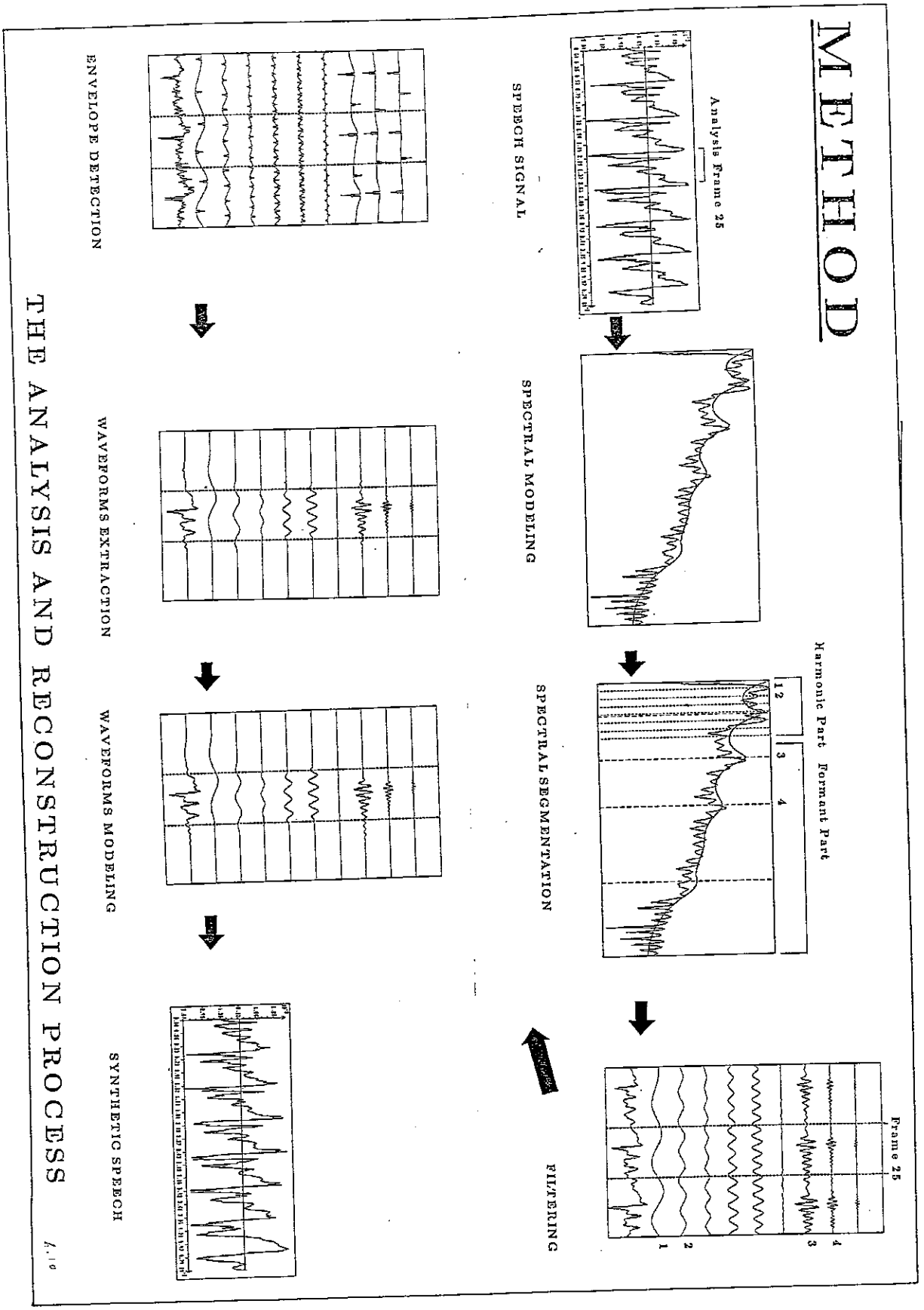

Figure 3.10: Synoptique du système d'analyse-synthèse.

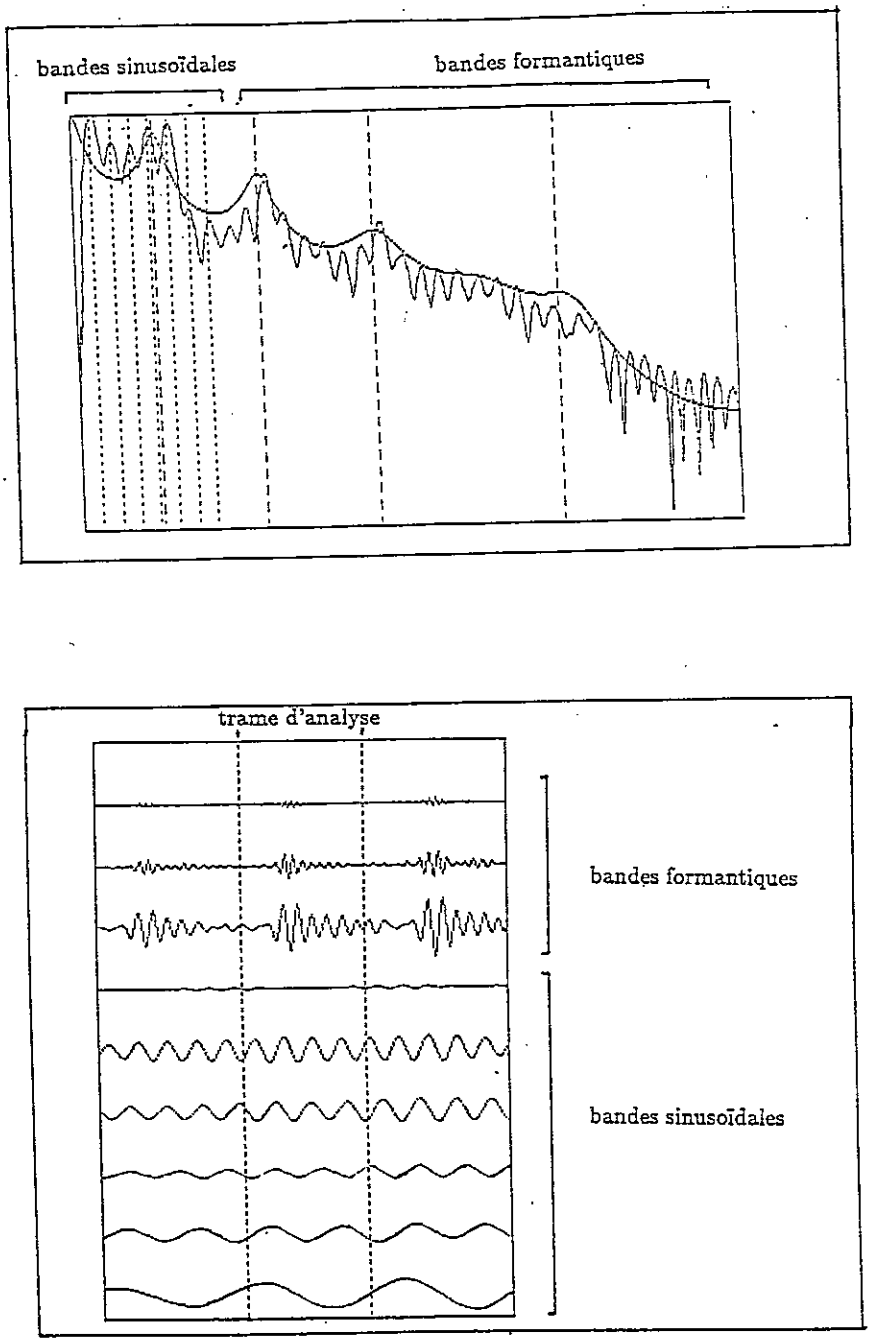

Figure 3.11: Modélisation (LPC et TFCT) et segmentation spectrale (haut). Filtrage par TFCT dans les régions ainsi définies (bas).

3.5.3 segmentation spectrale

La recherche des maxima spectraux peut emprunter à partir de l'analyse prédictive deux chemins différents: la détection sur l'amplitude du gain des maxima locaux, ou bien l'extraction explicite et le regroupement des racines du polynôme. C'est la première solution qui a été adoptée ici. Les deux méthodes donnent des résultats tout à fait comparables, mais la première est de loin la moins coûteuse en temps de calcul. La seconde méthode offre l'avantage d'une meilleure estimation des largeurs de bande, qui est le problème le plus difficile. Puisqu'il s'agit seulement de déterminer les régions spectralement dominantes, et non de déterminer avec précision les caractéristiques formantiques, le choix de la première méthode ne présente pas d'inconvénients et a été pratiquement préféré.

La précision d'estimation des maxima spectraux est liée au pas d'échantillonnage fréquentiel. Une interpolation parabolique entre les échantillons spectraux permet de raffiner cette estimation.

3.5.4 filtrage en bandes formantiques

méthode

Pour chaque trame d'analyse, un nouveau banc de filtres doit être défini à partir de la connaissance des maxima spectraux. La sommation des sorties du banc de filtres, trame par trame redonne le signal initial, lorsque la somme des gains $\tilde{h}_i$ $i = 1, \ldots, N_j$ de la $j^{ième}$ trame d'analyse est égale à l'unité.

$$\tilde{h}(\nu) = \sum_{i=1}^{N_j} \tilde{h}_i(\nu) = 1 \tag{3.81}$$

Une forme commode d'implémentation d'un banc de filtres *a priori* arbitraire utilise la transformation de Fourier à court terme.

Il s'agit à nouveau d'une analyse par trames successives. Cependant, trois types de trames doivent être distingués ici. D'abord les trames pour la modélisation spectrale, juxtaposées sans recouvrement, qui sont les *trames d'analyse* employées tout au long de l'analyse. Le filtrage implique également l'utilisation de trames durant lesquelles sera calculé le spectre de Fourier à court terme: ces *trames de transformée de Fourier à court terme* se recouvrent et leur durée définit la précision en fréquence de l'analyse (le pas entre deux échantillons spectraux). Enfin, pour éviter les effets de bord de l'analyse par recouvrement et addition, et pour la recherche des formes d'ondes, les calculs de transformation de Fourier à court terme doivent être menés sur une plage temporelle plus longue que le premier type de trame: ces *trames de filtrage* se recouvrent et sont centrées sur les trames d'analyse de façon à les rendre jointives.

Pour implémenter un banc de filtres linéaires (invariants dans le temps), la méthode la plus directe ne fait pas usage d'un filtre de synthèse. Dans l'interprétation par blocs, adoptée dans cette section, la méthode se déroule pour chaque trame de transformée de Fourier à court terme de la façon suivante:

- décalage du pointeur sur le signal d'entrée: ce décalage fixe le taux de recouvrement des trames. Rappelons que ce décalage doit être calculé en fonction des

caractéristiques spectrales et temporelles de la fenêtre employée à l'étape suivante;

- application d'une fenêtre d'analyse spectrale (par exemple fenêtre de Hamming, de Blackman-Harris) sur le signal, et constitution d'une trame de transformation de Fourier par complément avec des échantillons nuls;
- transformation de Fourier dicrète de la trame;
- multiplication du module des échantillons spectraux par le module du gain pour chaque filtre $\tilde{h}_i$ du banc;
- transformation de Fourier discrète inverse des échantillons spectraux ainsi modifiés;
- addition avec recouvrement, puis décalage du pointeur sur le signal de sortie. Ces deux dernières étapes se répètent autant de fois qu'il y a de filtres pour cette trame d'analyse. Le processus est répété depuis le premier point afin de traiter toute la trame de filtrage.

Cette méthode de filtrage, lorsque seuls les modules des échantillons spectraux sont modifiés, n'introduit pas de déphasage.

Le gain des filtres possibles est fixé d'une part par la fenêtre spectrale appliquée aux modules des échantillons spectraux, et d'autre part par la fenêtre temporelle appliquée sur les échantillons du signal. Le gain du filtre effectivement implémenté $\hat{h}_i$ est le produit de convolution du gain de la fenêtre temporelle w par le gain du filtre idéalement implémenté, ou fenêtre spectrale $\tilde{h}_i$:

$$\hat{h}_i(\nu) = \tilde{w}(\nu) * \tilde{h}_i(\nu) \tag{3.82}$$

La fenêtre temporelle introduit plus ou moins de diaphonie dans le filtrage en fonction de ses caractéristiques spectrales (amplitude des lobes secondaires). Il est préférable d'adopter ici une fenêtre dont le lobe principal est plus large (élargissant ainsi le module $\hat{h}_i$, et perdant de la précision sur la fréquence de coupure) mais les rebonds plus faibles, afin de diminuer la diaphonie.

Le choix de la fenêtre (forme, durée) et la taille de la transformation de Fourier discrète étant fixés, l'effet de cette convolution ne sera plus mentionné, malgré son importance. Un des inconvénients majeurs de la synthèse de filtres par transformée de Fourier est en effet la difficulté de calculer précisément les fréquences de coupure, à cause de la convolution fréquentielle périodique.

Le choix le plus simple pour les filtres h_i reste celui de filtres *idéaux*, de forme rectangulaire et de gain unité:

$$\tilde{h}_i(\nu) = 1$$

pour $\nu_{i_1} \leq |\nu| < \nu_{i_2}$

$$\tilde{h}_i(\nu) = 0$$

sinon.

Le domaine des fréquences est partitionné en N_j régions: aux limites de chaque région la coupure est aussi raide que l'autorise l'échantillonnage fréquentiel. La somme des gains vérifie bien 3.81.

L'observation d'un signal à travers un filtre, qui en limite la bande passante, introduit un *étalement temporel* de la même façon que la limitation temporelle d'un signal produit un étalement spectral.

limitation de la bande passante

Il est instructif de considérer l'effet de limitation de la largeur de bande sur un résonateur du second ordre $s(t)$ 3.42. En considérant une décomposition en parallèle du signal de parole, il s'agira de retrouver les paramètres de chaque section du second ordre en observant le signal dans les bandes de fréquence correspondant aux maxima spectraux. La fenêtre spectrale d'observation, autour de chaque maximum, entraîne une limitation de la largeur de bande de la section observée, à partir de la fréquence de coupure du filtre. Une limitation de la largeur des "jupes", c'est-à-dire du module du gain du résonateur sous $-6dB$ du sommet intervient donc. Pour une FOF, rappelons que c'est le paramètre β, lié au temps d'excitation qui règle, de façon assez indépendante de la largeur de bande, la largeur des jupes. Une coupure des jupes comme celle occasionnée par un filtre rectangulaire, aura donc intuitivement tendance à augmenter le temps d'excitation dans le domaine temporel.

Un filtre rectangulaire $\tilde{R}_{\nu_c}$, de gain unité, centré sur l'origine et de bande passante ν_c, possède comme réponse impulsionnelle:

$$r_{\nu_c}(t) = \frac{sin(\pi\nu_c t)}{\pi t} = \nu_c sinc(\nu_c t) \tag{3.83}$$

sinc est la fonction sinus cardinal. Il est bien connu que la limite lorsque ν_c tend vers l'infini des $r_{\nu_c(t)}$ vaut $\delta_0(t)$:

$$\delta_0(t) = \lim_{\nu_c \to \infty} \nu_c sinc(\nu_c t) \tag{3.84}$$

Ce qui correspond à l'observation du résonateur avec une fenêtre spectrale suffisamment large pour ne pas le perturber.

Si ν_c est fini, la réponse impulsionnelle s_r obtenue en considérant pour simplifier l'enveloppe de s centrée en zéro:

$$s_r(t) = s(t) * r_{\nu_c}(t) = \int s(\tau) r_{\nu_c}(t-\tau) d\tau \tag{3.85}$$

Si $\nu_c \gg \alpha$, r_{ν_c} est de courte durée par rapport à s, le développement de 3.85 suivant les moments successifs de r_{ν_c} donne, à l'ordre 1:

$$\begin{aligned} s_r(t) &= s(t) \int r_{\nu_c}(t) dt + e_1(t) & (3.86) \\ &\simeq s(t) \int r_{\nu_c}(t) dt & (3.87) \end{aligned}$$

Réciproquement, Si $\nu_c \ll \alpha$, s est de courte durée par rapport à r_{ν_c}, et le développement de 3.85 suivant les moments successifs de s donne, à l'ordre 1:

$$s_r(t) = r_{\nu_c}(t) \int s(t)dt + e_2(t) \tag{3.88}$$

$$\simeq r_{\nu_c}(t) \int s(t)dt \tag{3.89}$$

Dans le premier cas, s_r peut être approximée par une FOF, en réglant le paramètre β pour tenir compte de e_1 au début de la réponse temporelle (voir la figure 3.12).

Dans le second cas, l'enveloppe temporelle de s_r est dominée par le filtrage par r_{ν_c}, et l'exponentielle décroissante due à s disparaît sous les lobes secondaires de r_{ν_c} (voir la figure 3.12).

Notons que s et r_{ν_c} sont ici supposés tous les deux centrés de la même façon: si ce n'est plus le cas, dans le domaine temporel une modulation supplémentaire intervient, par une exponentielle complexe dont la fréquence vaut la différence entre les fréquences centrales de s et r_{ν_c}.

La réponse impulsionnelle d'une section parallèle observée dans une région spectrale limitée peut donc se comporter de façon très différente, en fonction de la largeur de bande d'observation.

En se plaçant dans le premier cas, le temps de montée de la réponse impulsionnelle sera intuitivement liée à la largeur du premier lobe, donc d'autant plus long que la fenêtre spectrale est de faible largeur.

introduction du temps d'excitation

Pour étudier le rapport entre le signal issu d'un modèle à formants en parallèle observé dans une région spectrale limitée et les FOF, on peut interpréter le temps d'excitation introduit dans 3.61 en terme de filtrage d'un résonateur. Il s'agit ici de rechercher le filtre qui, en permettant l'observation d'un résonateur dans une région spectrale limitée, aurait introduit le temps d'excitation sous la forme rencontrée dans $\Lambda(t)$. Soit $\Lambda(t)$ définie dans 3.61 et $e_s(t)$ l'enveloppe temporelle d'un résonateur:

$$e_s(t) = Ae^{-\alpha t} \tag{3.90}$$

pour $t \geq 0$.

$$\tilde{e}_s(\nu) = A \int_0^\infty e^{-(\alpha+2i\pi\nu)t} dt \tag{3.91}$$

soit h_o la forme d'onde qui a produit Λ en excitant e_s:

$$\Lambda(t) = h_o(t) * e_s(t) \tag{3.92}$$

On peut écrire dans le domaine fréquentiel:

$$\tilde{h}_o(\nu) = \frac{\tilde{\Lambda}(\nu)}{\tilde{e}_s(\nu)} \tag{3.93}$$

Donc en utilisant 3.65 et :

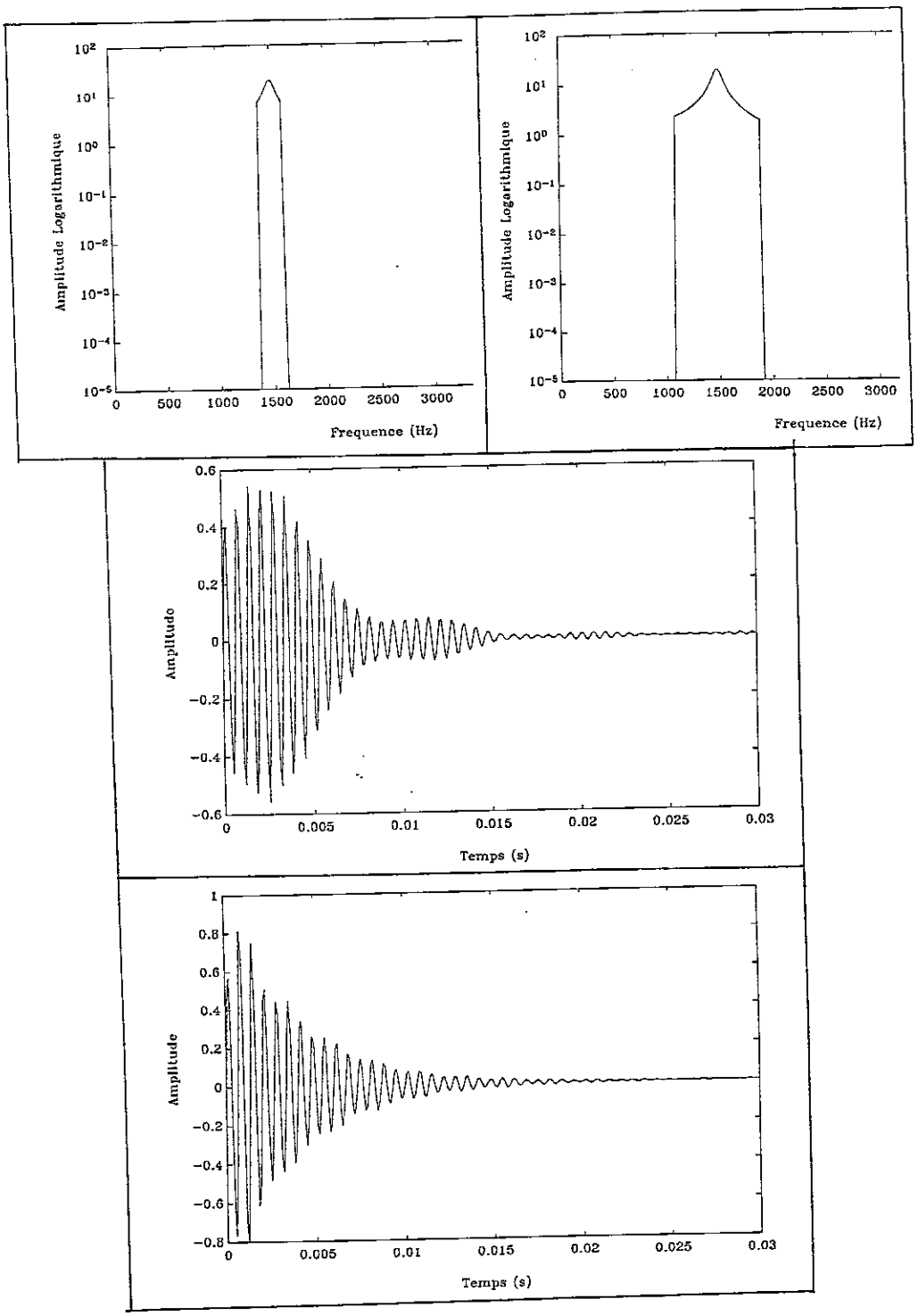

Figure 3.12: filtrage rectangulaire par transformée de Fourier d'un résonateur. Largeur de bande d'analyse 500 Hz, spectre d'amplitude logarithmique (haut à gauche) et signal temporel (milieu), puis 1000 Hz (haut à droite et bas).

$$\tilde{e}_s(\nu) = \frac{A}{\alpha + i2\pi\nu} \tag{3.94}$$

il vient:

$$\tilde{h}_o(\nu) = \frac{A\beta^2(e^{-(\alpha+i2\pi\nu)\pi/\beta} + 1)}{2((\alpha + i2\pi\nu)^2 + \beta^2} \tag{3.95}$$

dont le module vaut:

$$|\tilde{h}_o(\nu)| = \frac{A\beta^2\sqrt{e^{-2\alpha\frac{\pi}{\beta}} + 2e^{-\alpha\frac{\pi}{\beta}}cos(2\pi\nu\frac{\pi}{\beta}) + 1}}{2\sqrt{((\alpha^2 + 4\pi^2\nu^2)^2 + \beta^2(\beta^2 + 2\alpha^2 - 8\pi^2\nu^2))}} \tag{3.96}$$

Des hypothèses analogues à celles mentionnées plus haut sur l'ordre de grandeur des paramètres entraînent bien au voisinage de l'origine (fréquence centrale de la FOF) un module approximativement constant:

$$|\tilde{h}_o(\nu)| \simeq \frac{A\beta^2\sqrt{e^{-2\alpha\frac{\pi}{\beta}} + 2e^{-\alpha\frac{\pi}{\beta}} + 1}}{2\sqrt{\beta^4}} \tag{3.97}$$

$$= \frac{A(e^{-\alpha\frac{\pi}{\beta}} + 1)}{2} \tag{3.98}$$

Si l'on s'éloigne de l'origine, l'approximation 3.98 n'est plus valide et le module du gain $\tilde{h}_o(\nu)$ est donné par 3.96. Si l'on conserve $\beta^2 \gg \alpha^2$ et si $4\pi^2\nu^2 \gg \alpha^2$, ce module est approximativement égal à:

$$|\tilde{h}_o(\nu)| \simeq \frac{A\beta^2\sqrt{e^{-2\alpha\frac{\pi}{\beta}} + 2e^{-\alpha\frac{\pi}{\beta}} + 1}}{2\sqrt{(4\pi^2\nu^2 - \beta^2)^2}} \tag{3.99}$$

La figure 3.13 montre le gain du filtre qui produit une fof à partir d'un résonateur. Pratiquement, il est préférable d'utiliser un filtre plus simple, même si la forme exacte de l'excitation obtenue ne ressemble pas à celle d'une FOF.

influence du filtrage sur le modèle à formant

Il ressort de ce qui précède que, si l'on suppose la modélisation spectrale et la détection des maxima spectraux parfaites (c'est-à-dire que les maxima repérés correspondent exactement aux formants du modèle de production):

- le signal temporel obtenu ne correspond plus à celui de 3.60, mais à son observation à travers les filtres $R_{\nu c}$: des rebonds dûs à la convolution par la réponse impulsionnelle de ces filtres apparaissent .
- un temps d'excitation est alors introduit dans la réponse impulsionnelle de chaque formant dans le modèle parallèle, qui tend à la rendre symétrique contrairement au cas d'une section du second ordre. La forme de cette réponse impulsionnelle peut être approchée par celle d'une FOF telle que 3.61. Ces deux fonctions restent cependant théoriquement assez différentes.

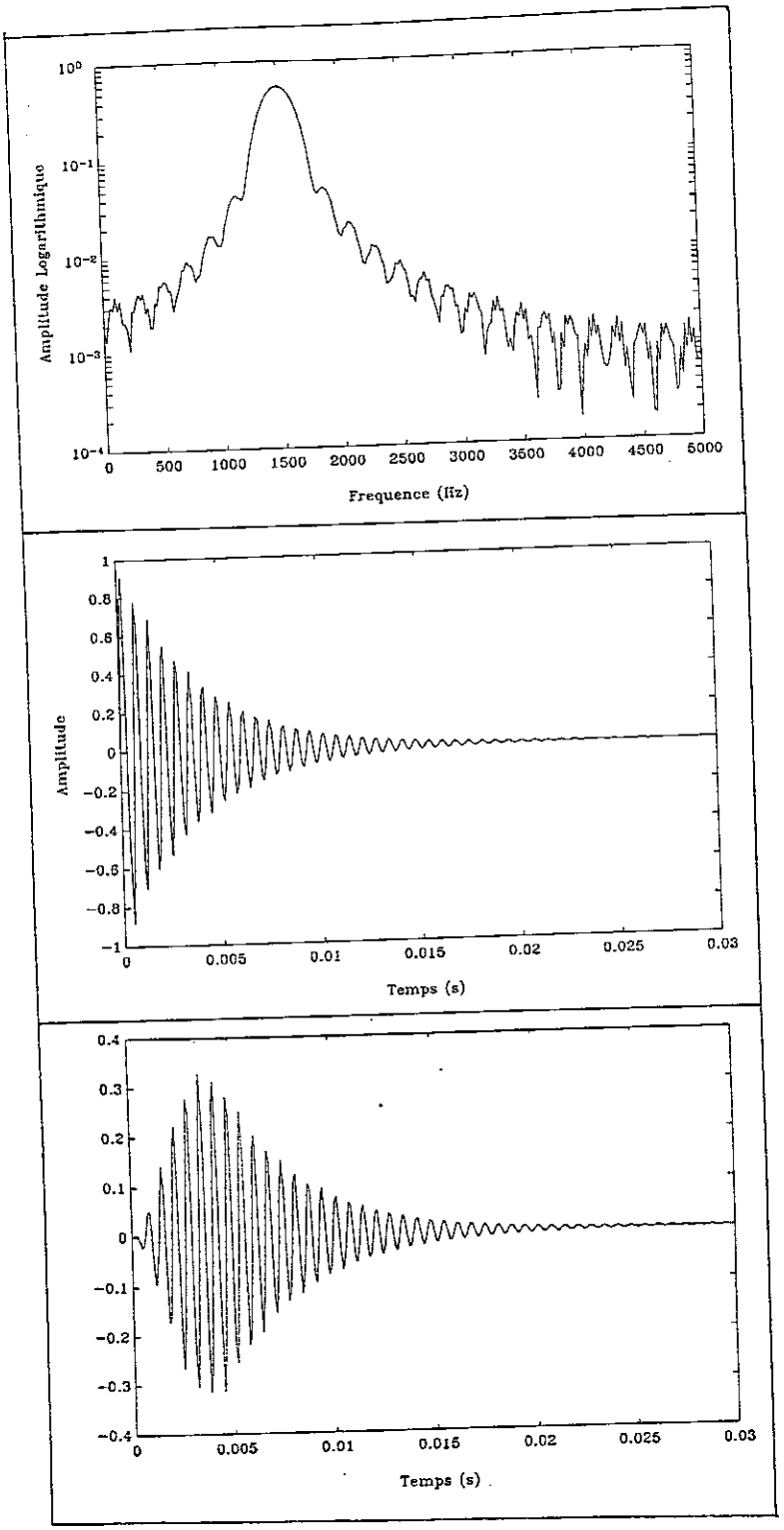

Figure 3.13: Spectre d'amplitude logarithmique (haut) du filtre d'observation qui a produit la FOF (signal du bas) à partir du résonateur (signal du milieu), obtenu par division spectrale.

3.5.5 segmentation temporelle

détection des formes d'ondes

Après la segmentation spectrale, le signal obtenu dans chaque bande de fréquence se présente comme la modulation d'une composante dominante approximativement sinusoïdale, dénommée porteuse, par une enveloppe variant plus lentement.

Il s'agit de retrouver dans ce signal filtré les réponses temporelles du modèle de production.

Idéalement, pour de la parole voisée, la fréquence de la porteuse se rapporte à la fréquence centrale du formant, sa phase à la valeur de la phase du formant à la fréquence centrale et à un terme linéaire dû à la fréquence centrale. L'enveloppe temporelle reflète d'une part les paramètres d'enveloppe spectrale (forme, largeur de bande, largeur des jupes) et d'autre part la périodicité de l'excitation. L'excitation se traduit par des maxima locaux de l'enveloppe, alors que les paramètres spectraux déterminent l'allure de la courbe entre deux maxima. Une segmentation de l'enveloppe temporelle par ses extrema permet dans ce cas idéal, avec les restrictions dues à 3.85, d'associer une forme d'onde au segment de signal caractérisé par un maximum. Ceci revient à considérer que la réponse impulsionnelle d'une section parallèle est observable dans le signal filtré, et deux cas sont possibles, suivant le filtre R_{ν_c}:

- les rebonds de r_{ν_c} sont apparents sur la courbe d'enveloppe, et à une réponse impulsionnelle formantique sont associées plusieurs formes d'ondes. On peut remarquer que cette propriété d'apparition des rebonds sur l'enveloppe provient de r_{ν_c}, mais aussi du choix de lissage de l'enveloppe.

- les rebonds de r_{ν_c} n'apparaissent pas sur la courbe d'enveloppe, et une seule forme d'onde est associée à chaque réponse impulsionnelle formantique. Ce cas reste bien sûr celui que l'on souhaiterait rencontrer le plus souvent.

Pour un signal aléatoire, bruit blanc filtré qui simule les fricatives, le signal temporel dans un bande apparaît localement de la même façon: la fréquence de la porteuse varie constamment, mais dans les limites imposées par le filtre d'observation. Des maxima locaux de l'enveloppe sont également présents, mais il est alors difficile de les relier à leur origine physique, sinon comme maxima locaux de l'excitation. Néanmoins, une segmentation de l'enveloppe temporelle reste possible, suivant ses extrema. Associer une forme d'onde à chaque maximum, comme dans le cas idéal précédent, revient à approcher localement (en temps, et en fréquence par le filtre d'observation) le signal aléatoire par un signal qui ne l'est pas. On rajoute ainsi une stationnarité locale, justifiable lorsque le voisinage sous-jacent est assez petit, comme il a été discuté plus haut.

Une impulsion isolée relève du premier cas idéal.

calcul de l'enveloppe

Les maxima obtenus, rapportés ensuite aux formes d'ondes dépendent de façon assez étroite du lissage de la courbe d'enveloppe. Le calcul de l'enveloppe Θ doit démoduler au mieux porteuse et enveloppe, tout en garantissant une courbe aussi lisse que possible:

les variations de l'enveloppe doivent rester lentes (d'un facteur au moins trois), devant celles de la porteuse.

Un calcul identique à celui proposé au second chapitre est utilisé ici, qui comprend les étapes suivantes:

- redressement double alternance:

$$\Theta 1_n = |s_n| \tag{3.100}$$

- recherche des maxima locaux, ou sommets, $s_{\hat{n}_i}$;

- liaison des sommets, par une fonction affine par morceaux:

$$\Theta 2_n = s_{\hat{n}_i} + \frac{(s_{\hat{n}_{i+1}} - s_{\hat{n}_{i+}})}{\hat{n}_{i+1} - \hat{n}_i}(n - n_i) \tag{3.101}$$

  pour $n_i \leq n < n_{i+1}$

- lissage de cette fonction par filtrage passe bas: un double résonateur l_{ν_Θ}, du 4^{ieme} ordre et non déphaseur, est utilisé. La fréquence de coupure ν_Θ est choisie inférieure à 800 Hz, de façon adaptée à la fréquence centrale f_c du filtre d'observation $\tilde{r}_{\nu_c}$:

$$\nu_\Theta = min(800, \frac{f_c}{k}) \tag{3.102}$$

  Expérimentalement, une bonne valeur pour k est 3, ce qui donne finalement pour le calcul de l'enveloppe:

$$\Theta_n = \Theta 2_n * l_{\nu_\Theta} \tag{3.103}$$

segmentation temporelle

Un ensemble de maxima et de minima est recherché sur l'enveloppe temporelle définie à la section précédente, qui permet de segmenter la bande considérée.

Il devient impossible dans le domaine temporel d'utiliser pour la segmentation des fonctions aussi simples que les filtres de segmentation spectrale, bien que le problème soit de même nature: observer les composantes présentes dans le signal indépendamment les unes des autres, et sur une durée limitée. La méthode adoptée ici consiste à partager la valeur de l'enveloppe en deux à l'endroit d'un minimum: les paramètres d'enveloppe provenant de l'estimation initiale seront recherchés de façon à valoir la moitié de la valeur du signal réel aux minima.

La segmentation du signal à l'aide de segments sinusoïdaux passant aux demi-valeurs pour les mimina, donne une reconstruction indiscernable de l'original, bien qu'une légère ondulation de la somme des fonctions de segmentation puissent apparaître.

3.5.6 estimation des paramètres: régions formantiques

estimation initiale des paramètres formantiques

Sur le signal segmenté peuvent s'estimer les paramètres des formes d'ondes.

- l'instant de référence, ou maximum de l'enveloppe d'une forme d'onde;
- le temps d'excitation;
- la rapidité d'amortissement;
- la fréquence de la porteuse;
- la phase de la porteuse;
- l'amplitude temporelle;

La position du maximum de la forme d'onde, ou instant de référence, est donnée par le maximum de l'enveloppe. La précision de cette estimation est limitée par la fréquence d'échantillonnage: les valeurs de la courbe d'enveloppe doivent être interpolées pour accroître cette précision.

Pendant le temps d'excitation l'enveloppe d'une forme d'onde est une arche montante de cosinus. L'estimation initiale de la demi-fréquence de ce cosinus, ou paramètre β est directe en faisant passer l'arche par deux points: le sommet de l'enveloppe à l'instant de référence et la demi-valeur de l'enveloppe au minimum précédant cet instant.

L'amortissement α, ou pulsation de la demi-largeur de bande, est obtenue de façon similaire en faisant passer une exponentielle décroissante entre le sommet de l'enveloppe temporelle à l'instant de référence et la demi-valeur de l'enveloppe au minimum suivant cet instant.

L'estimation initiale de la fréquence de la porteuse peut s'effectuer dans le domaine temporel comme décrit au chapitre précédent, ou dans le domaine spectral. Expérimentalement l'estimation effectuée lors de la segmentation de la courbe d'enveloppe spectrale est préférable, car suffisamment précise sans calcul supplémentaire.

La phase de la porteuse est estimée au voisinage de l'instant de référence. A partir de cet instant, l'instant du premier passage par zéro est recherché (avec interpolation) dans le sens du temps croissant, et la phase ($-\pi$ ou π) évaluée.

L'estimation initiale de l'amplitude temporelle est l'amplitude de l'enveloppe à l'instant de référence.

validation de l'estimation

L'estimation initiale des paramètres conduit à un signal de synthèse, qui doit être suffisamment proche du signal original, pour un certain critère. Le critère *naturel* pour une application en traitement de la parole est un critère perceptif: signal original et signal synthétique ne doivent pas être discernables, pour n'importe quel auditeur. Ce critère n'est évidemment pas formalisable en termes de traitement du signal, mais en termes psycho-acoustiques. Il ne peut donc être utilisé que sur la sortie du système, et non pendant l'élaboration de cette sortie.

Pour évaluer la qualité d'une représentation pendant son élaboration même, au critère perceptif est généralement substitué un critère plus opératoire, ordinairement celui des moindres carrés.

Un signal d'erreur ε_n est formé par différence entre le signal désiré s_n (ici le signal naturel) et le signal $\hat{s}_n$ issu du système dont on cherche à estimer les paramètres (ici une somme de formes d'ondes):

$$\varepsilon_n = s_n - \hat{s}_n \tag{3.104}$$

L'erreur quadratique totale ξ s'obtient en choisissant une fenêtre temporelle W pour les n.

$$\xi = \frac{1}{W} \sum_{n=n_0}^{n_0+W-1} \varepsilon_n^2 \tag{3.105}$$

$$= \frac{1}{W} \sum_{n=n_0}^{n_0+W-1} (s_n - \hat{s}_n)^2 \tag{3.106}$$

optimisation des paramètres

Pour affiner l'estimation initiale des paramètres il faut mettre en oeuvre un processus d'optimisation. Le signal dépend de façon non-linéaire des paramètres à optimiser et la méthode d'optimisation choisie relève de la théorie des systèmes adaptatifs [114].

Les paramètres de la porteuse ne sont pas à optimiser *a priori*. Seul les quatre paramètres d'enveloppe vont être considérés. Un vecteur des poids $\vec{p} = [p^1 p^2 p^3 p^4]$ est formé suivant les quatre directions de ces quatre paramètres:

$$\vec{p} = p^1\vec{i} + p^2\vec{j} + p^3\vec{k} + p^4\vec{l} \tag{3.107}$$

L'adaptation, en boucle fermée, consiste à minimiser un critère -ici l'erreur quadratique totale ξ- par rapport aux poids p^i, de façon itérative. En supposant la fenêtre W sur l'erreur fixée, le gradient normalisé de l'erreur $\nabla\xi(\vec{p})$, noté ∇ fonction du vecteur des poids $\vec{p}$ vaut:

$$\nabla = \frac{\nabla\xi(\vec{p})}{||\xi(\vec{p})||} \tag{3.108}$$

$$= \frac{1}{(\sum_{i=1}^{4}(\frac{\partial\xi}{\partial p^i})^2)^{1/2}}(\frac{\partial\xi}{\partial\vec{p}}) \tag{3.109}$$

$$= \frac{1}{(\sum_{i=1}^{4}(\frac{\partial\xi}{\partial p^i})^2)^{1/2}}(\frac{\partial\xi}{\partial p^1}\vec{i} + \frac{\partial\xi}{\partial p^2}\vec{j} + \frac{\partial\xi}{\partial p^3}\vec{k} + \frac{\partial\xi}{\partial p^4}\vec{l}) \quad (3.110)$$

le gradient est estimé en calculant la dérivée $\frac{\partial\xi}{\partial\vec{p}}$ par la méthode des différences centrales:

$$\frac{\partial\xi}{\partial\vec{p}} \simeq \frac{\xi(\vec{p}+\vec{\delta}) - \xi(\vec{p}-\vec{\delta})}{2||\vec{\delta}||} \quad (3.111)$$

Les valeurs initiales des poids sont données par l'estimation initiale précédente. La méthode de plus forte pente du gradient permet itérativement de se rapprocher du vecteur optimal. Si $\vec{p}_k$ est le vecteur des poids au pas k, une constante μ_k permet de trouver le vecteur $\vec{p}_{k+1}$ dans la direction du gradient ∇_k au pas k:

$$\vec{p}_{k+1} = \vec{p}_k + \mu_k(-\nabla) \quad (3.112)$$

La direction de recherche étant fixée par le gradient, la constante μ_k permet soit de s'approcher du vecteur optimal, soit de le dépasser. Si l'erreur $\xi(\vec{p}_{k+1})$ décroît, l'optimum n'est pas dépassé et en multipliant μ_k par une constante supérieure à 1 (par exemple en le doublant) un nouveau $\vec{p}_{k+1}$ peut être évalué. Si l'erreur ne décroît pas, l'optimum a été dépassé, une nouvelle itération $k+2$ commence, avec une nouvelle direction ∇_{k+1}, une nouvelle constante μ_{k+1} inférieure à μ_k et le nouveau vecteur $\vec{p}_{k+1}$.

Une limite sur le nombre d'itérations et/ou sur le critère d'erreur permettent d'arrêter la procédure d'optimisation.

Puisque l'optimisation est calculée pour chaque forme d'onde, la valeur de W est choisie comme le nombre de points entre les deux minima qui lui servent de limites.

Cette procédure se simplifie si l'on considère que l'instant de référence a été correctement estimé dans l'estimation initiale. Trois paramètres seulement restent à optimiser.

La version la plus élémentaire de l'optimisation consiste à ajuster l'amplitude temporelle de façon à égaler l'énergie du signal de synthèse et du signal naturel pendant la durée d'une forme d'onde

3.5.7 estimation des paramètres de la bande de base

traitement de la bande de base

La première bande formantique, au sens de premier maximum de l'enveloppe spectrale, sert à définir la bande de base. Le traitement direct de cette région par la même méthode que les autres ne se justifie plus et donne des résultats décevants, ce qui oblige à une redécomposition du signal dans la bande la plus grave.

La représentation de cette bande par des fonctions élémentaires conduit à une représentation sinusoïdale, ce qui permet un processus de traitement assez homogène à celui des bandes formantiques:

- obtention de la bande de base par filtrage dans la première région issue de l'analyse précédente.
- recherche du spectre d'amplitude dans cette région par transformation de Fourier.

- segmentation spectrale sur le spectre d'amplitude.
- filtrage dans les régions spectrales ainsi définies.
- détection des formes d'ondes.
- estimation des paramètres sinusoïdaux.

modélisation et segmentation spectrale

Les composantes que l'on se propose d'isoler dans cette région spectrale sont des sinusoïdes lentement modulées. La recherche est donc menée sur le spectre d'amplitude de la transformée de Fourier à court terme. Un nombre suffisant d'échantillons spectraux doit être disponible pour discerner ces composantes, par exemple pour un segment voisé avec un fondamental élevé. Les valeurs du spectre peuvent être interpolées pour assurer un précision accrue. On utilisera donc ici une transformée de Fourier rapide sur un relativement grand nombre d'échantillons (par exemple 1024), tout en conservant une fenêtre temporelle d'analyse suffisamment courte pour ne pas englober trop de périodes de la composante la plus basse.

Pour un segment voisé, les maxima du spectre correspondent aux harmoniques, et pour un signal non voisé aux maxima locaux de la densité spectrale.

filtrage

La méthode de filtrage employée est la même que celle discutée plus haut. Les filtres employés sont également des filtres rectangulaires, mais les artefacts dûs à la convolution par la fenêtre temporelle sont nettement plus sensibles ici. En effet, les bandes d'analyse sont plus graves, donc d'un ordre de grandeur qui se rapproche de celui de la fenêtre temporelle, et les largeurs de bande beaucoup plus petites. L'étalement dû au gain de la fenêtre est donc beaucoup plus important relativement aux dimensions spectro-temporelles des signaux filtrés, ainsi que la diaphonie.

Ces artefacts ne présentent pas d'inconvénients majeurs pour la suite du traitement.

segmentation temporelle

Après filtrage dans les régions harmoniques, les signaux temporels disponibles sont des sinusoïdes, évoluant assez lentement pour les segments voisés, mais qui peuvent s'avérer assez courtes pour des impulsions isolées (plosives par exemple) ou un signal bruité.

Le procédé de segmentation temporelle ne peut donc plus se baser sur l'enveloppe temporelle, puisqu'elle risque de ne pas évoluer notablement sur une trame, mais sur la périodicité. Les extrema détectés directement sur le signal donnent une description de chaque cycle des sinusoïdes.

Deux alternatives s'offrent pour segmenter: soit segmenter chaque bande séparément, en fixant un nombre de cycles par forme d'onde, soit privilégier une bande pour segmenter, et segmenter les autres bandes suivant les limites trouvées dans cette bande.

Le premier choix relativise les erreurs de segmentation dans une bande donnée, mais la quantité de formes d'ondes est directement proportionnelle à la fréquence de la

composante. Le sens de ces formes d'ondes n'est que d'interpoler au mieux (avec une précision relative constante) les paramètres sinusoïdaux, sans que cet excès d'information apporte de connaissances supplémentaires, sur le modèle de production par exemple.

Le second choix rend la bande de base dépendante de sa composante la plus grave. Par contre, un sens physique peut être attaché à la segmentation obtenue, et une notable réduction de la quantité de formes d'ondes, sans préjudice pour la qualité de la représentation.

Pour la parole voisée, la segmentation sur la bande grave permet idéalement d'isoler la contribution de la bande de base pour chaque période de voisement. La parole non voisée dans cette région spectrale se traduit essentiellement par les impulsions isolées dues aux plosives: l'énergie des bruits fricatifs est très faible en dessous d'environ $500Hz$. Ici encore segmenter suivant la composante la plus basse présente l'avantage de regrouper les basses fréquences d'une explosion. Une analyse temporelle plus fine n'apportera pas d'informations particulières sur la production pour justifier l'accroissement de la quantité de formes d'ondes.

Les fonctions de segmentation utilisées sont deux segments de sinusoïdes passant par le sommet de la forme d'onde, et par la demi-valeur au minimum suivant ou précédent, en fonction du segment.

estimation des paramètres sinusoïdaux

les paramètres à estimer pour la représentation sinusoïdale sont

- l'instant de référence;
- la fréquence;
- la phase;
- le paramètre d'enveloppe montante;
- le paramètre d'enveloppe décroissante;
- l'amplitude temporelle;

L'instant de référence donne ici également la phase.

La fréquence doit être estimée en comptant la fréquence moyenne contenue dans la forme d'onde: la fréquence obtenue par la segmentation spectrale n'est plus d'une précision relative suffisante pour pouvoir s'utiliser ici.

Les paramètres d'enveloppe s'estiment simplement en faisant passer les deux arches de sinusoïdes par le maximum et le minimum voulus.

L'amplitude temporelle s'obtient en égalant l'énergie de la forme d'onde synthétique et du segment de signal temporel correspondant.

3.5.8 frontières de trame

Les traitements présentés jusqu'à présent sont compris dans une trame d'analyse. Les formes d'ondes détectées permettent de s'en affranchir.

Les trames de filtrage permettent d'estimer les paramètres d'une forme d'onde qui excèdent les limites d'une trame d'analyse. Cependant, dans une trame donnée ne sont estimées que les formes d'ondes dont l'instant de référence appartient à la trame. Les dédoublements (la même forme d'onde vue par deux trames différentes) et les omissions sont ainsi évités, en supposant que les instants de maximum de la forme d'onde vue à travers deux trames, donc deux filtres différents soient identiques. Cette condition est réalisée si les largeurs de bandes des deux filtres d'observation sont d'un ordre de grandeur comparable.

Puisque les trames sont jointives, et de durée assez courte, les frontières de trames ne posent pas de problèmes particuliers, l'estimation des paramètres de chaque forme d'onde étant indépendante.

3.5.9 synthèse

A partir des paramètres formantiques et sinusoïdaux, la synthèse consiste simplement à calculer chaque forme d'onde et à effectuer la sommation.

Une sélection permet de diminuer la quantité de formes d'ondes. Expérimentalement, plus d'un tiers des formes d'ondes détectées sont supprimées sans dommage, en éliminant celles dont l'énergie est inférieure à un seuil. Ce sont bien sûr les formes d'ondes de l'aigu qui tendent à disparaître plutôt que celles du grave.

Ce type de synthèse est très économique, en tabulant les fonctions utilisées: sinus pour la porteuse et cosinus et exponentielle pour l'enveloppe. Le nombre de multiplications successives reste faible et l'arithmétique virgule fixe peut s'employer sans problème de précision pour les calculs.

Le signal d'erreur cumulé dans chaque bande d'analyse donne un signal résiduel. La somme du signal synthétique et du signal résiduel est égale au signal original.

3.6 applications

Des exemples de synthèse et puis d'analyse-synthèse vont être présentés, qui retracent en quelque sorte la démarche précédente. Notre ambition n'est pas de présenter une méthode de codage efficace, ce qui nécessiterait une étude soigneuse du rapport qualité/débit d'information pour le système. Il s'agit simplement de démontrer que la réalisation des principes évoqués précédemment conduit à des systèmes de synthèse et d'analyse/synthèse fonctionnels. Ainsi les systèmes réalisés proposent une validation de la représentation en formes d'ondes basée sur un modèle de production du signal de parole.

Les applications possibles qui suivent sont esquissées. Chacune mériterait un long développement qui n'a été qu'au plus partiellement réalisé jusqu'à présent.

Un intérêt particulier réside dans la possibilité de synthétiser de la parole à partir de paramètres acoustiques. Une des méthodes les plus répandues pour la synthèse de

parole à partir du texte utilise des règles sur l'évolution des paramètres acoustiques. L'usage possible de la représentation proposée dans ce cadre sera évoqué.

L'analyse spectro-temporelle fournit depuis un demi-siècle de précieux outils pour la recherche en phonétique (en particulier pour la phonétique acoustique). La visualisation du signal comme ensemble de formes d'ondes dans le plan temps-fréquence s'inscrit naturellement dans cette perspective.

La représentation proposée fournit des paramètres reliés au modèle de production. Une application utile pour l'étude et le traitement de la parole est la modification spectro-temporelle du signal. De par la localisation et l'indépendance des objets de la représentation, plusieurs types de traitements de cette nature sont très simples à mettre en oeuvre.

3.6.1 analyse-synthèse

synthèse par FOF

La synthèse par FOF a démontré son efficacité dans le cadre de la synthèse musicale. Pour les segments vocaliques, elle a également été utilisée en synthèse de parole. Un synthétiseur a été mis au point, qui permet le calcul de cinq formants en temps réel. Même en se limitant aux segments vocaliques, un tel synthétiseur se révèle particulièrement utile pour l'acoustique musicale, la pédagogie.

A l'aide d'une estimation non automatique de paramètres, des exemples de synthèse pour les segments non vocaliques ont été construits.

Sans modification du programme de synthèse par FOF, des fricatives sourdes de bonne qualité ont été produites, en appliquant des règles simples: les FOF, dont les paramètres spectraux ont été fixés d'après la densité spectrale du bruit fricatif à reproduire, ont été générées à des instants d'apparition qui suivent une loi aléatoire uniforme entre deux fréquences. Typiquement une fréquence d'apparition distribuée uniformément autour de 1 kHz, avec un variation maximale de 30 ou 40 % a donné des résultats satisfaisants.

Cette étude préliminaire a permis de se convaincre expérimentalement de la possibilité de représenter et de contrôler des segments non voisés par des FOF.

analyse-synthèse automatique de parole par FOF

Dans le même cadre a été réalisée l'analyse/synthèse automatique de phrases complètes. L'analyse est issue d'un système de détection des maxima spectraux par prédiction linéaire [20], joint à une analyse du fondamental.

La représentation des maxima spectraux sur chaque trame par des FOF, renouvelées à chaque période fondamentale délivre automatiquement un signal synthétique de bonne qualité, comparable par exemple à de la prédiction linéaire classique usant d'un nombre important de coefficients. Ce résultat expérimental confirme la validité de la représentation pour la parole, et la possibilité d'automatisation de la recherche des paramètres.

L'utilisation de ce système a permis d'en rencontrer certaines limites: l'excitation trop simple et commune à toutes les FOF (dépendant d'un analyse du fondamental), avec une excitation par période, est en grande partie responsable de la qualité

"métallique", de "l'effet bouteille" du signal synthétisé, comme en prédiction linéaire classique.

Même pour un signal voisé, chaque période fondamentale de certains locuteurs montre clairement deux ou plusieurs excitations au sein du même cycle vocalique (la forme de l'onde de débit glottique peut être suffisamment complexe pour exciter le conduit vocal à plusieurs instants). De plus, dans l'aigu du spectre le signal est rarement bien périodique, même lorsqu'il est bien voisé. Dans le cas d'une voix mal timbrée, voilée, ou à la limite chuchotée ce phénomène devient prédominant.

La nécessité de traiter chaque FOF de façon indépendante apparaît donc pour d'une part traiter un segment quelconque de parole, et d'autre part améliorer la qualité de synthèse.

Un léger défaut de représentation des premiers harmoniques est également apparu. Ainsi la simplicité de la représentation n'était pas compatible avec la complexité du signal dans le grave du spectre.

analyse-synthèse automatique de parole par formes d'ondes

Le système d'analyse-synthèse en formes d'ondes a tenté de résoudre les problèmes posés au paragraphe précédent: les formes d'ondes sont recherchées indépendamment dans le domaine temps-fréquence, et la bande de base fait l'objet d'un raffinement de représentation.

L'analyse-synthèse automatique de phrases prononcées par plusieurs locutrices et locuteurs, en voix normale, en voix faible, en voix chuchotée donne un signal de synthèse de très bonne qualité (voir figures 3.14, 3.15, 3.16). Cependant, des défauts apparaissent pour certains locuteurs et en présence de bruit. Les problèmes de qualité ou de robustesse de la représentation font l'objet d'une section suivante.

3.6.2 formes d'ondes et synthèse de parole

Un processus d'analyse-synthèse automatique offre pour la synthèse de parole l'avantage d'une estimation automatique des paramètres. Tant pour la synthèse par règles que pour la synthèse par unités concaténées [49] (diphones, syllabes, ...) cette possibilité est précieuse.

Dans le premier cas, la construction du système de règles peut ainsi être en partie automatisable, mais surtout vérifiable sur de nombreuses données réelles. Actuellement, le manque de système d'analyse complètement automatique pour les synthétiseurs à formants [48] reste un de leurs plus sérieux problèmes. Ce manque est compensé par la pertinence et la simplicité des paramètres mis en jeu.

Dans le second cas une description des unités de synthèse par des paramètres acoustiques doit permettre de lisser les frontières de concaténation. Ici le codage des unités prélevées dans de la parole réelle est automatique, au prix d'un manque de simplicité d'interprétation des paramètres utilisés, et donc d'une difficulté de manipulation des caractéristiques acoustiques du signal [108] [13].

Le développement d'un synthétiseur de parole pour la synthèse à partir du texte basé sur la représentation en formes d'ondes offre des perspectives intéressantes.

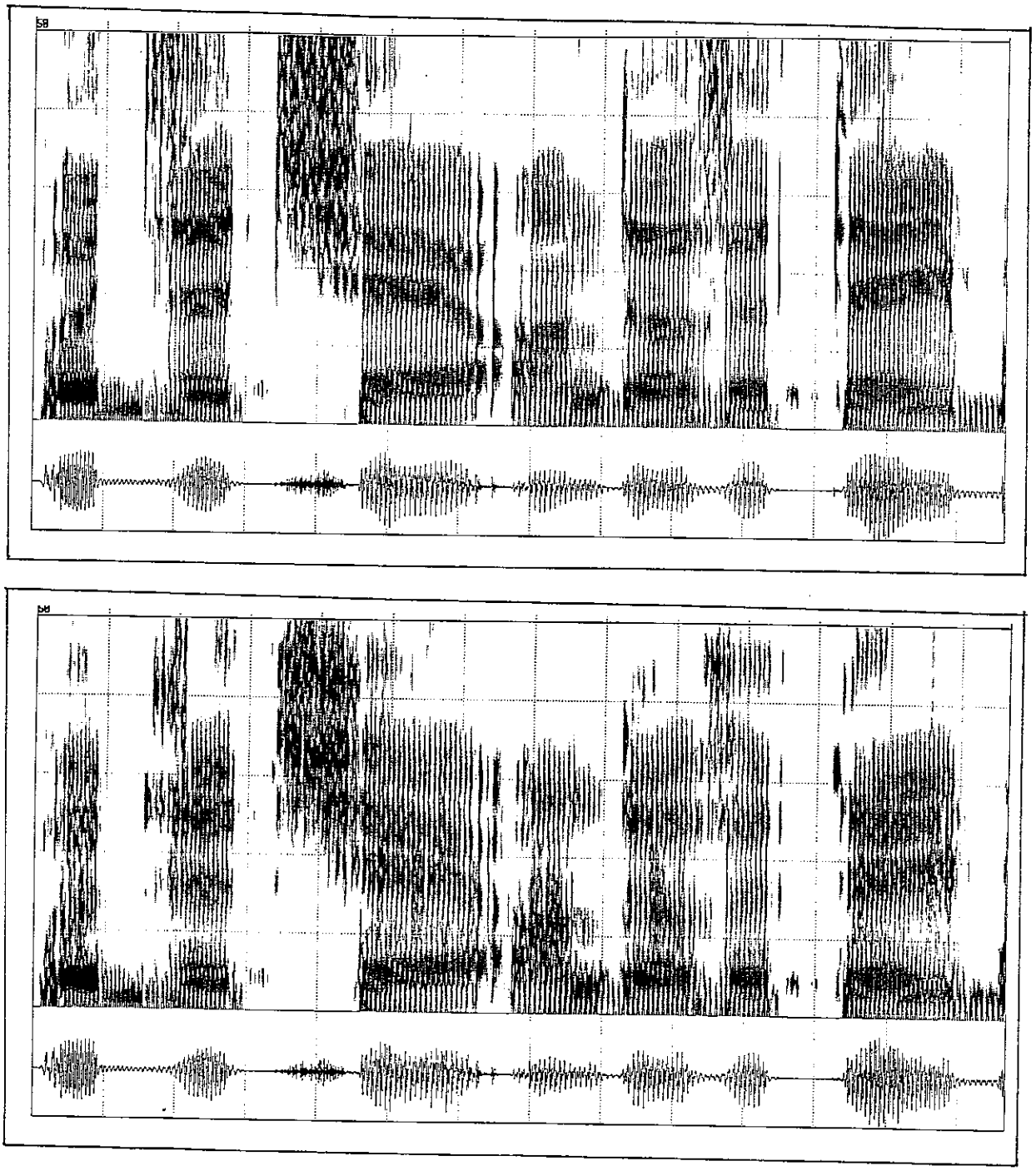

Figure 3.14: spectrogramme de "put the chair under the table", locuteur masculin, parole naturelle (haut) et synthétique (bas).

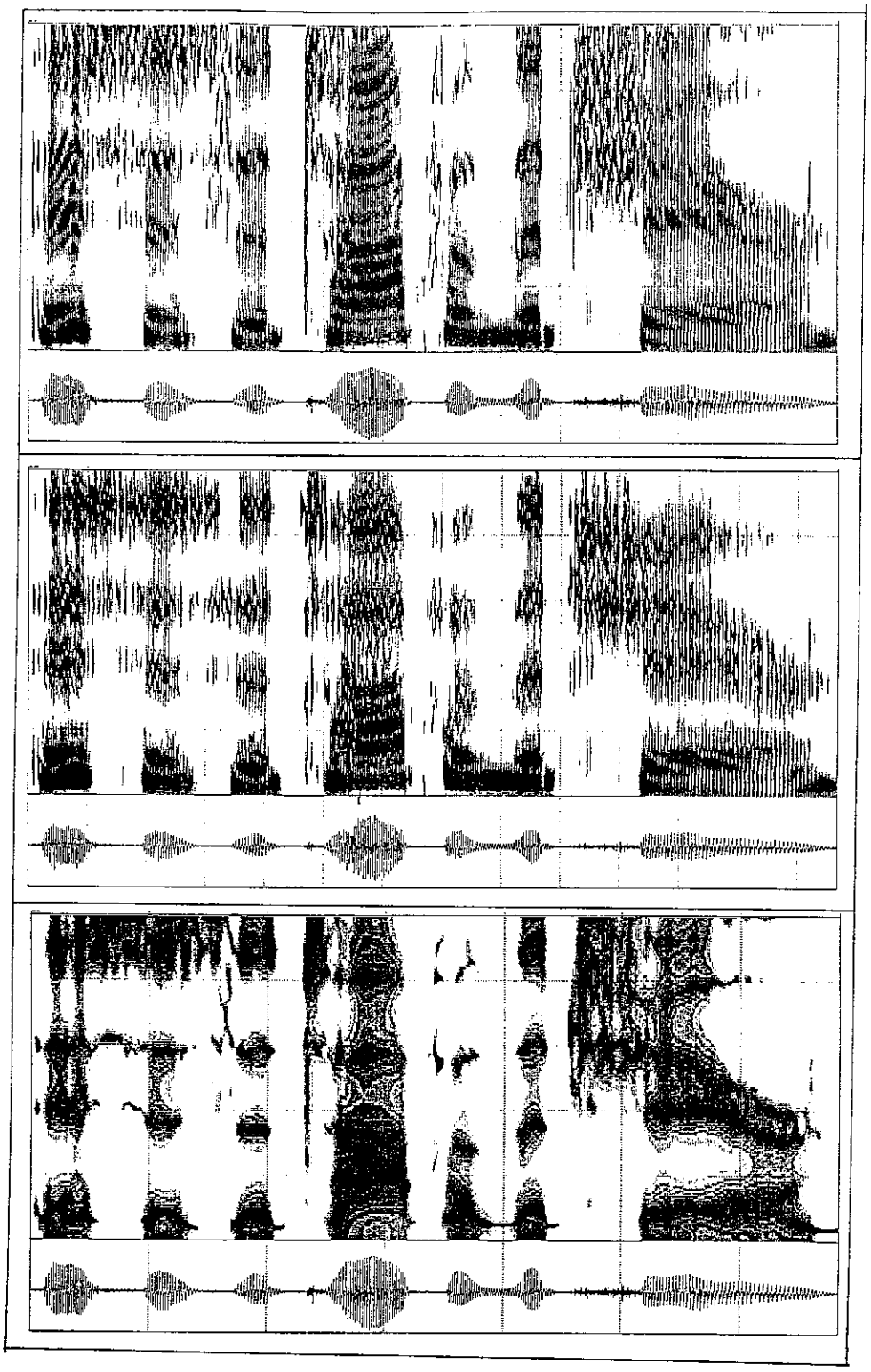

Figure 3.15: spectrogramme de "this is the top of the chair", locuteur féminin, parole naturelle (haut), synthétique (milieu), spectrogramme LPC (bas).

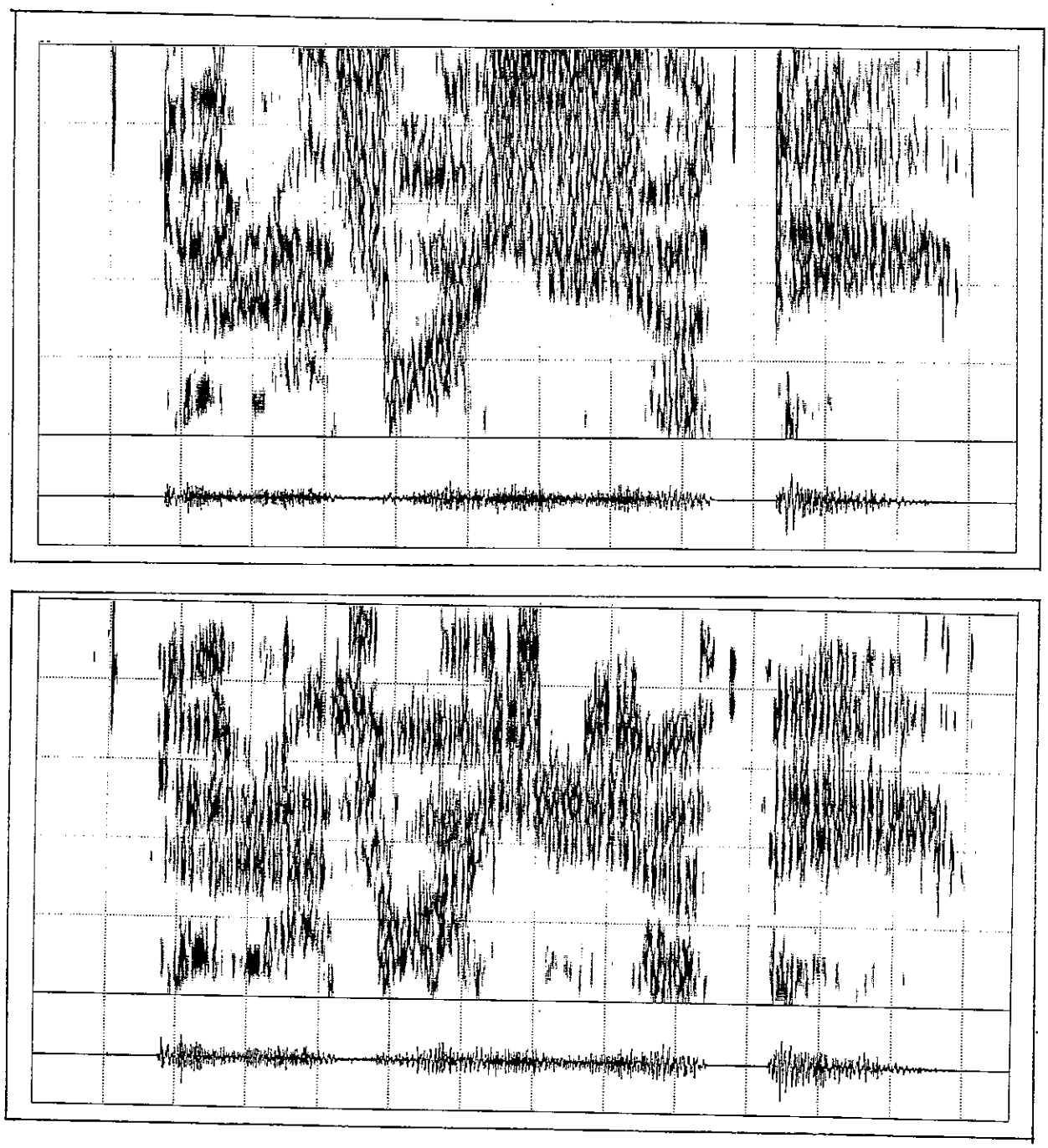

Figure 3.16: spectrogramme de "de la voix chuchotée", locuteur masculin, parole naturelle (haut), synthétique (bas).

3.6.3 visualisation dans le plan temps-fréquence

La visualisation du signal de parole dans le plan temps-fréquence reste un puissant moyen d'analyse. La visualisation directe des formes d'ondes dans ce domaine apporte un mode de représentation prometteur.

Pour la parole voisée, chaque cycle de voisement est bien visible dans chaque bande formantique. L'allure de la bande de base est apparente par ses composantes sinusoïdales. Les évolutions formantiques dues à la coarticulation apparaissent nettement, grâce à la modélisation spectrale initiale.

Pour les phénomènes temporels brefs, comme les explosions de plosives, la décomposition spectro-temporelle permet de visualiser l'évolution et la répartition de l'énergie dans le plan.

Pour un signal fricatif, les formes d'ondes sont regroupées spectralement autour des maxima de l'enveloppe spectrale. Par contre leur comportement temporel traduit leur nature aléatoire.

Le mélange de formes d'ondes quasi-périodiques et de formes d'ondes apparaissant aléatoirement peut se produire dans des régions fréquentielles différentes pour reconstituer par exemple des fricatives voisées (figures 3.17, 3.18).

3.6.4 modification du signal de parole

Les éléments sonores mis en évidence par la visualisation des formes d'ondes peuvent être manipulés de façon très simple. En effet chaque forme d'onde est synthétisée de façon indépendante, et peut être modifiée de façon indépendante.

Les phénomènes bien localisés dans le plan temps-fréquence peuvent ainsi être simplement modifiés, dans la mesure de leur représentation par des formes d'ondes.

Le modèle de production sous-jacent permet également de réaliser des transformations classiques:

- modification du fondamental: en multipliant par une constante ou une fonction les instants de référence. Les caractéristiques spectrales à court terme ne sont pas modifiées, mais ce procédé simple change la vitesse d'élocution.

- expansion-compression fréquentielle: en multipliant par une constante ou une fonction les fréquences centrales, les largeurs de bandes, les temps d'excitation. Le fondamental est inchangé, mais le timbre de la voix considérablement modifié.

- filtrage d'un domaine quelconque du plan: par multiplication par une constante ou une fonction des amplitudes.

- modifications du timbre: par multiplication par une constante ou une fonction des paramètres se rapportant au timbre: largeurs de bande, temps d'excitation, paramètres de la bande de base.

La réalisation d'un système complet de modification indépendante du fondamental et des durées, implique une génération ou une suppression de certaines formes d'ondes, [104].

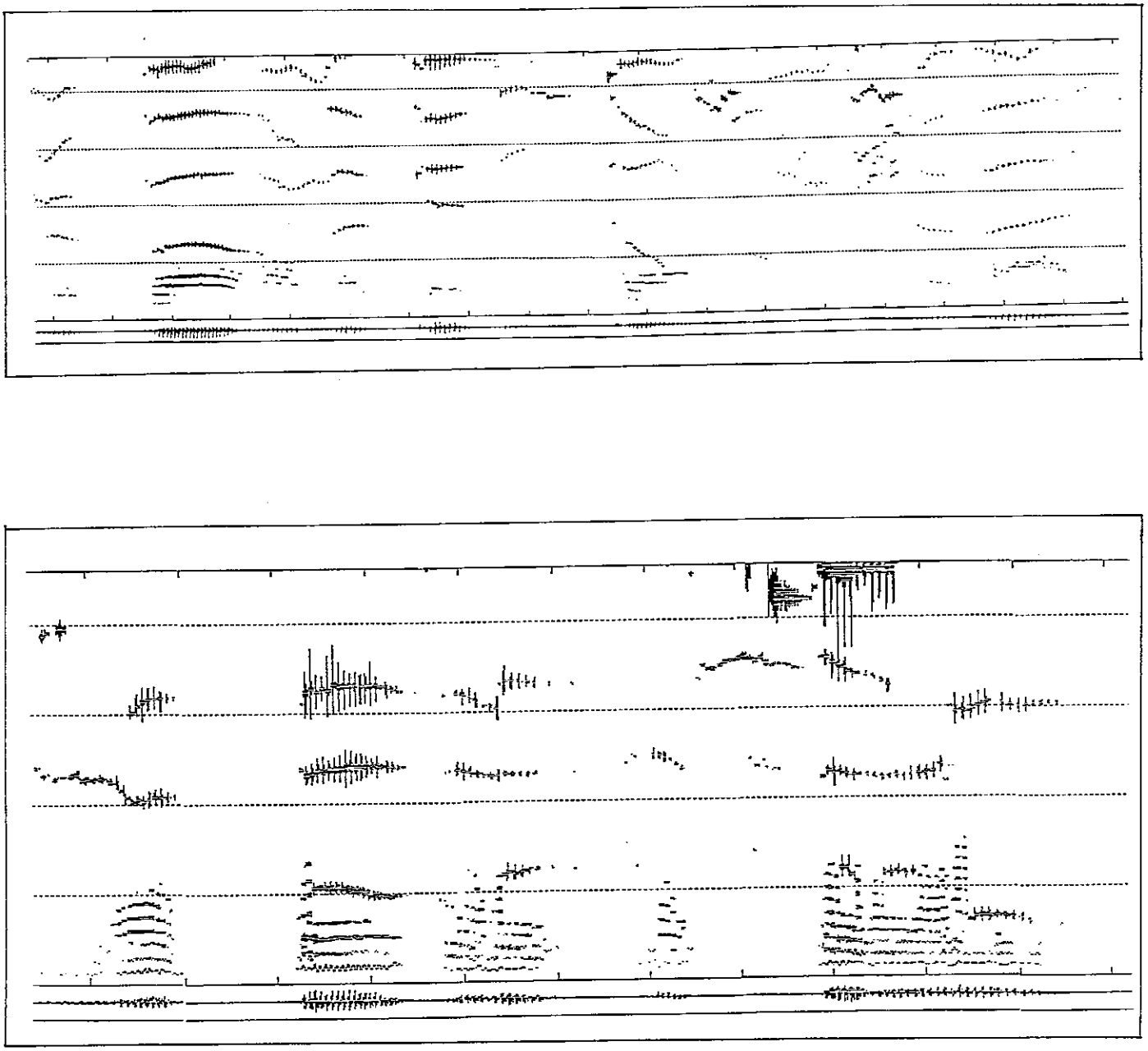

Figure 3.17: Affichage des formes d'ondes élémentaires dans le plan temps-fréquence: "Je vais en Afghanistan sur mon cheval", locuteur masculin (haut), et "this is the top of the chair", locuteur feminin (bas).

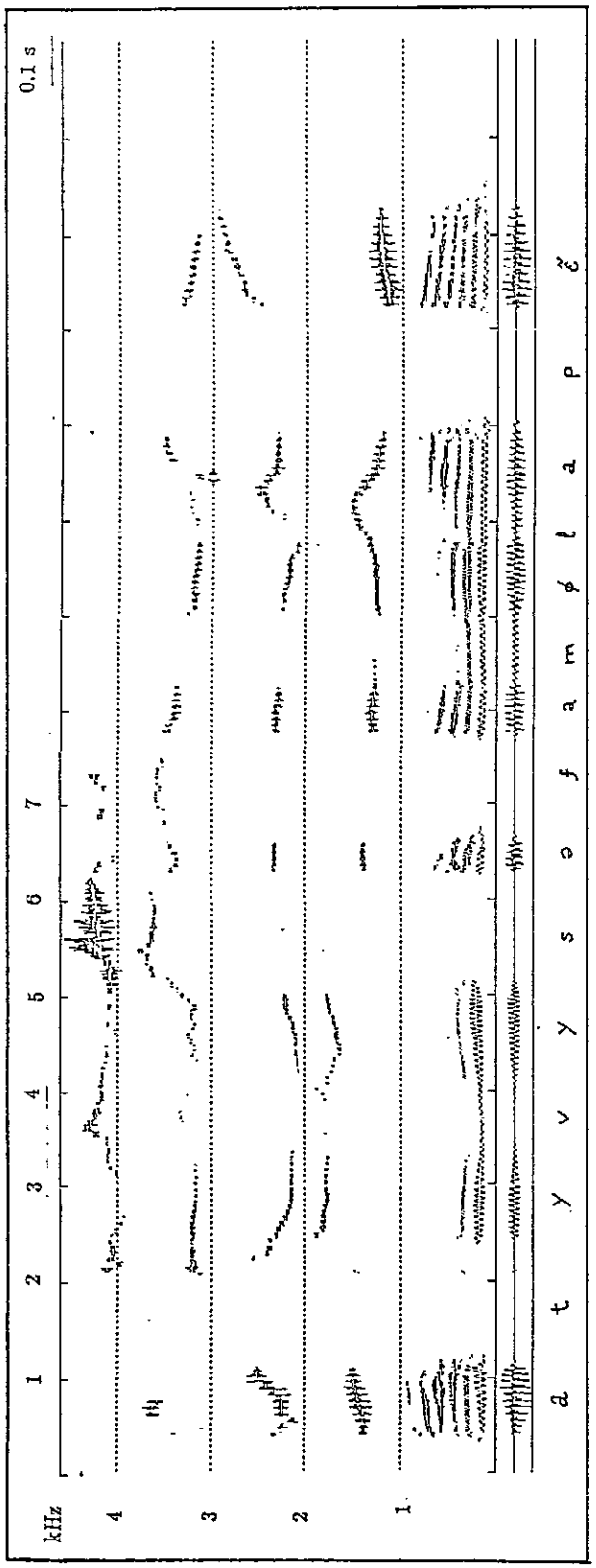

Figure 3.18: Affichage des formes d'ondes élémentaires dans le plan temps-fréquence: "as-tu vu ce fameux lapin", locuteur masculin

Le regroupement en formes d'ondes de l'information présente dans des évolutions rapides brèves et spectralement localisées du signal (explosion, cycle de voisement dans une région formantique, bruit de frication, ...) autorise de façon particulièrement simple leur modification, aussi localisée que le permettent les formes d'ondes. La perception semble très fine pour ce genre de phénomène, et la représentation en formes d'ondes bien adaptée par sa concision et sa précision: pas de notion fixe de trame ou de bande fréquentielle, mais une adaptation aux dimensions temporelles et spectrales.

3.7 discussion et conclusions

3.7.1 hypothèses

Le système de décomposition décrit possède plusieurs hypothèses sous-jacentes.

Les paramètres de chaque forme d'onde sont ici choisis constant pendant la durée d'une forme d'onde. En particulier la fréquence centrale d'un formant ne peut évoluer dans ce système au sein d'un cycle vocalique, bien que cela soit une situation fréquente. Cette remarque vaut également pour la largeur de bande.

La forme précise de l'enveloppe temporelle est excessivement simple par rapport à ce que l'on peut observer dans les signaux filtrés. Les rebonds explicitement introduits par la segmentation spectrale dans l'enveloppe temporelle ne peuvent pas être pris en compte s'il ne sont pas assez accusés pour donner naissance à une nouvelle forme d'onde.

La décroissance des réponses impulsionnelles doit être assez grande pour que le lissage de la courbe d'enveloppe temporelle n'empêche pas leur détection. En termes spectraux, les largeurs de bande doivent donc être assez importantes.

Au voisinage de la bande de base, le résultat de l'analyse est d'autant plus simple à interpréter que l'évolution spectrale du signal d'une trame à la suivante est régulière. Cette vitesse d'évolution n'entraîne pas un problème dans le calcul ou la qualité du résultat, mais une forme d'onde peut se décomposer en composantes sinusoïdales plutôt que garder son caractère formantique d'une trame sur l'autre.

Cette remarque peut s'étendre: la qualité de l'analyse est étroitement liée à son étape initiale de modélisation spectrale. La qualité de la modélisation, la précision des maxima spectraux, mais aussi la régularité de leurs évolutions restent à la base du procédé. Si le cas idéal des trajets formantiques bien dessinés n'est pas exigé ni nécessaire ici, plus l'on s'approche de cette situation, meilleurs sont les résultats (en terme de qualité et d'interprétation).

L'évolution des articulateurs du conduit vocal est supposée suffisamment lente pour considérer le signal comme stationnaire pendant la durée d'une forme d'onde. Cette durée est tout à fait variable, dans les limites de la bande fréquentielle d'analyse (au moins un cycle par forme d'onde).

3.7.2 critiques

robustesse

La confrontation du modèle préalable de signal à la réalité soulève des problèmes de robustesse.

L'erreur finale sur la sortie de la méthode est le cumul des erreurs effectuées à chaque étape. Le procédé présenté ici dépend en particulier de façon très étroite des résultats de la modélisation spectrale et du découpage spectral qui en résulte. Il est bien connu que la prédiction linéaire offre des résultats assez variables en fonction du locuteur: la parole d'un *bon* locuteur aura tendance à être considérablement mieux modélisée que celle d'un *mauvais locuteur*. Les critères de choix du locuteur ne sont pas formellement très définis, mais l'habitude de la synthèse permet de prévoir dans une certaine mesure l'aptitude d'une voix à la prédiction linéaire.

Cette variabilité des résultats de l'analyse prédictive en fonction du locuteur se propage dans toute l'analyse, et des résultats indiscutablement meilleurs sont obtenus pour certains.

La robustesse au bruit additionnel est faible, à cause de la modélisation spectrale *a priori*. Le bruit perturbe la détection des maxima spectraux, mais, ce qui est plus grave, le découpage et la modélisation temporelle par bande de fréquence sur une trame donnée a tendance à agglomérer le bruit présent dans la bande au signal. Un bruit blanc ajouté au signal sera transformé par l'analyse en bruit dans les régions formantiques. La simplicité de la modélisation temporelle se paie ici par une forte coloration du bruit additionnel. Ainsi, l'économie de modélisation dans les bandes formantiques (une forme d'onde représente une région spectro-temporelle bien plus grande que celle issue d'un échantillonnage comme ceux du premier chapitre) rend la méthode peu robuste à des perturbations fines et localisées comme un bruit de souffle.

Le regroupement suivant des régions spectrales privilégiées montre donc une tendance à "colorer" le signal analysé, en accord avec le modèle de production pour le signal de parole, et au détriment de la qualité de reproduction pour un bruit additionnel.

complexité de la méthode

La méthode proposée est coûteuse en quantité de calcul. N'étant pas entièrement portée sur du matériel rapide, la procédure d'analyse est donc lente. Cette lenteur est un inconvénient pour progresser dans l'étude de la méthode.

De même le système utilise des méthodes hétérogènes de calcul aux différentes étapes, et se prive ainsi de la beauté formelle d'une méthode unifiée, comme celles présentées au premier chapitre. Une méthode complexe et hétérogène reste beaucoup plus lourde, et difficile à interpréter. Aucune solution pour unifier le procédé de calcul des formes d'ondes n'est pour l'instant envisagée (envisageable?).

interprétation des résultats

Le résultat de l'analyse est un ensemble de formes d'ondes. Comme il a été évoqué au second chapitre, même si ce type de représentation peut sans doute constituer une base nouvelle pour l'étude de la perception du signal de parole, et même si l'utilisation en synthèse d'objets simples et puissants comme les formes d'ondes semble prometteur, la grande quantité de formes d'ondes représente une difficulté intrinsèque.

L'interprétation automatique des paramètres des formes d'onde en terme de production ne paraît pas tout à fait simple: retrouver le fondamental, les trajets formantiques, le voisement .... Visuellement évident, l'extraction de ces indices acoustiques présents

dans les formes d'ondes doit mettre en oeuvre des méthodes complexes de regroupement. Par rapport au second chapitre, un premier niveau de regroupement des formes d'ondes est cependant automatiquement réalisé, de façon préalable, par la modélisation spectrale.

En synthèse, le comportement complexe des formes d'ondes doit se concentrer en un ensemble concis de règles sur l'évolution de leurs paramètres. L'obtention automatique de ces règles reste un problème difficile.

3.7.3 conclusion

Le comportement du modèle classique linéaire de production du signal de parole a été étudié, dans le domaine temporel, dans l'optique d'une représentation en formes d'ondes.

Une revue des méthodes qui se rapportent à ce type de représentation a été présentée: modèles sinusoïdaux, à formants en parallèle, GMC et HMC, formant variant dans le temps.

Après une présentation de la synthèse par FOF, qui simule dans le domaine temporel un synthétiseur à formants en parallèle, des extensions de cette méthode pour la synthèse de parole ont été proposées. Il devient possible de synthétiser des segments non-vocaliques, et la bande de base fait l'objet d'un raffinement particulier (paramétrisation sinusoïdale).

Un système d'analyse-synthèse automatique basé sur les principes développés précédemment a été implémenté. Le signal synthétique obtenu est quasi-indiscernable de l'original, moyennant des précautions sur l'enregistrement.

Ce système permet de tester la validité de la représentation et ses limites, et les applications qui ont motivé ce travail, essentiellement pour la synthèse de parole et l'analyse du signal sont esquissées.

Bibliographie Chapitre 3

Bibliographie Chapitre 3

[1] C. d'Alessandro, X. Rodet 1987. *Fonctions d'ondes formantiques: extraction des paramètres et synthèse vocale* 16èmes Journées d'étude sur la parole de la Société Française d'Acoustique, Hammamet, Octobre 1987.

[2] C. d'Alessandro, J.S. Liénard 1988. *Decomposition of the speech signal into short-time waveforms using spectral segmentation* Proceedings of IEEE-ICASSP-88.

[3] C. d'Alessandro, 1988. *analyse synthèse de la bande de base par fonctions d'ondes élémentaires* 17èmes Journées d'étude sur la parole de la Société Française d'Acoustique, Nancy, Septembre 1988.

[4] C. d'Alessandro, X. Rodet, 1989. *Synthèse et analyse synthèse par fonctions d'ondes formantiques* Journal d'acoustique, Vol. 2, No. 2, Juin 1989.

[5] L. B. Almeida, F. M. Silva, 1984. *Variable-frequency synthesis: an improved harmonic coding scheme* Proceedings of IEEE-ICASSP-84.

[6] L. B. Almeida, J. M. Tribolet, 1983. *Nonstationary modeling of voiced speech* IEEE transactions on ASSP, Vol. ASSP-31, No. 3, Juin 1983.

[7] B. S. Atal, J. R. Remde, 1982. *A new model of LPC excitation for producing natural-sounding speech at low bit rates* Proceedings of IEEE-ICASSP-82.

[8] P. Badin, 1983. *Analyse de la parole- Synthèse à formants* Thèse de docteur-ingénieur, Institut National Polytechnique de Grenoble, 29 Mars 1983.

[9] M. Baumwolspiner, 1978. *Speech generation through waveform synthesis* Proceedings of IEEE-ICASSP-78.

[10] J. S. Bourgenot, C. Dechaux, 1975. *Codage de la parole à faible débit: le vocodeur CIPHON* Revue technique Thomson-CSF, Vol. 7, No. 4, Décembre 1975.

[11] E. Bruckert, M. Minow, W. Tetschner, 1983. *Three-tiered software and VLSI aid developmental system to read text aloud* Electronics, McGraw-Hill, 21 Avril 1983.

[12] G. Carayannis, C. Gueguen, 1976. *The factorial linear modelling: a Karhunen-Loève approach to speech analysis* Proceedings of IEEE-ICASSP-76.

[13] B. J. Carey, J. A. Howard, 1972. *A method for speech analysis by a wavefunction representation* IEEE-Conference record on speech communication and processing, 1972.

[14] F. Carton, 1974. *Introduction à la phonétique du Français* Bordas, Paris.

[15] F. Casacuberta, E. Vidal, 1987. *A nonstationary model for the analysis of transient speech signals* IEEE transactions on ASSP, Vol. ASSP-35, No. 2, Février 1987.

[16] F. J. Charpentier, M. G. Stella, 1986. *Diphone synthesis using an overlapp-add technique for speech waveforms concatenation* Proceedings of IEEE-ICASSP-86.

[17] L. Conturie, 1955. *Acoustique apliquée* Editions Eyrolles, Paris.

[18] F. Déchelle, 1984. *Programmation d'un algorithme de synthèse par fonctions d'ondes formantiques sur microprocesseur de traitement du signal TMS 320* Rapport de D.E.A. "traitement algorithmique de l'information", Université Paris VI, Septembre 1984.

[19] P. Delattre, 1965. *Comparing the phonetic features of English German Spanish and French* Julius Groos Verlag, Heidelberg.

[20] P. Depalle, 1984. *Analyse numérique des sons: codage par prédiction linéaire, extraction des formants* Rapport de D.E.A. d'acoustique appliquée, Université du Maine, Septembre 1984.

[21] G. De Poli, 1988. *Fore d'onda per la sintesi granulare sincrona* ATTI VII colloquio di informatica musicale, Rome, 23-26 Mars 1988.

[22] J. L. Duchet, 1981. *La phonologie* Presse Universitaires de France, Paris.

[23] L. Dolansky, 1960. *Choice of base signal in speech analysis* IRE transaction in Audio, Novembre-Décembre 1960.

[24] G. Duncan, M. A. Jack, 1986. *An improved formant tracking algorithm featuring adaptive pole enhancement* Proceedings of the institute of acoustics, Vol. 8, Part 71, Novembre 1986.

[25] G. Duncan, M. A. Jack, 1988. *Formant estimation algorithm based on pole focusing offering improved noise tolerance and feature resolution* IEE Proceedings, Vol. 135, No. 1, Février 1988.

[26] G. Fant, 1970. *Acoustic theory of speech production* Mouton, La Hague-Paris.

[27] G. Fant, J. Liljencrants, 1979. *Perception of vowels with truncated intraperiod decay envelopes* STL-QPSR 1/1979.

[28] G. Fant, 1979. *Glottal source and excitation analysis* STL-QPSR 1/1979.

[29] J. L. Flanagan, 1980. *Parametric coding of speech spectra* JASA, Vol. 68, No. 2, Aout 1980.

[30] H. Fujisaki, M. Ljungqvist, 1986. *Proposal of evaluation of models for the glottal source waveform* Proceedings of IEEE-ICASSP-86.

[31] H. Fujisaki, M. Ljungqvist, 1987. *Estimation of voice source and vocal tract parameters based on ARMA analysis and a model for the glottal source waveform* Proceedings of IEEE-ICASSP-87.

[32] D. H. Friedman, 1981. *Estimation of formant parameters by sum-of-poles modeling* Proceedings of IEEE-ICASSP 81.

[33] C. Gueguen, 1985. *Analyse de la parole par des méthodes de modélisation paramétrique* Annales de Télécomunications, Vol. 40, No 5-6, 1985.

[34] A. H. Gray, J. D. Markel, 1974. *A spectral-flatness measure for studying the autocorrelation method of linear prediction speech analysis* IEEE transactions on ASSP, Vol. ASSP-22, No. 3, Juin 1974.

[35] Y. Grenier, 1983. *Time-dependant ARMA modeling of nonstationary signals* IEEE transactions on ASSP, Vol. ASSP-31, No. 4, Avril 1983.

[36] D. W. Griffin, J. S. Lim, 1984. *Signal estimation from modified short-time Fourier transform* IEEE transactions on ASSP, Vol. ASSP-32, No. 2, Avril 1984.

[37] D. W. Griffin, J. S. Lim, 1985. *A new model-based speech analysis/synthesis system* Proceedings of IEEE-ICASSP-85.

[38] S. Grovel, J.S. Liénard, C. d'Alessandro, 1989. *Representation of the speech signal with elementary waveforms: a preliminary perceptual study* Proceedings of the 13th International Congress on Acoustics, Belgrade, Aout 1989.

[39] W. J. Hardcastle, 1976. *Physiology of speech production* Academic Press, Londre.

[40] P. Hedelin, 1981. *A tone-oriented voice-excited vocoder* Proceedings of IEEE-ICASSP-81.

[41] J. M. Heinz, K. N. Stevens, 1961. *On the properties of voiceless fricative consonants* JASA, Vol. 33, No. 5, Mai 1961.

[42] J. N. Holmes, 1975. *Low-frequency phase distortion of speech recordings* JASA, Vol. 58, No. 3, September 1975.

[43] G. W. Hugues, M. Halle, 1956. *Spectral properties of fricative consonants* JASA, Vol. 28, No. 2, Mars 1956.

[44] M. L. Honig, D. G. Messerschmitt, 1981. *Convergence properties of an adaptive digital lattice filter* IEEE Transaction on Circuit and system, Vol. CAS-28, No. 6, Juin 1981.

[45] D. Kewley-Port, 1983. *Time-varying features as correlate of place of articulation in stop consonants* JASA, Vol. 73, No. 1, Janvier 1983.

[46] D. Kewley-Port, D. B. Pisoni 1983. *Perception of static and dynamic acoustic cues to place of articulation in initial stop consonant* JASA, Vol. 73, No. 5, Mai 1983.

[47] D. H. Klatt, 1976. *Structure of a phonological rule component for a synthesis -by-rule program* IEEE transaction on ASSP, Vol. ASSP-24, No. 5, Octobre 1976.

[48] D. H. Klatt, 1980. *Software for a cascade/parallel formant synthesizer* JASA, Vol. 67, No. 3, Mars 1980.

[49] D. H. Klatt, 1987. *Review of text-to-speech conversion for English* JASA, Vol. 82, No. 3, Septembre 1987.

[50] M. Kunt, 1981. *Traitement numérique des signaux* Traité d'électricité, d'électronique et d'électrotechnique, vol. XX,Dunod, Paris.

[51] F. Laferrière, 1986. *Analyse synthèse et étude de règles acoustiques de production avec un synthétiseur à formants, avec application à l'étude des voyelles nasales du Français Montréalais* Thèse de Maitrise es science, Université du Québec, Mai 1986.

[52] J. Laroche, 1988. *Etude d'un système d'analyse/synthèse utilisant la méthode de Prony. Application aux instruments de musique de type percussif* Rapport interne I.R.C.A.M., Août 1988.

[53] J. Laroche, 1989. *A new analysis/synthesis system of musicals signals using Prony's method. Application to heavily damped percussive sounds* Proceedings of IEEE-ICASSP-89.

[54] Y. T. Lee, H. F. Silverman, 1986. *A model for nonstationary analysis of speech* Proceedings of IEEE-ICASSP-86.

[55] Y. T. Lee, 1987. *A general time-varying model for speech signals and estimation of its parameters* Thèse de Ph.D, Décembre 1987, publiée comme rapport technique LEMS-40, Laboratory for Engineering Man/Machine Systems, Brown University, Providence, Janvier 1988.

[56] Y. T. Lee, H. F. Silverman 1988. *On a general time-varying model for speech signals* Proceedings of IEEE-ICASSP-88.

[57] J. P. Lefevre, O. Passien, 1985. *Efficient algorithms for obtaining multipulse excitation for LPC coder* Proceedings of IEEE-ICASSP-85.

[58] I. Lehiste, éditeur, 1967. *Readings in acoustics phonetics* The M.I.T. Press, Boston.

[59] A. Leroi-Gourhan, 1964. *Le geste et la parole* Albin Michel, Paris.

[60] J. S. Liénard, F.K. Soong, 1984. *On the use of transient information in speech recognition* Proceedings of IEEE-ICASSP-84.

[61] M. G. Ljungqvist, 1986. *Speech analysis-synthesis based on modeling of voice source and vocal-tract characteristics* Thèse de PhD, Université de Tokyo, 20 décembre 1986.

[62] A. P. Lobo, W. A. Ainswort, 1988. *Variation of glottal pulse shape with fundamental frequency* 7th FASE Symposium, Edinbourg, Août 1988.

[63] R. J. McAulay, T. F. Quatieri, 1985. *Mid-rate coding based on a sinusoidal representation of speech* Proceedings of IEEE-ICASSP-85.

[64] R. J. McAulay, T. F. Quatieri, 1986. *Speech analysis/synthesis based on a sinusoidal representation* IEEE transaction on ASSP, Vol. ASSP-34, No. 4, Août 1986.

[65] R. J. McAulay, T. F. Quatieri, 1986. *Phase modelling and its application to sinusoidal transform coding* Proceedings of IEEE-ICASSP-86.

[66] R. J. McAulay, T. F. Quatieri, 1987. *Mixed-phase deconvolution of speech based on a sine-wave model* Proceedings of IEEE-ICASSP-87.

[67] S. S. McCandless, 1974. *An algorithm for automatic formant extraction using linear prediction spectra* IEEE transaction on ASSP, Vol. ASSP-22, No. 2, Avril 1974.

[68] J. I. Makhoul, L. K. Cosell, 1981. *Adaptative lattice analysis of speech* IEEE Transaction on Circuit and system, Vol. CAS-28, No. 6, Juin 1981.

[69] J. I. Makhoul, 1978. *A class of all-zero lattice digital filters: properties and application* IEEE Transaction on ASSP, Vol. ASSP-26, No. 4, Aout 1978.

[70] S. S. McCandless, 1974. *An algorithm for automatic formant extraction using linear prediction spectra* IEEE transaction on ASSP, Vol. ASSP-22, No. 2, Avril 1986.

[71] B. Malmberg, 1954. *La phonétique* Presses Universitaires de France, Paris.

[72] B. Malmberg, 1974. *Manuel de phonétique générale* Editions Picard, Paris.

[73] J. D. Markel, 1970. *On the interrelationships between a wave function representation and a formant model of speech* Thèse de PhD, University of California, Santa Barbara, Juillet 1970.

[74] J. D. Markel, A. H. Gray, 1976. *Linear prediction of speech* Springer-Verlag, Berlin.

[75] J. S. Marques, L. B. Almeida, 1987. *Quasi-optimal analysis for sinusoidal representation of speech* Onzième colloque GRETSI, Nice, Juin 1987.

[76] R. L. Miller, 1959. *Nature of the Vocal Cord Wave* JASA, Vol. 31, No. 6, Juin 1959.

[77] J. A. Moorer, 1977. *Signal processing aspects of computer music: a survey* Proceedings of the IEEE, Vol. 65, No. 8, Août 1977.

[78] G. Murillo Manzano, 1984. *Application de l'analyse par la synthèse à la génération des consonnes occlusives voisées du Français* Thèse de docteur-ingénieur, Institut National Polytechnique de Grenoble, 21 Juin 1984.

[79] S. E. Öhman, 1967. *Numerical model of coarticulation* JASA, Vol. 41, No. 2, 1967.

[80] A. V. Oppenheim, 1968. *Speech analysis-synthesis system based on homomorphic filtering* JASA, Vol. 45, No. 2, 1969.

[81] L. L. Pfeiffer, 1972. *The application of wavefunction analysis to single-speaker phoneme recognition* Thèse de PhD, University of California, Santa Barbara, Marst 1972.

[82] L. L. Pfeiffer, 1972. *Isolated-word phoneme recognition using features derived from wavefunction parameters* IEEE-Conference record on speech communication and processing, 1972.

[83] E. N. Pinson, 1963. *Pitch-synchronous time-domain estimation of formant frequencies and bandwidths* JASA, Vol. 35, Part 8, 1963.

[84] T. F. Quatieri, 1979. *Minimum and mixed phase speech analysis-synthesis by adaptive homomorphic deconvolution* IEEE transaction on ASSP, Vol. ASSP-27, No. 4, Aout 1979.

[85] T. F. Quatieri, R. J. McAulay, 1986. *Speech transformations based on a sinusoidal representation* IEEE transaction on ASSP, Vol. ASSP-34, No. 6, Décembre 1986.

[86] L. R. Rabiner, 1968. *Digital-formant synthesizer for speech synthesis* JASA, Vol. 43, 1968.

[87] L. R. Rabiner, R. W. Schafer, C. M. Rader, 1969. *The chirp z-transform algorithm and its application* Bell System technical journal, Vol. 48, Mai 1969.

[88] L. R. Rabiner, R. W. Schafer, C. M. Rader, 1969. *The chirp z-transform algorithm* IEEE transaction on Audio Electroacoustics, Vol. AU-17, Juin 1969.

[89] M. H. Razzam, 1979. *Modélisation mathématique, lissage et filtrage digital, d'un système de synthèse de parole* Thèse de troisième cycle, Université Paris XI, Juin 1979.

[90] J. C. Risset, 1967. *Sur l'analyse, la synthèse et la perception des sons, étudiées à l'aide de calculateurs électroniques* Thèse d'état, Université Paris XI, 29 Mai 1967.

[91] X. Rodet, C. Santamarina, 1975. *Synthèse, sur un micro-ordinateur, du signal vocal dans sa représentation amplitude-temps* 6èmes Journées d'étude sur la parole du Groupement des Acousticiens de Langue Française, Toulouse, Mai 1975.

[92] X. Rodet, 1977. *Analyse du signal vocal dans sa représentation amplitude-temps, synthèse de la parole par règles* Thèse d'état, Université Paris VI, Juin 1977.

[93] X. Rodet, J. L. Delatre, M. Razzam, 1979. *Construction du signal vocal dans le domaine temporel* 10èmes Journées d'étude sur la parole du Groupement des Acousticiens de Langue Française, Grenoble, Juin 1979.

[94] X. Rodet, J. L. Delatre, 1979. *Time domain speech synthesis by rules using a flexible and fast signal management system* Proceedings of IEEE-ICASSP-79.

[95] X. Rodet, 1980. *Time-domain formant-wave-function synthesis* Spoken language generation and understanding, J. C. Simon éditeur, D. Reidel publishing company, Dordrecht, Hollande.

[96] X. Rodet, P. Depalle, 1985. *Synthesis by rules: LPC diphones and calculation of formants trajectories* Proceedings of IEEE-ICASSP-85.

[97] X. Rodet, P. Depalle, G. Poirot, 1987. *Analyse et synthèse de la voix parlée et chantée par modélisation de l'enveloppe spectrale et de l'excitation* 16èmes Journées d'étude sur la parole de la Société Française d'Acoustique, Hammamet, Octobre 1987.

[98] X. Rodet, P. Depalle, G. Poirot, 1987. *Speech analysis and synthesis methods based on spectral envelopes and voiced/unvoiced functions* European Conference on Speech Technology, Edimbourg, Septembre 1987.

[99] A. E. Rosenberg,1971. *Effect of Glottal Pulse Shape on the Quality of Natural Vowels* JASA, Vol. 49, No. 2, 1971.

[100] S. Sagayama, F. Itakura, 1986. *Duality theory of composite sinusoidal modeling and linear prediction* Proceedings of IEEE-ICASSP-86.

[101] S. Saito, K. Nakata, 1985. *Fundamentals of speech signal processing* Academic Press, Tokyo.

[102] F. de Saussure, 1915. *Cours de linguistique générale* Payot, Paris, 1972.

[103] M. R. Schroeder, B. S. Atal, 1985. *Code-excited linear prediction (CELP): high-quality speech at very low bit rates* Proceedings of IEEE-ICASSP-85.

[104] S.Seneff, 1982. *System to independently modify excitation and/or spectrum of speech waveform without explicit pitch extraction* IEEE transaction on ASSP, Vol. ASSP-30, No. 4, Août 1982.

[105] S. D. Soli, 1981. *Second formant in fricatives: acoustic consequences of fricative-vowel coarticulation* JASA, Vol. 70, No. 4, Octobre 1981.

[106] S. D. Soli, 1981. *Second formant in fricatives: acoustic consequences of fricative-vowel coarticulation* JASA, Vol. 70, No. 4, Octobre 1981.

[107] M. M. Sondhi, 1975. *Measurement of the glottal waveform* JASA, Vol. 57, Janvier 1975.

[108] M. Stella, F. Charpentier, 1985. *Synthèse par diphones utilisant le codage prédictif multiimpulsionnel et un vocodeur de phase* 14èmes Journées d'étude sur la parole du Groupement des Acousticiens de Langue Française, 1985.

[109] K. N. Stevens, 1950. *Autocorrelation analysis of speech sounds* JASA, Vol. 22, No 6, Novembre 1950.

[110] K. N. Stevens, 1971. *Airflow and turbulence noise for fricative and stop consonants: static consideration* JASA, Vol. 50, Part 2, 1971.

[111] I.M. Trancoso, L. B. Almeida, J. M. Tribolet, 1985. *Pole-zero multipulse speech representation using harmonic modelling in the frequency domain* Proceedings of IEEE-ICASSP-85.

[112] I.M. Trancoso, L. B. Almeida, J. S. Rodrigues, J. S. Marques, J. M. Tribolet, 1988. *Harmonic coding- state of the art and future trends* Speech Communication, Vol. 7, No. 2, Juillet 1988.

[113] L. F. Willems, 1986. *Robust formant analysis* IPO annual progress report, No. 21, 1986.

[114] B. Widrow, S. D. Stearns 1985. *Adaptative signal processing* Prentice Hall, Englewood Cliffs, New-Jersey.

[115] B. Yegnanaryana, G. Duncan, 1988. *Formant extraction from group delay spectra: a novel approach with high resolution properties* Colloque "Speech Processing" de l'IEE, Janvier 1988.

[116] G. H. Yeni-Komshian, S. D. Soli, 1981. *Recognition of vowels from information in fricatives: perceptual evidence of fricatives-vowels coarticulation* JASA, Vol. 70, No. 4, Octobre 1981.

PERSPECTIVES

Le travail présenté s'est attaché à explorer la possibilité et l'intérêt de représenter le signal de parole par une somme de fonctions élémentaires. Cette représentation, par un ensemble discret d'événements spectro-temporellement localisés s'éloigne conceptuellement des méthodes habituelles de représentation, bien que certaines d'entre elles trouvent une interprétation naturelle dans ce cadre. La motivation première de ce travail reste la synthèse de parole, mais la représentation par formes d'ondes élémentaires a également été étudiée dans la direction des méthodes non-paramétriques de représentation et dans celle de l'analogie avec des modèles fonctionels du système auditif.

En premier lieu, les méthodes classiques non-paramétriques ont recues une interprétation du point de vue des fonctions élémentaires d'analyse, dans le domaine temporel. La transformée de Fourier et la transformée en ondelettes apparaissent comme une représentation du signal comme une somme de fonctions élémentaires particulières. En discrétisant le réseaux de points d'analyse dans le plan d'information (temps/fréquence, temps/échelle), une représentation du signal comme somme discrète de fonctions élémentaires est obtenue. La reconstruction du signal avec une précision voulue est possible, et les fonctions élémentaires d'analyse peuvent être choisies orthogonales. Des algorithmes efficaces en terme de quantité de calcul peuvent se déduire de méthodes de codage (filtres miroirs en quadrature) pour obtenir une décomposition/reconstruction en ondelettes orthogonales. Utiles en codage, ces représentations ne permettent pas directement de retrouver les phénomènes de production ou de mettre en lumière les événements perceptifs. Des modèles de production ou de perception doivent donc être joints à ce type de méthode pour interpréter, modifier ou générer le signal de parole. Par contre les contraintes théoriques pour une représentation exacte du signal sont clairement énoncées en terme de formes d'ondes élémentaires.

L'étude des relations entre représentation en formes d'ondes élémentaire et analyse du signal par des modèles fonctionnels du système auditif périphérique autorise quelques analogies. Une première étape de filtrage s'inspire de données psychoacoustiques, puis une détection des événements spectro-temporels importants, qui représentent le signal vu par un grand nombre de canaux simultanément, permettent une analyse du signal bien plus proche des données de la perception que l'analyse classique par trames temporelles successives. Une contradiction apparaît entre les dimensions spectro-temporelles dictées par l'étude du premier chapitre et celle choisie d'après des contraintes perceptives, induisant une perte de qualité lors de la resynthèse. La liaison entre les connaissances déduites de la perception et des méthodes non-paramétriques semble un sujet d'étude fécond. Le signal est ainsi appréhendé comme un ensemble discret d'événements

spectro-temporels obtenus par des procédés perceptivement justifiés. L'étude des regroupements de ces objets primitifs est un enjeu neuf et prometteur. Des recherches sur les processus mentaux de groupements des événements acoustiques, en perception de la parole, de la musique ou en psychoacoustique montrent l'intérêt potentiel de modes complexes de décompositions du signal acoustique dans un plan d'information. La manipulation de ce type d'objets et de processus implique leur représentation par des techniques informatiques sophistiquées: programation orientée objet, réseaux connexionistes par exemple.

La synthèse par formes d'ondes élémentaired est l'objet principal de ce mémoire. En s'appuyant sur les formes d'ondes formantiques, un modèle de production du signal de parole par formes d'ondes élémentaires peut être élaboré. Un traitement particulier des segments non-voisés, ou la possibilité d'introduire du bruit dans les segments voisés par une méthode homogène sont proposés. La bande de base, ou région grave du spectre nécessite un traitement particulier, paramétrisation sinusoïdale par formes d'ondes élémentaires. Ce modèle est proche de la synthèse à formants en parallèle, mais en se plaçant directement dans le domaine acoustique, sans distinction entre source de voisement et conduit vocal, par exemple. De plus les objet de base de la synthèse offre une indépendance et une puissance de traitement accrues, en particulier pour les phénomènes spectro-temporellement localisés. Pour valider le synthétiseur proposé, un système automatique d'analyse/synthèse a été construit. Il effectue une décomposition du signal de parole en formes d'ondes élémentaires basée sur un modèle spectral du signal de parole. Ce système délivre un signal synthétique perceptivement quasi-indiscernable de l'original. Néanmoins des problèmes de robustesse subsitent: ici encore le choix des formes d'ondes élémentaires est en contradiction avec ceux dictés par le premier chapitre. L'étude de la constitution comme ensemble de formes d'ondes élémentaires, obtenues automatiquement, des événements de production offre une approche pleine de promesses pour la synthèse du signal, tant par règles que par éléments concaténés. La possibilité d'une extraction automatique des paramètres entraîne celle de modifier le signal, de façon particulièrement fine, dans le plan temps fréquence. L'utilisation de ce type de synthétiseur dans un sytème de synthèse à partir du texte s'inscrit directement dans la perspective de ce travail.

CINQ ARTICLES

Cinq articles sont joints en annexe à ce mémoire:

- C. d'Alessandro & X. Rodet, 1989. *Synthèse et analyse synthèse par fonctions d'ondes formantiques* Journal d'acoustique, Vol. 2, No. 2, Juin 1989.
- C. d'Alessandro & J.S. Liénard 1988. *Decomposition of the speech signal into short-time waveforms using spectral segmentation* Proceedings of IEEE-ICASSP-88.
- C. d'Alessandro, 1988. *analyse synthèse de la bande de base par fonctions d'ondes élémentaires* 17ème Journées d'étude sur la parole de la Société Française d'Acoustique, Nancy, Septembre 1988.
- J. S. Liénard & C. d'Alessandro, 1989. *Wavelet transform and granular analysis of speech* in *Wavelets, Time-frequency methods and phase space*, J.M. Combes, A. Grossman et P. Tchamitchian editeurs, Springer-Verlag, Berlin, 1989.
- C. d'Alessandro, 1989. *Time-frequency modifications using an elementary waveform speech model* Proceedings of ESCA-EuroSpeech 89, Paris, Octobre 1989.

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].

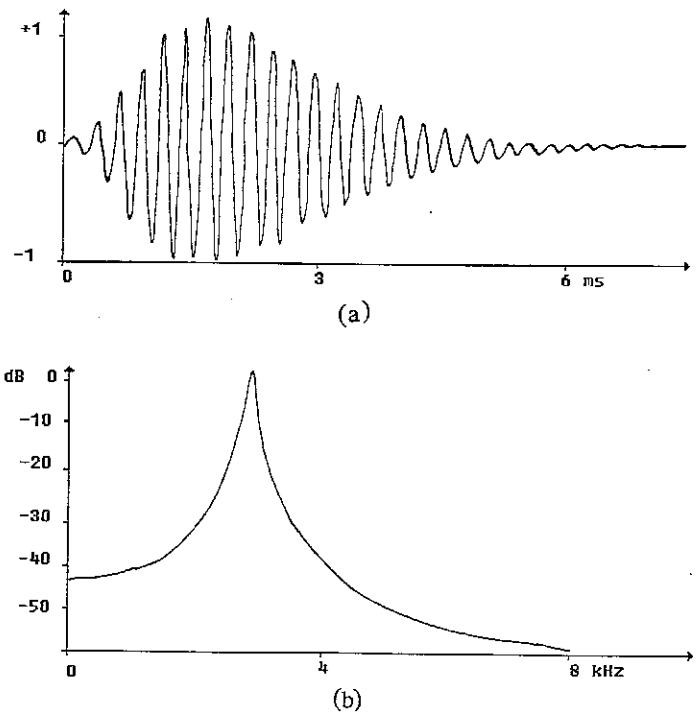

Fig. 1. — (a) Exemple de FOF : le signal temporel. (b) Exemple de FOF : le spectre d'amplitude logarithmique.

[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).

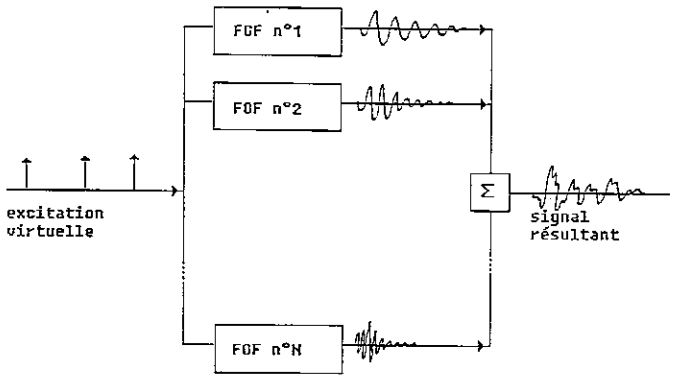

Fig. 2. — Structure d'un synthétiseur FOF parallèle.

[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

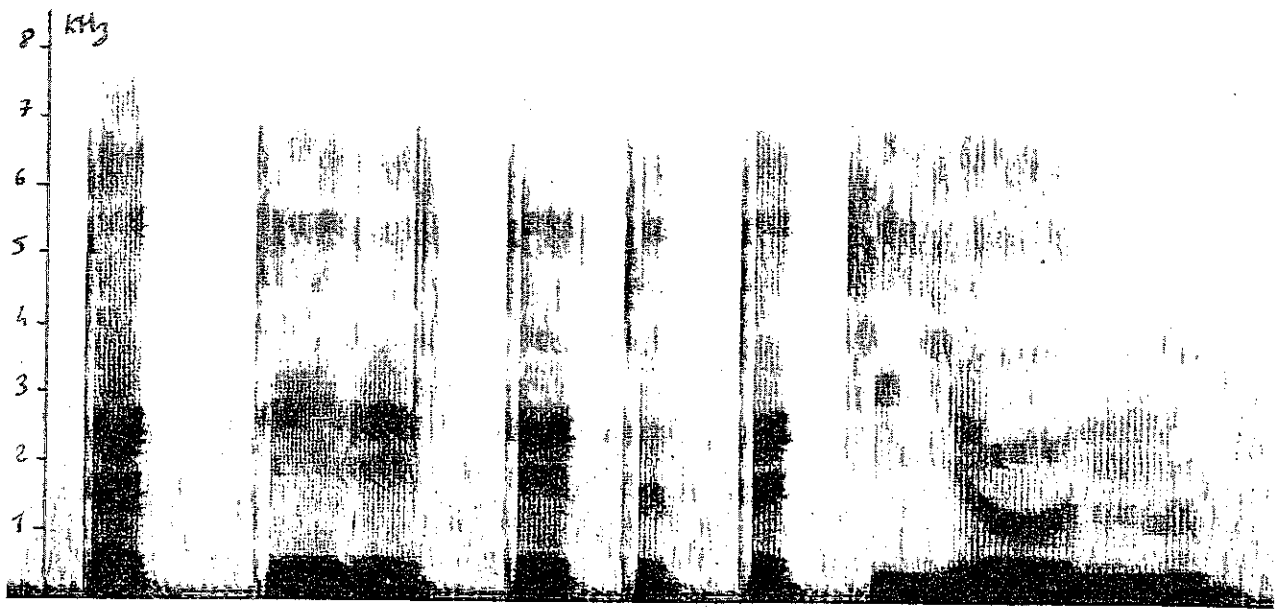

Fig. 3. — Sonagramme de la phrase « Tâter les têtes tatillonnes » parole naturelle.

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

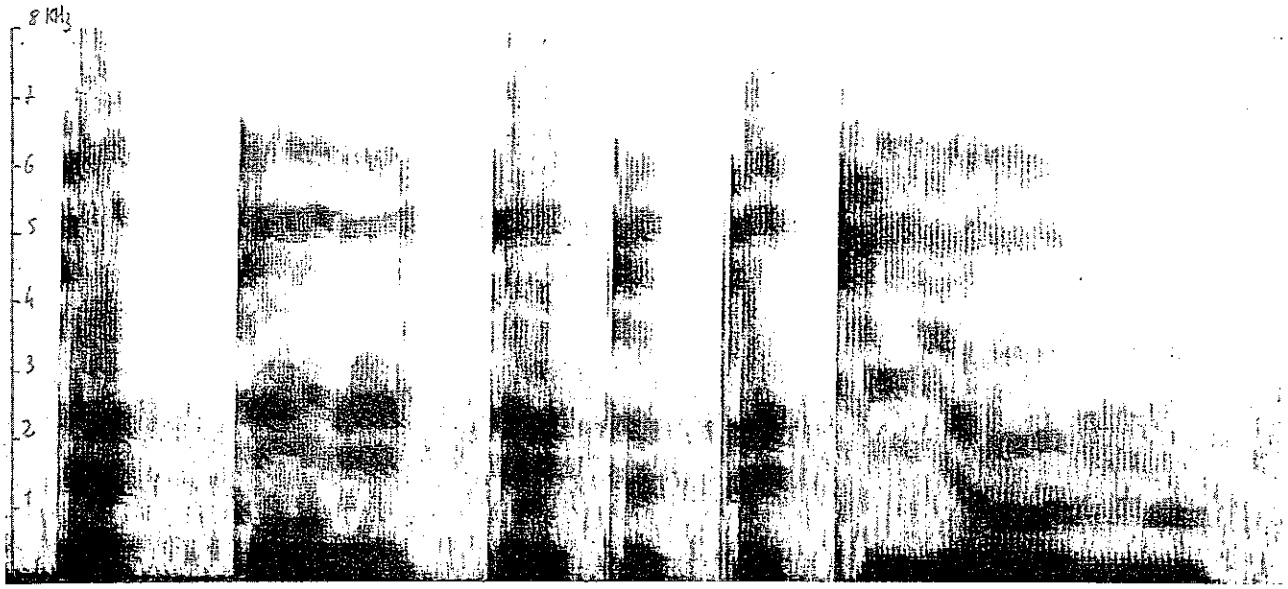

Fig. 4. — Idem : synthèse par FOF.

[Idem Formant-Wave-Function synthesis.]

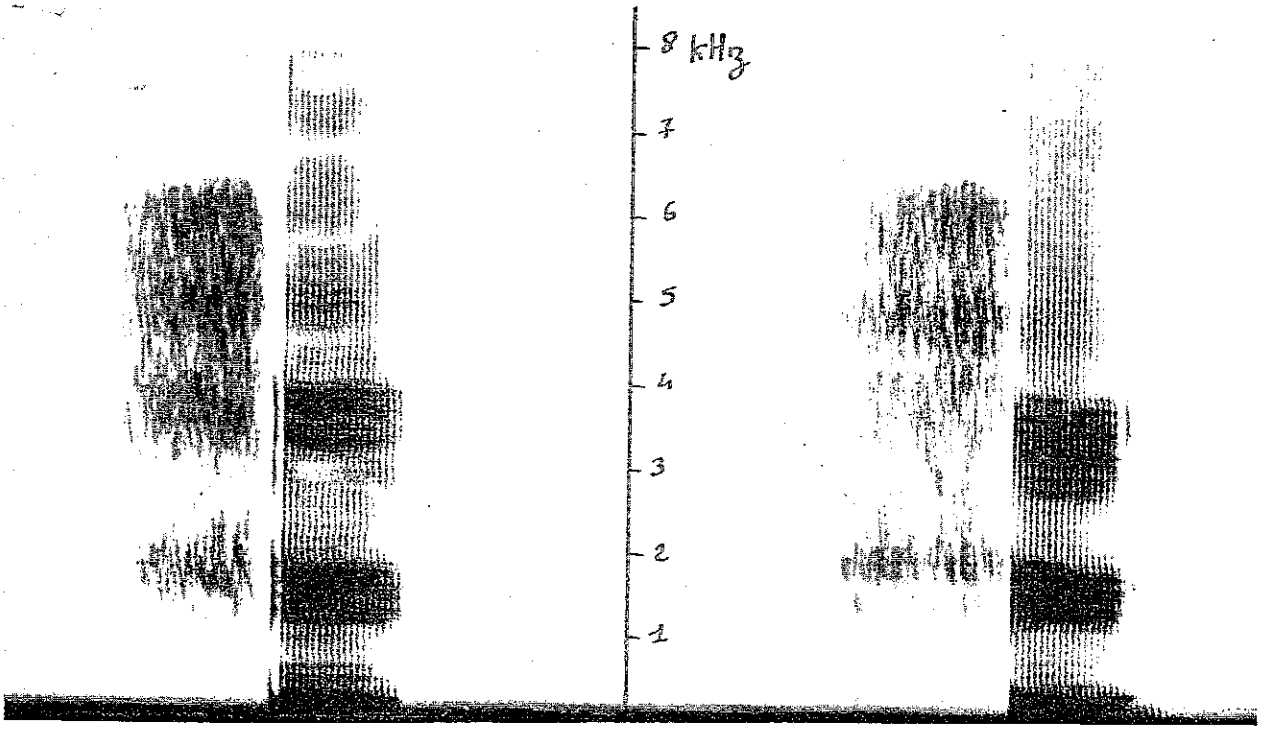

Fig. 5. — Sonagramme de la syllabe /fy/ naturelle puis synthétique (estimation manuelle des paramètres à partir du son naturel).

[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.

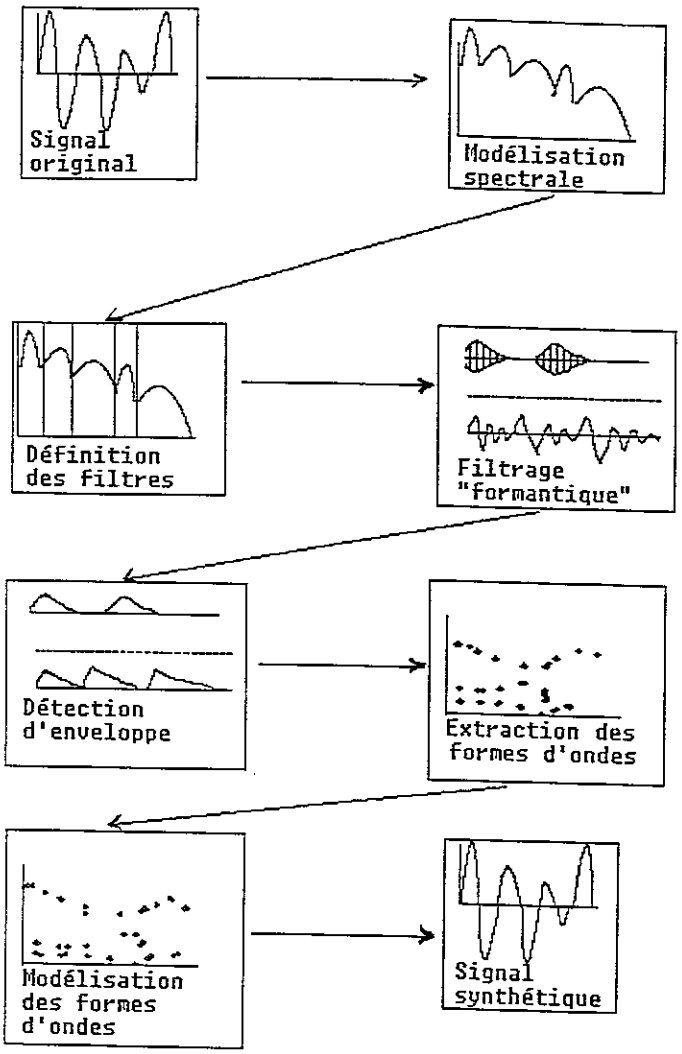

Fig. 6. — Processus d'analyse-synthèse.

[Analysis-synthesis process.]

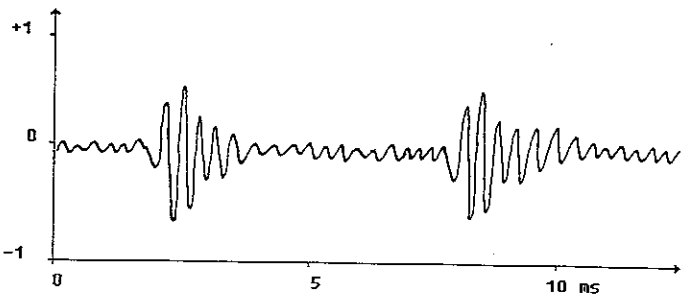

Fig. 7. — Signal issu d'un filtrage dans la région formantique.

[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.

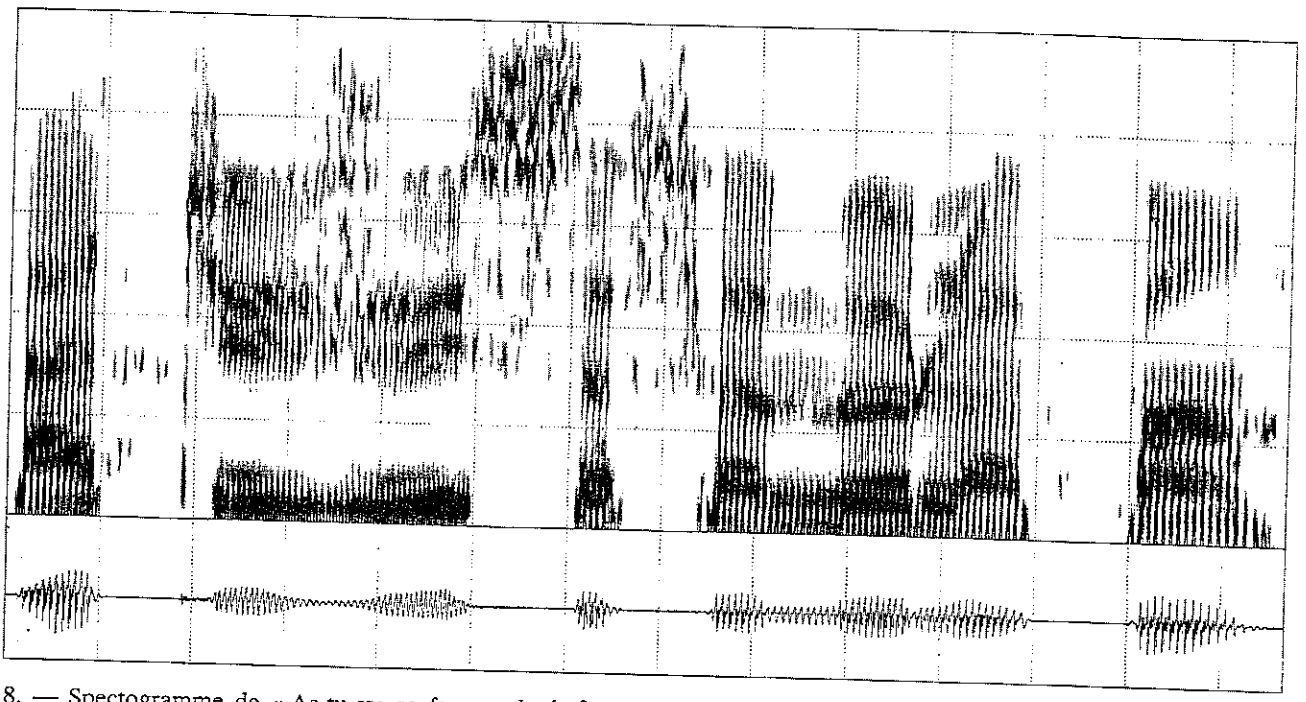

Fig. 8. — Spectogramme de « As-tu vu ce fameux lapin ? » parole naturelle.

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

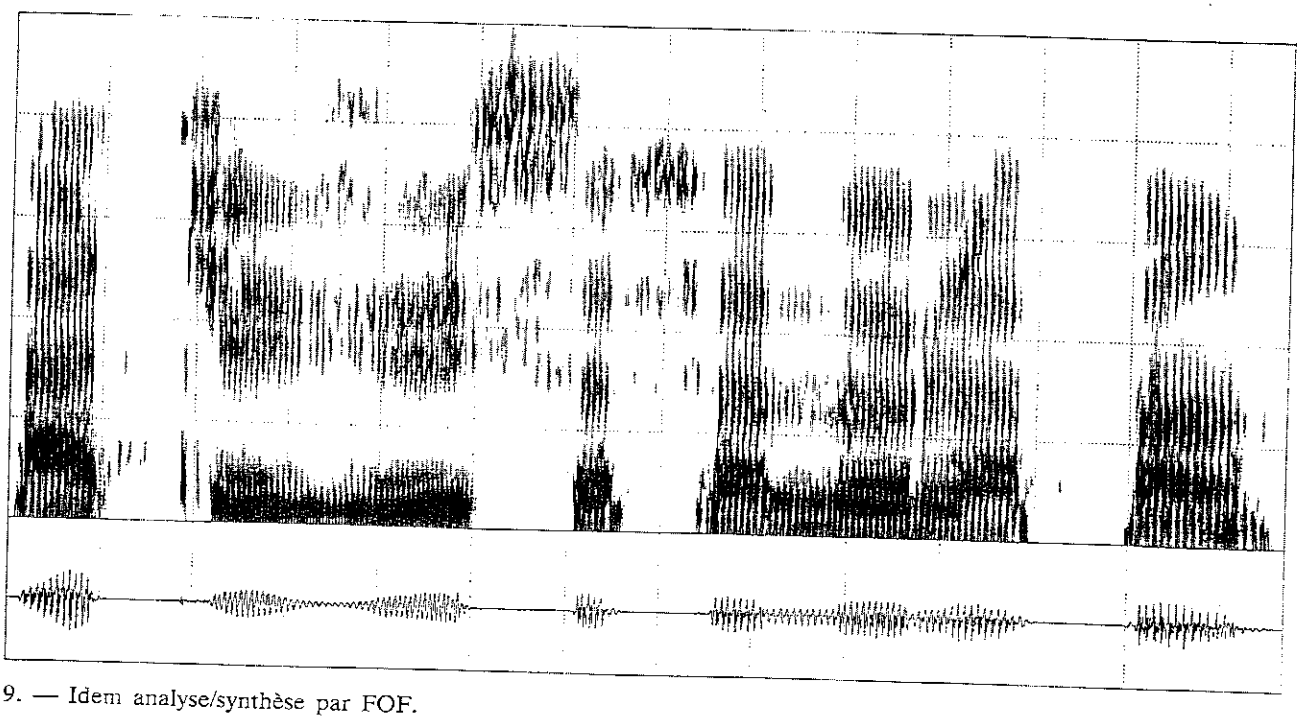

Fig. 9. — Idem analyse/synthèse par FOF.

[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).

S8.13

DECOMPOSITION OF THE SPEECH SIGNAL INTO SHORT-TIME WAVEFORMS USING SPECTRAL SEGMENTATION

Christophe d'Alessandro and Jean-Sylvain Liénard

LIMSI-CNRS, Orsay, France

ABSTRACT

Speech representation by a set of short-time, elementary waveforms appears as a new approach to speech processing. In the present study, the signal is pre-analysed frame by frame; the spectral envelope obtained for each frame is segmented into regions comprising a single peak. The signal is then filtered in each region, and the elementary waveforms are spotted in the time domain. The problem of grouping the waveforms in adjacent channels is thus circumvented. The resulting representation is satisfactory, as well as the signal reconstruction, except for some modelling problems remaining in the lowest part of the spectrum.

I - Introduction

This paper continues our work (ref 1) on a representation of the speech signal by a set of discrete elements which respect its acoustical and perceptive structures.

First, the temporal resolution of the analysis is given more importance than in traditional analyses. This option is shared by a recently developped analysis method (wavelet analysis, ref 2), which, after GABOR's work, decomposes the signal into well-localized time-frequency energy concentrations.

Second, we want to define elements within the signal ("grains", or elementary waveforms wfs) that contain all of the perceptual information, without, at this level, defining voicing or pitch explicitely. The concept of elementary waveform is close to X.RODET's FOF (Formant Wave Function), which has been used successfully for high-quality synthesis of singing voices (ref 3).

The "granular spectrogram" that results from this analysis will later be used to look for the classic elements of perception acoustics : voicing, pitch, formants, bursts etc. In order to make sure of the analysis relevance, we use to validate the decomposition by resynthesis. At the present stage of our study, we do not try to turn it into a coding method.

In the process presented in ref 1, the short-term spectral envelope of the speech signal was obtained using a zero-phase filterbank. The grains were defined through channel-by-channel modelling. After resynthesis, the quality obtained was excellent, but the representation was still redundant. Local grouping of adjacent channels yielded the desired representation; however some quality problems were encountered in the lower part of the spectrum during modelling and resynthesis.

We present here a somewhat different method based on spectral segmentation before temporal modelling. In this manner, we try to profit from a specificity of the speech signal, which has a spectrum composed of peaks : the interval between two valleys corresponds to the number of adjacent channels that were to be grouped in the former processing.

In section II we explain the principle of the analysis-synthesis model. In section III the system developed according to this principle is described. Some results are presented in section IV.

II - Overview and Production Model

In the traditional approaches (FFT, LPC, Cepstrum etc), one uses successive analysis windows, regularly spaced, of fixed length (often long enough to include several pitch periods). This way to perform the analysis tends to separate in the first place two different aspects of the signal: pitch (and the parameters related to the excitation), and spectral envelope. The excitation signal is then modelled, either in a rather crude way (the binary feature voiced/unvoiced is used in many systems), or more finely (for example, a sequence of pulses in the multipulse methods), but with a serious drawback, as far as interpretation capability is sought: there are no relations between the secondary pulses and the acoustic or perceptive structures.

CH2561-9/88/0000-0351 $1.00 © 1988 IEEE

As we wish to work out a method which provides a description of the signal preserving or emphasizing those structures, we will try to model the source and the spectral envelope in both spectral and time domains. The whole processing is summarized in fig 1.

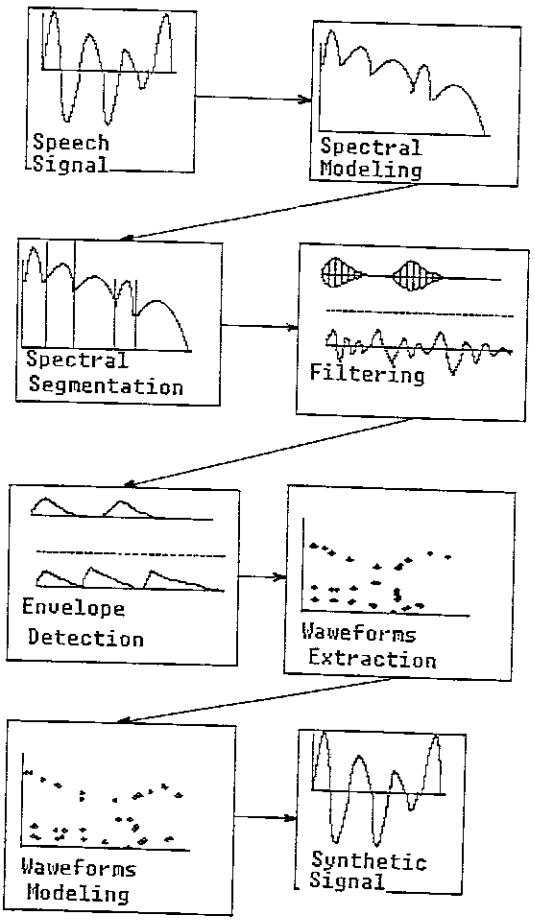

Fig 1 - The Analysis and Reconstruction Process.

The signal is first pre-analyzed, frame by frame, using a classical LPC algorithm; it is therefore modelled through an all-pole model. For each frame, the spectral envelope is segmented into regions, each containing one single envelope maximum. The signal is then filtered in each region, the elementary waveforms wfs are spotted between two successive minima of the time envelope, their parameters are evaluated (wf modelling), and reconstruction is achieved by summing the appropriate set of waveform models wfms.

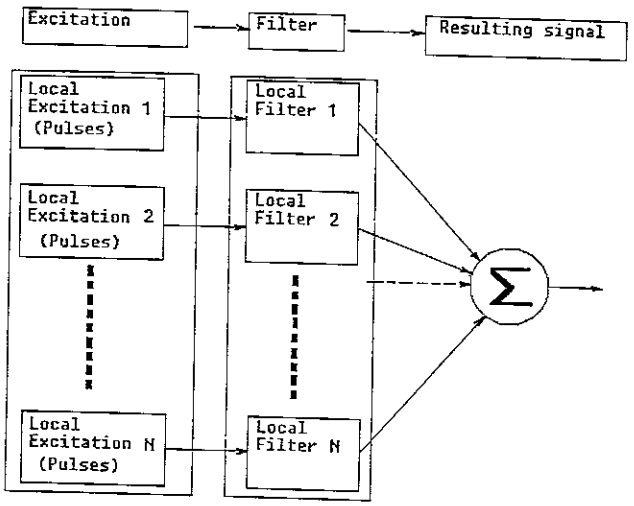

Fig 2 - The production model

An elementary waveform wf can be considered as the response of a spectrally local filter to a spectrally local excitation (fig 2). Under some conditions explained below, this local source can be modelled by a set of pulses, and the wfs can be identified with the impulse responses of the local filters - in the formant regions. Each "grain", or wf, represents a spectro-temporal event. The process can be formulated as follows.

$$s(t) = \sum_{i=0}^{N} \sum_{j=0}^{N_j} \alpha_{ij} \, (\delta_{ij} * p_i * f)(t)$$

s(t): speech signal
N : number of formants
N_j : number of pulses
α_{ij}: amplitude
p_i : impulse response of the spectral segmentation filter
δ_{ij} : unit impulse at instant τ_{ij}
f : impulse response of filter F given by the production model
* : convolution operator
i : index of formant bands: $1 \leq i \leq N$
j : index of the pulses in band i: $0 \leq j < N_i$

It should be mentioned that the model presented includes the classical source model, as well as the multipulse model. If a pulse appears exactly at the same time in all analysis bands, it provides a single pulse exciting the global filter F. We thus endorse the hypothesis - successfully used in multipulse analysis - that all types of excitation can be viewed as sequences of pulses. We just consider here a sequence of pulses to be localized in each formant region, in order to give a better account of its acoustical and perceptual contribution.

III - Experiments

Our method of analysis-synthesis allows for detection, modelling, and resynthesis in the time domain, of a set of elementary waveforms which will be defined below. The analysis proceeds as follows:

1) spectral modelling

Pre-analysis is done through the Adaptive Lattice algorithm described in ref 4. Successive frames are considered, every 6 ms. The effective length of the time window applied onto the signal is not explicit, but it can be evaluated to about 15 ms. Several maxima, here called "formants" for the sake of simplicity, are usually apparent in each frame.

2) Segmenting the spectral envelope

The formant regions are simply defined between two successive minima of the spectral envelope.

3) filtering in the formant region

The original signal is filtered by short-term Fourier analysis-synthesis (ref 5) in each of the formant regions previously defined. For each frame, N partial signals are therefore obtained; the sum of all of these is equal to the original signal. The chosen filters (rectangular or triangular shape of the transfer function) do not introduce any phase distorsion. The filtering itself is done on a long segment of the original signal (50 ms), surrounding the frame considered, in order not to create any edge effect.

4) temporal envelope peak detection

The temporal envelope of each partial signal is then calculated according to the process described in ref 1; the minimum and maximum values are extracted, and the segment found between two successive minima is processed as the main part of one of the expected wfs, provided that its reference instant (amplitude maximum) appears within the 6 ms frame interval (fig 3). The sum of the detected waveforms is again equal to the partial signal considered. Each elementary waveform represents a local peak in the time-frequency domain.

5) waveform modelling

As shown in II, it appears to be possible to consider the elementary waveforms as the impulse responses of the local filters described above.

We can decompose the F filter into parallel sections. Locally, in proximity

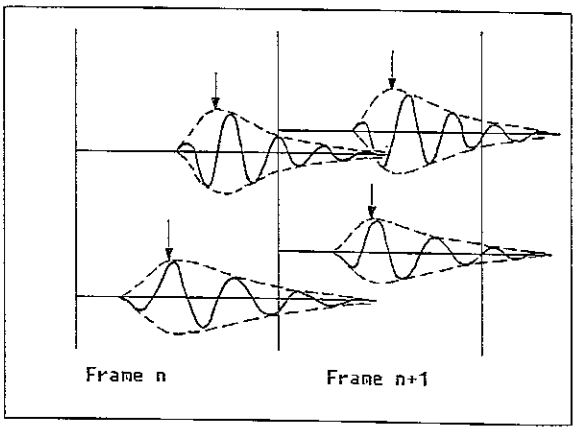

Fig 3 - Waveform behavior at frame boundaries

to a formant, F can be approximated to a second order section having an impulse response which is a sine wave with an exponentially decreasing envelope. If the frequency sectioning function is a rectangular one centered on the formant, the impulse response is modified, but is still able to be modelled in terms of the product of a sine wave and a window.

In order to avoid having an infinitely long waveform, we have introduced an attenuation function, the object of which is to limit effects of overlapping waveforms.

We suppose the following hypotheses:

- the central frequency is constant over the whole length of the waveform.

- the precise form of the envelope is not extremely important.

- an impulse response decreases rapidly. This facilitates its detection by peak-picking of the signal envelope.

The frequency analysis gives the central frequency of each waveform, the temporal analysis gives the reference instant and the time locking of the carrier oscillation, as well as the envelope parameters (amplitude, attack and decay durations).

IV. Results

The above described system was tested for various male and female voices. Fig 4 shows the beginning of a French sentence uttered by a male speaker, after analysis and evaluation of the main wf parameters. Each detected elementary waveform is represented by a diamond-shaped dot, of heigth proportional to the logarithm of the wf amplitude. Compared to a similar document in ref 1, the grouping of the wfs on the spectral peaks is obviously satis-

factory, in voiced segments as well as in fricative ones. The rapid temporal events (stop releases) also seem to be well represented.

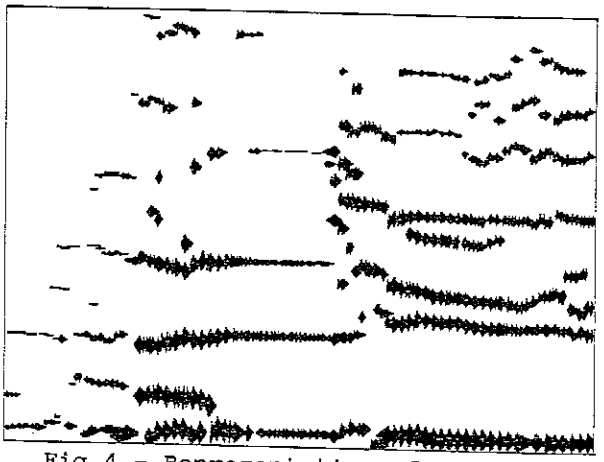

Fig 4 - Representation of a French sentence (beginning of "As-tu vu ce fameux lapin ?", male speaker) as a set of wfm parameters in the time-frequency domain.

From the perceptive point of view, there is no or little loss of quality if the lower part of the spectrum is not taken into account. However, some detection and modelling problems can be found in the part of the spectrum under the F1 region, where parameter estimation errors become perceptually important.

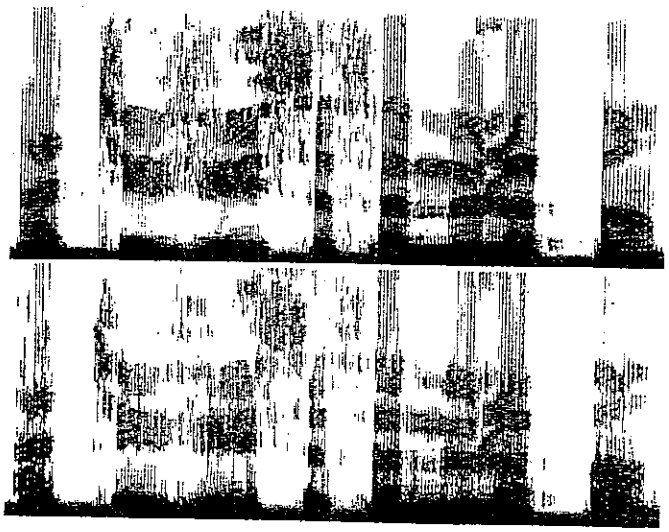

Fig 5 - Conventional spectrograms (top: original; bottom: synthetic) of the entire sentence.

IV - Conclusion

In this paper we have presented a new manner of analysing the speech signal and of modelling it as a set of elementary waveforms. The goal is the same as in ref 1, i.e. to obtain a description of the signal in terms of entities representative according to the production or perception point of view. Yet the difference lies in the fact that we profit from the structure of the speech signal - poles or spectral maxima - to predetermine the spectral regions where the elementary waveforms should be searched for. The representation we obtain is adequate and can be used as a basis for research into acoustic perceptive structures. Resynthesis yields a signal that is perceptually very close to the original - with the exception of the lowest region of the spectrum (band including Fo) in which some modelling problems remain.

Ref 1 and this paper represent two fairly different points of view; the former is perception-oriented, where the latter uses some characteristics of the speech production system. The premisses are, however, the same. In both cases, as well as in the wavelet approach, a better mastery of the compromise between time and frequency resolutions is sought, and early signal structure decisions are avoided.

VI - References

1 - Liénard, J.S. "Speech Analysis and Reconstruction Using Short-time, Elementary Waveforms". ICASSP-87, Dallas.

2 - Kronland-Martinet, R.,Morlet, J. and Grossmann, A. "Analysis of Sound Patterns Through Wavelet Transforms". To appear in the International Journal of Pattern Recognition and Artificial Intelligence, special issue on Expert Systems and Pattern Analysis.

3 - Rodet, X. "Time Domain Formant-Wave Function Synthesis", in "Spoken Language Generation and Understanding", J.C.Simon ed., D.Reidel Publishing Co, Dortrecht, Holland, 1980.

4 - Makhoul, J. and Cosell, L. "Adaptive Lattice Analysis of Speech". IEEE Trans. on Circuits and Systems, 28-6, June 1981.

5 - Crochiere, R.E. "A weighted Overlap-Add Method of Short-Time Fourier Analysis-Synthesis", IEEE Trans. on ASSP, 28-1, Feb 1980.

ANALYSE-SYNTHESE DE LA BANDE DE BASE PAR FORMES D'ONDES ELEMENTAIRES

C.d'Alessandro

LIMSI-CNRS: BP 30 F-91406 ORSAY Cedex

ABSTRACT

This paper is a continuation of our work on the representation of speech signal by a set of well-localized time-frequency energy concentrations (elementary waveforms). We present here a method and a system for analysis-synthesis of the speech "baseband" (which will be defined below). The quality problems which were encountered in the lower part of the spectrum during modelling and synthesis are thus circumvented. We use the same kind of method for the processing of both formantic areas and "baseband". After an introduction in section 1, we describe the synthesis formulae in section 2 and the system developed according to these principles in section 3. Some conclusions are presented in section 4.

1 INTRODUCTION

Le traitement automatique de la parole reste tributaire de la représentation préalable du signal qui en est le support acoustique. Parmi les nombreuses méthodes disponibles, l'analyse en formes d'ondes élémentaires se présente comme un moyen neuf, et prometteur dans la mesure où il vise à une représentation permettant une reconstruction parfaite du signal et manipulant des objets pertinents tant du point de vue de la perception que de celui de la production [Liénard 87].

Dans un papier précédent, une méthode de représentation du signal de parole en fonctions d'ondes élémentaires a été développée, en se basant sur une décomposition en parallèle de la fonction de transfert du conduit vocal [d'Alessandro 87]. Les fonctions d'ondes élémentaires apparaissent comme des contributions bien localisées dans le plan spectro-temporel, et permettent ainsi de rendre compte des phénomènes de production (formants, excitations du conduit vocal, explosions ...) de façon explicite, par un ensemble discret d'éléments. L'exploitation de la structure particulière du signal de parole guide la recherche des formes d'ondes élémentaires dans les régions de maximum d'énergie spectrale ("formants", au sens large) et temporelle ("impulsions", au sens large) et permet l'obtention de paramètres perceptivement pertinents [d'Alessandro 88].

Notre système d'analyse-synthèse en fonctions d'ondes élémentaires (s'appuyant sur les fonctions d'ondes formantiques) permettait une bonne représentation du signal de parole, sans perte de qualité du point de vue perceptif, sauf dans la "bande de base" (région spectrale jusqu'au premier formant inclus) où des problèmes de modélisation apparaissaient. Par une démarche semblable à celle adoptée dans les vocodeurs à bande de base, nous présentons ici une nouvelle méthode pour décomposer la bande de base du signal en formes d'ondes élémentaires utilisant un processus d'analyse-synthèse analogue à celui employé précédement et s'appuyant sur une représentation sinusoïdale.

2 REPRESENTATIONS

2.1 représentation formantique

Le modèle linéaire classique de production du signal vocal suppose le filtrage d'une certaine fonction d'excitation par un filtre linéaire évoluant dans le temps.

$$s(t) = e(t) * R(t)$$

- e(t): signal d'excitation.
- R(t): réponse impulsionnelle du filtre.
- s(t): signal résultant.

Dans ce qui suit on suppose le signal stationnaire (sur une tranche de temps assez courte) et donc le filtre de réponse impulsionnelle R invariant. Ce filtre, associé au conduit vocal, peut être décomposé en n sections parallèles, chacune d'elle représentant une résonance (ou formant). Dans le domaine temporel il est ainsi possible d'identifier le signal de parole avec la somme des réponses de chaque section au signal d'excitation. En première aproximation, si celui-ci est constitué d'une série d'impulsions idéales, on peut écrire:

$$s(t) = \sum_{j=1}^{m} \sum_{i=1}^{n} \delta_0(t - t_j) * R_i(t)$$

où R_i représente la réponse impulsionnelle de la $i^{ième}$ section, et $\delta_0(t - t_j)$ une impulsion d'excitation à l'instant t_j.

Si l'on assimile de plus les sections parallèles à des résonateurs du second ordre [Klatt 80], alors:

$$R_i(t) = G_i e^{-\alpha_i t} sin(\omega_i t + \phi_i)$$

où α_i règle la largeur de bande (à -6 dB du sommet), G_i le gain à la résonance, ω_i la fréquence centrale du $i^{ième}$ résonateur et ϕ_i sa phase.

soit:

$$s(t) = \sum_{j=1}^{m} \sum_{i=1}^{n} \delta_0(t - t_j) * (G_i e^{-\alpha_i t} sin(\omega_i t + \phi_i))$$

Pour une représentation en formes d'onde, on peut de plus rendre indépendantes les excitations des différentes sections, ce qui affine le compromis entre précision fréquencielle et précision temporelle en le localisant, et estimer les paramètres pour chaque réponse impulsionnelle; le pavé spectro-temporel où l'on suppose le signal stationnaire est ainsi délimité par la forme d'onde:

$$s(t) = \sum_{i=1}^{n} \sum_{j_i=1}^{m_i} \delta_0(t - t_{j_i}) * (G_{j_i} e^{-\alpha_{j_i} t} sin(\omega_{j_i} t + \phi_{j_i}))$$

Les fonctions d'ondes élémentaires choisies sont donc ici identiques aux fonctions d'ondes formantiques [Rodet 80].

2.2 représentation sinusoïdale.

Pour la partie grave du spectre (en deçà du premier formant), l'utilisation d'un signal d'excitation trop simple pose de sérieux problèmes de qualité. Pour pallier à ce défaut de nombreux modèles d'excitation ont été proposés, en particulier la représentation sinusoïdale [McAulay 86]:

$$s(t) = \sum_{i=1}^{k} A_i sin(\omega_i t + \phi_i)$$

Le nombre k de sinusoïdes ainsi que l'amplitude A_i, la fréquence ω_i et la phase ϕ_i évoluent dans le temps et doivent donc être estimés sur une tranche de temps pendant laquelle le signal est quasi-stationnaire. Une alternative aux différentes méthodes proposées pour cette estimation est l'utilisation de formes d'ondes élémentaires pour représenter chaque segment de sinusoïde, pendant la durée desquelles on suppose le signal stationnaire:

$$s(t) = \sum_{i=1}^{k} \sum_{j_i=1}^{l_i} \delta_0(t - t_{j_i}) * (A_{j_i} env_{j_i}(t) sin(\omega_{j_i} t + \phi_{j_i}))$$

l'enveloppe temporelle choisie $env_{j_i}(t)$ doit permettre la reconstitution de la sinusoïde initiale; il s'agira par exemple de segments sinusoïdaux.

$$env_{j_i}(t) = 1/2(1 + cos(\beta_{j_{i_1}} t))$$

pour $0 \leq t < \pi/2\beta_{j_{i_1}}$

$$env_{j_i}(t) = 1/2(1 + cos(\beta_{j_{i_2}}(t - \pi/2\beta_{j_{i_1}}) + \pi/2))$$

pour $\pi/2\beta_{j_{i_1}} \leq t < \pi/2\beta_{j_{i_2}} + \pi/2\beta_{j_{i_1}}$

Les β_i sont calculés de façon à conserver un nombre de cycles constant dans chaque forme d'onde (qui est alors une ondelette au sens de [Goupillaud 85]).

2.3 représentation par formes d'ondes

La représentation par forme d'onde complète s'appuie sur une segmentation spectrale préalable, qui permet de dégager les régions de maximum d'énergie, auxquelles est appliquée une représentation en formes d'ondes élémentaires de type "formantiques" (au delà du premier maximum) ou "sinusoïdales" (en deçà du premier maximum) (fig. 1).

$$s(t) = (\sum_{i=1}^{n} \sum_{j_i=1}^{m_i} \delta_0(t - t_{j_i}) * (A_{j_i} env_{j_i}(t) sin(\omega_{j_i} t + \phi_{j_i}))) +$$

$$(\sum_{a=1}^{k} \sum_{b_a=1}^{l_a} \delta_0(t - t_{b_a}) * (G_{b_a} e^{-\alpha_{b_a} t} sin(\omega_{b_a} t + \phi_{b_a})))$$

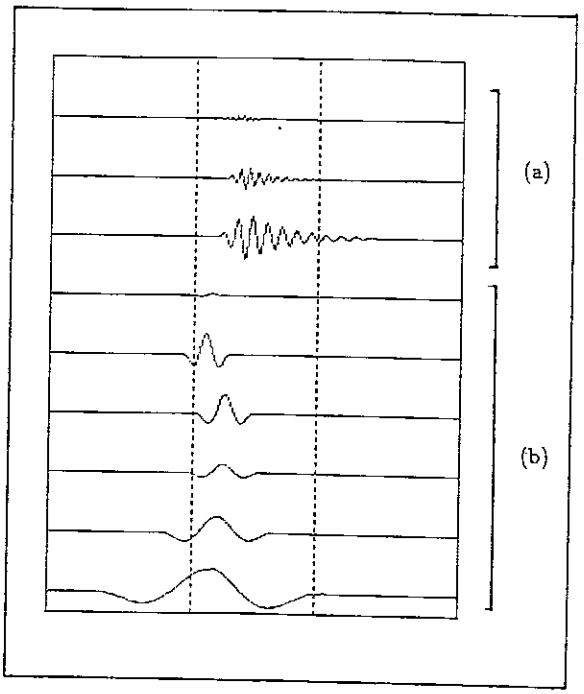

Figure 1: Modèles (a) "formantiques", (b) "sinusoïdaux" de formes d'ondes élémentaires

3 PROCESSUS D'ANALYSE-SYNTHESE

Le processus d'analyse-synthèse en fonctions d'ondes formantiques a déjà été décrit. Rappelons qu'à la suite d'une modélisation de l'enveloppe spectrale, un filtrage à phase nulle permettait d'obtenir des signaux à bande large centrés sur les maxima spectraux. Les formes d'ondes élémentaires etaient ensuite détectées grâce à l'enveloppe temporelle de ces signaux, puis modélisées comme précédemment pour la synthèse.

Pour le traitement de la bande de base, un procédé analogue est mis en œuvre, trame par trame (les trames sont de durée assez courte, 6 ms, pour que l'on puisse supposer le signal quasi-stationnaire):

- Modélisation de l'enveloppe spectrale par prédiction linéaire.
- Détection des maxima spectraux, associés aux formants, et définition de la "bande de base", comme la région spectrale jusqu'au premier maximum inclus.
- Calcul du module de la transformée de Fourier d'une tranche de signal centrée sur la trame.

- Recherche des maxima spectraux, associés aux harmoniques pour de la parole voisée, et segmentation spectrale autour de ces maxima. Tout ce qui suit ne concerne évidemment plus que la bande de base (fig. 2).
- Filtrage à phase nulle dans chacune des bandes ainsi définies, pour obtenir des signaux à bande étroite (fig. 3).
- Dans chaque bande, détection des formes d'onde (qui appartiennent à la trame si leur maximum y appartient) par recherche des maxima du signal.
- Synthèse, par les formules vues précédemment, après estimation des paramètres d'amplitude, de fréquence, de phase et d'enveloppe de chaque forme d'onde.

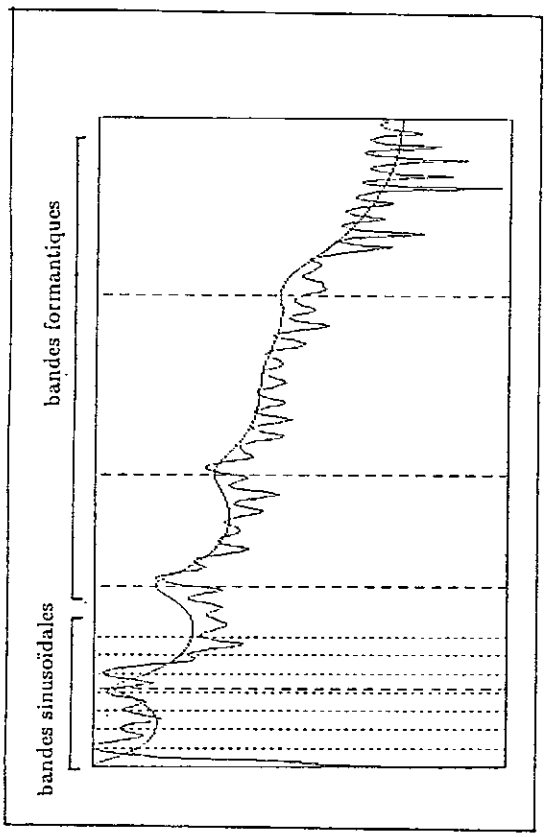

Figure 2: Modélisation et segmentation spectrale.

De par la largeur des bandes spectrales, les signaux obtenus dans chaque bande d'analyse sont des segments de sinusoïdes. Pour la parole voisée, il s'agit bien sur des premiers harmoniques, et l'enveloppe de ces signaux évolue beaucoup plus lentement que celle des signaux issus des bandes formantiques. Par contre, il est de toute première importance d'estimer précisément leur phase et leur fréquence (la vitesse d'évolution de la phase dépend évidemment de la fréquence).

Le choix du modèle de forme d'onde élémentaire adopté (nombre de cycles de la sinusoïde constant quelle que soit la fréquence) est motivé par ce souci de résolution spectro-temporelle, et non par un examen de l'enveloppe temporelle qui perd ici de son importance.

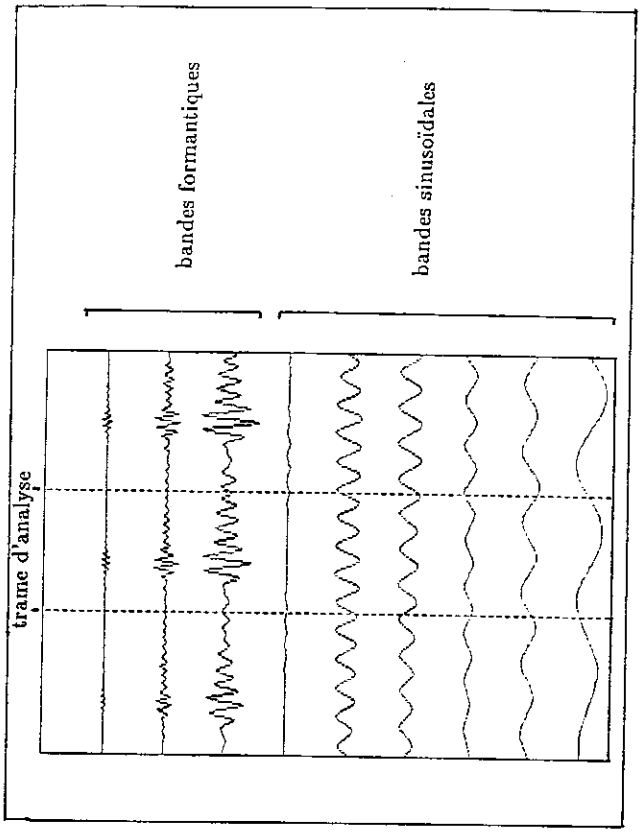

Figure 3: Filtrage dans les bandes précédemment définies.

Ainsi, le critère de segmentation d'une forme d'onde n'est plus basé sur une considération d'amplitude, comme c'est le cas pour les signaux formantiques en bande large, mais plutôt sur une considération de périodicité locale.

De même que dans les bandes d'analyse formantique, Ce processus donne des résultats satisfaisant tant pour de la parole voisée (on détecte alors des segments d'harmoniques) que pour la parole non voisée: le caractère local de la détection des formes d'onde permet en effet de reproduire des signaux transitoires trés brefs ou des signaux aléatoire.

4 RESULTATS

Le système réalisé permet le traitement automatique d'un segment de parole, et a été testé pour diverses voix tant féminines que masculines. La qualité de synthèse est excellente (pas ou très peu de différence avec l'original), mais doit maintenant faire l'objet de tests systématiques.

Le procédé s'appuie sur le modèle linéaire classique de production de la parole (de par la segmentation sur l'enveloppe spectrale) mais il autorise une analyse d'une grande finesse spectro-temporelle tout en ne délivrant qu'un jeu discret d'objets porteurs de toute l'information contenue dans le signal.

L'affichage des formes d'ondes dans le plan temps/fréquence permet de visualiser le résultat obtenu, qui paraît particulièrement intéressant, en ce sens que la plupart des caractères perceptivement pertinents (formants, pitch, voisement, explosions ...) sont représentés par un ensemble réduit de formes d'ondes (fig. 4, fig. 5, fig. 6).

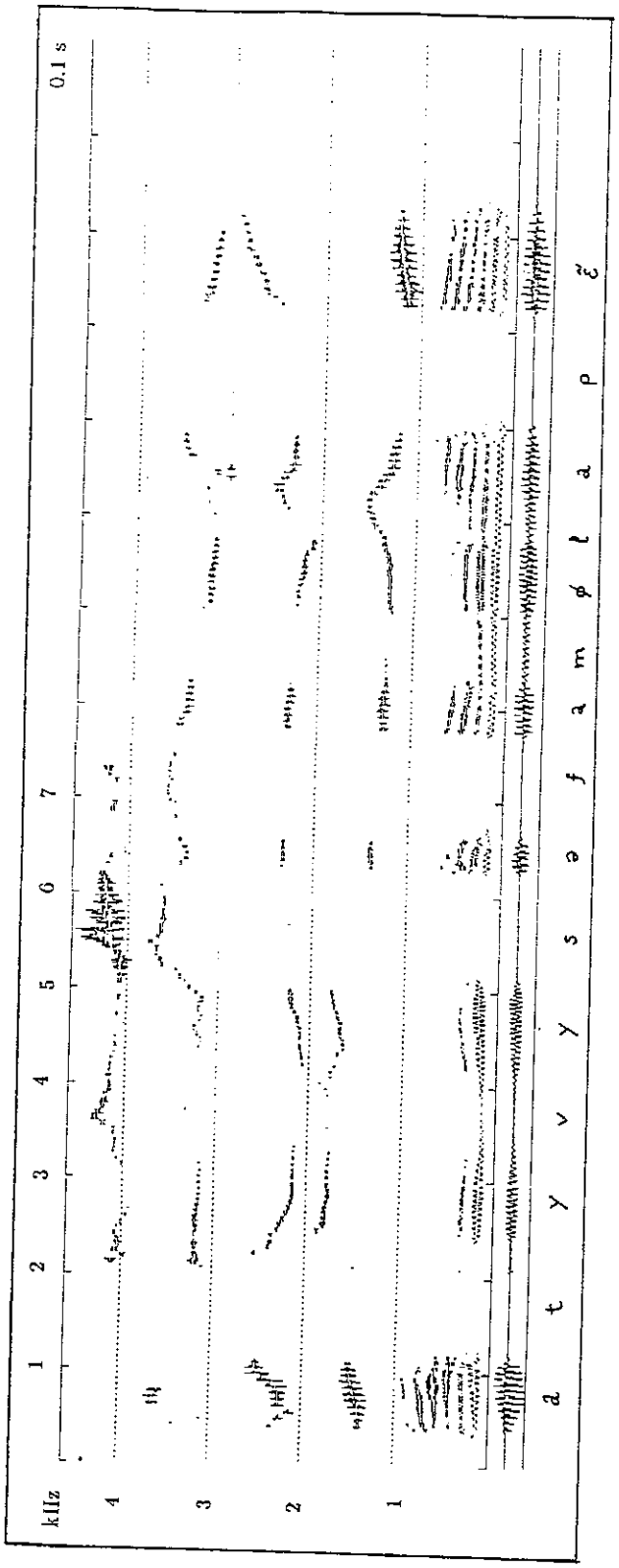

Figure 4: Affichage des formes d'ondes dans le plan temps/fréquence.

5 CONCLUSIONS

Nous avons présenté les fondements et la réalisation d'un système de représentation du signal de parole en formes d'ondes élémentaires. De part le choix opéré sur les formes d'ondes à rechercher, qui dérive du modèle classique de production du signal vocal, un traitement différent doit être appliqué aux différentes régions spectrales: une méthode spécifique semble en effet nécessaire pour rendre compte de la partie grave du spectre, et se trouve ici développée.

Il s'agit désormais, en raprochant cette méthode de modèles de perception, de l'appliquer à l'analyse de la parole. Les paramètres qu'elle fournit semblent également utiles pour une variante de la synthèse à formants en parallèle.Le débit d'information obtenu, qui reste à estimer de façon systématique, paraît offrir de bonnes potentialités en vue du codage: un gain semble en effet possible par rapport au codage sinusoïdal de part l'agglomération de plusieurs harmoniques dans une seule forme d'onde.

Ce travail représente le fruit de nombreuses discussions avec M^{rs} J.S.Liénard & X.Rodet, que l'auteur tient à remercier ici.

REFERENCES

[Liénard 87] Liénard, J.S. "Speech Analysis and Reconstruction Using Short-Time, Elementary Waveforms". IEEE-ICASSP 87, Dallas.

[d'Alessandro 87] d'Alessandro, C. & Rodet, X. "Fonctions d'ondes formantiques: extraction des paramètres et synthèse vocale". $16^{ième}$ JEP, 1987 Hammamet.

[d'Alessandro 88] d'Alessandro, C. & Liénard, J.S. "Decomposition of the Speech Signal into Short-Time Waveforms Using Spectral Segmentation". IEEE-ICASSP 88, New-York.

[Klatt 80] Klatt, D. "Software for o cascade/parallel formant synthesizer". JASA vol. 67(3), Mar. 1980.

[Rodet 80] Rodet, X. "Time Domain Formant-Wave-Function Synthesis". in "Spoken Language Generation and Understanding", J.C. Simon ed., D.Reidel publishing compagny, Dordrecht.

[McAulay 86] McAulay, R. & Quatieri, T. "Speech Analysis/Synthesis Based on a Sinusoidal Representation". IEEE trans. on ASSP, vol. ASSP 34 no. 4 1986.

[Goupillaud 85] Goupillaud, P., Grossmann, A. & Morlet, J. "Cycle-Octave and Related Transforms in Seismic Signal Analysis". Geoexploration 1985.

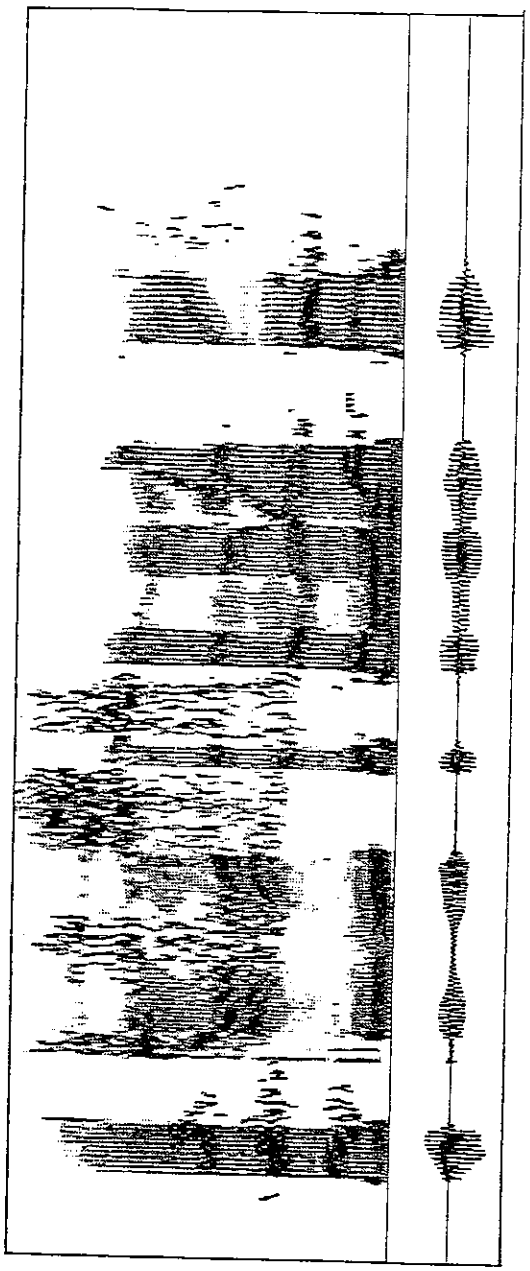

Figure 5: Spectrogramme du signal naturel correspondant à la figure 4.

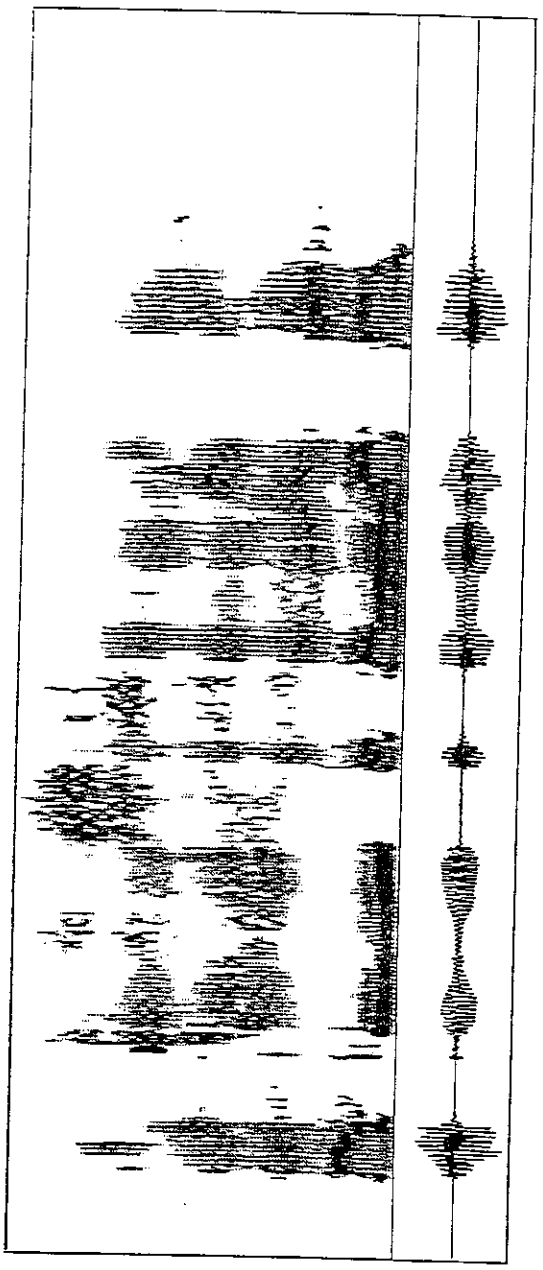

Figure 6: Spectrogramme du signal synthétique correspondant à la figure 4.

WAVELETS AND GRANULAR ANALYSIS OF SPEECH

J.S.Liénard and C. d'Alessandro

LIMSI-CNRS
BP 30, 91406 Orsay Cedex, France

1 - Very short time analysis of speech

The speech signal comes from the convolution of a source signal - due to the vibration of the vocal cords or to the airflow through a narrowing of the vocal tract - with the impulse response of the vocal tract. Both constituents rapidly change over time, and one usually considers that the phonetic information in the signal is mainly related to the evolution of the two or three first resonances of the vocal tract, called "formants" F1, F2, F3. The vocal cords vibrating frequency, Fo, is closely related to a perceptive quality of sounds called "pitch".

In order to extract the phonetic information the signal is considered to be stationary over a time interval long enough to include several pitch periods, and short enough to capture the evolution of the spectral envelope. The usual tradeoff yields some 25 ms for the width of the analysis window, and 10 ms for the interval between successive windows. So, despite the fact that everybody agrees about the relevance of classical spectrographic analysis - which uses bandpass filters 300 Hz wide, with impulse responses as short as a few ms - the information extracted from the signal for transmission or recognition is altered from the start in the time dimension. Some rapid transitions in consonants are smoothed or even erased, the sound structure disappears, the possible perceptual interaction between segmental and prosodic information is deliberately discarded.

The "granular" analysis we present hereafter aims at decomposing the signal into a set of discrete elements associated with energy concentrations in the time-frequency coordinates, with some emphasis on the time resolution (in the 1 to 2 ms range). In the voiced segments (vowels, some consonants) those grains correspond to the resonance maxima of each proper mode of the vocal tract, at each pitch period. In the noise segments (consonants such as "s" or "ch") the grains are randomly distributed in some region of the time-frequency plane. Finally the bursts found at the onset of some sounds like "p" or "t" give birth to one or several grains precisely located at the same instant, following a silence. In our view of speech analysis the notions of voicing, pitch, formants, for which no method gives a completely satisfactory answer, cannot be directly extracted from the signal, but should result from a structural study of the grain distribution. For instance the

signal will be declared as "voiced" when, locally, comparable grains appear at regular intervals.

This analysis is based on hypotheses about the temporal coding of the acoustical wave by the human auditory apparatus. Consequently it is tempting to implement an auditory model in order to check them. However, it is difficult to validate and interpret the results of such models, because they implicitly take into account some further processing by the brain, which cannot be modeled or understood as yet. Thus we choose to implement the granular analysis in a way such that objective or subjective verification is permitted through a reconstruction process (ref 1).

The analysis process is composed of two steps. The first one aims at decomposing the signal into a series of narrowband signals, covering the frequency band of speech, i.e. from 70 to 5000 Hz. The second step consists in modeling each of them into a series of successive elementary waveforms. At the present time the first step only can be related to the theory of wavelet analysis.

2 - Decomposing the speech wave into a set of narrowband signals

In order to reconstruct the signal by a mere addition of its narrowband components it is necessary that all of the filters respond with the same phase, have the same slopes, and have their gains properly adjusted with respect to the distribution of the center frequencies along the frequency scale. We implemented a recursive filter structure, used twice with time-reversal in order to cancel any delay or phase distortion. The result is a zero-phase filter, the order of which is twice the initial filter order. For the basic unidirectional filter we choose a simple resonator (second order), so that the resulting filter is of order four (slopes at infinity tend toward -12 dB/octave, slopes around the cutoff frequencies depend on the quality factor, Q).

The distribution of the center frequencies along the frequency scale is one of the filterbank parameters. We used several tunings, ranging from linear to logarithmic, with an intermediary choice (Bark scale) close to what is known of ear physiology (tendency toward the linear scale in the low frequencies, toward the logarithmic scale in the high frequencies). The gains are automatically adjusted with respect to the number and distribution of the filters. Usually the number of filters is between 12 and 32, the bandwidths range from 100 to 600 Hz, the quality factor remains within the 1 to 10 range.

The filtering process is illustrated on Fig 1, which shows the decomposition of a series of impulses into 16 linearly distributed channels, as well as their additive reconstruction. Except for some noise due to the poor bandpass limitation of the signal, it is clear that the reconstruction is satisfactory.

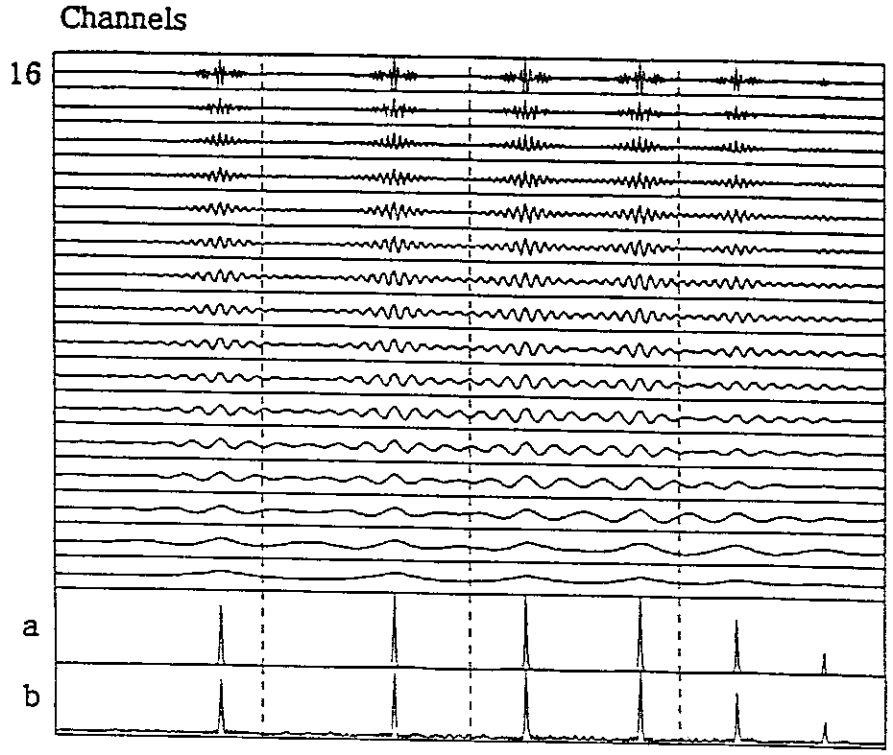

Fig 1 - Decomposition of a series of pulses (a) with a zero-phase filterbank, and signal reconstructed (b) by summation of the 16 output signals. Time scale 10 ms between vertical dotted lines (valid for all the figures in the present paper).

Fig 2 shows the analysis of a speech signal, with some differences in the filterbank parameters, and a different representation of the output signals : only the positive parts of each signal are represented, after logarithmic compression of the amplitude. This representation exemplifies the synchronization phenomena occurring among adjacent channels when several filters capture the same signal component. Here again, reconstructing the signal through summation of the outputs yields a signal very close to the original. When listening to both, no difference can be heard.

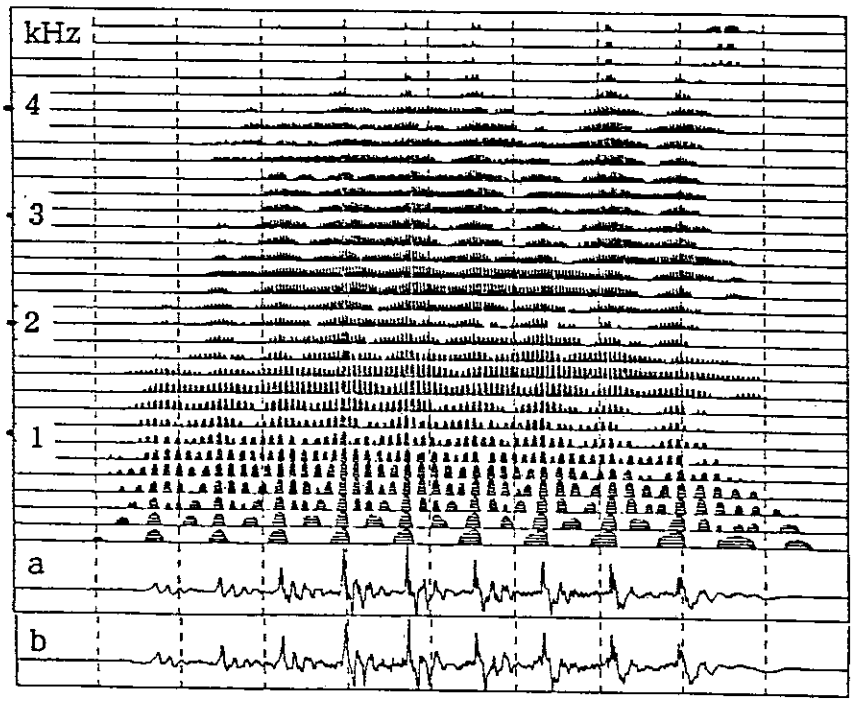

Fig 2 - Vowel "a" analysed with a 32-channel linear filterbank; a) original, b) reconstructed signal.

Basically our analysis process consists of the convolution of the signal with a symmetrical "wavelet" made of the bidirectional impulse response of the filter. Comparison with the Morlet and Grossmann wavelet theory yields some remarks, which can be classified as analogies and differences. The first similarity is to be found in the need to a better mastery the tradeoff between time and frequency resolutions. In both cases a better time resolution is expected in the high frequency part of the spectrum, while the frequency precision is expected in the low frequency part. Another important similarity lies in the additive reconstruction possibility.

As for the differences, the formal expressions given by the wavelet theory are obviously an advantage, thanks to the insights and guarantees they provide. On the other hand, the logarithmic frequency scale imposed by the wavelet shape conservation in its compressions and dilations may not be perfectly adapted to our psycho-physiological needs, and is not mandatory in our approach. In most of our experiments the equivalent "wavelets" composed of the bidirectional impulse responses of the filters were closer to the Gabor type than to the Morlet-Grossmann type. Finally the last noticeable difference lies in the implementation of a recursive IIR filter, which allows the computations to be achieved very quickly, several order of magnitude faster than the wavelet analysis. The filter used has a low Q factor, and does not cause any stability problem.

3 - Modeling the output signals into discrete elements

If speech processing could be reduced to decomposing the signal into n narrowband signals, we would have gained nothing; the information rate of the signal would simply be multiplied by n. We are actually looking for a decomposition into a set of discrete elements, or grains, which we call Elementary Waveform Models, or wfms (fig 3). This decomposition has been validated for singing voice synthesis (ref 2), but our problem is the opposite, i.e. how to go from the signal to the list of grains ?

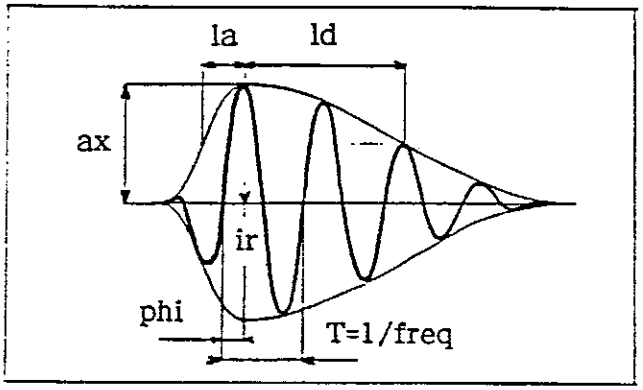

Fig 3 - Waveform model, with attack and decay shaped by raised sinusoids.

A first way to deal with this probleme is to decompose each narrowband signal into a string of wfms. For this we spot the extrema of the envelope, associate to each maximum a prototypical waveform whose parameters have been adjusted so that the sum of two successive wfms makes up a good approximation of the signal for the zone being modeled. This process produces a set of wfms for each channel (fig 4a). Reconstruction of the signal by regenerating the wfms and summation of all of the channels furnishes a signal perceptually very close to the original.

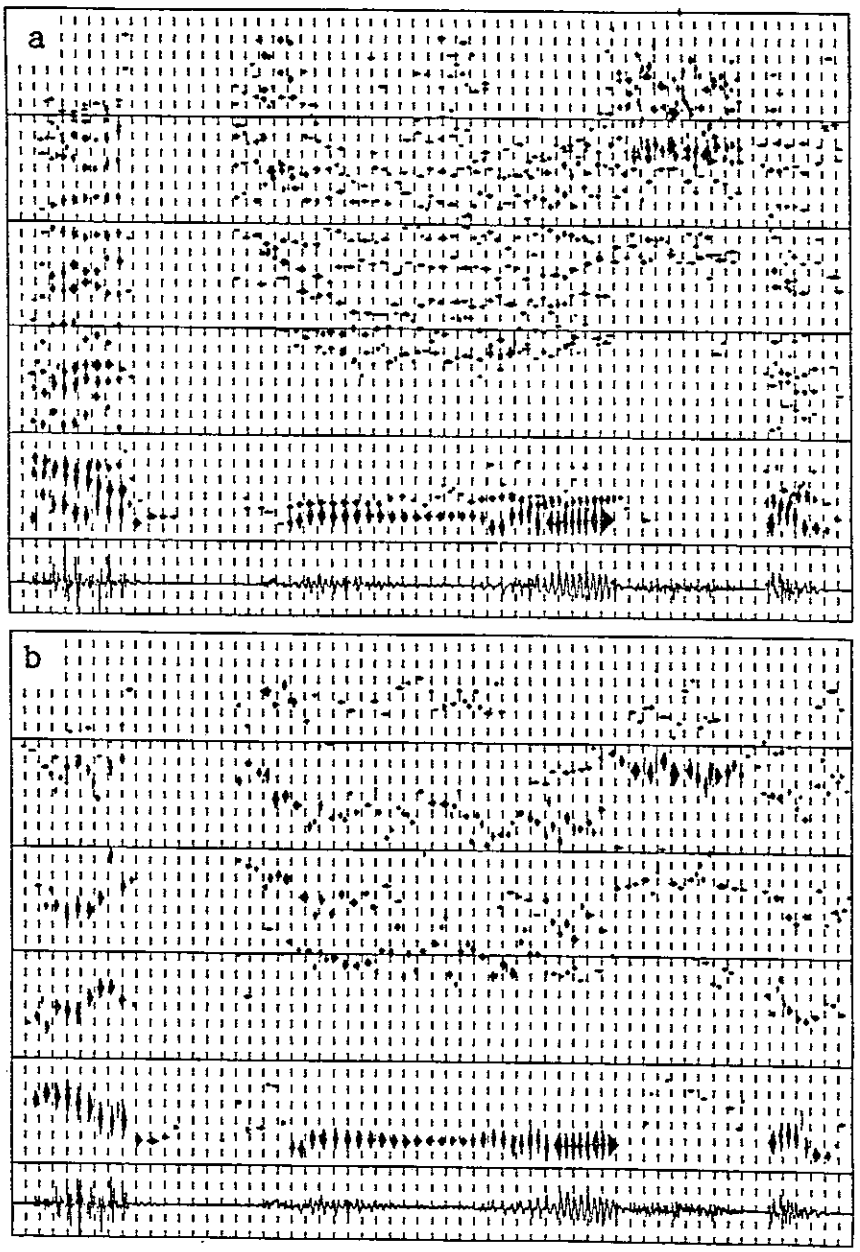

Fig 4 - Speech segment /atyvysə/ displayed as a set of wfms in the time-frequency plane. (a) : channel by channel modeling. (b) : same, after iterative grouping of adjacent channels.

Even though this represents a considerable economy compared to the set of narrowband signals produced by the filterbank, the channel-by-channel modeling process still is highly redundant. As the filters we use are not extremely selective, one given constituent of the signal has been analysed by several filters, resulting in several wfms in adjacent

channels. It is only when the sum of all of the wfms in those adjacent filters is calculated that that the original constituent - grain - can be regenerated. We have therefore created a local grouping procedure for the narrowband signals surrounding any local maximum of the envelope in the time-frequency space. This procedure greatly reduces the total number of wfms obtained (fig 4b), and should ensure their invariance against different filterbank configurations, but some spotting problems appear in the low frequency range where the pitch period and the center frequency of the analysis channel are close to each other.

Since the grouping procedure described above is costly and not perfect, a third approach has been elaborated which makes use of the specific structure of the speech signal (ref 3). It consists of filtering a short segment of the signal (some 50 ms, in order to avoid any boudary effect) in the regions of spectral prominence, as evaluated by a classical LPC analysis, rather than in fixed, permanently defined frequency bands. The new regions are frequently reestimated (every 6 ms in our experiments). Each filtered signal is then segmented and modeled by the same procedure, giving the desired grains without the need of a grouping procedure. This approach gives satisfactory results if it is adapted to the lower part of the spectrum (modeling the first harmonics) (ref 4).

4 - Conclusion

Our short term analysis, as well as wavelet analysis, has the desire to dominate the compromise between time and frequency resolutions. Both processes may be seen as filtering, or as the convolution of the signal with a particular, symmetric, Gaussian shaped kernel. But beyond this common desire, we are trying to model the signal into a set of discrete elements, or grains, which are supposed to be perceptually pertinent. This aim contributes to defining an inverse problem for which, at present, wavelet theory has no answer.

5 - References

1 - J.S.Liénard : "Speech Analysis and Reconstruction Using Short-Time, Elementary Waveforms", IEEE-ICASSP, Dallas, 1987.

2 - X.Rodet : "Time-Domain Formant-Wave-Function Synthesis", Computer Music Journal, vol 8, 3, 1985.

3 - C. d'Alessandro and J.S.Liénard : "Decomposition of the Speech Signal into Short-Time Waveforms Using Spectral Segmentation", IEEE-ICASSP, New York, 1988.

4 - C. d'Alessandro : "Analyse-Synthèse de la bande de base par formes d'ondes élémentaires", 17e Journées d'Etude sur la Parole de la SFA, Nancy, 1988.

TIME-FREQUENCY MODIFICATIONS USING AN ELEMENTARY WAVEFORM SPEECH MODEL

Christophe d'Alessandro
LIMSI-CNRS, BP 133, F-91403 Orsay Cédex, FRANCE

ABSTRACT

An elementary waveform speech model (EWSM) is defined and some capabilities are demonstrated for the modification of localized time-frequency events. The elementary waveforms allow for modelling the local spectro-temporal maxima of energy inside the speech signal by simple mathematical functions. EWSM parameters are estimated using a frame by frame processing: spectral modelling and segmentation using short-time Fourier transform and LPC spectrum, Fourier filtering according to this segmentation, waveforms spotting in each channel waveform modelling with simple functions. The EWSM parameters are relevant according to the classical theory of speech production, and their modifications yield well-localized time-frequency transformations, including frequency compression/expansion, pitch, formant, noise modifications.

1 INTRODUCTION

In this paper we discuss the ability of a new speech signal representation method for time-frequency modifications. *Global* modifications, like frequency expansion/compression, pitch and duration mofications are realized in a very simple way, as well as more unusual *local* modifications allowed by the properties of our representation: such as pitch period, formants, burst, fricative noise modification for instance. The later problem is also treated here in a very simple way, though it remains a difficult task for other representation methods.

We show elsewhere [1] that expansion of the speech signal into a discrete sum of time-frequency well-localized elementary waveforms can be achieved at least from three viewpoints:

- non-parametric methods, short-term Fourier transform and wavelets transform for instance, can receive an elementary waveform interpretation, crossing the classical filterbank analysis and block analysis interpretations [2] [3]. Exact representations and theoretical results are thus available, but some difficulties remain in order to establishing relationship with a speech production or perception model.
- granular analysis [4] based on analogies between auditory models and spectro-temporal analysis. Here, extraction of speech production parameters, formants or pitch for instance is a very interesting problem, but the same kind of difficulties than in auditory modelling are encountered: this point is at present time under study.
- model-based speech elementary waveform decomposition is a continuation of formant waveform synthesis [5]: an elementary waveform speech model (EWSM) can be derived from the classical acoustic model for speech production. The EWSM parameters are thus directly relevant, as speech production parameters. Automatic parameter estimation allows for using this model in the field of speech synthesis or modification.

We will only present and discuss the third approach, for a sake of simplicity, though the first and second ones are able to perform the same type of processing: only interpretation of waveforms parameters from a speech production viewpoint remains more or less difficult in these different cases.

Section 2 introduce the EWSM. Elementary waveforms formulas as well as speech production events viewed trough waveform representation are described.

The automatic analysis/synthesis process, based on spectral segmentation is explained in section 3.

Section 4 deals with the modifications and gives some examples.

In section 5 a conclusion is proposed.

2 ELEMENTARY WAVEFORMS SPEECH MODEL

The EWSM for speech representation is an extension of parallel formant model, in the time domain (figure 1).

The main differences between EWSM and parallel formant model is the lack of excitation/filter distinction in the first case: excitation is only virtual. Thus distinction between source and filter is avoided and the model is clearly located in acoustic domain.

For ideal voiced speech, an elementary formant waveform will be associated to each pitch period, in each formant area. The baseband, defined as the area below the first formant, where the contribution of the glottal airfow waveform is dominant, requires a special treatment: an elementary sinusoïdal parameterization of this contribution is performed.

For ideal unvoiced speech (frication noise), a previous study [6] has experimentaly shown that random generation of elementary waveforms is able, under certain conditions, to produce a noise spectrally equivalent to filtered white noise.

For an actual speech signal, one can easily mix these two ideal cases to produce, for instance, voiced fricatives, stops, or noisy voices.

Thus, two types of elementary waveforms allow for synthesis of both voiced, unvoiced and mixed speech: the next section presents justifications and formulas to choose elementary waveform models.

2.1 formant waveforms

According to the classical acoustic theory of speech production, voiced speech is obtained in the time domain by convolution of an excitation waveform $e(t)$ with the impulse response of a filter $R(t)$ associated to the vocal tract.

$$s(t) = e(t) * R(t) \quad (1)$$

If R is supposed linear and time-invariant, and if excitation is reduced to a train of pulses, parallel decomposition of equation 1 is written in time domain:

$$s(t) = \sum_{j=1}^{m} \sum_{i=1}^{n} R_i(t, t_j) \quad (2)$$

where R_i represent the impulse response of the i^{th} parallel section, at time t_j. For a second order section in equation 2, associated with *formants*, the impulse response is:

$$R_i(t) = G_i e^{-\alpha_i t} sin(\omega_i t + \phi_i) \quad (3)$$

where α_i sets bandwidth, G_i amplitude, ω_i central frequency, and ϕ_i phase of the i^{th} formant.

equations 3 and 2 present the behaviour of parallel formant synthesis, with pulse-like excitation in time domain.

We extend this model in two directions: first equation 3 is extended by using a more general formant waveform model proposed by [7], which introduces a smooth attack, and second equation 2 is extended by defining an independant excitation for each formant waveform: it is thus possible to synthesize both periodic and random signals:

$$s(t) = \sum_i G_i \Lambda_i(t) e^{-\alpha_i t} sin(\omega_i t + \phi_i) \quad (4)$$

Λ is a step function, with a cosine rising segment, beginning at reference instant t_i:

$$\Lambda_i(t) = 0 \; for \; t \leq t_i \quad (5)$$

$$\Lambda_i(t) = \frac{A}{2}(1 - cos(\beta(t - t_i))) \; for \; t_i < t \leq t_i + \frac{\pi}{\beta} \quad (6)$$

$$\Lambda_i(t) = 1 \; for \; t > t_i + \frac{\pi}{\beta} \quad (7)$$

equation 4 describes a discrete set of formant waveforms, located at points (t_i, ω_i) in time-frequency plane.

2.2 sinusoïdal parameterization of baseband waveform

For baseband synthesis, using formant waveforms is no more justified, and we propose a short-term sine waveform parameterization , close to [8]. The elementary waveforms are sinusoïdal segments, and the baseband signal is described with a formula close to equation 4:

$$s(t) = \sum_i G_i \Lambda_i(t) sin(\omega_i t + \phi_i) \quad (8)$$

where G_i represents the amplitude, ω_i the frequency, ϕ_i the phase and Λ_i the envelope of the sinusoïdal waveform. Λ_i is a temporal window, made of a rising and a decaying sine for example centered at reference instant t_i.

The complete EWSM combine the two types of waveform, using equations 4 and 8.

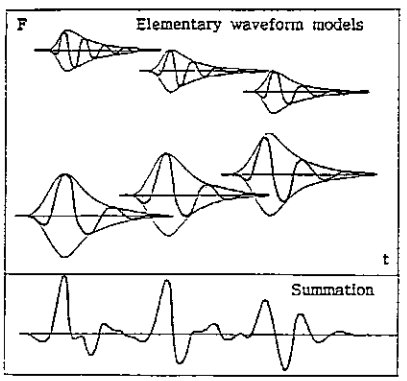

Figure 1: EWSM representation of a voiced segment, by summation of elementary waveforms (principle, from [1] ©IEEE-87)

2.3 EWSM representation of articulatory events

Waveform is the basic element of this representation: thus, an articulatory event is organized as a little set of waveforms.

In the time domain, voiced speech is composed of voicing periods. Each voicing period is composed of formant waveforms sharing a similar reference instant, ideally one in each formant area, and of sine waveforms sharing a similar reference instant, ideally one for each harmonic in the baseband. In frequency domain, voiced speech is composed of formants: a formant is viewed as a set of waveforms sharing similar central frequencies, ideally one waveform for each voicing period. The baseband is decomposed into harmonics: an harmonic is viewed as a set of waveforms sharing similar central frequencies, ideally one waveform for each voicing period.

Unvoiced speech is composed of randomly distributed formant waveforms and sine waveforms, according to the statistics of desired noise (more concentrated in the formant areas, if any).

Burst of stops are composed of a little number of very short-time waveforms, located at the burst instant, and reflecting its spectral composition.

Unvoiced fricatives or noisy voice are obtained by mixing the voiced and the unvoiced case.

For speech modification, the main point is that waveforms parameters are close to production parameters, they represent formant parameters, or voicing period or burst parameters etc., and that each waveform is a basic element which can be treated independently.

3 ANALYSIS/SYNTHESIS PROCESS

An automatic system for EWSM parameter estimation from actual speech has been developed. This system is based on a spectral wideband LPC model in formant area and spectral narrowband STFT model in baseband. Thus spectral local maxima are detected. Spectral segmentation and filtering in these areas give back time domain signals, and temporal segmentation using local temporal maxima allows for detection of natural elementary waveforms. Waveforms parameters are then estimated, and the sum of all synthetic elementary waveforms is the reconstructed signal.

Figure 2 summarize the analysis/synthesis process.

4 TIME-FREQUENCY MODIFICATIONS

The output of the analysis stage, and the input of the synthesis stage, is a set of elementary waveforms described by their parameters. Hence, performing spectro-temporal localized modifications comes to modifying those parameters. This modification is simple to understand, owing to the acoustic relevance of the parameters.

4.1 examples of global modifications

4.1.1 pitch and duration modification

Pitch modification is achieved without explicit pitch extraction. EWSM predict that for ideal voiced speech only one waveform appear for each voicing period. Pitch modification is obtained by modifying only one parameter (the reference instant) for formant waveforms, and by modifying two parameters (the reference instant and the frequency) for sine waveforms. Phase interpolation is achieved by the overlap-add process for sine waveforms. A duration modification occurs with pitch modification. A time domain treatment is used for duration modification alone, wich is not specific to our method [9]. Combining both allows pitch modification without any time distortion.

4.1.2 frequency expansion/compression

Frequency scale expansion/compression is achieved by modification of a single parameter, central frequencies, both for formant and sine waveforms.

4.2 examples of local modifications

Spectro-temporal local modifications of the speech signal are straightforward and simple to understand on the EWSM parameter, provided that the waveforms involved in the modification are well labeled. Thus, the main problem is to assign a set of waveforms to the particular acoustic or articulatory event under study. Automatic waveform labelling is beyond the scope of this paper, and we just attempt to show here the ability of the method for spectro-temporal localized modification. Waveform labelling was manually performed, by visual inspection of EWSM analysis results. Figure 3 is an exemple of such a representation.

4.2.1 formant modification

Amplitude, bandwidth, central frequencie, phase, temporal attack are explicit parameters of the EWSM. Hence, formant modifications are achieved in a very straightforward way. Figure 4.a.b.c gives an exemple of vowel change. The second formant central frequency is shifted down for all the /α/ to obtain /a/.

4.2.2 noise modification

Modifying the spectro-temporal behaviour of fricative noise is achieved in the same way in time-frequency plane. In figure 4.d.e noise is cut in /s/ and /f/ to obtain /t/ and /p/. In figure 4.f, voicing is drawn out from a /v/, and a little amount of noise is added to obtain a /f/.

5 CONCLUSION

The ability of a new spectro-temporal model-based speech representation for localized modifications has been demonstrated.

Modifications were performed on natural speech trough a high-quality analysis-synthesis system, hence naturalness was preserved.

This method provides a powerfull tool for speech modification, specially suited for phonetic, psychoacoustic and speech synthesis experiments.

REFERENCES

[1] d'Alessandro, C. "Représentation du signal de parole par une somme de fonctions élémentaires". These de doctorat en science, Université Paris VI, April 89 (in french).

[2] Flandrin, P. "Time frequency and time scale". IEEE Fourth Annual ASSP workshop on Spectrum estimation and Modeling, Minneapolis, August 1988.

[3] Combes, J.M., Grossman, A. & Tchamitchian, P. (ed.) "Wavelets, Time-frequency methods and phase space". Springer-Verlag, Berlin, 1989.

[4] Liénard, J.S. "Speech analysis and reconstruction using short-time, elementary waveforms". Proceedings of IEEE-ICASSP-87.

[5] d'Alessandro, C. & Liénard, J.S. "Decomposition of the Speech Signal into Short-Time Waveforms Using Spectral Segmentation". Proceedings of IEEE-ICASSP 88.

[6] d'Alessandro, C. & Rodet, X. "Synthèse et analyse-synthèse par fonctions d'ondes formantiques". Journal d'Acoustique, No. 2, No.2, June 1989 (in french).

[7] Rodet, X. "Time Domain Formant-Wave-Function Synthesis". in "Spoken Language Generation and Understanding", J.C. Simon ed., D.Reidel publishing compagny, Dordrecht, 1981.

[8] McAulay, R. & Quatieri, T. "Speech Analysis/Synthesis Based on a Sinusoidal Representation". IEEE trans. on ASSP, Vol. ASSP-34, No. 4, 1986.

[9] Neuburg, E.P. "Simple pitch-dependant algorithm for hight-quality speech rate changing." JASA, Vol. 63, No. 2, 1978.

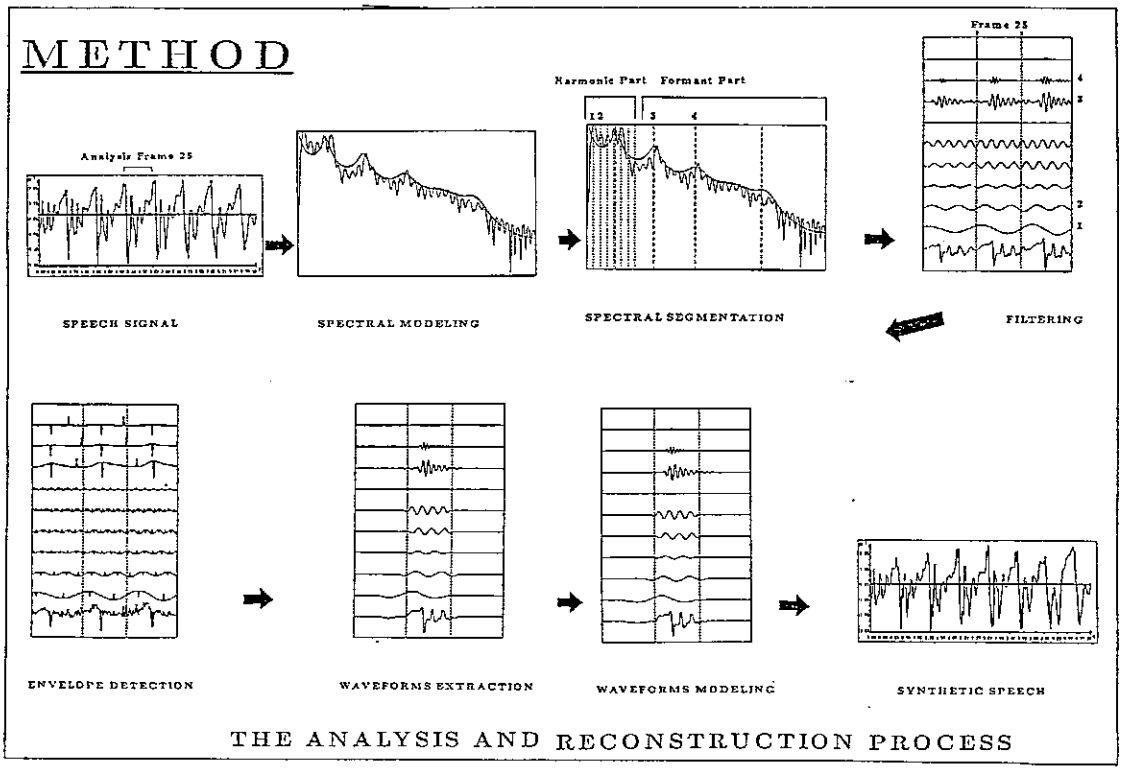

Figure 2: analysis-synthesis process

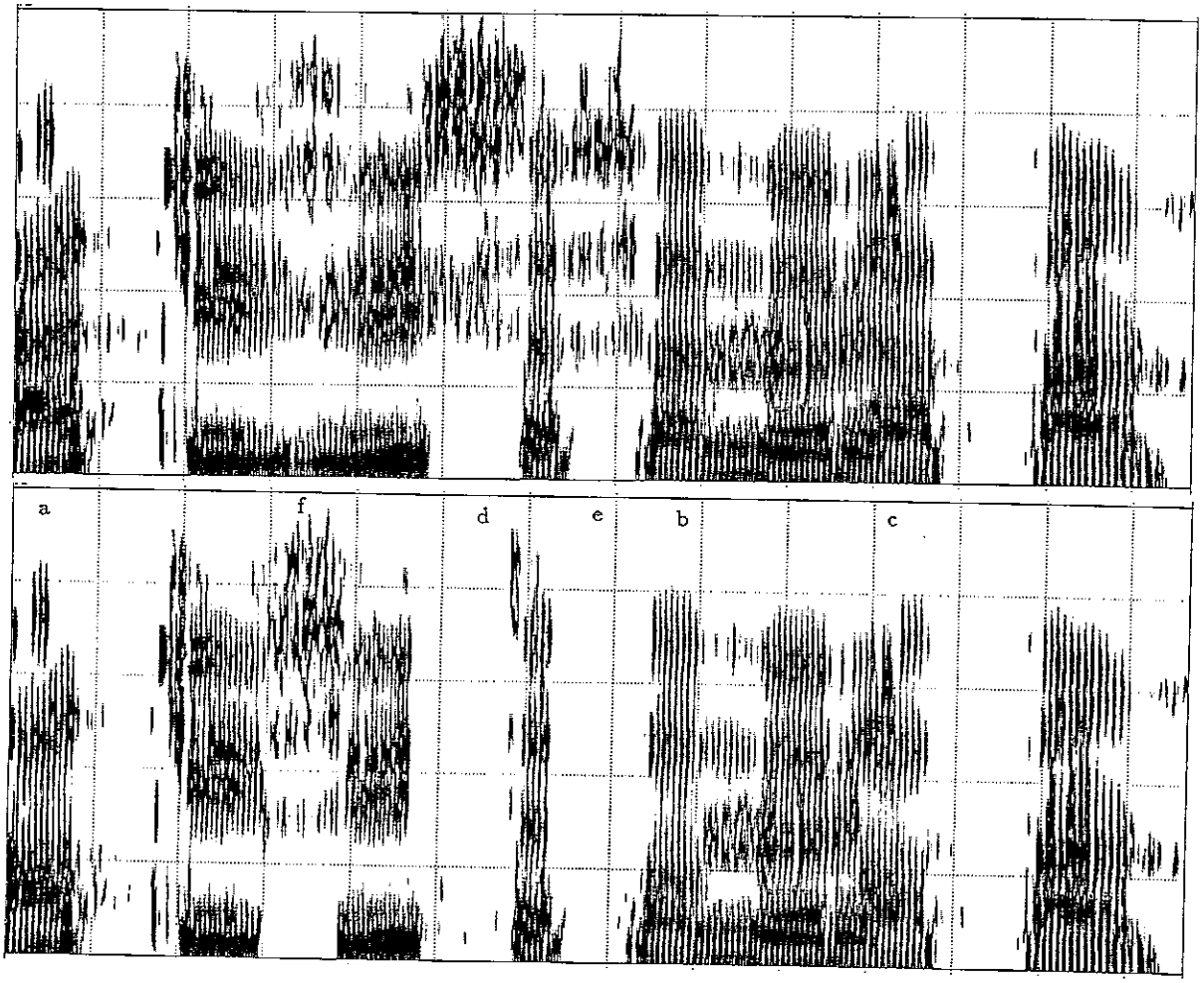

Figure 4: male voice speaking "as tu vu ce fameux lapin ?".
Top: original speech. Bottom: modifyied speech (see text).

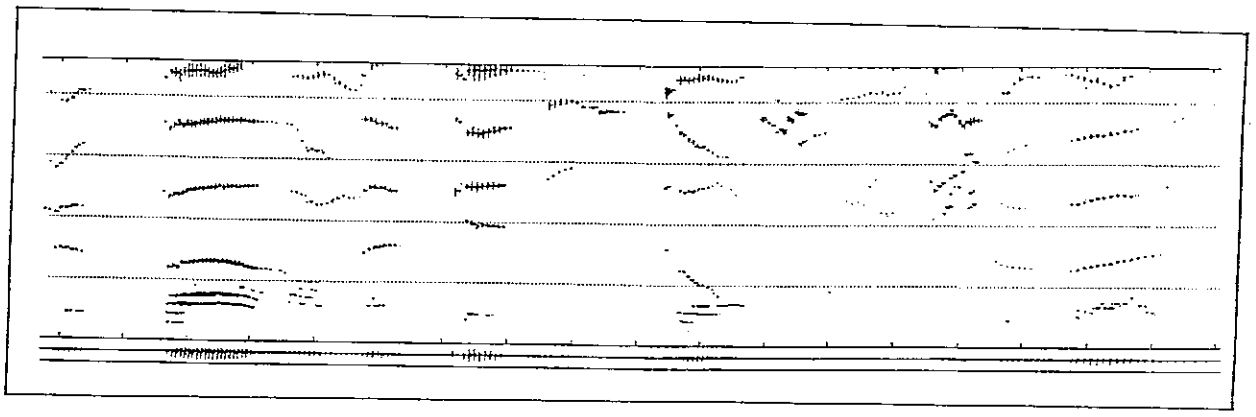

Figure 3: waveforms spotting in the time-frequency plane:
male voice speaking "je vais en Afganistan sur mon cheval".