Représentations temps-fréquence du signal de parole

Speech, Time-Frequency Representations

Christophe d'ALESSANDRO
LIMSI-CNRS,
BP 133, 91403 Orsay Cedex,
France

Docteur de l'Université Paris VI, en 1989. Chargé de recherche au CNRS depuis 1989, et responsable du thème « Analyse et synthèse de la parole » au LIMSI, laboratoire propre du CNRS. Principaux domaines d'intérêt : synthèse de la parole, représentation par ondelettes, perception et analyse de la voix chantée.

Christian DEMARS
LIMSI-CNRS,
BP 133, 91403 Orsay Cedex,
France

Docteur de 3e cycle, Université Paris VI, en 1972. Chercheur associé au LIMSI, laboratoire propre du CNRS, depuis 1983. Principaux domaines d'intérêt : représentation temps-fréquence instantanée, représentations temps-fréquence, analyse de la parole.

RÉSUMÉ

Le propos de cet article est de présenter une bibliographie récente sur l'utilisation des méthodes de représentation temps-fréquence en analyse et en traitement automatique de la parole. Les méthodes sont classées en trois grandes familles : méthodes dérivées de la production, méthodes d'analyse du signal, méthodes modélisant la perception. Après ce panorama, quelques rapides conclusions sur l'état actuel de l'utilisation de ces méthodes, et quelques perspectives sont tentées.

MOTS CLÉS

Représentation temps-fréquence, parole.

ABSTRACT

This paper presents a review on the use of time-frequency representations in the fields of speech analysis and automatic speech processing. Three main groups of methods are considered : speech production based methods, general signal analysis methods, auditory-based methods. After this review, some short conclusions on their current use, and on some possible future evolutions are proposed.

KEY WORDS

Time-frequency representations, speech.

1. Introduction

1.1. PROPOS

La nature du signal de parole conduit, pour le représenter, à l'utilisation d'une description de type *temps-fréquence*. En effet, l'évolution dans le temps de son contenu *spectral* est la condition nécessaire pour qu'un signal acoustique puisse véhiculer un code linguistique.

Cet article esquisse un panorama sur l'utilisation des méthodes temps-fréquence, dans le contexte particulier du signal de parole. Ce contexte comprend des contraintes issues de la production et de la perception de la parole,

ainsi que les besoins spécifiques du traitement automatique de la parole quant à la représentation acoustique. Le terme *représentations temps-fréquence* se comprend ici dans une acception assez vaste pour inclure diverses formes de représentations, et diverses notions qui se rattachent à celles de fréquence et de temps. Les deux principales catégories de lecteurs potentiels de ce travail de rassemblement sont les spécialistes du traitement de la parole d'une part et les spécialistes du traitement du signal d'autre part. Les premiers y trouverons sans doute des références à des travaux peu connus. Les seconds y trouveront le reflet de leurs préoccupations dans le domaine du traitement de la parole, ainsi qu'une présentation (élémentaire) des problèmes liées aux représentations temps-fréquence dans ce domaine. L'exposé des différentes méthodes sera très succint. Par contre nous espérons que cet article paraîtra assez explicite et complet, en offrant au lecteur :

1. l'essentiel de la bibliographie que les auteurs ont été capables de réunir sur les représentations temps-fréquence ayant eu au moins une application au traitement de la parole ;

2. une tentative d'organisation raisonnée de cette bibliographie, suivant des axes présentés dans cette introduction ;

3. une conclusion plus personnelle sur l'utilisation actuelle des différents types de représentations.

Les références bibliographiques sur les méthodes de représentation temps-fréquence sont aujourd'hui très abondantes. Dans cet article, nous avons volontairement réduit le nombre de références, en écartant des méthodes importantes, mais qui n'ont pas été appliquées à la parole. De même, pour les méthodes citées, un choix a été opéré afin de ne pas alourdir la bibliographie : il faut considérer cet article comme un ensemble de *clefs* pour aborder une bibliographie plus vaste. On trouvera dans les articles cités des compléments bibliographiques, et en particulier l'ensemble de nos sources bibliographiques dans [56].

La suite de cette introduction rappelle les différents domaines d'utilisation des représentations temps-fréquence du signal de parole, puis propose des axes de classification pour organiser ce vaste ensemble de méthodes. La section 2 décrit les représentations temps-fréquence qui utilisent un modèle du signal de parole, ou un modèle fonctionnel de production de la parole. La section 3 est consacrée aux représentations temps-fréquence qui ne font pas référence à un modèle a priori. La section 4 discute des représentations temps-fréquence obtenues par modélisation de l'analyse du signal acoustique dans le système auditif périphérique. La dernière section conclut.

1.2. UTILISATION DES REPRÉSENTATIONS TEMPS-FRÉQUENCE EN TRAITEMENT DE LA PAROLE

Il n'est pas inutile de préciser les domaines d'utilisation des représentations temps-fréquence du signal de parole. Le lecteur peu familier de l'étude de la parole trouvera dans l'ouvrage classique de Flanagan, dans la présentation plus succinte de Liénard, ou dans le récent ouvrage collectif de Calliope [31], [71], [121] des introductions à cette discipline. On peut tenter, de façon quelque peu arbitraire, un regroupement des activités en traitement automatique de la parole autour de cinq axes principaux :

1. Reconnaissance de la parole. Une représentation temps-fréquence du signal intervient pour paramétriser le signal acoustique. Une revue récente des tendances actuelles en reconnaissance automatique de la parole se trouve dans l'article de Mariani [138]. À l'exception de la reconnaissance par des systèmes experts en lecture de spectrogrammes, la quantité de données utilisées pour la paramétrisation acoustique est plutôt faible. les vecteurs (qui représentent une trame) d'analyse sont constitués généralement de 8 à 20 coefficients, comportant des informations spectrales et d'autres informations (énergie totale de la trame par exemple). Environ une centaine de vecteurs par seconde semble nécessaire et suffisant.

2. Analyse acoustique. Le problème est ici de mesurer des paramètres ou des indices acoustiques. La qualité de la mesure (résolution, biais, etc.) et la lisibilité des paramètres ou des indices acoustiques sur les coefficients ou les paramètres d'analyse sont fondamentaux, sans que le débit ou le temps de calcul soient obligatoirement des données critiques. L'analyse par synthèse est largement répandue, ce qui implique des représentations capables de reconstituer le signal.

3. Codage. Les méthodes temps-fréquence peuvent contribuer à réduire le débit du signal. Pour évaluer ces méthodes, il faut rapporter la qualité de la parole codée au coût du codage, en terme de débit, de puissance de calcul, de résistance au bruit et aux erreurs. De nombreuses revues partielles sur le codage existent. Des revues plus générales se trouvent dans le livre de Jayant & Noll [103] ou dans [31]. La possibilité d'interprétation physique des coefficients ou des paramètres d'analyse n'est pas impérative.

4. Traitement et modification du signal. Certaines applications impliquent de traiter ou de modifier le signal acoustique capté. Citons par exemple : la modification de la fréquence de voisement, des formants, des durées segmentales, la conversion de la voix d'un locuteur en la voix d'un autre locuteur, le rehaussement de parole noyée dans du bruit, la séparation de plusieurs sources, la compensation des particularités de la prise de son... Le lien avec un modèle de production apparaît fondamental dans la plupart des cas : pour modifier un paramètre acoustique (fréquence fondamentale, fréquence formantiques etc.), par exemple, il faut utiliser des méthodes qui en font un paramètre explicite. De même, des contraintes perceptives sont souvent incluses dans ce type de méthode (comme la courbe de sélectivité fréquentielle de l'oreille, par exemple).

5. Synthèse de la parole. Afin de générer automatiquement à partir d'un texte un signal de parole, des méthodes de codage, de modification et de description acoustique de la parole sont nécessaires. Ce problème emprunte donc en partie aux trois points précédents. Comme pour la reconnaissance, la représentation du signal n'est qu'un des éléments d'un système complet de synthèse, qui contient aussi des composantes phonétiques, phonologiques, lexica-

les, syntaxiques, voire sémantiques et pragmatiques. Une revue récente sur la synthèse de parole, pour l'anglais, se trouve dans l'article de Klatt [110].

1.3. CLASSIFICATION

1.3.1. Un triple point de vue sur le signal de parole

Pour son analyse ou son traitement par un algorithme, le signal acoustique apparaît d'abord comme un objet *mathématique*. Ce signal est conditionné par *l'appareil vocal* qui l'a produit, et par le *système auditif* auquel il est destiné. On ne peut faire abstraction de ce triple point de vue pour représenter le signal de parole. Ces trois points supposent des signaux classés par ordre de généralité décroissante, signaux vocaux pour les méthodes utilisant un modèle du signal de parole, signaux acoustiques dans la bande audible pour les méthodes utilisant des analogies auditives, signaux sans hypothèses particulières quant à leur provenance ou à leur nature pour les méthodes générales.

1.3.2. Fréquence, phase, échelle

Un axe important de classification des méthodes temps-fréquence est le sens précis attaché à la variable *fréquentielle*. La notion de frequence se rattache à la répétition d'un motif dans le temps : au mouvement périodique, dont l'archétype est le signal sinusoïdal. Au moins trois sortes de variables « fréquentielles » se déduisent de trois aspects de la sinusoïde :

1. Une première généralisation de la sinusoïde est de considérer un signal $x(t)$ d'énergie finie comme une somme d'exponentielles complexes, par transformation de Fourier (notée $\tilde{x}$) :

$$\tilde{x}(\nu) = \int x(t)\, e^{-2i\pi\nu t}\, dt \quad \text{et} \quad x(t) = \int \tilde{x}(\nu)\, e^{2i\pi\nu t}\, d\nu\,.$$

La *fréquence* ν au sens de Fourier, variable duale du temps, permet de repérer les composantes physiques du signal.

2. Une autre façon de généraliser le comportement d'une sinusoïde est d'exprimer le signal à l'aide d'une unique exponentielle complexe, modulée en amplitude et en phase. On peut, par exemple écrire le signal de parole $x(t)$ sous la forme de la partie réelle de son *signal analytique* x_a, complexe, déduit par suppression des fréquences négatives :

$$x(t) = \mathrm{Re}\ [x_a(t)] = \mathrm{Re}\ [\mathrm{A}(t)\, e^{-i\varphi(t)}]\,. \qquad (1)$$

De φ, la phase instantanée, on déduit la fréquence instantanée ν_i par dérivation par rapport au temps :

$$\nu_i = \frac{1}{2\pi}\frac{d\varphi}{dt}\,. \qquad (2)$$

L'enveloppe instantanée A possède le sens physique d'enveloppe temporelle du signal. La fréquence instantanée, ou vitesse de la phase, donne la pente de la courbure locale en temps du signal. Cette fréquence ne coïncide pas avec la fréquence au sens de Fourier en dehors du cas des signaux monochromatiques.

3. Un troisième aspect de la sinusoïde est sa périodicité : invariance par répétition d'une forme à des instants différents. Une généralisation de cet aspect est de rechercher une invariance de forme à des échelles temporelles et fréquentielles d'observation différentes. Le signal est alors représenté par une somme de fonctions obtenues par translation (aux instants τ) et dilatation temporelle (par les facteurs a) d'une fonction prototype o (c_0 est une constante) :

$$x(t) = \frac{1}{c_0}\iint p_{\tau,a}\,\frac{1}{\sqrt{a}}\,o\left(\frac{t-\tau}{a}\right)\frac{d\tau\,da}{a^2}$$

et

$$p_{\tau,a} = \frac{1}{\sqrt{a}}\int o^*\left(\frac{t-\tau}{a}\right)x(t)\,dt$$
$$= \sqrt{a}\int \tilde{o}(a\nu)\,\tilde{x}(\nu)\,e^{2i\pi\nu\tau}\,d\nu\,.$$

Les méthodes temps-échelle s'interprètent comme des méthodes à résolution fréquentielle relative constante.

Ces trois types fondamentaux de grandeur fréquentielle sont obtenus par des analyses non causales, qui impliquent la connaissance du signal sur une durée infinie (clairement pour les points 1 et 3, et dans le calcul de la transformée de Hilbert dans le point 2). D'autres sortes de fréquence ont donc été proposées, soit en considérant des analyses « à court terme », ou des « analyses glissantes », soit en proposant des procédures pratiques pour les obtenir, comme le comptage des passages du signal par des seuils, la détection de synchronies.

1.3.3. Date, temps

Un autre axe de classification est le sens de la variable « temps ». Deux classes de méthodes se distinguent par leur façon d'envisager l'évolution temporelle du contenu spectral. Pour les méthodes *adaptatives*, le temps est un ensemble de dates, ou instants de référence. A chaque date est associée une description spectrale du signal, localement considéré comme un signal stationnaire. Les méthodes *évolutives* considèrent le temps comme une des variables de la représentation du signal, sans hypothèse de stationnarité locale. Ce sera par exemple la variable d'une famille de fonctions sur laquelle le signal est décomposé.

1.3.4. Décomposition, distribution, paramétrisation

Deux voies sont possibles pour représenter un signal, sans référence à un modèle du signal. Le signal peut être représenté par la *distribution* spectro-temporelle de son contenu énergétique ou d'autres grandeurs physiques, comme ses propriétés instantanées. Il peut également être représenté par son comportement en regard d'une famille de fonctions, en utilisant une *décomposition :* comme somme de fonctions affectées de coefficients d'analyse.

Une troisième possibilité, lorsqu'on utilise un modèle du signal, est de le représenter par les paramètres du modèle, c'est la *paramétrisation*. Il faut noter que le terme « para-

métrisation » a déjà été employé ici avec un sens différent, car il désigne également l'étape d'obtention des vecteurs représentant le signal acoustique en reconnaissance de parole, avec ou sans modèle.

2. Modèles temps-fréquence de la parole

2.1. MODÈLE LINÉAIRE DE PRODUCTION DE LA PAROLE

Le signal de parole possède des propriétés remarquables dues au système particulier qui l'a produit, l'appareil phonatoire. L'ouvrage classique de Fant permet d'aborder la théorie acoustique de production de la parole [65]. La plupart des modèles temps-fréquence de la parole sont construits sur un modèle linéaire fonctionnel très simple de production de la parole décrit par Markel & Gray dans [140]. Le propos de ce paragraphe est de rappeler ce modèle, afin de situer les différentes méthodes de la section dans cette perspective commune. Les ordres de grandeurs spectraux-temporels de la production de parole seront également évoqués, afin de préciser la résolution utile dans ce domaine.

Trois composantes acoustiques interviennent dans la production du signal vocal : une source d'énergie acoustique, sa transformation par le conduit vocal, puis le rayonnement de cette énergie acoustique hors de la bouche et des narines. Trois types de sources sonores se combinent ou interviennent séparément pour produire de la parole : 1. la vibration des cordes vocales ; 2. l'écoulement turbulent de l'air au passage d'une constriction dans le conduit vocal, qui génère un bruit ; 3. une rapide occlusion dans le conduit vocal dont le relâchement génère une impulsion acoustique. Il est commode de simplifier le modèle acoustique linéaire de production en réunissant dans un même filtre les contributions de la glotte (U_g), du conduit vocal (V), et du rayonnement (L), et en réduisant l'excitation à un train périodique d'impulsions P, pour la parole voisée. Si l'on considère des signaux discrets, par transformée en z on obtient :

$$\begin{aligned} S(z) &= P(z)\, U_g(z)\, V(z)\, L(z) \\ &= P(z)\, H(z) \end{aligned}$$

une simplification identique est obtenue pour la parole non voisée en remplaçant l'excitation par une source de bruit blanc N :

$$\begin{aligned} S(z) &= N(z)\, V(z)\, L(z) \\ &= N(z)\, H(z) \end{aligned}$$

donc en regroupant les deux cas dans la source E :

$$S(z) = E(z)\, H(z)\,. \qquad (3)$$

Pour la parole voisée l'excitation possède un caractère périodique, et des propriétés particulières duent à la forme de l'onde de débit glottique (ou de façon équivalente de son spectre). Dans ce modèle simplifié, la propriété globale (train périodique d'impulsions ou bruit blanc) de l'excitation est comprise dans la source de spectre plat E, alors que les propriétés particulières de cette excitation, se joignent à l'action du conduit vocal et du rayonnement dans le filtre H. Ce filtre linéaire évolue dans le temps.

Les mouvements de l'appareil vocal induisent les caractéristiques et les ordres de grandeur des dimensions spectro-temporelles de la parole, qu'il n'est peut-être pas inutile de rappeler ici. La *fréquence fondamentale*, ou *fréquence de voisement*, notée F0, varie entre environ 80 et 150 Hz pour une voix d'homme, entre 150 et 250 Hz pour une voix de femme, et peut atteindre 400 Hz pour une voix d'enfant. Cette source d'excitation possède une pente spectrale globale d'environ – 12 dB/octave. Le conduit vocal agit comme un filtre acoustique : en dessous d'environ 4 kHz on considère une propagation monodimensionnelle (par ondes planes) de l'énergie acoustique. Cette cavité possède des résonances, et des anti-résonances (lors de l'utilisation de la dérivation nasale) dénommées *formants* et anti-formants. Dans le cas de la voyelle neutre (le « *e* » du Français), le conduit vocal peut être assimilé à un tube uniforme : le premier formant possède une fréquence centrale de 500 Hz, le second 1 500 Hz, le troisième 2 500 Hz etc. Dans le cas d'autres conformations, les fréquences centrales du premier formant varie entre 200 et 800 Hz, du second entre 600 et 2 000 Hz etc. Leurs largeurs de bande (à – 3 dB du sommet) varient entre quelques dizaines et quelques centaines de Hz. Les organes articulateurs (velum, langue, lèvres, mâchoires) provoquent des évolutions rapides de la conformation du conduit vocal, et donc de ses propriétés acoustiques. La vitesse d'évolution des fréquences centrales formantiques peut atteindre 1 kHz en 60 ms (16,7 kHz/seconde). Le terme de rayonnement représente la conversion de l'onde de débit aux lèvres et aux narines en onde de pression à une certaine distance de la tête. L'action de cette conversion peut être en première approximation assimilée à un filtrage de préaccentuation : spectralement, une pente d'environ + 6 dB/octave en résulte, induisant une pente globale (en tenant compte de la pente spectrale due à la source) d'environ – 6 dB/octave pour le signal de parole (ce qui limite de fait sa bande passante).

Le propos des représentations temps-fréquence qui font référence aux modèles de production de la parole est essentiellement d'identifier les paramètres liés aux sources d'excitation et au filtre linéaire évoluant dans le temps associé au conduit vocal, et les modèles qui vont suivre seront examinés à la lumière de cette décomposition.

2.2. MÉTHODES PARAMÉTRIQUES

2.2.1. Prédiction linéaire

La prédiction linéaire (LPC : Linear Predictive Coding) du signal est une technique largement utilisée en traitement de la parole [140] [85] [132] [131] [19].

Une simplification supplémentaire du modèle linéaire de production (3) amène à choisir le filtre H = 1/A comme un filtre tout-pôles :

$$S(z) = \frac{E(z)}{A(z)} \quad \text{avec} \quad A(z) = \sum_{i=0}^{M} a_i\, z^{-i}\,.$$

Le filtre obtenu par ce modèle linéaire simplifié de production est équivalent à un filtre prédictif, ou modèle auto-régressif du signal. En effet, la connaissance d'un certain nombre p d'échantillons jusqu'à un instant donné $n-1$ permet de prédire l'échantillon suivant, noté $\hat{s}_n$, avec l'erreur (ou le résiduel) de prédiction ε_n :

$$s_n = \hat{s}_n + \varepsilon_n \simeq \hat{s}_n = \beta_1 s_{n-1} + \beta_2 s_{n-2} + \cdots + \beta_p s_{n-p} . \quad (4)$$

La prédiction linéaire suppose que le filtre H est un filtre tout-pôle : il apparaît effectivement qu'un nombre de pôles suffisant permet d'analyser correctement les segments de parole qui exigeraient un filtre pôles-zéros, d'après la théorie acoustique (les voyelles nasales par exemple).

L'erreur de prédiction ε_n s'interprète en terme de source d'excitation e dans (3). La modélisation la plus simple de la source d'excitation en prédiction linéaire peut donc simplement être un générateur d'impulsion périodique ou un générateur de bruit blanc, avec une décision voisé/non voisé, qui interdit le mélange de ces deux sources.

2.2.2. Modèles ARMA, modèles d'excitation

Les défauts de qualité de l'analyse prédictive sont de deux natures :

1. simplification du filtre H (filtre tout-pôles) ;
2. simplification de l'excitation.

Un filtre H pôles-zéros s'obtient par un modèle ARMA (Auto-Régressif à Moyenne Ajustée) du signal. Des modèles ARMA adaptatifs [143] [79] ont été appliqués en synthèse de parole. Des modèles ARMA évolutifs, les coefficients du modèle ARMA dépendant du temps, par décomposition sur une base de fonctions [87] [88], ont été proposés par Grenier, en particulier pour la synthèse de parole également.

Par ailleurs, un modèle ARMA intervient en codage ADPCM (adaptative differential pulse code modulation). Un algorithme normalisé de codage à 32 kbit/s est décrit dans une norme du CCITT [36].

De nombreuses solutions ont été envisagées pour améliorer le modèle d'excitation, peut-être le problème le plus important pour améliorer la qualité de la représentation. Plusieurs familles parmi ces solutions s'attachent à coder le résiduel, en ne faisant plus de séparation entre parole voisée et non voisée. Ce résiduel est obtenu comme un ensemble d'impulsions adapté au signal dans le codage multi-impulsionnel introduit par Atal & Remde (Multi-Pulse Excited Linear Predictive Coding, MPLPC [20] [180] [118]), ou à l'aide d'un dictionnaire de résiduels (Code Excited Linear Predictive Coding, CELPC, codage introduit par Schroeder et Atal [172] [115]).

D'autres améliorations de l'excitation en LPC conservent la distinction entre signal voisé/non voisé, mais en utilisant des modèles réalistes d'onde de débit glottique pour l'excitation voisée, comme ceux de Rosemberg ou de Fant [66] [165]. Cette solution est plus satisfaisante du point de vue de la modélisation acoustique, pour la synthèse par exemple.

La LPC semble actuellement dominer pour le codage à bas et moyen débit. En reconnaissance, il est fréquent d'utiliser des paramètres acoustiques dérivés de la LPC, comme les coefficients LPC, les coefficients LAR : Log Area Ratio [140], les coefficients LSP : Line Spectrum Pairs [191] etc. En analyse acoustique, phonétique acoustique, synthèse, la LPC est largement utilisée pour estimer les formants [130] et les caractéristiques de la source de voisement. Cette méthode de séparation source/filtre permet par ailleurs de modifier le signal, en manipulant de façon séparée l'enveloppe spectrale (formants) et l'excitation [112] [164].

2.3. CEPSTRE

Le modèle linéaire source/filtre présente des composantes spectrales combinées de façon multiplicative. La méthode du *cepstre*, présentée par Oppenheim [149], permet de les séparer de façon additive, et ainsi d'estimer filtre et excitation [40]. Le cepstre réel est défini comme la transformée de Fourier inverse du logarithme du module de la transformée de Fourier d'un signal. Si, d'après 3 :

$$\log\left(\left|\tilde{S}(\nu)\right|\right) = \log\left(\left|\tilde{E}(\nu)\,\tilde{H}(\nu)\right|\right)$$
$$= \log\left(\left|\tilde{E}(\nu)\right|\right) + \log\left(\left|\tilde{H}(\nu)\right|\right)$$

on note $s(\kappa)$ le cepstre de $s(t)$, c'est-à-dire la transformée de Fourier inverse de l'expression précédente. Le cepstre dépend de la variable κ ou *quéfrence* qui est homogène à un temps. La transformation de Fourier est linéaire, donc :

$$s(\kappa) = e(\kappa) + h(\kappa) .$$

Alors que dans le domaine fréquentiel cette propriété d'additivité des contributions du filtre et de la source n'est pas utilisable, car elles se mélangent pour chaque fréquence, dans le domaine quéfrentiel la séparation est possible, si toutefois les échelles de temps de la source et du filtre sont assez différentes. Le conduit vocal possède une contribution fréquentiellement assez lisse, qui donne un cepstre basse-fréquence : les échantillons cepstraux correspondant au conduit vocal sont près de l'origine des quéfrences. Réciproquement, le spectre correspondant à la source d'excitation varie rapidement en fréquence, tant pour une excitation voisée que pour une excitation bruitée : le cepstre orrespondant sera haute-quéfrence. La séparation des deux composantes peut donc s'effectuer par une fenêtre quéfrentielle, ou *liftre* passe-bas pour le conduit vocal et *liftre* passe-haut pour la source d'excitation dans le domaine quéfrentiel, à condition que le premier formant soit assez éloigné du fondamental.

Le spectre de phase n'est pas considéré dans la version réelle du cepstre, et il est nécessaire pour la synthèse de faire des hypothèses supplémentaires : systèmes à phase minimum par exemple. Une autre forme de cepstre, le cepstre complexe, utilise le logarithme complexe de la transformée de Fourier, au lieu du logarithme réel du spectre d'amplitude.

Le cepstre est utilisé pour la reconnaissance, comme paramétrisation spectrale [86] [53] [106] [17] : coefficients MFCC (Mel Frequency Cepstrum Coefficients), coefficients LFCC (Linear Frequency Spectrum Coefficients), coefficients LPCC (Linear Predictive Cepstrum Coefficients). Le cepstre permet de détecter les formants, et F0 pour l'analyse acoustique ou la modification du signal [147] [94] [156].

2.4. MODÈLES SINUSOÏDAUX

Pour la parole voisée, l'excitation e est périodique. On peut donc la considérer comme une somme de composantes sinusoïdales. Deux méthodes principales ont développé cette représentation de deux façons différentes : le codage sinusoïdal et le codage harmonique.

2.4.1. Codage sinusoïdal

La représentation sinusoïdale de McAulay et Quatieri généralise la décomposition du signal comme somme de segments sinusoïdaux [127]. L'excitation $e(t)$, est exprimée sous forme d'une somme de sinusoïdes :

$$e(t) = \sum_{l=1}^{L(t)} a_l(t)\, e^{\,i\left(2\pi \int_{t_l}^{t} \nu_l(\tau)\, d\tau + \varphi_l\right)}$$

Le nombre de segments sinusoïdaux $L(t)$ dépend du temps, ainsi que les amplitudes a_l, les fréquences ν_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 $H(t, \nu)$, évoluant dans le temps :

$$H(t, \nu) = M(t, \nu)\, e^{i\phi(t, \nu)}$$

le signal $s(t)$ résultant du modèle complet s'écrit :

$$s(t) = \sum_{l=1}^{L(t)} A_l(t)\, e^{i\Psi_l(t)}$$

avec :

$$A_l(t) = a_l(t)\, M(t, \nu_l(t))$$

et :

$$\Psi_l(t) = 2\pi \int_{t_l}^{t} \nu_l(\tau)\, d\tau + \varphi_l + \phi(t, \nu_l(t))\,.$$

L'estimation des amplitudes, fréquences et phases, pour chaque composante, utilise la transformée de Fourier à court terme, ce qui est justifié pour la parole voisée. Pour la parole non voisée, une représentation sinusoïdale est valide, car le spectre à court terme reste proche de la densité spectrale de puissance du bruit. L'analyse se déroule trame par trame, en raccordant d'une trame à l'autre les segments sinusoïdaux. Le codage sinusoïdal délivre un signal perceptivement indiscernable de l'original, et représente donc une méthode de codage [127] [128]. Des modifications de fréquence fondamentale et de durée sont possibles en séparant sur chaque segment sinusoïdal les contributions de l'excitation et du filtre [129] [157].

2.4.2. Codage harmonique

Une autre extension de la décomposition harmonique de la parole voisée a été proposée par Almeida & Tribolet : le *codage harmonique* [13] [14] [192]. En généralisant le comportement spectral d'un train périodique d'impulsions à un train quasi-périodique d'impulsions, on peut définir des harmoniques généralisés comme composantes spectrales de $e(t)$, de rang k :

$$e(t) = \sum_{k} A(t)\, e^{ik\Phi(t)}\,.$$

Les amplitudes et les phases varient lentement dans le temps. A chaque harmonique généralisé de l'excitation, correspond un harmonique généralisé du signal voisé par application de $h(t, \tau)$. L'estimation des paramètres utilise une méthode proche de celle définie pour le codage sinusoïdal, mais nécessite de connaître la fréquence fondamentale. Le codeur complet utilise d'une part le codage harmonique, et d'autre part transmet un signal résiduel pour traiter des parties non voisées et corriger les erreurs des parties voisées.

Par ailleurs, le codage de l'excitation LPC par un ensemble de sinusoïdes a été proposé par Hedelin (Tone Oriented voice-excited LPC) [95]. Une représentation sinusoïdale composite qui est une méthode de prédiction linéaire modifiée a été introduite par Carayannis & Gueguen [32] [169].

2.5. MODÈLES FORMANTIQUES

2.5.1. Synthèse à formants

Dans la décomposition source/filtre du modèle acoustique de production, le filtre est caractérisé par les maxima de sa fonction de transfert, les formants. L'idée de représenter la parole par ses formants (comprenant trois paramètres, la fréquence centrale, la largeur de bande et l'amplitude) est apparue très tôt [159].

Deux formes particulières de synthétiseurs à formants coexistent : synthétiseur en série lorsque la fonction de transfert est une cascade de résonateurs, par exemple le synthétiseur de Klatt [109], et en parallèle lorsque les résonateurs sont disposés en parallèle, par exemple le synthétiseur d'Holmes [97]. En synthèse, le modèle en série possède moins de paramètres et semble peut-être moins souple d'utilisation, car il n'offre pas de contrôle sur les amplitudes des formants, bien qu'il soit théoriquement plus justifié pour les voyelles. La disposition des formants en parallèle introduit des interférences entre formants voisins, qui peuvent dégrader le signal désiré.

La synthèse à formant nécessite des sources d'excitation glottique et bruitée réalistes. Les paramètres de ces modèles sont généralement réactualisés à des intervalles temporels fixes, ou de façon synchrone au fondamental : ils n'évoluent pas pendant une durée de l'ordre de la période fondamentale.

Un modèle de représentation du signal de parole par des formants évoluant dans le temps a été proposé par Lee, extension évolutive de la représentation formantique en parallèle adaptative [116] [117].

La représentation formantique a été utilisée essentiellement, en synthèse (Klatt [110] présente une revue très complète du sujet) et en modification de la parole [164]. Les trajets formantiques sont également des paramètres pour la reconnaissance. Bien que l'économie remarquable des paramètres formantiques soit un argument important pour le codage, cette méthode est peu employée dans ce domaine : l'estimation automatique des formants et des paramètres d'excitation glottique reste délicate malgré l'emploi de techniques complexes.

2.5.2. Représentation par formes d'ondes

La représentation d'un signal comme somme de signaux élémentaires dans le plan temps-fréquence a initialement été étudiée par Gabor [80] [21]. Proche de la représentation formantique, dans le domaine temporel, des méthodes de représentation par formes d'ondes (méthodes GMC (Gaussian Modulated Cosine) de Markel [139], HMC (Hanning Modulated Cosine), ou de synthèse (Baumwolspiner [23]) ont été proposées. Les paramètres sont estimés par filtrage (formantique, ou en bandes d'octaves), puis détection des formes d'ondes sur les signaux filtrés, une application des modèles GMC et HMC au décodage acoustico-phonétique a été développée par Carey et Pfeifer [33] [152].

La synthèse par Formes d'Ondes Formantiques de Rodet [163], est une synthèse à formant en parallèle, mais considérée du point de vue temporel. Un modèle du signal de parole par formes d'ondes formantiques et sinusoïdales a été proposé pour la modification de la parole par d'Alessandro [6] [7]. Ce modèle s'appuie sur une décomposition fréquentielle suivant les maxima de l'enveloppe spectrale (formant) et du spectre de Fourier à court terme (harmoniques) en basse fréquence, puis sur une décomposition temporelle dans chaque bande d'analyse.

Enfin, un renouveau d'intérêt semble se manifester pour la décomposition des signaux, dont la parole, comme somme de signaux exponentiels [59] [151] [35].

3. Représentations non paramétriques

Cette section est consacrée aux représentations temps-fréquence construites sans référence à un modèle du signal. Ce vaste domaine est organisé ici en trois sections : les distributions énergétiques, les distributions liées à la phase instantanée, les décompositions.

La principale application des distributions énergétiques est l'analyse visuelle de diagrammes temps-fréquence représentant la parole : le prototype de ces représentations est donc le spectrogramme.

Les distributions liées à la phase instantanée offrent une alternative aux distributions énergétiques pour l'estimation d'un spectre fréquentiel : plusieurs applications à l'analyse de la parole utilisent la robustesse de cette alternative.

La décomposition du signal sur une famille de fonctions est très largement utilisée dans tous les domaines du traitement automatique de la parole. La forme la plus répandue est bien sûr l'analyse de Fourier, cependant, plusieurs autres familles possèdent des propriétés mathématiques intéressantes.

3.1. DISTRIBUTIONS ÉNERGÉTIQUES

La puissance instantanée $e(t)$ (resp. densité fréquentielle d'énergie $E(\nu)$) d'un signal est le carré de son module $e(t) = |s(t)|^2$ (resp. de son spectre $E(\nu) = |\tilde{S}(\nu)|^2$). Les distributions d'énergie dans le plan temps-fréquence doivent donc être des fonctions $P(t, \nu)$ qui représentent la densité spectro-temporelle d'énergie, c'est-à-dire dont la somme dans un domaine du plan temps fréquence représente la fraction d'énergie que possède le signal dans ce domaine. Des revues générales récentes des distributions temps-fréquence se trouvent dans les travaux de Cohen et Flandrin [45] [74]. La problématique de la distribution d'énergie en temps-fréquence est présentée dans les travaux de Ackroyd [1], et Lacoume & Kofman [113] [114].

3.1.1. Spectrogrammes

L'examen visuel de la distribution d'énergie dans le plan temps-fréquence est un outil fondamental pour l'analyse acoustique descriptive de la parole. Le prototype de la représentation énergétique est le spectrogramme, historiquement la première représentation temps-fréquence utilisée pour l'analyse visuelle de la parole. La première forme de spectrogramme, analogique, a fait partie du mobilier habituel des laboratoires d'acoustique ou de phonétique jusqu'à une date récente. Par cette antériorité historique, et cette diffusion considérable, une expertise très importante s'est développée pour interpréter les spectrogrammes. Un grand nombre de concepts et de techniques modernes, pour le traitement automatique de la parole ou en phonétique, sont directement issus des propriétés de cette analyse.

La version moderne du spectrogramme s'obtient en visualisant le spectre de puissance d'une transformée de Fourier à court terme :

$$S_x(t, \nu) = \left| \int x(\tau)\, w(t-\tau)\, e^{-2i\pi\nu\tau} d\tau \right|^2 .$$

La fenêtre w fixe la résolution spectrale et temporelle d'analyse. Deux types de fenêtre se sont dégagées en pratique, l'analyse en bande étroite (résolution fréquentielle d'environ 50 Hz), qui met en évidence la structure harmonique du signal voisé, et l'analyse en bande large (résolution fréquentielle d'environ 300 Hz), qui met en évidence la structure formantique du signal voisé, et les évolutions temporelles rapides. En général les spectrogrammes sont calculés par transformée de Fourier rapide, en utilisant 256 à 1 024 coefficients et 1 à 2 ms entre chaque analyse.

Il est bien connu que les spectrogrammes ne peuvent pas offrir une grande résolution simultanément en temps et en fréquence, à cause de la relation réciproque entre le support temporel et la bande passante de la fenêtre d'analyse. Cette « incertitude » fréquence-temps a été étudiée initialement par Gabor [80], une discussion plus récente a été proposée par Tsao [194], et Cohen & Lee dans le cas du spectrogramme [46]. Par l'utilisation de deux types de résolution, il n'est pas certain que cela soit un réel problème pour l'analyse de la parole. Sa capacité à suivre des variations formantiques rapides (comme celles rappelées au paragraphe 2.1) a fait l'objet d'une polémique récente [185] [201], dont il ressort que la résolution de l'analyse spectrographique est acceptable dans ce cas. De même que pour les autres distributions énergétiques, des termes d'interférence entre composantes spectrales ont été mis en évidence par Jeong & Williams [104] pour le spectrogramme : ce problème potentiel ne semble pas s'être posé pour les lecteurs de spectrogrammes.

L'analyse spectrographique a été utilisée directement pour la reconnaissance de parole, en tentant de modéliser la connaissance d'un lecteur de ce spectrogramme par un système expert (par Zue & Lamel [208]). Cette méthode est actuellement moins efficace que d'autres, mais garde un intérêt explicatif important. Par ailleurs des paramétrisations dérivées du spectrogramme sont souvent utilisées en reconnaissance : par exemple un spectrogramme en bande large, évalué toutes les 8 ou 10 ms, avec regroupement des canaux d'analyse sur une échelle Bark (voir par exemple, en analyse, le spectrographe auditif de Carlson & Granström [34]). En analyse et en synthèse, le spectrogramme est très largement utilisé pour l'estimation de paramètres acoustiques descriptifs.

3.1.2. Distribution de Wigner-Ville

La distribution de Wigner-Ville, présentée à la même période que le spectrogramme par Ville [198], connaît un regain d'intérêt depuis une dizaine d'années en analyse du signal (voir par exemple les articles de Claasen & Mecklenbräuker [42], de Flandrin & Escudié [72]). Elle est définie par :

$$W_x(t, \nu) = \int x^*(t - \tau/2)\, x(t + \tau/2)\, e^{-2i\pi\nu\tau}\, d\tau .$$

L'argument en faveur de cette représentation est sa résolution simultanée en temps et en fréquence, qui ne connaît pas les limitations du spectrographe.

De plus, la distribution de Wigner-Ville possède des propriétés théoriques qui la placent au cœur de la problématique des représentations non paramétriques bilinéaires, puisqu'il suffit qu'une représentation bilinéaire C soit invariante par translations temporelles et fréquentielles pour qu'elle s'exprime en fonction de la représentation de Wigner-Ville et d'un noyau spectro-temporel K :

$$C_x(t, \nu) = \iint W_x(\tau, \alpha)\, K(\tau - t, \alpha - \nu)\, d\alpha\, d\tau .$$

Toutes les distributions citées ici, le spectrogramme en particulier, peuvent s'exprime sous cette forme.

On peut la considérer comme une distribution d'énergie dans le plan-temps fréquence, puisque la sommation de la distribution en temps (resp. en fréquence) permet de retrouver la densité d'énergie fréquentielle (resp. la puissance instantanée). Cependant, cette interprétation énergétique est discutable à cause de l'existence de termes d'interférences importants entre les différentes composantes fréquentielles, apparents sur les diagrammes temps-fréquence, et de la présence de valeurs négatives locales sur la distribution.

Ces propriétés peu souhaitables peuvent être atténuées par utilisation du signal analytique à la place du signal réel, et par lissage. Une forme de lissage particulièrement intéressante met en œuvre un noyau séparable en temps et en fréquence, comme produit d'une fenêtre temporelle w et d'une fenêtre spectrale W, $K(t, \nu) = \nu(t)\, W(\nu)$, pour obtenir la distribution Pseudo Wigner-Ville Lissée :

$$PWVL_x(t, \nu) = \int w(\tau) \times \left[\int \nu(u - t)\, x^*(u - \tau/2)\, x(u + \tau/2)\, du \right] e^{-2i\pi\nu\tau}\, d\tau .$$

La distribution de Wigner-Ville a été proposée comme outil d'analyse de la parole par de nombreux auteurs [47] [38] [39] [200] [10] [197] [123] [161] [162]. Cependant, les formats d'affichage des diagrammes temps-fréquence utilisent des échelles temporelles et fréquentielles beaucoup plus fines que le spectrographe (c'est-à-dire des durées très brèves, de l'ordre d'un phonème, et/ou des bandes de fréquence étroite, de l'ordre de la largeur de bande d'un formant). Dans ces conditions, cette représentation montre clairement la supériorité de sa résolution, mais ne permet pas d'accéder aux évolutions plus globales du signal, comme le mouvement des formants entre plusieurs segments. Des expériences ont été également menées pour utiliser la distribution en reconnaissance [38] [39] [41]. Cette distribution n'a peut-être pas encore reçu toute l'attention quelle mérite en traitement de la parole.

3.1.3. Distributions de Page et de Margenau-Hill-Rhiaczek

La distribution de Page, ou « spectre instantané de puissance », est définie par [150] :

$$P(t, \nu) = \frac{\partial}{\partial t} \left| \int_{-\infty}^{t} x(\tau)\, e^{-2i\pi\nu\tau}\, d\tau \right|^2 .$$

La distribution de Margenau-Hill-Rhiaczek [160] [96] [107], appelée « densité d'énergie complexe » par Rhiaczek, est définie par :

$$R(t, \nu) = \tilde{x}_a^*(\nu)\, e^{-i\,2\pi\nu t}\, x_a(t) .$$

Ces deux distributions ne sont pas forcément positives pour toute valeur du temps et de la fréquence. Elles ont été comparées sur quelques exemples de parole à celles de Wigner-Ville par Cohen et Pickover [47]. Les formants

sont difficiles à extraire, surtout lorsque les fréquences centrales évoluent. Quand plus d'une composante fréquentielle est présente des termes d'interactions apparaissent pour la distribution de Rhiaczek. Cette dernière représentation a été utilisée pour l'analyse de données électrophysiologiques auditives [15] [63].

3.2. DISTRIBUTIONS UTILISANT LA PHASE INSTANTANÉE

La phase instantanée est bien adaptée pour l'analyse de signaux à bande étroite, bien que l'interprétation physique de la fréquence instantanée diffère de la notion intuitive de fréquence [183] [137] [92] [153]. En combinant phase instantanée et analyse de Fourier, on obtient des spectres qui offrent une bonne résolution, et plus de robustesse dans l'estimation de paramètres acoustiques que le spectre de puissance. Des revues de ce type de méthodes en parole se trouvent dans les travaux de Berthomier [27] et de Seggie [177].

3.2.1. Représentation du signal analytique

Il est clair que le signal analytique est un concept très général qui intervient dans de nombreux traitements. Ce paragraphe présente les applications qui l'utilisent de façon assez exclusive.

L'utilisation directe du signal analytique comme représentation permet de tracer un diagramme spiralé qui a été utilisé pour identifier visuellement des voyelles, mais semble-t-il sans résultats convaincants, par Lerner [119].

La fréquence instantanée d'une version filtrée passe-bas du signal, ou l'enveloppe instantanée, permettent d'extraire F0 [188] [158] [16] [25]. De même les fréquences formantiques sont accessibles par la fréquence instantanée dans une bande de fréquence autour du formant. La fréquence instantanée et l'enveloppe instantanée permettent de segmenter le signal en unités phonémiques [195] [178].

La modification d'amplitude instantanée et de phase instantanée a été proposée pour le rehaussement de parole en présence de bruit ou de réverbération [144] [199] [43].

3.2.2. Spectre de fréquence instantanée

Cette représentation utilise la phase de la transformée de Fourier à court terme : c'est la distribution de sa dérivée par rapport au temps. Si le spectre à court terme, avec une fenêtre w, vaut :

$$X(t, \nu) = a(t, \nu)\, e^{i\varphi(t, \nu)} = \int x(\tau)\, w(t-\tau) \exp^{-2i\pi\nu\tau} d\tau$$

alors la représentation a pour expression :

$$\nu_i(t, \nu) = \frac{\partial}{\partial t} \varphi(t, \nu) .$$

La distribution de fréquence instantanée permet de détecter des lignes spectrales, avec des performances meilleures qu'en utilisant le spectre d'amplitude. Dans une version à bande large, on obtient des diagrammes temps-fréquence où les formants sont très nets. Cette distribution a servi de paramétrisation en reconnaissance (Friedman [76] [77]). Une version à bande étroite permet de détecter les harmoniques de la parole voisée (harmonogrammes de Charpentier), pour l'extraction de F0 [37].

La distribution des phases de la transformée en cosinus discrète permet d'améliorer le codage par l'introduction de trajectoires, comme en codage sinusoïdal (Fjalbrant & Mekuria [69] [70]).

3.2.3. Spectre de temps de groupe

Le spectre de temps de propagation de groupe est la dérivée de la phase par rapport à la fréquence d'une transformée de Fourier à court terme :

$$\tau_i(t, \nu) = \frac{\partial}{\partial \nu} \varphi(t, \nu) .$$

Dans le produit spectral entre source d'excitation et filtre, la phase est la somme des arguments de chacun des facteurs, d'où les propriétés additives du temps de groupe qui est sa dérivée. De même, le temps de groupe d'une cascade de résonateurs est la somme des temps de groupe de chaque résonateur.

Le spectre de temps de groupe se calcule à partir de la transformée de Fourier à court terme, en utilisant une fenêtre temporelle, de durée inférieure à une période de voisement et le signal analytique. L'estimation des formants par le temps de groupe semble supérieure en résolution à celle obtenue par d'autres méthodes, surtout dans les cas difficiles : voix aiguës, formants proches, d'après les travaux de Yegnanarayana & coll. [61] [203] [145]. Une variante de cette estimation de formants par le spectre de phase a été proposée en LPC [202]. Des paramétrisations du signal pour la reconnaissance basées sur le temps de groupe ont été mises en œuvre par Itakura *et al.* [101] [181]. Le temps de groupe a été également utilisé pour le rehaussement de parole noyée dans du bruit [204].

3.2.4. Distribution temps-fréquence instantanée

En utilisant le signal analytique, une distribution temps-fréquence instantanée $T(r, \nu)$ est définie comme la somme de l'énergie (carré de l'enveloppe instantanée A) dans un domaine centré sur cette fréquence instantanée et cet instant [26].

On obtient alors dans le plan temps-fréquence instantanée une représentation de type spectrographique, nommée lambdagramme. Les lambdagrammes, dont l'interprétation n'est pas évidente en terme formantiques, ont été testés comme paramétrisation en reconnaissance par Demars [58] [57].

3.3. DÉCOMPOSITIONS PHYSIQUES DU SIGNAL

Les méthodes linéaires de représentation décomposent le signal sur une famille de fonctions. Nous distinguerons d'une part les décompositions physiques qui permettent d'étendre le sens physique de la fréquence, c'est-à-dire qui dépendent d'un paramètre localement interprétable

comme une fréquence, et d'autre part les décompositions qui ne permettent pas cette interprétation, ou décompositions mathématiques. Le prototype de la décomposition physique est la transformation de Fourier. La transformation en ondelettes en est le prolongement, en changeant la notion de fréquence locale en notion d'échelle.

3.3.1. Transformation de Fourier à court terme

Un sens local temporellement pour la notion de fréquence a initialement été proposé par Gabor [80] [22] [102]. Une transformée de Fourier à court terme pour définir le spectre fréquentiel du signal sur une plage temporelle localisée a été définie dans la continuité de ces travaux, en particulier par Portnoff [154] [155], Allen [51] et Crochière [11].

A partir du signal à une dimension $x(\tau)$ on définit un signal à deux dimensions $x^w(t, \tau) = x(\tau)\, w(t-\tau)$ qui représente, à t fixé, le signal x vu à travers la fenêtre d'analyse w centrée en t. La transformée de Fourier à court terme est la transformée de Fourier de x^w par rapport à la seconde variable τ :

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

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

On retrouve $x(\tau)$ en sommant les signaux à court terme sur tous les instants d'analyse :

$$x(\tau) = \int x^w(t, \tau)\, dt$$

donc, on reconstitue le signal original par transformée de Fourier inverse des signaux à court terme et sommation en temps :

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

Outre cette interprétation « par blocs », la transformée de Fourier à court terme possède une autre interprétation. On peut réécrire l'analyse comme une convolution :

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

ce qui permet d'interpréter la transformation de Fourier à court terme comme le filtrage linéaire, par le filtre (passe-bas) de réponse impulsionnelle $w(\tau)$ de $x(\tau)$ modulé par $e^{-2i\pi\nu\tau}$: $w(\tau)$ porte ainsi le nom de filtre d'analyse. La formule de synthèse correspond à la sommation d'un ensemble d'oscillateurs commandés, dans le temps et pour chaque fréquence, par les sorties des filtres passe-bas précédents.

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

La transformation de Fourier à court terme possède une contrepartie en terme de signaux échantillonnés en temps et en fréquence par utilisation de la transformation de Fourier discrète. La transformation de Fourier à court terme offre de très nombreuses variantes (par exemple la version modifiée de Griffin & Lim [89] [90]). Elle est très largement répandue pour l'analyse et le traitement de la parole, en particulier par le spectrographe, obtenu par visualisation du spectre de puissance à court terme. Par l'interprétation en banc de filtre, la transformation de Fourier à court terme est très largement utilisée dans les applications qui se rapportent à un filtrage, à cause de l'efficacité de la transformée de Fourier rapide. Un grand nombre d'autres méthodes temps-fréquence sont calculées par cette même transformée : distribution de Wigner-Ville, temps de groupe, fréquence instantanée, cepstre, représentations sinusoïdales, etc. La paramétrisation acoustique en reconnaissance utilise fréquemment cette représentation, soit directement (banc de filtres), soit indirectement (cepstre...). Les méthodes de codage sont largement dépendantes de la transformation de Fourier à court terme (codage sinusoïdal, harmonique, vocodeurs de phase, vocodeurs à canaux, codages par transformées...). C'est un puissant moyen de modification du signal vocal : modification des durées, de la mélodie, filtrage rehaussement de signaux bruités, déréverbération, etc.

3.3.2. Transformation par ondelettes

La transformée en ondelettes, ou méthode temps-échelle, représente le signal comme une somme de fonctions déduites par contraction temporelle d'une fonction prototype. Une revue de ce type de méthode se trouve dans [48], à la suite des travaux de Grossmann et Morlet [91] et de Daubechies [52], dans le domaine de l'analyse du signal, et de Meyer en analyse fonctionnelle [142].

Ce sont des méthodes à résolution fréquentielle relative constante, comme celle proposée antérieurement par Youngberg & Boll [206], et non à résolution constante. La nécessité d'observer un signal, comme la parole, à des échelles de résolutions différentes a été reconnue très tôt : voir par exemple les différentes résolutions d'analyse du spectrographe analogique (sonagraphes) des années quarante, les analyses acoustiques par filtres en tiers d'octaves.

L'utilisation d'échelle au lieu de fréquence dans une transformation linéaire permet dans une certaine mesure de contourner la limite de résolution due à l'incertitude Temps-Fréquence. En effet, si un phénomène a une bande passante importante, il sera observé à plusieurs échelles d'analyse, et donc à la fois finement en temps et en fréquence, à des échelles différentes.

La transformation peut recevoir deux interprétations, de même que la transformation de Fourier à court terme. La transformée en ondelettes apparaît comme un filtrage linéaire, par un banc de filtres linéaires possédant la propriété $\Delta\nu/\nu = \text{Cte}$, 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 $o(t)$. Par contre, l'interprétation par blocs de la transformée de Fourier à court terme n'a pas d'équivalent puisque la notion de durée d'analyse dépend de la fréquence. La structure de la transformée par ondelettes se retrouve dans le codage en sous-bandes utilisant des filtres miroirs en quadrature d'Esteban et Galland [64], utilisés pour le codage de parole. Ce type de codage non paramétrique (par banc de filtres), a été proposé antérieurement aux travaux sur les ondelettes, mais sont maintenant interprétés dans le cadre mathématique général de l'analyse multi-résolution, ou transformation en ondelettes discrètes, par les travaux de Mallat [133] [134]. La structure pyramidale, par bande d'octave, de la transformation a été exploitée pour construire des algorithmes assez efficaces de calcul. En prenant plusieurs bandes d'analyse par octave, et en visualisant les amplitudes et/ou phases des coefficients d'ondelettes, on obtient une représentation visuelle de type spectrographique (voir l'article de Kronland-Martinet [111]). L'échelle logarithmique en fréquence se rapproche des échelles fréquentielles auditives, comme il sera discuté dans la partie suivante, ce qui est en faveur de cette représentation. Un système d'analyse synthèse proche de l'analyse par ondelettes sous l'aspect de la décomposition du signal comme somme de fonctions élémentaires localisées a été proposé par Liénard [122]. L'analyse en ondelettes a été adaptée à une échelle auditive par d'Alessandro et Beautemps [8] [9].

3.4. DÉCOMPOSITIONS MATHÉMATIQUES DU SIGNAL

Les décompositions de cette section se situent à la limite du propos de cet article. Il s'agit des décompositions du signal sur des familles de fonctions optimales suivant différents critères. Ces décompositions ne permettent donc pas d'analyser des composantes physiques du signal, mais offrent des représentations mathématiques théoriquement satisfaisantes, par exemple des transformées orthogonales (une étude sur ces transformées orthogonales se trouve dans l'ouvrage de Ahmed [3]).

3.4.1. Décomposition sur une base de fonctions sphéroïdales aplaties

Les fonctions sphéroïdales aplaties $\psi_i(t)$, étudiées par Slepian & Pollack [186] [187] sont définies comme les fonctions propres d'un opérateur. Elles sont solutions des équations (avec les valeurs propres λ_i) :

$$\lambda_i \psi_i(t) = \int_{-T/2}^{T/2} \frac{\sin(2\pi B(t-\tau))}{\pi(t-\tau)} \psi_i(\tau)\, d\tau$$

où T est un intervalle temporel, B un intervalle fréquentiel. Ces fonctions forment un système complet et orthonormé pour les fonctions à bandes limitées dans [– B/2, B/2], et un système complet et orthogonal pour les fonctions de carré sommable sur [– T/2, T/2]. Ce sont des fonctions à bandes limitées, les meilleures fonctions de base lorsqu'on cherche à approcher des signaux dont le produit BT est connu par un nombre minimum de vecteurs de base indépendants. Ces fonctions réalisent la meilleure concentration simultanée d'énergie en temps et en fréquence.

L'application de cette décomposition en analyse de la parole a été esquissée par Roy [166]. Elles interviennent dans le calcul de modèles ARMA évoluants dans le temps de Grenier [87]. Une version discrète du développement sur une base de fonctions sphéroïdales aplaties a été étudiée pour le cryptage de la parole par Sridharan & coll. [189] [190].

3.4.2. Transformation de Karhunen-Loève

La transformée de Karhunen-Loève est une transformation orthogonale, le signal est développé sur une base de fonctions F_i [30] [103] :

$$x(n) = \sum_{i=0}^{N-1} C_i F_i(n)$$

qui sont les fonctions propres de la matrice de covariance. La décomposition de Karhunen-Loève est optimale pour plusieurs critères, elle est souvent prise comme base de comparaison pour d'autres transformations, dites sous-optimales : 1. c'est une transformation orthogonale ; 2. les coefficients C_i sont décorrélés ; 3. elle concentre le plus d'énergie sur peu de coefficients d'indices inférieurs, et ainsi l'erreur de troncature est plus faible qu'avec les autres transformations. Par contre, le principal inconvénient est que le calcul des valeurs propres de la matrice de covariance et des vecteurs propres associés doit être recommencé pour chaque tranche de signal, et qu'il n'y a pas d'algorithme rapide dans le cas général. Les applications à la parole, codage et encryptage, restent donc limitées [189] [190].

3.4.3. Transformation en cosinus discrète

Elle est définie par :

$$G_x(0) = \sqrt{2/M} \sum_{m=0}^{M-1} x(m)$$

$$G_x(k) = 2/M \sum_{m=0}^{M-1} x(m) \cos\left(\frac{(2m+1)k\pi}{2M}\right)$$

pour $k = 0, 1, \ldots (M-1)$.

Elle est largement utilisée [207] [2] [73] [3], [193] car les performances de cette transformée sont comparables à celles de Karhunen-Loève pour le critère du taux de distorsion : c'est une transformée sous-optimale. Une structure hiérarchique à deux niveaux a été utilisée pour la reconnaissance par Jallibrant & Mekuria [70]. Pour diminuer les effets de bloc en codage sans augmentation du débit, des transformées orthogonales à recouvrement ont été étudiées par Malvar & coll., qui sont liées à la transformée en cosinus [135] [136].

3.4.4. Transformation de Walsh-Hadamard

Elle est définie par [120] [24] :

$$Z_x = 1/N \sum_{i=0}^{N-1} x_i \, WAL(n, i).$$

WAL(n, i) étant les fonctions de Walsh, famille de fonctions à valeurs dans [– 1, + 1], et définies sur des intervalles convenablement choisis dans [1, N]. Il existe un algorithme rapide pour le calcul de la transformée de Walsh, sans multiplication. Cette transformée a été utilisée pour le codage, l'encryptage de parole [184]. Plusieurs applications à la reconnaissance ont été présentées [67] [196] [148].

3.4.5. Transformation d'Hermite modifiée

Les polynômes d'Hermite sont obtenus par dérivations successives de e^{-t^2}, et constituent une famille de fonctions orthogonales [141] [30]. La transformée d'Hermite modifiée fait intervenir une famille de polynômes orthogonaux, modifiés pour être normalisés et pour s'appliquer au cas discret. Le calcul de cette transformée ne demande que 2 N multiplications réelles (ou des divisions) pour un signal de N échantillons, et elle admet une transformée inverse qui utilise pour son calcul la transformée directe. Elle a été proposée par Akansu et Haddad [4] [5] [93] pour le codage adaptatif de la parole au moyen d'un nouveau schéma pour la transmission des signaux digitaux : SATC (Statistical Adaptative Transform Coding). Celui-ci utilise la statistique (moyenne, variance) des coefficients de la transformation pour adapter le codage aux changements statistiques du signal d'entrée. Cette transformée semble moins bonne en qualité que la transformée en cosinus discrète, mais la charge de calcul est plus faible.

3.4.6. Transformation de Tchebitcheff

Cette transformation utilise le développement du spectre d'un signal sur une base de polynômes de Tchebicheff. L'approximation d'une fonction par des polynômes de Tchebicheff est connue comme la meilleure au sens que l'erreur maximale est minimisée [44]. Le but est d'obtenir directement un espacement des canaux spectraux non uniforme, suivant les zéros des polynômes. L'échantillonnage du spectre concentre les canaux sur les basses fréquences, ce qui est assez avantageux, mais les espace trop en haute fréquence, et une technique complémentaire est utilisée pour rapprocher l'espacement des canaux de celui des bandes critiques. Cette représentation a été proposée pour l'analyse-synthèse, et dans une tâche assez restreinte de reconnaissance de phonème par Corr et Smith [50].

4. La modélisation auditive comme représentation

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

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 (échelles fréquentielles Mel, Bark), puis plus récemment en incorporant des modèles sophistiqués du système auditif périphérique et de perception de la parole. 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. En analyse acoustique ou en codage, l'introduction de contraintes auditives fait l'objet d'un net regain d'intérêt actuellement. Les mécanismes d'analyse acoustique mis en œuvre dans le système auditif ont souvent offert une source d'inspiration pour l'analyse du signal.

Trois étapes sont généralement considérées dans les modèles fonctionnels du système auditif périphérique, qui se rapportent à trois types de traitement et à trois divisions anatomo-physiologiques :

1. la transformation du signal acoustique présenté au pavillon en mouvement de la membrane basilaire : il est clair aujourd'hui que cette étape de filtrage comporte un caractère actif ;

2. la transformation du mouvement de la membrane basilaire en activité électrique dans le nerf auditif, sous forme de décharges électriques des neurones, ou potentiels d'action ;

3. l'interprétation de l'ensemble des décharges nerveuses par le système nerveux central. Cette étape est mal connue, mais diverses solutions ont été proposées pour délivrer une information de type spectral à partir de l'activité présente dans le nerf auditif.

Il est hors du propos de cette section de rappeler l'état des connaissances anatomiques et physiologiques sur la perception auditives, mais le lecteur intéressé pourra consulter [18] [12] [54] [108] [205] [167] [173].

Nous avons choisi quelques modèles du système auditif périphérique comme support pour cette section, ceux de : Caelen [28], Carlson et Granström [34], Cooke [49] [170], Dolmazon [18] [29] [60], Delgutte [55] [18], Ghitza [83] [82] [85], Goldhor [84], Lyon [124] [125] [126], Seneff [176] [175], Shamma [182].

4.1. FILTRAGE

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 le renforcement d'une large zone de fréquences (le pavillon se comporte comme un cornet auditif et le conduit auditif comme un résonateur, qui privilégie ainsi certaines plages fréquentielles). 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.

La modélisation de l'oreille externe et de l'oreille moyenne apparaît, dans la plupart des modèles fonctionnels du système auditif, 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 actif. Dans [28], un modèle actif est donné par un filtre variable commandé par un processus d'adaptation et par sa propre sortie (processus réflexe). Ce caractère actif est négligé dans la plupart des modèles fonctionnels, qui ne s'intéressent qu'à la modélisation cochléaire.

L'essentiel des propriétés spectro-temporelles du système auditif périphérique se concentre donc dans l'oreille interne. La cochlée possède une organisation tonotopique : une certaine fréquence a tendance à exciter préférentiellement une certaine zone le long de la membrane basilaire, position répartie de l'aigu vers le grave (de la fenêtre ovale vers l'apex). A cause de 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.

Le filtrage ainsi réalisé répond à une échelle d'allure logarithmique au-delà d'environ 800 Hz (analyse en 1/3 d'octaves), et d'allure linéaire en dessous (analyse en bande étroite, d'environ 100 Hz de largeur de bande). Il est courant de simuler cette analyse fréquentielle par un banc de filtres de largeur de bande 1 Bark [174], échelle issue de la psycho-acoustique, mais assez proche des échelles auditives physiologiques. Cette échelle fréquentielle est liée à la mesure des bandes critiques.

L'échelle 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 asymétrique, avec une coupure franche de l'aigu et une coupure beaucoup plus douce des graves [105].

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. Il a été démontré que le filtrage est actif et non linéaire.

Une application directe du filtrage cochléaire est le spectrographe auditif de Carlson & Granström, qui utilisant des filtres répartis sur une échelle Bark (largeur de bande de 1 Bark), et affiche l'amplitude sur une échelle dérivée des courbes d'isosonie (sensibilité de l'oreille à l'amplitude en fonction de la fréquence) [34]).

Dans d'autres modèles le gain des filtres suit fidèlement un modèle de courbe d'accord, et présente donc une asymétrie et une forme assez complexe, en suivant une échelle approximativement logarithmique [83], ou Bark [176].

Shamma prend en compte le spectre de phase, en particulier un saut de phase à la résonance qui rend compte de données physiologiques [182].

Des modèles introduisent un second filtre [49], hypothèse aujourd'hui presque abandonnée, pour rendre compte de la grande sélectivité du filtrage cochlaire et de certaines non-linéarités. On pense maintenant que c'est le caractère actif du filtrage qui induit cette sélectivité et ces non-linéarités, ainsi que le couplage entre membranes dans la cochlée, qui est pris en compte dans certains modèles [29].

Ghitza, dans une étude expérimentale sur l'utilisation d'un modèle d'oreille pour une tâche de reconnaissance semble montrer que des filtres passe-bande simples conviennent aussi bien que des filtres déduits des études physiologiques dans ce cas [82] ; l'apport du modèle d'oreille résiderait alors dans les traitements ultérieurs. Il est cependant clair que pour l'analyse acoustique et le codage, c'est la pondération du filtrage par des données auditives, au moins l'échelle fréquentielle, qui est déterminante.

4.2. TRANSDUCTION ENTRE MEMBRANE BASILAIRE ET NERF AUDITIF

Les rapports entre l'ensemble de potentiels d'action observés dans le nerf auditif et le signal excitateur montrent que le filtrage au niveau du nerf auditif n'est pas linéaire [168] : inhibition à deux tons (qui renforce les contrastes fréquentiels), sons subjectifs, sons de combinaison, adaptation à court terme (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, et les contrastes temporels sont renforcés), seuils de déclenchement et seuils de saturation des fibres en fonctions de l'intensité d'excitation.

Les potentiels d'action générés, suivant une certaine statistique, présentent la forme d'impulsions électriques, et ne dépendent que d'une phase d'excitation. La forme de l'impulsion est constante pour une fibre donnée pour tous les niveaux d'excitation. Ce phénomène est souvent modélisé par un redressement simple alternance.

La population des fibres nerveuses n'est pas homogène : un petit sous-ensemble de fibres semble posséder un taux de décharges spontanées faible, et donc un seuil d'intensité élevé, ce que certains modèles prennent en compte (par exemple Delgutte [55]). De plus, il est probable que certaines fibres jouent un rôle particulier de marquage de certains événements acoustiques.

Dans [83], [82], [81], la détection des fronts montants par un ensemble de niveaux, 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.

Dans [124], [125], [126] l'enveloppe des signaux issus de chaque bande d'analyse sont redressée (les parties négatives sont supprimées), et filtrée passe-bas (fréquence de coupure 1 kHz). Un mécanisme sophistiquée de contrôle automatique du gain, avec un 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.

Dans [176], [175], une compression non linéaire de

l'enveloppe temporelle des signaux filtrés est réalisée en utilisant deux mécanismes de contrôle de gain automatique répondant à l'équation :

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

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 ($\simeq 3$ ms) et le second un temps plus long ($\simeq 40$ ms). Un redressement simple alternance suit.

Dans [49] la transduction s'effectue en utilisant un modèle physiologique de génération et de migration d'agents électro-chimiques (ou *neutro-transmetteurs*) dans les cellules ciliées [173]. L'entrée du système est l'enveloppe du signal disponible après le traitement initial, et la sortie représente une courbe d'enveloppe après adaptation par le mécanisme sophistiquée des réservoirs multiples.

Dans [55] une détection d'enveloppe suit le filtrage, puis des non-linéarités sans mémoire représentent le taux de décharge en fonction du niveau d'excitation afin de 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. Un élément d'adaptation à court terme intervient ensuite, utilisant une constante de temps de 30 ms et une autre de quelques ms. Les phénomènes non linéaires comme la suppression à deux tons, et les sons subjectifs ne sont pas pris en compte.

4.3. EXTRACTION D'INDICES SPECTRAUX

La dernière étape traitée généralement par les modèles d'oreille est celle de l'interprétation de la représentation obtenue au niveau du nerf auditif : il s'agit de passer du signal représentant un ensemble de potentiels d'action à des indices acoustiques de type spectraux, évoluant dans le temps.

Deux mécanismes principaux semblent exister pour estimer ces indices spectraux. Le premier mécanisme est lié à l'organisation tonotopique du système : la mesure de l'énergie présente à un lieu donné (c'est-à-dire à une fréquence donnée) donne une première forme de spectre. C'est un spectre analogue à un spectre de puissance à court terme classique, sur une échelle auditive et avec l'introduction de non-linéarité.

Le second mécanisme est le 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 statistique des décharges d'une fibre dans une période du signal excitateur suit approximativement la forme de cette période. On peut donc obtenir un spectre par la mesure de synchronisation des phases instantanées, donc par l'évolution temporelle des phases instantanées dans une population de fibres : ce spectre s'apparente à un spectre de fréquence instantanée.

Un troisième mécanisme pour interpréter l'ensemble des décharges nerveuses, proposé par Shamma [182], s'appuie sur le changement de phase entre les fibres voisines, et s'apparente plutôt à un spectre de temps de groupe.

Le spectrographe auditif [34], outre l'analyse spectrale géographique qui résulte de la répartition d'énergie le long de la membrane basilaire, ajoute une analyse temporelle par recherche des fréquences dominantes dans chaque filtre d'analyse. L'amplitude est fonction du nombre de filtres dominés par la même fréquence. Ce modèle a été testé en resynthèse par Hukin & Damper [98].

Dans [83], [82], [81], nous avons vu que le taux de décharges est obtenu par le taux de passage par des seuils. En répartissant l'inverse des intervalles temporels entre passages d'un ensemble d'intervalles récents (les 20 derniers) sur 100 échantillons fréquentiels, une fréquence est obtenue pour chaque niveau de seuil de chaque fibre. La somme pour une fibre de toutes les fréquences à tous les niveaux de seuils 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 sommant des contributions de toutes les fibres : l'intensité à une fréquence mesure ainsi implicitement le nombre de régions dont le comportement est synchronisé. Il s'agit donc d'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.

Dans [142], [125], [126] 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'exploitation des relations entre les réponses des diverses fibres. Le modèle peut délivrer en sortie une représentation graphique, ou *neurogramme*, comparable à un spectrogramme.

Dans le modèle de [176], [175], après l'étape initiale du 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 2 ms (500 Hz) et 16 ms (62 Hz). Une variante de l'autocorrélation permet ainsi de calculer la fréquence fondamentale. Le modèle considère un codage temporel dans des bandes fréquentiellement déterminées (tonopie). 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, a fait l'objet de tests comme paramétrisation acoustique pour la reconnaissance par Hunt & Lefèvre [99], [100]. 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).

Dans le modèle de Cooke [49], 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. Il faut noter que la mesure d'une fréquence par les passages par des seuils d'amplitude du signal (passages par zéro, par exemple) est une technique qui a été proposée par divers auteurs pour le rehaussement de parole dans le bruit, la détection du fondamental, la reconnaissance [146], [75], [62].

Le modèle de Delgutte [55] 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.

5. Conclusions

Le propos de cet article est de présenter les transformées temps-fréquences qui ont fait l'objet d'au moins une application en parole. Ce panorama a été construit autour des applications à la parole, sans souci de cohérence théorique dans la description des représentations temps-fréquence. La conclusion tente quelques remarques plus personnelles.

5.1. LES MÉTHODES DOMINANTES ET LES DOMAINES

La première évidence, au vu du nombre important de méthodes proposées est qu'il n'existe pas une motivation ni une approche unique pour la représentation du signal de parole ! Il n'y a donc aucun sens à rechercher la meilleure méthode en général, mais celle qui répond le mieux à un problème donné. On peut par contre dresser le constat de l'utilisation actuelle des méthodes.

De façon quantitative, les deux méthodes dominantes sont dérivées de la prédiction linéaire et de la transformée de Fourier à court terme. La première méthode offre un modèle simple mais efficace de production. La seconde méthode, sans lien avec un modèle, rend dans une certaine mesure compte de la perception en respectant la notion de fréquence.

Le codage représente un cas particulier, car il n'utilise pas forcément de représentation temps-fréquence. Les formes économiques de codage, de qualité limitée (pour la restitution de messages dans les jouets, voitures, aéroports etc.) utilisent de préférence un modèle de la parole : la prédiction linéaire, sous différentes formes est très répandue. Pour le cryptage et le codage il est fait usage largement de décompositions, mais pratiquement pas de distributions temps-fréquence.

En reconnaissance, la prédiction linéaire, et les méthodes à base de transformée de Fourier à court terme (banc de filtre, cepstre) sont majoritaires, avec très souvent l'introduction de pondérations auditives. Il semblerait que la tendance actuelle se tourne plus vers la prise en compte de critères auditifs que vers l'utilisation de modèles de la parole. Les décompositions sont parfois utilisées, surtout la transformée en cosinus qui est proche de la transformée de Fourier. Les distributions le sont rarement.

L'analyse acoustique est le domaine de choix pour les distributions. La représentation de Wigner-Ville, qui donne une unité théorique aux distributions énergétiques, possède des avantages certains qui n'ont sans doute pas été exploités à leur mesure à l'heure actuelle. Le spectrogramme (et l'analyse sous-jacente par transformée de Fourier à court terme) a accumulé une masse considérable d'expertise, et il reste largement inégalé du point de vue de la diffusion, y compris dans des disciplines scientifiques ou médicales voisines. Les autres types de distributions, moins courantes, connaissent actuellement un intérêt certain, ainsi que l'utilisation de données auditives pour l'analyse. Les méthodes basées sur un modèle, en premier lieu la prédiction linéaire, sont des outils largement répandus. Les décompositions mathématiques sont naturellement d'un faible intérêt pour l'analyse acoustique.

Pour la modification et la synthèse de la parole, les méthodes basées sur un modèle sont les mieux adaptées. Cependant toutes les méthodes qui permettent de manipuler des composantes ou des paramètres acoustiques sont utiles.

5.2. RÉSOLUTION

Une représentation de la parole doit rendre compte des événements de production qui sont perceptivement importants. Cette relation production/représentation/perception est de toute première importance, comme le montrent les exemples de la prédiction linéaire (modèle simplifié de production qui favorise les formants, perceptivement dominants) ou la transformation de Fourier à court terme (analyse spectrale à court terme, perceptivement plausible, et qui fait apparaître les paramètres acoustiques de production de la parole).

Un problème important qui se déduit du point précédent est celui de la résolution d'analyse, qui varie selon le point de vue choisi (perception ou production). Les méthodes linéaires sont assujéties au compromis de résolution fréquence-temps. Pour contourner ce compromis, on peut utiliser une méthode paramétrique, une méthode bilinéaire, une méthode multirésolution. Dans le premier cas, le problème est la qualité du modèle et son adéquation à la

réalité ; dans le second, l'apparition de terme croisés « parasites », bien qu'ils puissent être utilisés ; dans le dernier cas, il faut que les événements acoustiques à analyser apparaissent à des échelles différentes pour que l'analyse multirésolution apporte plus de renseignements : c'est une analyse linéaire. Une autre façon de contourner le problème de la résolution est de combiner des analyses différentes, pour obtenir plusieurs points de vue sur le même signal. Actuellement, les méthodes adaptatives sont largement plus utilisées que les méthodes évolutives.

5.3. ROBUSTESSE

Aucune représentation n'est aujourd'hui capable de rendre compte de l'invariance perceptive remarquable du signal de parole par rapport à sa variabilité acoustique : la voyelle /a/ prononcée par divers locuteurs dans divers environnements, à diverses forces de voix (...) sera perçue comme un /a/ par la majorité des auditeurs, alors que le signal acoustique peut être extrêmement différent. Ce n'est évidemment pas un problème que le traitement du signal peut résoudre seul, puisque des niveaux supérieurs de traitement sont mis en jeu.

Cependant, il est probable que les améliorations futures des représentations temps-fréquence passeront par l'introduction de méthodes aptes à traiter ce paradoxe. De telles méthodes aujourd'hui s'inspirent de la modélisation auditives pour accroître la robustesse de l'estimation spectrale. Les modèles d'oreille combinent analyse tonotopique et analyse temporelle dans ce but, en utilisant des méthodes plutôt inexactes localement (comme l'estimation de la fréquence par des passages par zéro), mais dont la moyenne sur un grand nombre de mesures simultanées prend un sens. L'invariance perceptive du signal est mieux prise en compte lorsque l'on utilise des échelles fréquentielle proches des échelles perceptives.

D'autres critères importants pour le succès d'une représentation restent l'économie du codage et la rapidité des calculs.

Remerciements : En terminant ce panorama, les auteurs tiennent à remercier J. S. Liénard pour avoir initié et encouragé les travaux sur la représentation temps-fréquence de la parole au LIMSI, ainsi que pour leurs remarques et critiques, les trois experts anonymes qui ont examiné cet article.

Manuscrit reçu le 4 juillet 1991.

BIBLIOGRAPHIE

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

[2] N. Ahmed, T. Natarajan, K. R. Rao, 1974. *Dicrete cosine transform.* IEEE Transactions on Computers, January 1974, pp. 90-93.

[3] N. Ahmed, 1975. *Orthogonal transforms for digital signal processing.* Springer-Verlag, Berlin, 1975.

[4] A. N. Akansu, 1987. *The MHT, a new transform for SATC of speech signals.* Ph. D. dissertation, Polytech. Univ., Brooklyn, NY, June 1987.

[5] A. N. Akansu, R. A. Haddad, 1990. *On assymetrical performance of discrete cosine transform.* IEEE-ASSP, Vol. 38, No. 1, January, pp. 154-156.

[6] C. D'Alessandro, J. S. Lienard, 1988. *Decomposition of the speech signal into shirt-time waveforms using spectral segmentation.* Proceedings of IEEE-ICASSP-1988.

[7] C. D'Alessandro, 1990. *Time-frequency speech transformation based on an elementary waveform representation.* Speech Comm. Vol. 9, No. 6-7, December 1990, pp. 419-431.

[8] C. D'Alessandro, D. Beautemps, 1991. *Transformation en ondelettes sur une échelle fréquentielle auditive.* Actes du 13e colloque GRETSI, 16-20 septembre 1991.

[9] C. D'Alessandro, D. Beautemps, 1991. *Justification perceptive du spectrographe auditif.* Proceedings of XIIth ICPhS, Aix-en-Provence, 19-24 août 1991.

[10] J. F. Allard, C. Valiere, R. Bourdier, 1988. *Broadband signal analysis with the smoothed pseudo-Wigner distribution.* J.A.S.A. 83 (3), March 1988, pp. 1041-1044.

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

[12] J. B. Allen, 1985. *Cochlear modelling.* IEEE ASSP magazine, janvier 1985.

[13] L. B. Almeida, F. M. Silva, 1984. *Variable-frequency synthesis : an improved harmonic coding scheme.* Proceedings of IEEE-ICASSP-1984.

[14] L. B. Almeida, J. M. Tribolet, 1983. *Nonstationary modelling of voiced speech.* IEEE Transactions on ASSP, Vol. ASSP-31, No. 2, June 1983.

[15] R. A. Altes, 1978. *Possible reconstruction of auditory signals by the central nervous system.* (Abstract). J.A.S.A., Vol. 64, Suppl. I, S137.

[16] T. V. Ananthapadmanabha, B. Yegnarayana, 1975. *Epoch extraction of voiced speech.* IEEE-ASSP, Vol. 33, No. 6, December 1975, pp. 562-570.

[17] Y. Ariki, S. Mizuta, N. Nagata, T. Sakai, 1989. *Spoken-word recognition using dynamic features analysed by two-dimensional cepstrum.* IEEE Proceedings, Vol. 136, Pt. I, No. 2, April 1989, pp. 133-140.

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

[19] B. S. Atal, S. L. Hanauer, 1971. *Speech analysis and synthesis by linear prediction of the speech wave.* J.A.S.A., Vol. 50, No. 2, pp. 637-655.

[20] 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-IC ASSP-1982.

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

[22] M. J. Bastiaans, 1985. *On the sliding-window representation in digital processing.* IEEE ASSP, Vol. 33, No. 4, August 1985, pp. 868-873.

[23] M. Baumwolspineer, 1978. *Speech generation through waveform synthesis.* Proceedings of IEEE-ICASSP-1978.

[24] K. G. Beauchamp, 1984. *Applications of Walsh and related functions.* Academic Press, Orlando, Floride.

[25] C. Berthomier, 1979. *Calcul analogique de la fréquence du fondamental.* Proceedings of the 9th ICPhS Copenhagen, 6-11 August 1979, p. 260.

[26] C. Berthomier, 1983. *Instantaneous frequency and energy distribution of a signal.* Signal processing, 5, 1983, pp. 31-45.

[27] C. Berthomier, 1983. *Fréquence instantanée et représentation du signal de parole.* Actes du Séminaire « Traitement du signal de parole » ENST, Paris, 15-16 décembre 1983, pp. 119-124.

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

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

[30] S. J. Campanella, G. S. Robinson, 1971. *A comparison of orthogonal transformation for digital speech processing.* IEEE Trans. Commun. Technol., pp. 1045-1050. December 1971.

[31] Calliope, 1989. *La parole et son traitement automatique,* Masson, Paris.

[32] G. Carayannis, C. Gueguen, 1976. *The factorial linear modelling : a Karhumen-Loève approach to speech analysis.* Proceedings of IEEE-ICASSP-76.

[33] B. H. Carey, J. A. Howard, 1972. *A method for speech analysis by a wavefunction representation,* IEEE-Conference record on speech communication and processing, 1972.

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

[35] F. Casacuberta, E. Vidal, 1987. *A nonstationary model for the analysis of transient speech signals.* IEEE Transactions on ASSP, Vol. ASSP-35, No. 2, February 1987.

[36] CCITT, 1984-1988. *G series, Blue book, recommendation G721.* 32 kbit/s adaptative differential pulse code modulation (ADPCM).

[37] F. J. Charpentier, 1986. *Pitch detection using the short-term phase spectrum.* IEEE-ICASSP 86, Tokyo, pp 113-116.

[38] D. Chester, 1982. *The Wigner distribution and its application to speech recognition and analysis.* Ph. D. University microfilm inc. 1982.

[39] D. Chester, F. J. Taylor, 1984. *The Wigner distribution in speech processing applications.* Journal of the Franklin Institute, Pergamon Press, Vol. 318, No. 6, December 1984, pp. 415-430.i.

[40] D. G. Childers, D. P. Skinner, R. C. Kemerait, 1977. *The Cepstrum : a guide to processing.* Proceedings of the IEEE, Vol. 65, No. 10, October 1977, pp. 1428-1443.

[41] S. C. Chen, X. Yang, 1988. *Speech recognition with high recognition rate by smoothed spaced pseudo Wigner-Ville distribution (SSPWD) and overlap slide window spectrum methods.* ICASSP, New York, 11-14, April 1988, S5.3, pp. 191-194.

[42] T. A. C. M. Claasen, W. F. G. Mecklenbräuker, 1980. *The Wigner distribution.* A tool for time-frequency signal analysis.
Part 1 : Continuous-time signals. Philips Journal of Research, Vol. 35, No. 3, 1980.
Part 2 : Discrete-time signals. Philips Journal of Research, Vol. 35, No. 4-5, 1980.
Part 3 : Relations with other time-frequency signal transformations. Philips Journal of Research, Vol. 35, No. 6, 1980.

[43] P. M. Clarkson, S. Baghat, 1989. *Real time speech enhancement system using envelope expansion technic.* Electronics Letters, 17th August 1989, Vol. 25, No. 17, pp. 1186-1188.

[44] E. Cohen, 1981. *A spline approach to speech analysis/synthesis.* ICASSP 81 Atlanta, pp. 362-365.

[45] L. Cohen, 1989. *Time-Frequency distributions-A Review.* Proceedings of the IEEE, Vol. 77, No. 7, July 1989, pp. 941-981.

[46] L. Cohen, C. Lee, 1990. *Instantaneous bandwith for signals and spectrograms.* ICASSP 90, Albuquerque, pp. 2451-2454.

[47] L. Cohen, C. A. Pickover, 1986. *A comparison of joint time-frequency distributions for speech signals.* IEEE Inter. Symp. on Circuits and Systems, pp. 42-45.

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

[49] 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, December 1986.

[50] P. H. Corr, F. J. Smith, 1988. *A Chebushev transform : a new transform with particular application to speech.* Colloquium organised by PGE10 (Circuit, theory and design) on « Speech processing » at Savoy Place, 19, January 1988 by IEEE, Electronics Division.

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

[52] I. Daubechies, 1990. *The wavelet transform, time-frequency localization and signal analysis.* IEEE Transactions on IT, Vol. IT-36, No. 5, September 90, pp. 961-1005.

[53] S. B. Davis, P. Mermelstein, 1980. *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences.* IEEE Transactions on ASSP, Vol. ASSP-28, No. 4, August 1980.

[54] 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.

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

[56] C. Demars, 1990. *Représentation temps-fréquence et paramétrisations d'un signal.* Éléments de monographie. Notes et Documents LIMSI : 90-10, November 1990.

[57] C. Demars, J. L. Gauvain, 1983. *Application de la fréquence instantanée à la reconnaissance de mots isolés.* Actes du séminaire « Traitement du signal de parole » de la Sociéré Française d'Acoustique, ENST, Paris, décembre 1983, pp. 216-224.

[58] C. Demars, J. L. Gauvain, 1985. *Applications des L-distributions à la reconnaissance de la parole.* 14e JEP, Paris, juin 1985, pp. 283-287.

[59] L. Dolansky, 1960. *Choice of base signal in speech analysis.* IRE Transactions in Audio, November-December 1960.

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

[61] G. Duncan, B. Yegnarayana, H. A. Murthy, 1989. *A nonparametric method of formant estimation using group delay spectra.* IEEE-ICASSP 89, pp. 572-575.

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

[63] J. J. Eggermont, G. M. Smith, 1990. *Characterizing auditory neurons using the Wigner and Rihacek distributions : a comparison.* J.A.S.A., Vol. 87, No. 1, January 1990, pp. 246-259.

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

[65] G. Fant, 1970. *Acoustic theory of speech production Mouton,* La Hague-Paris.

[66] G. Fant, 1979. *Glottal source and excitation analysis.* STL-QPSR 1-1979.

[67] F. K. Feldman, T. Haque, 1991. *Development of Walsh linear coding and its application to speech recognition.* Speech Comm., Vol. 10, No. 1, 1991, pp. 91-97.

[68] T. F. Jallbrant, 1985. *A wide band approach to adaptive transform coding of speech signals.* A TMS320 signal processor implementation. Proceedings of the International Symposium on Circuits and Systems, Kyoto, 1985, pp. 321-324.

[69] T. F. Jallbrant, F. Mekuria, 1987. *Frequency domain phase derivatives used for data reduction in an ATC-system.* Digital signal processing 1987, V. Cappellini, A. G. Constantanides (Eds.) Elsevier Science Publishers B.V. (North-Holland), 1987, pp. 289-293.

[70] T. F. Jallibrant, F. Mekuria, 1989. *A hierarchical two-level analysis structure for use in speech coding and recognition.* IEEE-ICASSP 1989, May 1989, Glasgow, pp. 766-769.

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

[72] P. Flandrin, B. Escudie, 1985. *Principe et mise en œuvre de l'analyse temps-fréquence par transformation de Wigner-Ville.* Traitement du Signal, Vol. 2, No. 2, 1985, pp. 143-151.

[73] P. Flandrin, 1988. *Time-frequency and time-scale.* ASSP-Workshop on Spectrum Estimation and Modeling. 3-5, August 1988, Minneapolis, Minnesota, pp. 77-80.

[74] P. Flandrin, 1989. *Représentation temps-fréquence des signaux non stationnaires.* Traitement du Signal, Vol. 6, No. 2, 1989, pp. 89-101.

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

[76] D. H. Friedman, 1985. *Instantaneous frequency density vs time : an interpretaion of the phase structure of speech.* ICASSP 1985, Tampa 29.10.1-4, pp. 1121-1124.

[77] D. H. Friedman, 1987. *Formulation of a vector distance measure for the instantaneous-frequency distribution (IFD) of speech.* ICASSP 1987, pp. 1748-1752.

[78] H. Fujisaki, M. Ljungqvist, 1986. *Proposal of evaluation of models for the glottal source waveform.* Proceedings of IEEE-ICASSP-1986.

[79] 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-1987.

[80] D. Gabor, 1946. *Theory of communication.* Journal of the IEE, No. 93-146, Londres, pp. 429-441.

[81] 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, December 1986.

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

[83] 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-1985.

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

[85] C. Gueguen, 1985. *Analyse de la parole par des méthodes de modélisation paramétrique.* Annales de Télécommunications, Vol. 40, No. 5-6, 1985.

[86] A. H. Gray, J. D. Markel, 1976. *Distance measures for speech processing.* IEEE Transactions on ASSP, Vol. ASSP-24, No. 5, October 1976, pp. 380-391.

[87] Y. Grenier, 1983. *Time-dependant ARMA modeling of nonstationary signals.* IEEE Transactions on ASSP, Vol. ASSP-31, No. 4, April 1983.

[88] Y. Grenier, 1984. *Time-frequency analysis using time-dependant ARMA models.* Proceedings of IEEE-ICASSP 1984, San Diego, paper 41B5.

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

[90] D. W. Griffin, J. S. Lim, 1985. *A new model-based speech analysis/synthesis system.* Proceedings of IEEE-ICASSP-1985.

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

[92] M. S. Gupta, 1975. *Definition of instantaneous frequency and frequency measurability.* American Journal of Physics, Vol. 43, No. 12, December 1975, pp. 1087-1088.

[93] R. A. Haddad, A. N. Akansu, 1988. *A new orthogonal transform for signal coding.* IEEE Transactions on ASSP, Vol. 36, No. 9, September 1988, pp. 1404-1411.

[94] P. Halle, 1983. *Techniques cepstrales améliorées pour l'extraction d'enveloppe spectrale et détectiondu pitch.* Actes du séminaire « Traitement du signal de parole » de la Société Française d'Acoustique, ENST, Paris, December 1983, pp. 83-93.

[95] P. Hedelin, 1981. *A tone-oriented voice-excited vocoder.* Proceedings of IEEE-ICASSP-1981.

[96] R. D. Hippenstiel, P. M. De Oliveira, 1990. *Time varying spectral estimation using the instantaneous power spectrum (IPS).* IEEE ASSP, Vol. 38, No. 10, October 1990, pp. 1752-1759.

[97] J. Holmes, 1983. *Research report, formant synthesizer : cascade or parallel.* Speech Comm., Vol. 2, No. 4, pp. 251-273.

[98] R. W. Hukin, R. I. Damper, 1989. *Testing an auditory model by resynthesis.* Proceedings of ESCA-Eurospeech'89, Paris, September 1989, pp. 243-246.

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

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

[101] F. Itakura, T. Umezaki, 1987. *Distance measure for speech recognition based on the smoothed group delay spectrum.* IEEE-ICASSP 1987, 29.1.1-29.1.4, pp. 1257-1260.

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

[103] N. S. Jayant, P. Noll, 1984. *Digital coding of waveforms.* Englewood Cliffs, N. J. Prentice Hall, 1984.

[104] J. Jeong, W. J. Williams, 1990. *On the cross-terms in spectrograms.* Proceedings of IEEE-ICASSP 1991, pp. 1565-1568.

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

[106] J. C. Junqua, H. Wakita, 1989. *A comparative study of cepstral lifters and distance measures for all pole models of speech in noise.* IEEE-ICASSP 89, Glasgow, pp. 476-478.

[107] P. Johannesma, A. Erten, B. Cranen, L. Van Erning, 1981. *The phonochrome a coherent spectro-temporal representation of sound.* Hear. Res., Vol. 5, pp. 123-145.

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

[109] D. H. Klatt, 1980. *Software for a cascade/parallel formant synthesizer.* JASA, Vol. 67, No. 3, March 1980, pp. 971-995.

[110] D. H. Klatt, 1987. *Review of text-to-speech conversion for English.* JASA, Vol. 82, No. 3, September 1987.

[111] 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, december 1988.

[112] H. Kuwabara, 1984. *A pitch-synchronous analysis-synthesis system to independently modify formant frequencies and bandwidth for voiced speech.* Speech Comm. Vol. 3, pp. 211-220.

[113] J. J. Lacoume, W. Kofman, 1975. *Description des processus non stationnaires par la représentation temps-fréquence : Applications.* GRETSI, Nice, 16-21 juin 1975, pp. 95-101.

[114] J. L. Lacoume, W. Kofman, 1975. *Étude des signaux non stationnaires par la représentation en temps et en fréquence.* Annales des Télécom., 30, No. 7-8, 1975, pp. 231-238.

[115] A. Le Guyader, M. Massaloux, M. Zurcher, 1983. *A robust and fast CELP coder at 16 kbit/s.* Speech Comm., Vol. 7, No. 2, pp. 217-226.

[116] Y. T. Lee, H. F. Silverman, 1988. *On a general time-varying model for speech signals.* IEEE-ICASSP 88, New York, 11-14 April 1988, S2.9, pp. 95-98.

[117] Y. T. Lee, 1987. *A general time-varying model for speech signals and estimation of its parameters.* Thèse de Ph. D., December 1987, publiée comme rapport technique LEMS-40, Laboratory for Engineering Man/Machine Systems, Brown University, Providence, January 1988.

[118] J. P. Lefevre, O. Passien, 1985. *Efficient algorithms for obtaining multipulse excitation for LPC coder.* Proceedings of IEEE-ICASSP-1985.

[119] R. M. Lerner, 1959. *A method of speech compression.* Doctoral Thesis, M.I.T., August 1959.

[120] J. Lifermann, 1979. *Les méthodes rapides de transformation du signal : Fourier, Walsh, Hadamard, Haar.* Masson, Paris, 1979.

[121] J. S. Lienard, 1977. *Les processus de la communication parlée.* Masson, Paris.

[122] J. S. Lienard, 1987. *Speech analysis and reconstruction using short-time, elementary waveforms.* Proceedings of IEEE-ICASPP-1987.

[123] D. Lowe, M. J. Tomlinson and R. K. Moore, 1986. *The Wigner distribution as a speech signal processing tool.* Proceedings of the Institute of Acoustics, Autumn Conference on Speech and Hearing, Winermere, 28-30, November 1986.

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

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

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

[127] R. J. McAulay, T. F. Quatieri, 1986. *Speech analysis/synthesis based on a sinusoidal representation.* IEEE Transactions on ASSP, Vol. ASSP-34, No. 4, August 1986.

[128] R. J. McAulay, T. F. Quatieri, 1986. *Phase modelling and its application to sinusoidal transform coding.* Proceedings of IEEE-ICASSP-1986.

[129] R. J. McAulay, T. F. Quatieri, 1987. *Mixed-phase deconvolution of speech based on a sine-wave model.* Proceedings of IEEE-ICASSP-1987.

[130] S. S. McCandless, 1974. *An algorithm for automatic formant extraction using linear prediction spectra.* IEEE Transactions on ASSP, Vol. ASSP-22, No. 2, April 1974.

[131] J. I. Makhoul, L. K. Cosell, 1981. *Adaptive lattice analysis of speech.* IEEE Transactions on Circuit and System, Vol. CAS-28, No. 5, June 1981.

[132] J. I. Makhoul, 1975. *Linear prediction : a tutorial review.* Proceedings of IEEE, 1975, pp. 561-580.

[133] S. G. Mallat, 1989. *A theory for multiresolution signal decomposition : the wavelet representation.* IEEE Transactions on PAMI, Vol. PAMI 31, 1989, pp. 674-693.

[134] S. G. Mallat, 1989. *Multiresolution approximation and wavelets.* Trans. Am. Math. Soc., Vol. 135, 1989, pp. 69-88.

[135] H. S. Malvar, D. H. Staelin, 1989. *The LOT : transform coding without blocking effects.* IEEE-ASSP, Vol. 37, No. 4, April 1989, pp. 553-559.

[136] H. S. Malvar, 1990. *Lapped transforms for efficient transform/sibband coding.* IEEE-ASSP, Vol. 38, No. 6, June 1990, pp. 969-978.

[137] L. Mandel, 1974. *Interpretation of instantaneous frequency.* American Journal of Physics, Vol. 42, October 1974, pp. 840-846.

[138] J. Mariani, 1990. *Reconnaissance automatique de la parole : progrès et tendances.* Traitement du Signal, Vol. 7, No. 4, pp. 239-266.

[139] 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, July 1970.

[140] J. D. Markel, A. H. Gray, 1976. *Linear prediction of speech.* Springer-Verlag, Berlin.

[141] J. B. Martens, 1990. *The Hermite transform-theory.* IEEE-ASSP, Vol. 38, No. 9, September 1990, pp. 1595-1606.
The Hermite transform-applications. IEEE-ASSP, Vol. 38, No. 9, September 1990, pp. 1607-1618.

[142] Y. Meyer, 1990. *Ondelettes et opérateurs.* Hermann Ed., Paris.
The Hermite transform-applications. IEEE-ASSP, Vol. 38, No. 9, September 1990, pp. 1607-1618.

[143] H. Morikawa, H. Fujisaki, 1986. *A speech analysis-synthesis system based on the ARMA model and its evaluation.* IEEE-ICASSP 1986, pp. 1253-1256.

[144] J. Mourjopoulos, J. K. Hammond, 1983. *Modelling and enhancement of reverberant speech signals using an envelope convolution model.* IEEE-ICASSP 1983, pp. 1144-1147.

[145] H. A. Murthy, K. V. Madhu Murthy, B. Yegnarayana, 1990. *Formant extraction from phase using weighted group delay fonction.* Electronics Letters, 9th, November 1989, Vol. 25, pp. 1609-1611.

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

[147] A. M. Noll, 1967. *Cepstrum Pitch Determination.* The Journal of Acoustical Society of America, Vol. 41, No. 2, pp. 293-309.

[148] H. Ohga, H. Yabuuchi, E. Tsuboka *and al.,* 1982. *A Walsh-Hadamard transform LSI for speech recognition.* IEEE Trans. Consumm. Elect., Vol. CE-28, pp. 263-270.

[149] A. V. Oppenheim, 1969. *Speech analysis-synthesis system based on homorphic filtering.* J.A.S.A., Vol. 45, No. 2, 1969, pp. 458-465.

[150] C. H. Page, 1952. *Instantaneous power spectra.* Journal of applied physics. Vol. 23, No. 1, January 1952, pp. 103-106.

[151] S. Parthasarathy, D. W. Tufts, 1987. *Signal modeling by exponential segments and application in voiced speech analysis.* Proceedings of IEEE-ICASSP 87, pp. 645-648.

[152] L. L. Pfeifer, 1972. *Isolated-word phoneme recognition using features derived from wavefunction parameters.* IEEE Conference Record on Speech Communication and Processing. Newton, 24-26 April 1972, pp. 93-96.

[153] B. Picinbono, W. Martin, 1983. *Représentation des signaux par amplitudes et phases instantanées.* Ann. Télécom., Vol. 28, No. 5-6, 1983, pp. 179-190.

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

[155] M. R. Portnoff, 1981. *Time-frequency of speech based on short-time Fourier analysis.* IEEE Transactions on ASSP, Vol. ASSP-29, No. 3, June 1981.

[156] T. F. Qautieri, 1979. *Minimum and mixed phase speech analysis-synthesis by adaptive homomorphic deconvolution.* IEEE Transactions on ASSP, Vol. ASSP-27, No. 4, August 1979.

[157] T. F. Quatieri, R. J. Mcaulay, 1986. *Speech transformations*

based on a sinusoidal representation. IEEE Transactions on ASSP, Vol. ASSP-34, No. 6, December 1986.

[158] C. H. Rader, 1964. *Vector pitch detection.* J.A.S.A. 36(C) 1963 L 1964.

[159] L. R. Rabiner, 1968. *Digital-formant synthesizer for speech synthesis.* JASA, Vol. 43, 1968.

[160] A. W. Rihaczek, 1968. *Signal energy distribution in time and frequency.* IEEE Transactions on Information Theory, Vol. IT-14, No. 3, May 1968, pp. 369-374.

[161] M. D. Riley, 1987. *Beyond quasi-stationarity : designing time-frequency representations for speech signals.* IEEE-ICASSP 87, pp. 657-660.

[162] M. D. Riley, 1989. *Speech time-frequency representations.* Kluwer Academic publishers, Boston, 1989.

[163] X. Rodet, 1980. *Time-domain formant-wave-function synthesis Spoken language generation and understanding.* J. C. Simon éditeur, D. Reidel publishing company, Dordrecht, Hollande.

[164] X. Rodet, P. Depalle, 1985. *Synthesis by rule : LPC diphones and calculation of formant trajectories.* Proceedings of IEEE-ICASSP 85, Tampa, pp. 736-739.

[165] A. E. Rosenberg, 1971. *Effect of glottal pulse shape on the quality of natural vowels.* J.A.S.A. Vol. 49, No. 2, 1971.

[166] B. Roy, 1979. *Contribution à l'étude de la dimension d'un signal approximativement limité en temps et en fréquence.* Thèse de docteur ingénieur INPG, Grenoble 1979.

[167] 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, August 1979.

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

[169] S. Sagayama, F. Itakura, 1986. *Duality theory of composite sinusoidal modeling and linear prediction.* Proceedings of IEEE-ICASSP-86.

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

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

[172] 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, pp. 937-940.

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

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

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

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

[177] D. Seggie, 1987. *The application of analytic signal analysis in speech processing.* Proceedings of The Institute of Acoustics, Vol. 8, Part. 7 (1986), pp. 82-85.

[178] D. Seggie, 1987. *Analysis of speech signal envelope-frequency relationships.* European Conference on Speech Technology, Edinburg, September 1987, Vol. 1, pp. 38-41.

[179] D. Seggie, 1987. *The use of temporal frequency in speech signal analysis.* Proceedings of The Eleventh International Congress of Phonetic Sciences, August 1-7, 1987, Tallinn, Estonia, U.S.S.R., pp. 364-367.

[180] S. Singhal, B. S. Atal, 1984. *Improving performance of multi-pulse LPC coders at low bit rates.* Proceedings of IEEE-ICASSP 84, San Diego, paper 1.3.

[181] H. Singer, T. Umezaki, F. Itakura, 1990. *Low bit quantization of the smoothed group delay spectrum for speech recignition.* IEEE-ICASSP 90, Albuquerque, pp. 761-764.

[182] S. A. Shamma, 1985. *Speech processing in the auditory system I : The representation of speec sounds in the responses of the auditory nerve. II : Lateral inhibition and the central processing of speech evoked activity in the auditory nerve.* J.A.S.A. Vol. 78, No. 5, November 85, pp. 1612-1621, 1622-1632.

[183] J. Shekel, 1953. *« Instantaneous » frequency.* Proceedings of the I.R.E., April 1953, p. 548.

[184] F. Y. Y. Shum, A. R. Elliott, W. O. Brown, 1973. *Speech processing with Walsh-Hadamard transforms.* IEEE, Transactions Audio. Electroacoustics, Vol. AU-21, No. 3, June 1973, pp. 174-179.

[185] 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, pp. 63-86.

[186] D. Slepian, H. O. Pollak, 1961. *Prolate spheroidal wave functions, Fourier analysis and uncertainty-part I, part II, part III.*
The Bell system technical journal, January 1961 part I : pp. 43-63, part II : pp. 65-84.
The Bell system technical journal, July 1962 part III : pp. 1295-1336.

[187] D. Slepian, 1978. *Prolate spheroidal wave functions, Fourier analysis and uncertainty, V :* The discrete case. Bell Syst. Tech. J., Vol. 57, No. 5, pp. 1371-1430.

[188] M. M. Sondhi, 1964. *Equivalence of « vector » and autocorrelation pitch detectors.* J.A.S.A. 36(C) 1964 L 1964.

[189] S. Sridharan, E. Dawson, B. Goldburg, 1990. *Speech encryption in the transform domain.* Electronics Letters, 10th May 1990, Vol. 26, No. 10, pp. 655-657.

[190] S. Sridharan, E. Dawson, B. Goldburg, 1990. *Speech encryption using discrete orthogonal transforms.* ICASSP 90, Albuquerque, pp. 1647-1650.

[191] N. Sugamura, H. Fujisaki, 1986. *Speech analysis and synthesis method developped at ECL and NTT.* From LPC to LSP. Speech Comm., Vol. 5, pp. 199-298.

[192] 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.

[193] J. Rribolet, R. E. Cochiere, 1979. *Frequency domain coding of speech.* IEEE-ASSP, October 1979, pp. 512-530.

[194] Y. H. Tsao, 1984. *Uncertainty principle in frequency-time methods.* J.A.S.A. Vol. 75, No. 5, 1984, pp. 1532-1540.

[195] A. Tsopanoglou, J. Mourjopoulos, G. Kokkinakis, 1989. *Continuous speech phoneme segmentation method based on the instantaneous frequency.* EUROSPEECH 89, Paris, 26-28 September 1989, pp. 67-70.

[196] J. Tyler, 1986. *Speech recognition system using Walsh analysis and dynamic programming.* Microprocessors and microsystems, Vol. 10, No ; 8, October 1986, pp. 427-431.

[197] E. F. Velez, R. G. Absher, 1989. *Transient analysis of speech signals using the Wigner time-frequency representation.* ICASSP, May 89, Glasgow, pp. 2242-2245.

[198] J. Ville, 1948. *Théorie et applications de la notion de signal analytique.* Câbles et transmissions, Vol. 2, 1948, pp. 61-74 (hors commerce).

[199] 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, pp. 279-283.

[200] 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.

[201] W. Wokurek, 1991. *Comments on « On the spectrographic representation of rapidly time varying speech »*. Computer speech and language, Vol. 5, No. 1, January 91, pp. 1-10.

[202] B. Yegnanarayana, 1978. *Formant extraction from linear-prediction phase spectra*. J.A.S.A. Vol. 63, No. 5, pp. 1638-1640.

[203] 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.

[204] B. Yegnanarayana, H. A. Murthy, V. R. Ramachandran, 1990. *Speech enhancement using group delay functions*. ICLSP 90, Kobé, 18-22 novembre 90, pp. 301-304.

[205] 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.

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

[207] R. Zelinski, P. Noll., 1977. *Adaptative transform coding of speech signals*. IEEE-ASSP, Vol. ASSP-25, No. 4, August 1977, pp. 299-309.

[208] V. W. Zue, L. F. Lamel. *An expert spectrogram reader : a knowledge-based approach to speech recognition*, IEEE-ICASSP 86, pp. 1197-1200, Tokyo.