ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ À
L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

COMME EXIGENCE PARTIELLE
À L'OBTENTION DE LA MAÎTRISE EN GÉNIE ÉLECTRIQUE
M.Ing.

PAR
CHARPENTIER, Christophe

CONCEPTION D'UNE MÉTHODE ROBUSTE DE RECONNAISSANCE DE LA PAROLE POUR UN SYSTÈME EMBARQUÉ

MONTRÉAL, LE 14 JANVIER 2008

CE MÉMOIRE A ÉTÉ ÉVALUÉ

PAR UN JURY COMPOSÉ DE :

M. Gheorghe Marcel Gabrea, directeur de mémoire
Département de génie électrique à l'École de technologie supérieure

M. Christian Gargour, président du jury
Département de génie électrique à l'École de technologie supérieure

M. Jean-Marc Lina, membre du jury
Département de génie électrique à l'École de technologie supérieure

IL A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 11 JANVIER 2008

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

REMERCIEMENTS

Je tiens à adresser mes remerciements à mon directeur de mémoire, Mr Gheorghe Marcel Gabrea, pour m'avoir accordé sa confiance en me permettant de réaliser ce projet. Sa patience, sa compréhension, ainsi que ses conseils avisés m'auront été d'une grande aide dans l'accomplissement de ma tâche.

J'adresse également mes salutations à tous les membres de ma famille, qui auront du se passer de ma présence pendant de trop longues années, mais dont le soutien régulier n'a jamais failli. Je dois admettre que vous m'avez également un peu manqué. Je remercie aussi les nombreuses personnes rencontrées durant mon séjour à Montréal, qui m'auront permis de passer d'agréables moments, et sans qui j'aurais peut-être eu plus de difficultés à atteindre mes objectifs. Merci encore à Félicia, ma collègue et amie, pour avoir toujours su me motiver à aller au bout de mon travail, et à David pour son aide dans la préparation de ma soutenance.

Enfin, je dédie ce mémoire à mon grand-père, Jean, à qui j'aurais été si fier de présenter mon travail, mais qui nous a quittés trop tôt pour cela…

CONCEPTION D'UNE MÉTHODE ROBUSTE DE RECONNAISSANCE DE LA PAROLE POUR UN SYSTÈME EMBARQUÉ

CHARPENTIER, Christophe

RÉSUMÉ

Malgré sa présence grandissante dans les applications de la vie quotidienne, la reconnaissance robuste de la parole reste un défi à relever lorsqu'il s'agit de l'appliquer à partir de systèmes aux ressources limitées, même pour un vocabulaire très simple (reconnaissance de chiffres connectés à travers un canal téléphonique). Plutôt que d'utiliser une approche traditionnelle fondée sur les modèles de Markov cachés, qui nécessite un calcul intensif, et qui ne fonctionne d'ailleurs pas toujours bien dans des conditions acoustiques défavorables, la méthode présentée dans ce document se base sur un modèle de construction globale, identique pour l'ensemble des chiffres du vocabulaire, pour réduire la complexité de la tâche de reconnaissance tout en maintenant un bon niveau d'efficacité. Après une phase de segmentation découpant le signal en une succession d'unités acoustiquement homogènes, un processus de reconstruction se charge d'insérer ces segments dans le modèle global, en se fondant sur certaines contraintes et propriétés propres au vocabulaire concerné, pour arriver à déterminer au final le candidat le plus susceptible de correspondre au mot inconnu. La reconnaissance de chiffres aussi bien isolés que connectés est alors permise avec un taux de réussite satisfaisant, au moyen d'une quantité de calculs très réduite, tout comme l'espace mémoire requis.

CONCEPTION D'UNE MÉTHODE ROBUSTE DE RECONNAISSANCE DE LA PAROLE POUR UN SYSTÈME EMBARQUÉ

CHARPENTIER, Christophe

ABSTRACT

Despite its growing presence in many daily applications, robust speech recognition remains a challenge, when used on embedded systems, even for a very simple vocabulary (recognition of connected digits through a telephone channel). Rather than using a traditional approach based on the hidden Markov Models, which requires intensive computing, and does not always work well in adverse acoustical conditions, the method presented in this paper is based on an overall construction model, invariant for all the digits in the vocabulary, to reduce the complexity of the recognition task while maintaining a good level of efficiency. After a segmentation phase, during which the signal is cut in a succession of acoustically homogeneous units, a reconstruction process put these segments, using constraints and properties set by the structure of the vocabulary, to achieve the selection of the candidate that corresponds most likely to the unknown unit. The recognition of isolated digits, as well as connected, is permitted with a good success rate, using a very limited number of calculations, and a low amount of required memory.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

Page

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

DAV	Détection d'activité vocale
DTW	Dynamic time warping
HMM	Hidden Markov model
LLR	Log likelihood ratio
LPC	Linear predictive coding
LPCC	Linear prediction cepstral coefficient
MHF	Multi hybrid filter
TFD	Transformée de fourier discrète
ZC	Zero crossing

INTRODUCTION

La transmission de la parole fait appel à deux mécanismes bien distincts (Furui, 1989) : tout d'abord le dispositif de production dont le but est de générer l'onde acoustique contenant les informations à transmettre, puis le mécanisme de perception dont le rôle est de recevoir ce signal et d'en extraire les paramètres importants, pour les envoyer ensuite au cerveau. Ce dernier joue finalement le rôle du dispositif de compréhension chargé de décoder les informations. Ces deux dispositifs font appel à des processus fort complexes, mettant en jeu de nombreux organes du corps humain, allant du simple cartilage aux os, en passant par les muscles, chaque élément jouant un rôle très important pour un échange parfait de l'information. Au niveau de la production du son, la juxtaposition de tous ces éléments entraîne la possibilité de créer des sons très variés, que ce soit en termes de composantes fréquentielles, d'intensité ou de hauteur. C'est de la complexité de l'appareil de production de la parole que le langage humain tire toute sa richesse. Au niveau de la perception, l'objectif semble plus simple : décoder les informations contenues dans l'onde sonore. Le dispositif associé à cette tâche est pourtant bien plus alambiqué, notamment en raison de la nécessité de transformer la vibration de l'air en information compréhensible par le cerveau. Si plusieurs millénaires d'évolution ont été nécessaires à l'être humain pour maîtriser parfaitement la transmission orale de l'information, les chercheurs en traitement du signal n'ont bénéficié que de quelques décennies pour tenter de dompter la grande complexité de la parole et des mécanismes qui lui sont associés, en vue de permettre son utilisation dans les relations homme-machine.

Avec l'essor des technologies de l'information, les applications faisant appel au signal de parole se sont en effet multipliées, et jouent maintenant un rôle prépondérant dans notre quotidien. Des systèmes de télécommunication aux processus à commande vocale, en passant par les applications multimédia, l'être humain est régulièrement amené à interagir avec les machines que ce soit pour être compris, pour transmettre une information ou en recevoir. Toutes ces nouvelles utilisations de la parole ont été rendues possibles grâce à la concentration des chercheurs sur trois principaux axes de traitement du signal vocal : le

codage, la synthèse et la reconnaissance. Alors que le premier est notamment très important pour la téléphonie mobile, où la limitation des canaux de transmission impose de représenter les informations vocales au moyen d'un nombre réduit de données, le second est encore relativement peu utilisé, compte tenu de son intérêt limité et des difficultés que l'on éprouve à générer une voix synthétique la moins artificielle possible. En revanche, la reconnaissance de la parole est un domaine en pleine expansion depuis plusieurs années, grâce au développement de systèmes électroniques de plus en plus élaborés, offrant des gains significatifs de puissance et d'espace mémoire, et permettant l'utilisation, pour les applications destinées au grand public, de méthodes jusque là principalement cantonnées aux séances d'essais en laboratoires. Néanmoins, ces avancées technologiques ont surtout permis aux outils tels que les logiciels de dictée vocale de se démocratiser, tandis que de nombreuses autres applications sont embarquées sur des systèmes aux ressources encore trop limitées pour permettre une reconnaissance de qualité dans des conditions adaptées à leur utilisation.

Parmi ces systèmes embarqués, nous pouvons citer les processeurs de traitement du signal présents notamment dans les téléphones mobiles. Apparus au début des années 80 et spécialisés dans les opérations de traitement du signal, ces processeurs sont parfaitement adaptés aux systèmes mobiles en raison de leur faible consommation et de leur propension à effectuer avec une grande efficacité les opérations de multiplication-addition, très présentes dans le traitement de la parole. Les téléphones cellulaires, quant à eux, ont de nombreuses raisons de faire appel à la reconnaissance de la parole, notamment pour la numérotation vocale des numéros de téléphone, composés de suites de chiffres, voire de noms préenregistrés. La réussite de telles tâches est néanmoins fortement limitée par les capacités réduites des processeurs en question, et si l'on arrive à obtenir de bons résultats lorsque les conditions de reconnaissance sont idéales, la qualité se dégrade rapidement quand on élargit l'utilisation de l'application à un grand nombre de locuteurs ou à des environnements bruités. La faute en incombe principalement à l'utilisation de méthodes de reconnaissance qui ne sont pas adaptées au système sur lequel elles sont implémentées, ou à la tâche demandée, comme c'est le cas de celles basées sur les modèles de Markov cachés ou sur la

programmation dynamique, qui ont effectivement été conçues à l'origine pour fonctionner sur des systèmes fixes capables de leur fournir notamment des dictionnaires de référence imposants et des vitesses de calcul « phénoménales ». Sur des systèmes embarqués, ces méthodes sont donc bridées, et il serait plus intéressant de développer des méthodes de reconnaissance adaptées à une tâche spécifique, en tenant compte des ressources disponibles.

L'objectif de ce mémoire sera donc de concevoir une méthode de reconnaissance de la parole destinée à reconnaître des chiffres, tant isolés que connectés, pouvant fonctionner aisément sur un système embarqué, c'est-à-dire répondant aux contraintes d'espace mémoire et de temps d'exécution limités. L'intérêt ici sera de proposer un algorithme adapté au vocabulaire considéré, à savoir les chiffres allant de 0 à 9, prononcés en anglais, afin d'optimiser au maximum son fonctionnement selon les caractéristiques du système, tout en obtenant un taux de réussite convenable, pour un grand nombre de locuteurs différents. Pour ce faire, nous reprendrons certains éléments intéressants des autres méthodes habituellement utilisées, en les ajustant à notre environnement de travail, de manière à réduire le plus possible la quantité de calculs nécessaires à la reconnaissance. Ce document sera donc organisé de façon à suivre le cheminement d'une chaîne de traitement de la parole, en partant de l'onde sonore jusqu'au processus de décision finale de la reconnaissance, retenant à chaque fois les notions et propriétés importantes pour la conception finale de notre méthode.

Le premier chapitre parlera de l'étude du processus de production de la parole. Pour réaliser notre système de reconnaissance, il serait effectivement vain, et quelque peu utopique, d'essayer de reproduire le plus fidèlement possible l'appareil auditif, tant les éléments impliqués sont complexes et possèdent de nombreuses caractéristiques encore non maîtrisées. L'objectif étant de concevoir une méthode capable de fonctionner sur un système relativement peu puissant et travaillant avec un vocabulaire réduit, notre intérêt ne se situe pas dans une étude approfondie de l'appareil de perception de la parole. En revanche, il peut être très intéressant de pouvoir tirer parti des caractéristiques spécifiques

du signal généré par le corps humain pour arriver à en extraire des paramètres contenant le plus d'informations possibles. C'est pour cette raison qu'il nous est nécessaire d'étudier, au moins brièvement, la source de ce signal, c'est-à-dire l'appareil vocal humain, et de déterminer de quelle façon sa structure nous aidera à comprendre le signal que nous avons à traiter. Les connaissances acquises sur le signal de parole seront alors mises à profit au cours du chapitre 2, lors de la présentation des outils généralement utilisés dans les systèmes traitement du signal pour d'extraire des renseignements précis sur l'information véhiculée par l'onde sonore. On cherchera notamment à déterminer une façon de représenter cette information au moyen d'un nombre restreint de données, en se basant pour cela sur certaines caractéristiques propres au signal vocal.

La seconde grande partie de notre étude sera consacrée à la tâche de reconnaissance vocale. Le chapitre 3 nous permettra tout d'abord de nous familiariser avec ce domaine, avant de s'attarder sur la description du fonctionnement des principales méthodes de reconnaissance. Nous aurons ainsi l'opportunité d'observer leurs caractéristiques, pour ensuite discuter de leurs avantages et inconvénients lorsqu'elles sont utilisées dans des systèmes embarqués. Tenant compte des conclusions de cette discussion, la méthode conçue pour ce mémoire sera alors présentée au cours du chapitre 4. Nous prendrons soin de décrire chacun des éléments importants qui la composent et qui font sa particularité, avant d'exposer dans le chapitre 5 les résultats que nous avons obtenus au cours de nombreuses séries de tests destinés à vérifier son efficacité dans trois domaines importants : l'espace mémoire requis, la rapidité d'exécution, et la qualité de la reconnaissance.

CHAPITRE 1

LE PROCESSUS DE PRODUCTION DE LA PAROLE

1.1 Introduction

Au cours de ce premier chapitre, nous allons décrire brièvement les mécanismes entrant en jeu lors du fonctionnement de l'appareil vocal humain, du moment où le son est créé, jusqu'au moment où il sort de la bouche, en commençant par une rapide présentation des organes utilisés. S'en suivra ensuite une description du mécanisme de production et des différents types de sons ainsi générés. Nous terminerons enfin ce par une ouverture vers le domaine du traitement du signal, en montrant comment il est possible de transposer les éléments présentés lors des deux premières parties dans le domaine électrique, pour faciliter le traitement du signal vocal.

1.2 L'appareil vocal humain

Ce que l'on appelle l'appareil vocal humain est constitué des poumons, de la trachée, du larynx, du pharynx, et enfin des cavités nasales et orales. Connectés les uns aux autres, ils forment une sorte de tube représentant l'appareil de production de la parole. La figure 1.1 présente un schéma de cet ensemble d'organes. Cet appareil peut être divisé en trois sous-ensembles : le premier, composé des poumons et de la trachée, alimente le second, constitué du larynx, dont la fonction est d'exciter le dernier ensemble, représenté par le conduit vocal. Chaque sous-système a donc une fonction bien précise dans le mécanisme de la phonation. Décrivons rapidement le fonctionnement de chacun d'entre eux, en vue de mieux comprendre leur rôle dans la production de la parole.

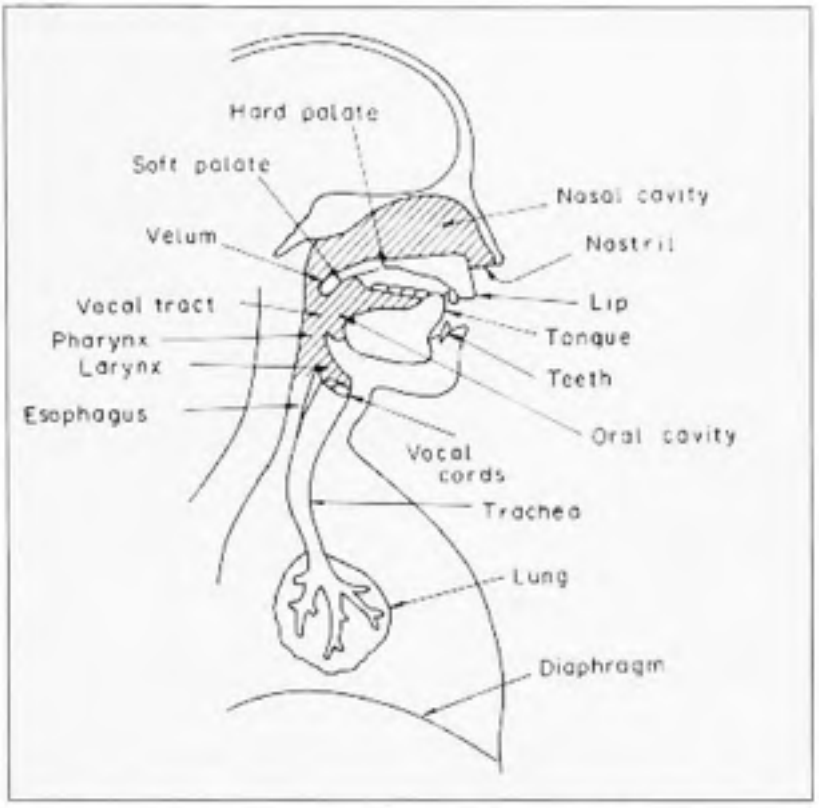

Figure 1.1 ***Schéma de l'appareil vocal humain.***
(Tiré de Sadaoki Furui, 1985)

1.2.1 Les poumons et la trachée

Dans la parole, le rôle des poumons est de fournir l'air nécessaire à la production d'un son. Cet air est ensuite transporté, au travers de la trachée, jusqu'au reste de l'appareil vocal. En résumé, l'appareil respiratoire joue le rôle de fournisseur d'énergie auprès du mécanisme de production de la parole. Cela n'a toutefois pas d'importance dans notre étude du traitement du signal de parole car finalement, la force de l'expiration d'air influe uniquement sur la puissance du son qui sera produit, et nom sur ses caractéristiques fréquentielles.

1.2.2 Le larynx

Deuxième élément important de l'appareil de production de la parole, le larynx est un ensemble de muscles et de cartilages mobiles, entourant une cavité située au niveau de la

partie supérieure de la trachée. La figure 1.2 nous présente une description du larynx, vu sous deux coupes différentes.

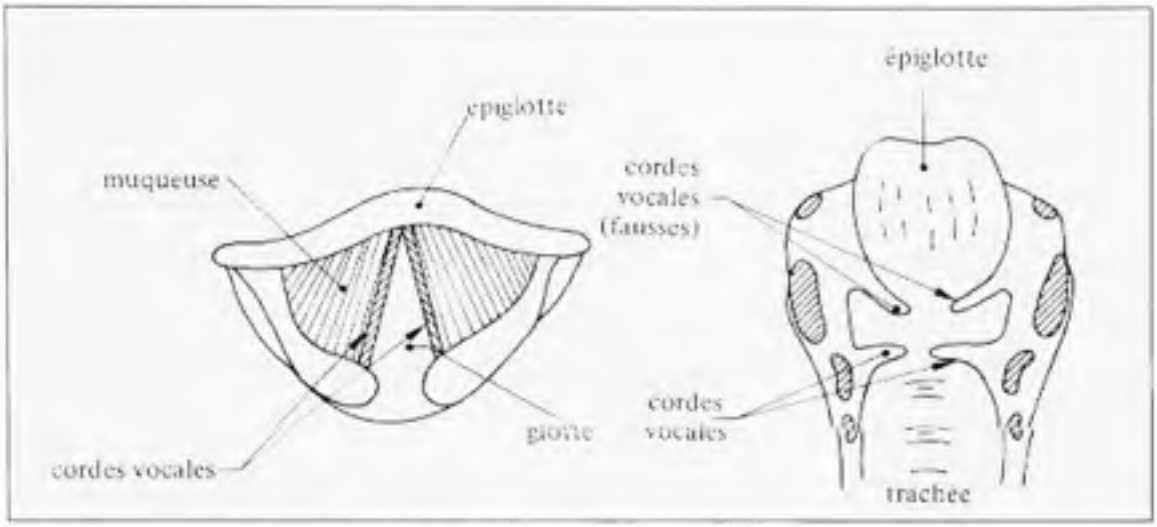

Figure 1.2 ***Schéma descriptif du larynx.***
(Tiré de Boite, 1987)

Cet ensemble de cartilages héberge deux lèvres symétriques horizontales, faites de ligaments et de muscles, appelées cordes vocales. Placées en travers du larynx et vibrant au contact des masses d'air en provenance de la trachée, les cordes vocales peuvent se fermer ou s'écarter pour déterminer une ouverture triangulaire appelée glotte. L'air peut ainsi y passer librement lors de la respiration, de la voix chuchotée, ou des sons sourds, tandis que les sons de nature périodique sont produits grâce aux vibrations des cordes vocales resserrées (Boite, 1987). Le larynx est l'organe essentiel de la voix, sa présence différenciant l'être humain de nombreuses autres espèces animales. Nous verrons par la suite son influence quant à la nature du son produit.

1.2.3 Le conduit vocal

Ce terme désigne généralement tout ce qui se situe entre les cordes vocales et les lèvres. Considéré comme une succession de tubes ou cavités acoustiques de sections différentes, le conduit vocal est composé du pharynx, de la cavité orale et de la cavité nasale. Toutes ces

cavités jouent un rôle très important dans la forme finale du son produit. Elles sont en quelque sorte une caisse de résonance de l'appareil de production de la parole.

Le pharynx, débutant au croisement de la trachée et de l'œsophage et s'étendant jusqu'au voile du palais, se contente juste de faire circuler le son en provenance de la glotte, mais sa forme peut être modifiée en bougeant d'autres éléments du conduit vocal. La cavité orale est constituée d'éléments extrêmement mobiles tels que la langue, la mâchoire, le palais et les lèvres qui, en modifiant la forme de la caisse de résonance, modifient les sons qui en sortent. La cavité nasale est immobile, mais peut être connectée ou déconnectée du reste du conduit vocal en modifiant la position du voile du palais.

Voilà qui clôt notre description sommaire de l'appareil de production de la parole. Toutefois, plus de précisions sur les caractéristiques des différents organes présentés ici peuvent être trouvées dans les ouvrages de Parsons (1986) et Rowden (1992). Concernant la perception auditive, il existe de nombreux ouvrages accordant une rapide description des éléments mis en jeu dans ce dispositif. Pour de plus amples explications, il sera ainsi possible de se référer à l'étude très détaillés réalisée par Pickles (1988), mais les références de Boite(1987) et Furui (1989) nous présentent une étude plus condensée sur les points importants que sont la physiologie de l'oreille, ses relations avec le cerveau ainsi que les phénomènes de sélection et masquage fréquentiel.

1.3 La production d'un son

Maintenant que nous connaissons les principaux acteurs du processus de production de la parole, nous allons décrire rapidement leur fonctionnement ainsi que la façon dont ils influent sur la nature du son produit.

Comme nous l'avons expliqué précédemment, tout ce processus a pour origine l'air expulsé des poumons. Mais la véritable création du signal de parole a lieu au niveau des cordes vocales. C'est à cet endroit que, sous l'effet des vibrations, la pression de l'air est en

quelque sorte transformée en une suite de pulsations quasi-périodiques ou aléatoires. Une première distinction très importante est donc faite quant aux différents types de sons produits : les sons générés par une vibration pseudo-périodique des cordes vocales sont dits *voisés*, tandis que les sons obtenus à partir d'un simple écoulement de l'air au travers de la glotte sont dits *sourds* ou *non voisés*.

L'onde ainsi obtenue est ensuite modulée en fréquence en passant au travers du conduit vocal. Sous l'effet résonateur des différentes cavités, et en fonction de la position des organes tels que les dents, la langue, les lèvres, la bouche ou le voile du palais, différents types de sons peuvent alors être produits pour un même type de pulsation. De très légères variations de l'agencement de cet ensemble d'organes peuvent entraîner de grosses modifications dans la signification des mots, et engendrer notamment d'autres distinctions entre les sons : on opposera ainsi les *voyelles* aux *consonnes*, les sons *plosifs* aux *fricatifs* et les *occlusifs* aux *liquides*.

Toutes ces catégories de sons, qui peuvent ensuite être divisées en sous-catégories, représentent autant de petites unités phonétiques différentes permettant au langage humain d'être si riche. Ces unités sont appelées phonèmes, et leur nombre peut varier d'une langue à l'autre; on recense notamment 36 phonèmes différents pour la langue française, tandis que l'anglais en comporte 42. Étant donné que nous ne ferons pas appel aux phonèmes lors de l'élaboration de notre méthode de reconnaissance vocale, nous ne nous pencherons pas plus profondément sur ces éléments. Il est toutefois possible de se référer en à l'ouvrage de Flanagan (1972) pour obtenir une étude phonétique assez complète. Kenneth Church (1987) présente également une étude très complète de l'utilisation du vocabulaire phonétique dans la reconnaissance de la parole. En revanche, les caractéristiques globales des sons produits par l'appareil vocal nous intéressent fortement. Nous allons donc revenir plus en détail sur ce qui différencie un son voisé d'un son non voisé.

1.3.1 Les sons voisés

Les sons voisés résultent donc de l'excitation du conduit vocal par des impulsions périodiques de pression liées aux oscillations des cordes vocales. La fréquence fondamentale du signal, également appelée pitch, est alors déterminée par la fréquence de vibration des cordes vocales, alors que son intensité est liée à la pression de l'air en amont du larynx. Selon Boite (1987) la fréquence du fondamental varie en fonction des individus, allant de 80 à 200 Hz pour une voix masculine, de 150 à 450 Hz pour une voix féminine, et de 200 à 600 Hz pour une voix d'enfant .

Nous pouvons observer l'allure temporelle d'un son voisé sur la figure 1.3 ainsi que sa représentation fréquentielle sur la figure 1.4.

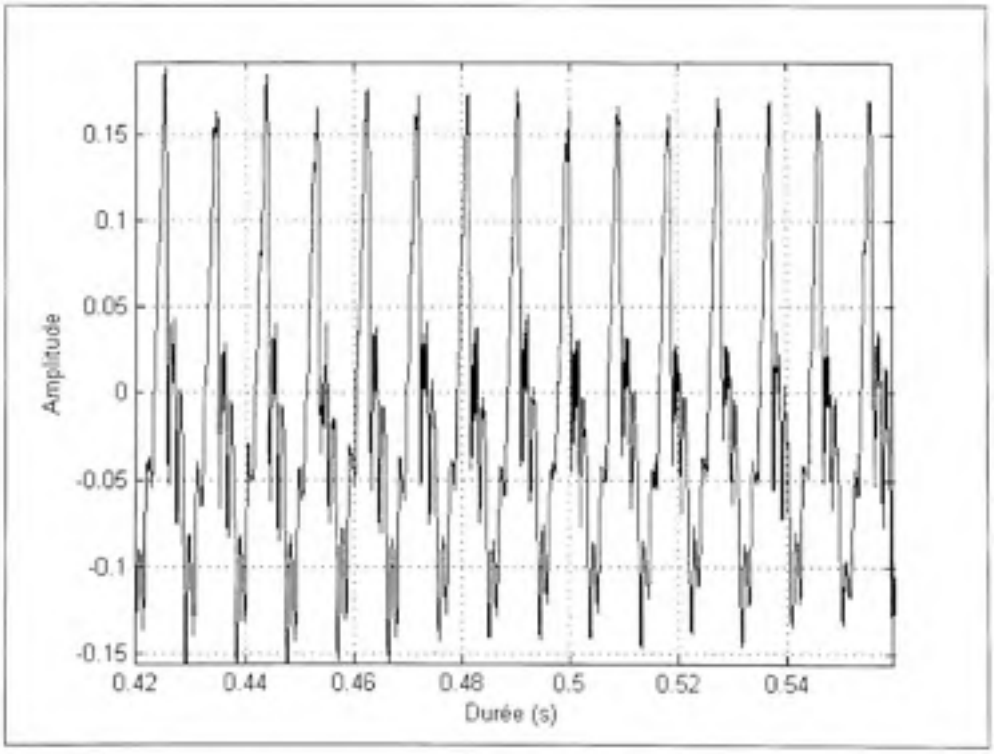

Figure 1.3 ***Représentation temporelle du son voisé « a ».***

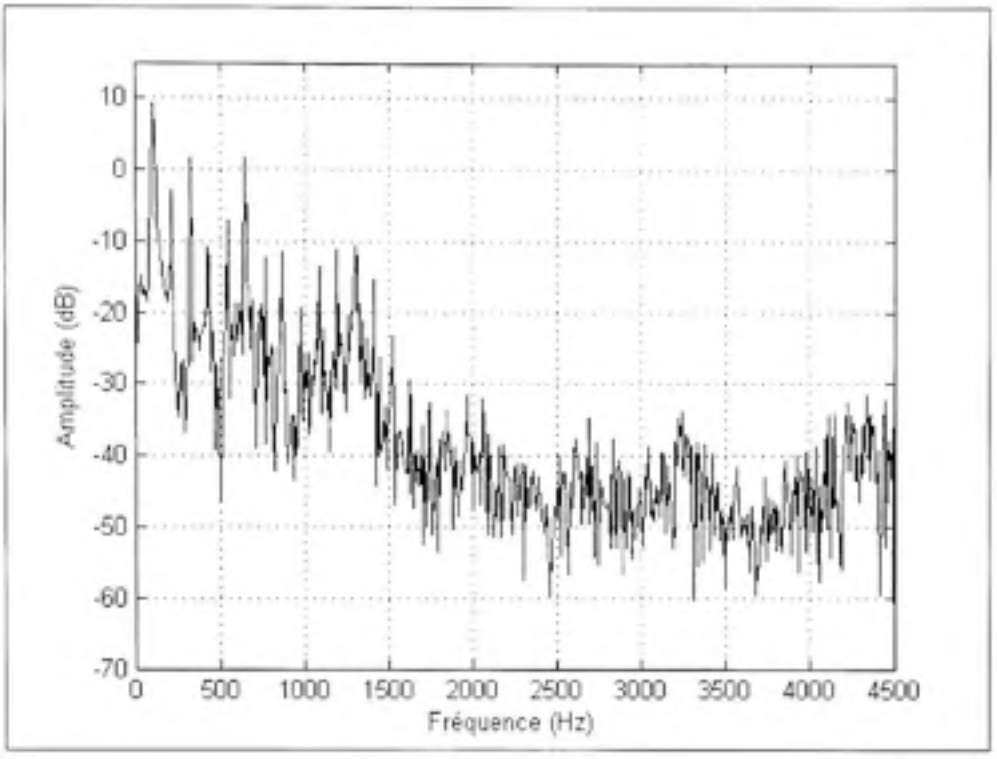

Figure 1.4 ***Représentation fréquentielle du son voisé « a ».***

Il apparaît donc clairement qu'un son voisé est quasi-périodique, et que son spectre affiche un premier pic d'amplitude à la fréquence du fondamental, suivis d'autres pics appelés formants, et correspondant aux fréquences propres du conduit vocal. L'ensemble de ces raies va représenter le timbre de la voix, et peut nous donner beaucoup d'indications sur le son produit ou le locuteur qui le prononce.

1.3.2 Les sons non voisés

Les sons non voisés sont produits lorsque les cordes vocales ne vibrent pas. Le son est alors uniquement du à un frottement de l'air dans le conduit vocal, et peut être assimilé à un bruit blanc filtré par la transmittance du conduit vocal. Les figures 1.5 et 1.6 présentent les représentations temporelle et spectrale d'un son non voisé.

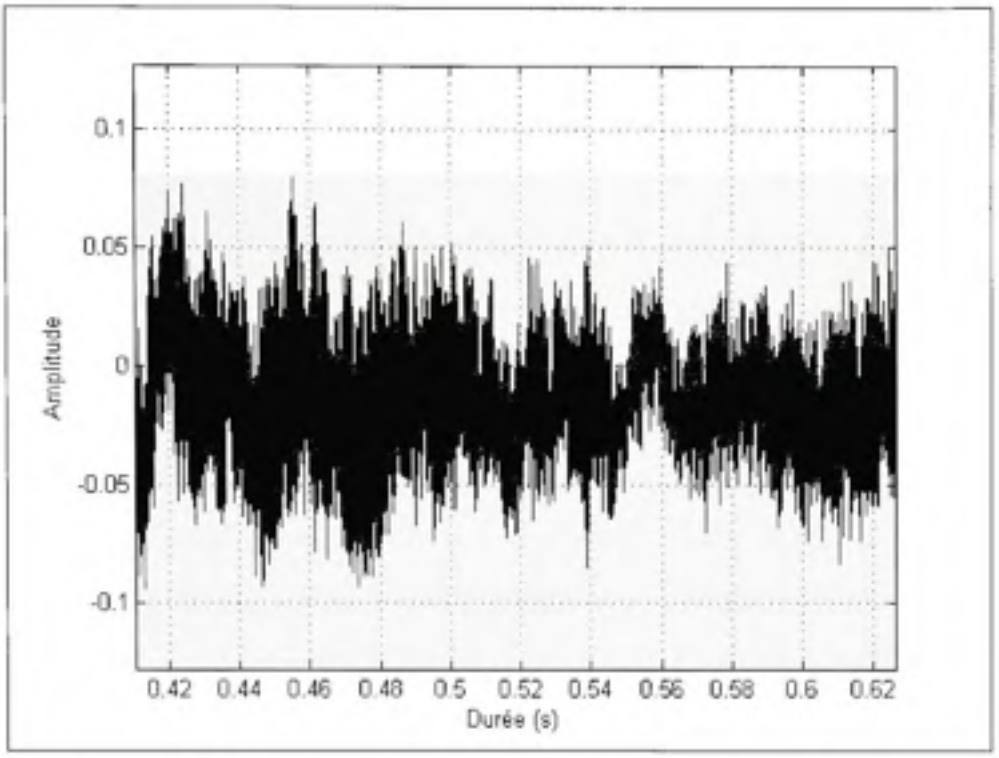

Figure 1.5 ***Représentation temporelle du son non voisé « s ».***

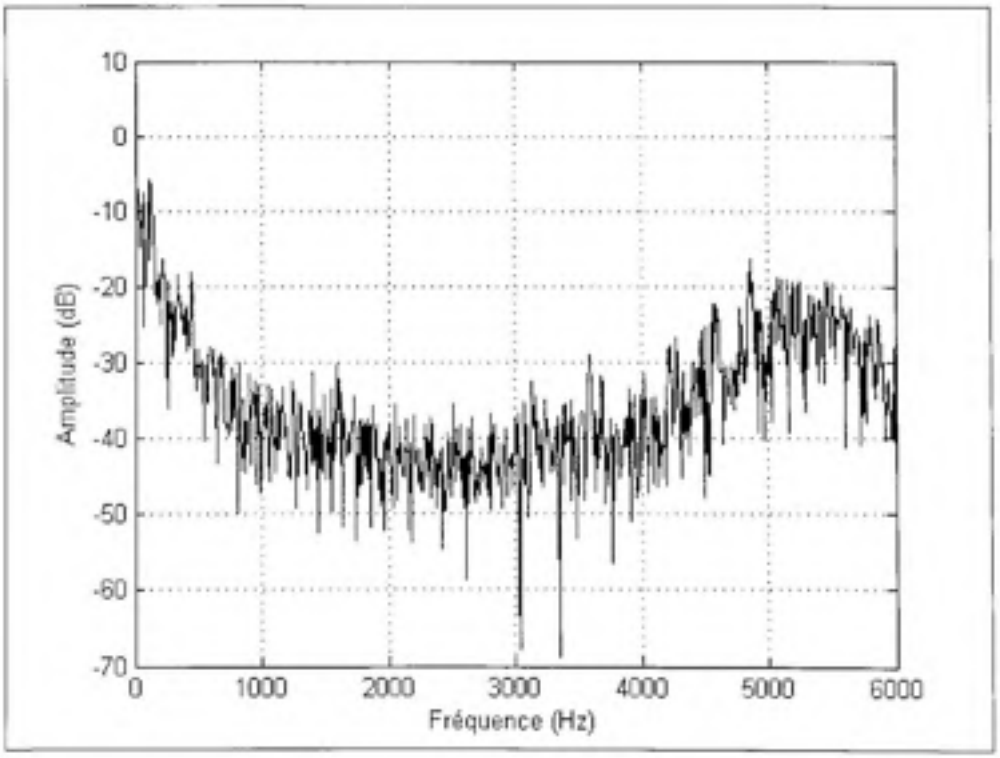

Figure 1.6 ***Représentation fréquentielle du son non voisé « s ».***

Ainsi, un son non voisé ne présente aucune structure périodique, donc pas de fréquence fondamentale, et voit son énergie nettement répartie dans les hautes fréquences. Pour cette raison, il est relativement difficile de le distinguer d'un bruit quelconque.

Différencier deux sons de catégories différentes s'annonce donc relativement aisé, la présence ou non d'une fréquence fondamentale nous indiquant très rapidement la catégorie à laquelle appartient le mot. En revanche, il sera bien plus difficile de différencier deux sons d'une même catégorie. Une minutieuse analyse du signal vocal sera donc nécessaire afin d'arriver à extraire des paramètres suffisamment précis pour distinguer des différences minimes entre deux sons de même « catégorie ». Durant la partie suivante, nous expliquerons brièvement en quoi la structure si particulière du processus de production de la parole, que nous venons de détailler, peut nous aider à traiter facilement le signal qu'il génère.

1.4 Modélisation du mécanisme de production de la parole

De ce que nous avons observé précédemment, nous pouvons retenir que chaque son de parole est le fruit des actions successives des poumons, du larynx et du conduit vocal, tous agissant indépendamment les uns des autres pour permettre à l'être humain de pouvoir produire une très grand quantité de sons différents. Cette relative indépendance des sources et de leurs transformations est justement à la base de la théorie acoustique de la production de la parole, longuement détaillée par Flanagan (1972). Cette théorie considère les termes de source, souvent non linéaires, et un filtre linéaire qui transforme le signal de source en modifiant son enveloppe spectrale. Nous pouvons ainsi résumer le cheminement du son, depuis son origine jusqu'à sa sortie du corps humain, et assimiler chacun des organes de production de la parole à des éléments plus « mécaniques », comme présenté sur la figure 1.7.

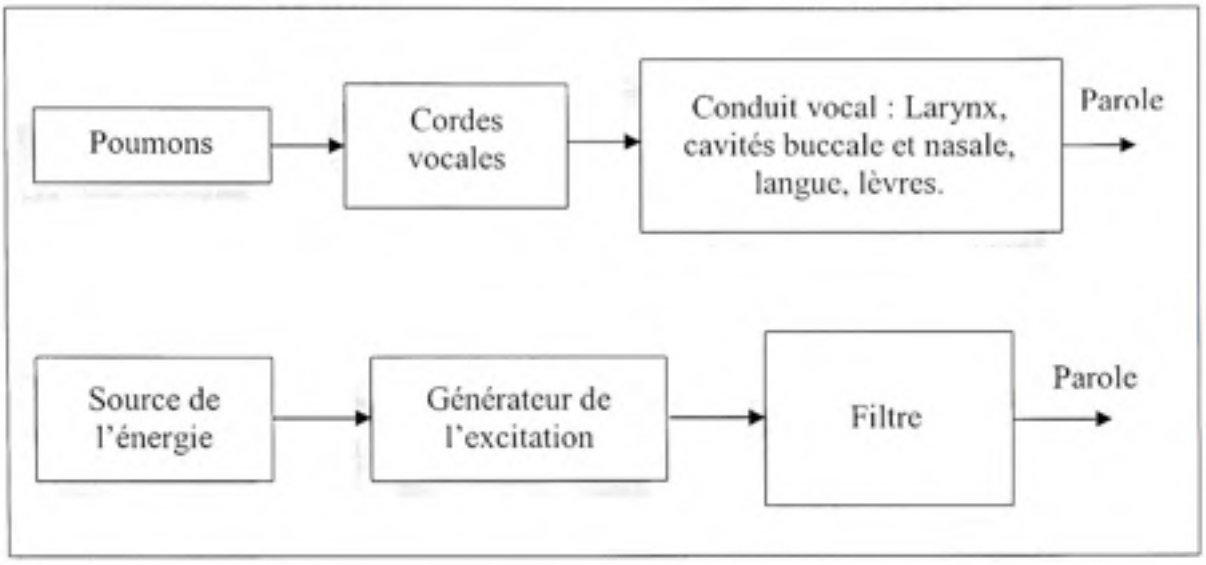

Figure 1.7 ***Schématisation du processus de production de la parole.***

La source sonore peut donc être modélisée par une suite d'impulsions périodiques, pour les sons voisés, ou un bruit blanc, pour les sons non voisés, excitant un filtre dit *tous-pôles*, dont les éléments représentent les caractéristiques du conduit vocal. Concernant les caractéristiques du son engendré, nous pouvons dire que la source d'excitation fournit le fondamental, et le conduit vocal engendre un phénomène de résonance qui crée des formants. Nous verrons qu'une telle modélisation nous sera fortement utile par la suite, lorsqu'il s'agira de tirer partie des propriétés spécifiques du signal vocal pour en extraire des informations de façon la plus efficace possible. Ce sera le sujet du second chapitre.

CHAPITRE 2

LE TRAITEMENT NUMÉRIQUE DU SIGNAL VOCAL

2.1 Préambule

Le premier chapitre ayant permis de définir le comportement global du signal vocal, il convient maintenant de s'intéresser aux différentes façons d'en extraire des informations exploitables. L'objectif de ce chapitre est de s'appuyer sur les caractéristiques particulières du signal de parole afin de déterminer la méthode d'extraction de paramètres la plus appropriée aux besoins du système développé au cours de ce mémoire. Pour ce faire, nous décrirons les différentes étapes d'une chaîne de traitement du signal telle qu'utilisée généralement dans les systèmes de reconnaissance de la parole, en détaillant chacun des processus de transformation qui conduisent à une forme exploitable de l'information véhiculée par le signal.

Au cours du chapitre 1, le signal de parole a été présenté comme un déplacement d'air transportant simultanément des informations linguistiques et émotionnelles, engendré par un ensemble de muscles et d'organes. A partir de ce point, et pour les besoins de notre étude, il sera considéré comme une forme d'onde contenant tout un jeu d'informations sur le message qu'elle transporte. Le terme x(n) représentera, par conséquent, le $n^{ième}$ échantillon de parole ayant pénétré dans le système. Une telle représentation du signal est obtenue à l'aide d'un Convertisseur Analogique-Numérique (CAN), qui sélectionne la valeur du signal analogique d'entrée à chaque instant $T_n = T_{n-1} + T_e$, T_e représentant la période d'échantillonnage exprimée en secondes. Plus généralement, on parlera en terme de fréquence d'échantillonnage, donnée par la formule 2.1 :

$$F_e = \frac{1}{T_e} \tag{2.1}$$

À partir de là, il est possible de représenter la chaîne de traitement du signal vocal au moyen de la figure 2.1, tirée du livre de Junqua et Haton (1996), qui donne un bon aperçu global des différentes tâches à effectuer avant d'entrer dans le processus de reconnaissance.

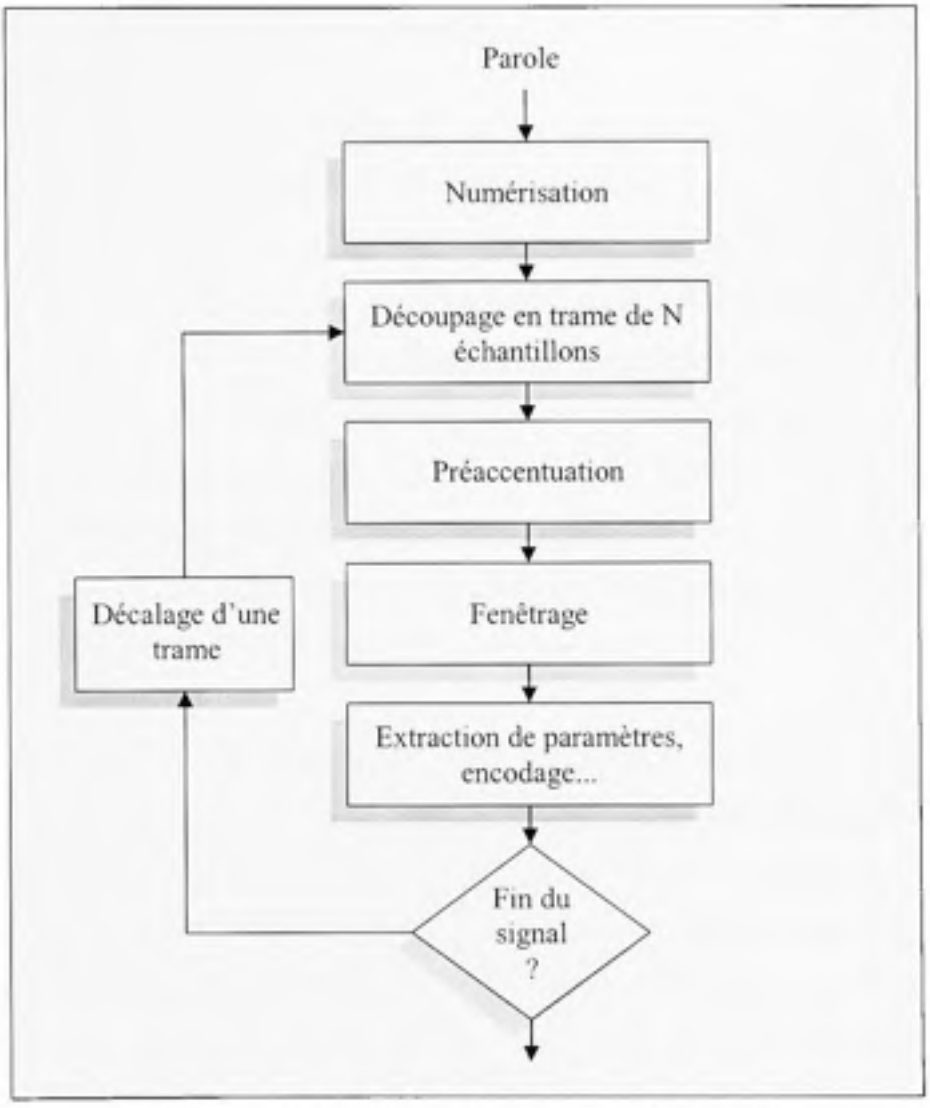

Figure 2.1 ***La chaîne de traitement du signal.***

L'architecture de ce chapitre s'appuiera donc sur ce schéma, en détaillant au cours de la première partie les processus de prétraitement du signal que sont le découpage du signal en trames constituées d'une quantité finie d'échantillons, la préaccentuation et le fenêtrage. La seconde partie fournira quant à elle une description des principaux outils utilisés dans les applications de traitement du signal pour extraire des informations significatives sur le

signal. Enfin, la dernière partie sera consacrée à la présentation des principales méthodes de paramétrisation du signal, destinées à représenter les caractéristiques de la parole sous une forme condensée, en conservant le maximum d'informations possibles.

2.2 La chaine de prétraitement du signal

Tel que mentionné au cours du chapitre 1, lors du processus de production de la parole, l'appareil vocal évolue constamment, de façon très rapide, et la moindre petite variation de sa forme peut entraîner une modification très importante du son généré. Si cette particularité rend le langage humain si riche, elle est également à l'origine de la complexité du signal de parole, mais surtout de son caractère non-stationnaire. Ainsi, pour obtenir une représentation précise de l'état du signal à un instant donné, il est important de travailler sur des tranches temporelles très courtes, bien que suffisamment longues pour contenir assez de renseignements sur l'information contenue dans le signal. Généralement, on préfère donc travailler sur des périodes allant de 10 à 30 ms, durant lesquelles le signal est considéré comme étant quasi stationnaire (Boite, 1987).

Ainsi, une fois le signal numérisé, la première action d'un système de traitement de la parole est de regrouper les échantillons entrants en un bloc de taille finie, appelé trame. Par la suite, nous noterons N la taille de la trame, en nombre d'échantillons, et toutes les opérations effectuées, dites alors « court terme », ne concerneront que cette trame. Le reste du signal sera traité en décalant la trame de façon répétée, d'un nombre d'échantillons L, L était inférieur à N, jusqu'à la fin du signal. C'est ensuite sur ces trames de parole que peuvent s'appliquer les opérations destinées à préparer le signal d'entrée pour la phase d'extraction des informations. Cette partie est généralement constituée de deux blocs, que nous allons décrire brièvement.

2.2.1 Préaccentuation du signal

Passage très fréquemment rencontré dans les applications de traitement de la parole, la préaccentuation a pour effet d'accentuer la partie haute du spectre du signal. En effet, la forte concentration énergétique des basses fréquences, observée dans la plus grande majorité des spectres de parole, est considérée comme une nuisance car, lors de l'analyse, elle minimise l'importance de l'énergie du signal aux moyennes et hautes fréquences. Ce problème est souvent partiellement résolu en passant le signal dans un filtre de transmittance $1-\alpha.z^{-1}$, c'est le processus de préaccentuation. Concrètement, le signal préaccentué est donc donné par la formule 2.2 (Kunt, 1984).

$$y(n) = s(n) - \alpha \times s(n-1) \tag{2.2}$$

Où s(n) correspond au n^{ième} échantillon du signal d'entrée, et α est une constante dont la valeur est généralement fixée entre 0,9 et 1. De rares systèmes optent toutefois pour une valeur recalculée à chaque nouvelle trame, à l'aide de la formule 2.3.

$$\alpha_{opt} = \frac{\sum_{n=0}^{N-2} s(n) \times s(n+1)}{\sum_{n=0}^{N-1} s^2(n)} \tag{2.3}$$

Néanmoins, bien que ce procédé apparaisse dans la plupart des systèmes de traitement du signal, de nombreuses applications de reconnaissance vocale ne l'utilisent pas. La principale raison étant que, si le rehaussement de l'énergie des hautes fréquences ne modifie que très peu la structure des sons non voisés, dont les principales composantes spectrales sont situées dans cette zone, cela affecte en revanche très fortement la plupart des sons voisés, dont les caractéristiques majeures sont représentées par les premiers formants. Vergin et O'Shaughnessy (1995) ont démontré à quel point cela peut nuire aux taux de reconnaissance, et présentent même une méthode permettent de calculer un coefficient α

adapté à la catégorie du signal, pour accroître ainsi fortement la qualité de la reconnaissance. Toutefois, bien que très efficace, un tel procédé s'avère coûteux en nombre de calculs, et n'est donc pas forcément adapté au système développé au cours de ce mémoire.

2.2.2 Fenêtrage du signal

Indispensable au traitement du signal, la création de blocs de N échantillons n'en reste pas moins porteuse d'effets néfastes pour le spectre du signal. En effet, le découpage brutal subi aux extrémités de la trame fait apparaître des discontinuités engendrant l'apparition de composantes hautes fréquences indésirables. Il est ainsi nécessaire d'appliquer à la trame d'échantillons ce que l'on appelle une fenêtre de pondération, dont le but est d'adoucir les discontinuités en réduisant le poids des échantillons situés aux extrémités, par rapport à ceux situés au centre. Qu'elles soient de type triangulaire, comme la fenêtre de Bartlett, de forme convexe comme celle de Kaiser, ou bien sinusoïdale comme celles de Hamming, Von Hann et Blackman, l'objectif ici n'est pas de toutes les décrire, mais de présenter celle qui semble la plus adaptée à nos besoins. Toutefois, nous ne pouvons que conseiller les ouvrages de Christian Gargour (2001) et Tamal Bose (2004) pour obtenir de plus amples informations sur les fenêtres.

La fenêtre de pondération la plus utilisée dans le domaine de la reconnaissance vocale est la fenêtre de Hamming, que l'on peut réaliser à l'aide de la formule 2.4.

$$W(l) = \begin{cases} k_1 - k_2 \cos\left(\frac{2\pi l}{L-1}\right) + k_3 \cos\left(\frac{4\pi l}{L-1}\right) & 0 \le l \le L-1 \\ 0 & autrement \end{cases} \qquad (2.4)$$

Où $k_1 = 0{,}54$, $k_2 = 0{,}46$ et $k_3 = 0$. Précisons que les fenêtres de Von Hann et de Blackman sont obtenues à l'aide de la même formule, en modifiant simplement les valeurs des

coefficients k_1, k_2 et k_3. De forme semblable, on leur préfère néanmoins la fenêtre de Hamming car elle n'annule pas le signal aux extrémités, permettant ainsi une meilleure transition entre les trames successives de signal. On peut observer la forme de la fenêtre de Hamming sur la figure 2.2.

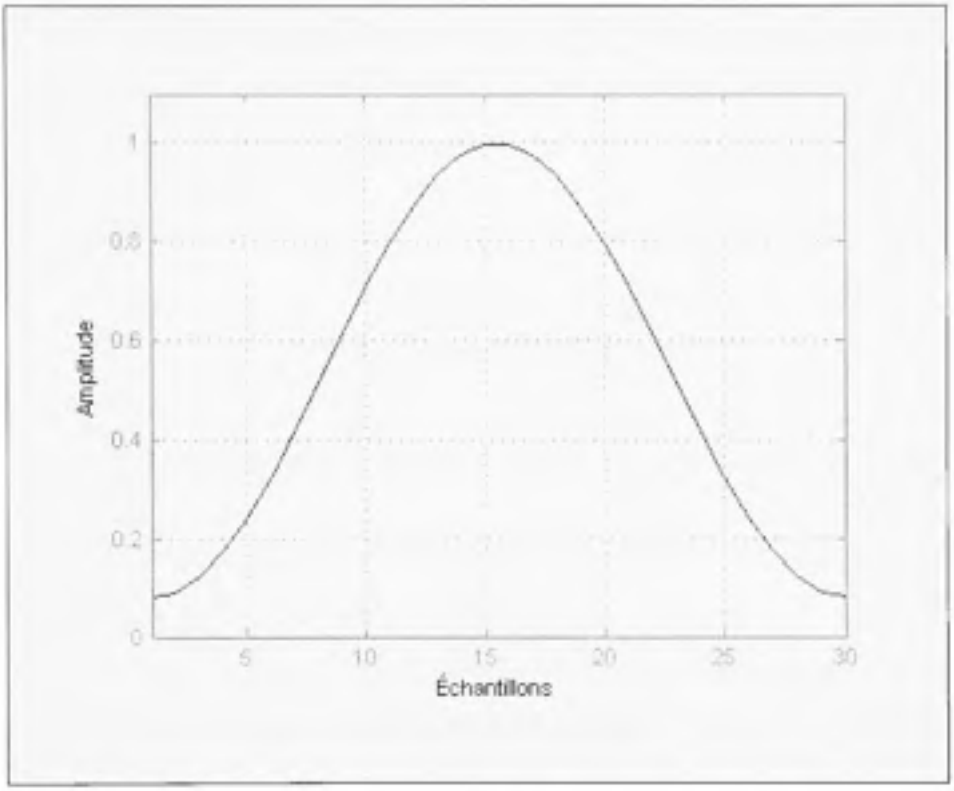

Figure 2.2 ***Forme temporelle de la fenêtre de Hamming.***

Harris (1978), lors de son étude de l'utilisation des fenêtres pour l'analyse spectrale par Transformée de Fourier Discrète (TFD) présente les TFD de nombreuses fenêtres, permettant ainsi de visualiser leurs effets sur le spectre des signaux qu'elles pondèrent. L'observation de la TFD d'une fenêtre rectangulaire, à la figure 2.3, permet en effet de mesurer la faible atténuation entre le lobe principal et le lobe secondaire, ainsi que les chutes d'atténuation de 6 dB par octave pour les lobes suivants, caractéristiques d'une fonction présentant des discontinuités.

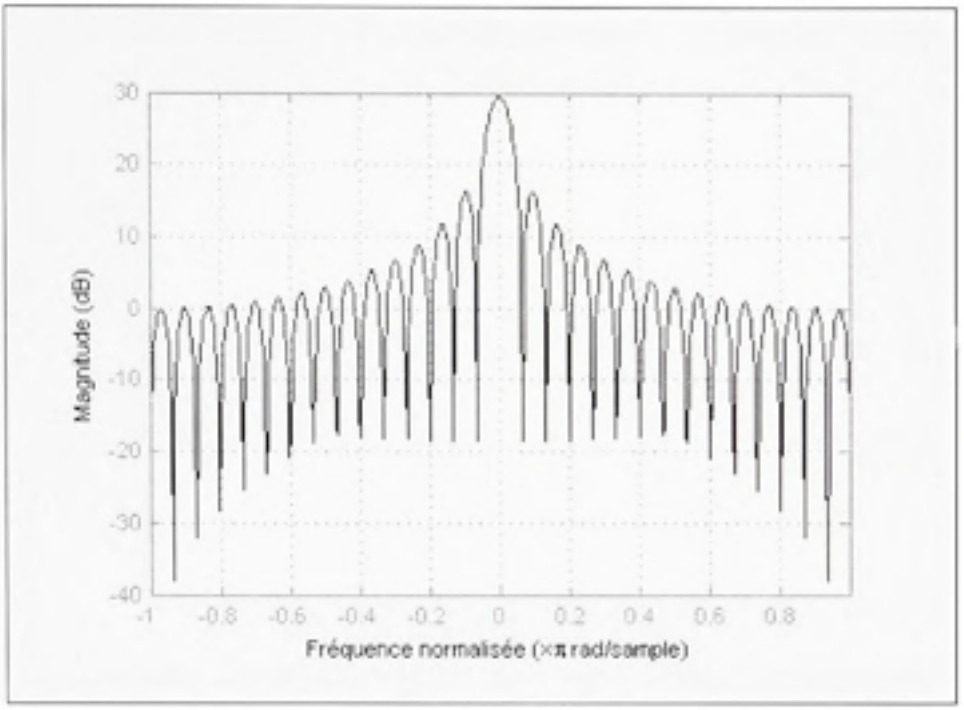

Figure 2.3 ***Transformée de Fourier Discrète d'une fenêtre rectangulaire.***

En revanche, la TFD d'une fenêtre de Hamming, à la figure 2.4, nous présente bien une atténuation nettement plus forte entre le lobe principal et le lobe secondaire. Notons également que la faible discontinuité aux extrémités de la fenêtre entraîne une atténuation des lobes suivants bien différente de la fenêtre rectangulaire.

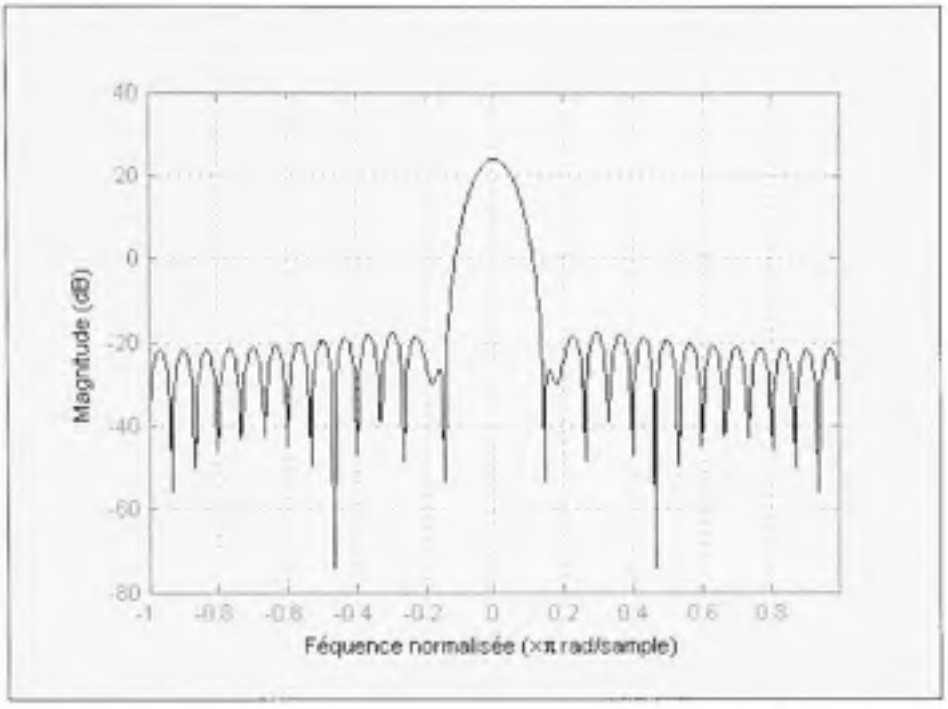

Figure 2.4 ***Transformée de Fourier Discrète de la fenêtre de Hamming.***

Ces caractéristiques font que les spectres des signaux pondérés à l'aide de cette fenêtre sont nettement moins affectés par le procédé qu'avec une fenêtre rectangulaire, et l'information contenue dans le signal n'est donc pas altérée. C'est pour cette raison que l'on retrouve la fenêtre de Hamming dans la plupart des systèmes de reconnaissance vocale, domaine où une représentation parfaite du signal original est exigée.

Cette première phase de traitement du signal permet donc de préparer idéalement le signal en vue de la phase d'extraction de paramètres décrivant l'information qu'il contient, phase dont nous décrivons plusieurs méthodes fréquemment utilisées, au cours de la partie suivante.

2.3 Les outils de traitement du signal

Pour répondre aux besoins des nombreuses applications du traitement de la parole, telles que le codage, la synthèse ou la reconnaissance de la voix, il existe un très grand nombre d'outils d'analyse, allant du simple calcul d'énergie jusqu'à la modélisation complète d'un

conduit vocal. Au cours de cette partie, nous détaillerons principalement les outils nécessaires à la conception de notre méthode de reconnaissance, ou tout du moins certaines notions qui nous seront fort utiles par la suite.

2.3.1 Énergie du signal

L'énergie d'un signal est une caractéristique liée à la quantité d'information représentée. Plus concrètement, l'énergie sera très faible en période de silence, et élevée en présence d'activité vocale, dépendamment de la catégorie et de l'amplitude du son émis. On la détermine au moyen de la formule 2.5.

$$E = \sum_{k=1}^{N} x^2(n) \tag{2.5}$$

2.3.2 Taux de passage par zéro

Le taux de passage par zéro est notamment très utile pour différencier une zone voisée d'une zone non voisée. En effet, une des principales différences entre ces deux zones est la fréquence avec laquelle le signal change de signe entre deux échantillons successifs. On détermine donc le pourcentage de changements de signe, par rapport au nombre total d'échantillons, au moyen de la formule 2.6 (Boite, 1987).

$$ZC = \frac{1}{2N} \sum_{k=0}^{N-1} \left[\left| signe\left[x(k)\right] - signe\left[x(k-1)\right] \right| \right] \tag{2.6}$$

Étant donné qu'une zone non voisée voit son énergie principalement répartie dans les hautes fréquences, les valeurs de ZC seront bien plus élevées pour un son non voisé que pour un son voisé.

2.3.3 Étude spectrale

La façon la plus répandue d'étudier « l'intérieur » d'un signal est d'observer ses composantes spectrales. Les figures 1.3 et 1.6, qui nous présentent les répartitions harmoniques respectives d'un son voisé et d'un son non voisé, nous montrent bien à quel point une analyse fréquentielle peut nous permettre d'observer idéalement les caractéristiques d'un signal, et notamment de conclure quant à la périodicité ou non de celui-ci. L'outil le plus utilisé est la transformée de Fourier, qui donne le contenu harmonique d'un signal, c'est-à-dire la répartition de ses composantes fréquentielles ainsi que leurs amplitudes (Kunt, 1984). Pour un signal discret, la définition de la transformée de Fourier continue est donnée par la formule 2.7.

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n) \times e^{-j\omega n} \tag{2.7}$$

La transformée de Fourier court-terme, qui est la transformée de Fourier d'un signal de taille finie est donnée par la formule 2.8.

$$X(e^{j\omega}) = \sum_{n=0}^{N-1} x(n) \times e^{-j\omega n} \tag{2.8}$$

Étant donné que nous travaillons dans le domaine numérique, nous utiliserons plutôt la transformée de Fourier Discrète (TFD), que l'on obtient à l'aide de la formule 2.9.

$$X(k) = \sum_{n=0}^{N-1} x(n) \times e^{-j2\pi nk/N} \tag{2.9}$$

Précisons également qu'il est possible de retrouver le signal x(n) à partir de sa TFD, en utilisant la formule 2.10.

$$x(n) = \frac{1}{N}\sum_{k=0}^{N-1} X(k) \times e^{j2\pi nk/N} \tag{2.10}$$

2.3.4 Autocorrélation du signal

L'autocorrélation d'un signal est la corrélation de ce signal avec lui-même, et son estimation est donnée par la formule suivante 2.11 (Boite, 1987).

$$\phi_{xx}(k) = \frac{1}{N-k}\sum_{i=0}^{N-k} x(i) \times x(k+i) \tag{2.11}$$

On parle généralement de l'autocorrélation en termes de coefficients d'autocorrélation, qui peuvent être calculés selon la formule 2.12.

$$\rho_x(k) = \phi_{xx}(k)/\phi_{xx}(0) \tag{2.12}$$

Concrètement, la fonction d'autocorrélation mesure la ressemblance entre un signal et des versions décalées de ce même signal. Pour un signal périodique, la ressemblance est donc maximale lorsque le décalage est égal à un multiple de la période. En cas de signal non-périodique, le résultat obtenu est un signal aléatoire, comme le montrent les figures 2.5 et 2.6 qui présentent respectivement les coefficients d'autocorrélation des sons « a » et « s » utilisés au chapitre 1.

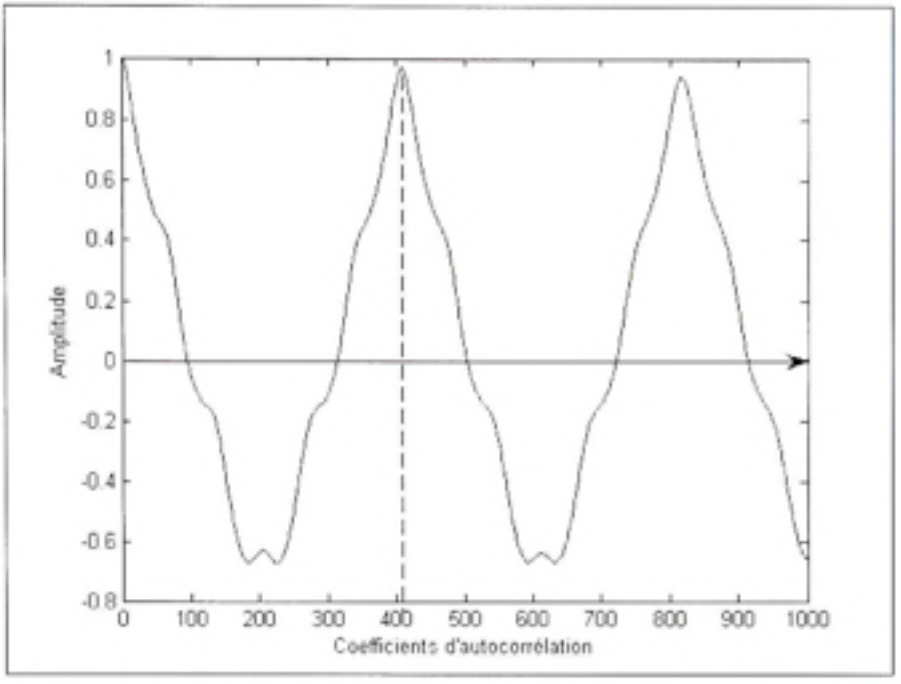

Figure 2.5 ***Coefficients d'autocorrélation du son voisé « a ».***

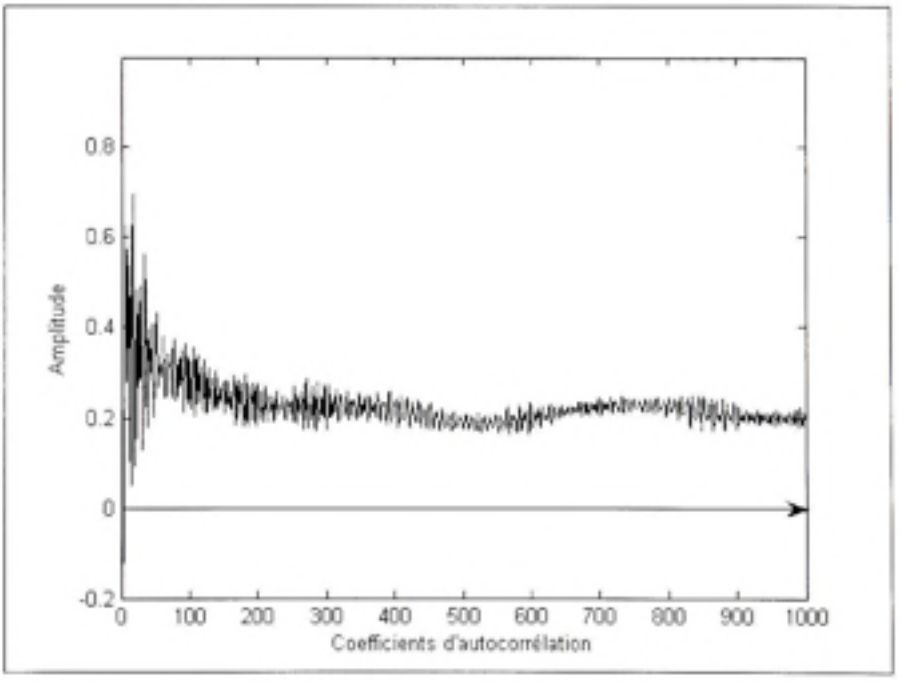

Figure 2.6 ***Coefficients d'autocorrélation du son non voisé « s ».***

Si l'autocorrélation est principalement utilisée pour la détection de la fréquence fondamentale (en multipliant la distance, en termes d'échantillons, entre deux pics d'amplitude maximale, par la fréquence d'échantillonnage), elle est très utile dans le domaine de la reconnaissance vocale lorsqu'il s'agit de différencier une zone voisée d'une zone non-voisée. Cette méthode nécessite toutefois d'importantes ressources, que ce soit en terme de puissance de calcul qu'en terme de stockage des coefficients.

2.3.5 Densité spectrale de puissance

La densité spectrale de puissance, qui nous présente la répartition fréquentielle de l'énergie d'un signal est déterminée, approximativement, à partir de la transformée de Fourier, par la formule 2.13 (Boite, 1987).

$$S(e^{j\omega}) = \left| X(e^{j\omega}) \right|^2 \tag{2.13}$$

Il est également possible d'obtenir la densité spectrale de puissance, à partir de l'autocorrélation du signal, selon la formule 2.14.

$$S(e^{j\omega}) = \sum_{n=0}^{N-1} \phi_{xx}(n).e^{-j\omega n} \tag{2.14}$$

Cette partie nous aura permis de présenter plusieurs outils généraux, utilisés également dans de nombreux autres domaines, mais fournisseurs d'informations importantes sur l'état du signal. Il existe toutefois de nombreuses autres façons d'extraire des informations du signal de parole, telles que les transformées en ondelettes, ou l'étude par banques de filtres. Néanmoins ces notions ne seront pas utilisées par la suite de ce mémoire, puisqu'elles sont notamment très coûteuses en calculs. C'est pourquoi nous ne les détaillerons pas, mais conseillerons de se reporter aux ouvrages spécialisés pour obtenir de plus amples informations sur ces outils. Il reste cependant deux autres notions sur lesquelles il est important de s'attarder : la modélisation autorégressive et le traitement homomorphique,

très utiles pour obtenir des renseignements sur l'état du signal vocal, mais dont l'intérêt majeur pour notre application est de permettre sa représentation au moyen d'un nombre restreint de paramètres. La prochaine partie sera donc consacrée à la description de ces méthodes, et des calculs devant être mis en œuvre pour paramétrer le signal.

2.4 Les outils de paramétrisation du signal vocal

Destinée à représenter les caractéristiques du signal vocal au moyen d'un nombre réduit de données, la paramétrisation est une étape nécessaire à de nombreux systèmes de traitement du signal, des appareils de télécommunication aux applications de reconnaissance de la parole. Allègement des calculs, diminution de la quantité de données à transmettre et réduction de la taille mémoire nécessaire sont autant d'avantages permis par le codage de la parole. Il existe de nombreuses façons de paramétrer le signal, la plupart d'entre elles se basant sur certaines propriétés spécifiques au signal vocal pour en extraire un maximum d'informations, en un minimum de données. C'est notamment le cas du codage par prédiction linéaire (souvent appelé LPC, pour *Linear Prediction Coding*), qui est une conséquence de la modélisation autorégressive de la parole, et qui peut ensuite donner lieu à d'autres méthodes de paramétrisation dérivées de ses coefficients.

2.4.1 Modélisation autorégressive du signal vocal

Une caractéristique très intéressante du signal vocal, mise en évidence sur la figure 1.7, est que sa production peut être assimilée à l'excitation d'un filtre par une source pseudo-périodique ou aléatoire. En d'autres termes, nous pouvons considérer le signal vocal comme étant la convolution de la source u(n) et de la réponse impulsionnelle du filtre représentant le conduit vocal, h(n), comme présenté sur la formule 2.15.

$$x(n) = u(n) \otimes h(n) \tag{2.15}$$

Si, à première vue, cette donnée semble jouer un grand rôle dans le domaine de la synthèse de la parole, étant donné qu'il suffit « simplement » d'injecter une source adéquate dans un filtre pour générer de la voix, elle est également très utile pour le codage et la reconnaissance de la parole. En effet, la possibilité de modéliser le conduit vocal par les coefficients d'un filtre permet de posséder des informations très précises sur le signal à partir d'un nombre de données très restreint. Nous allons donc décrire les étapes conduisant à une telle modélisation simplifiée de l'organe de production de la parole (Boite, 1987). Ramenée dans le domaine discret, la formule 2.15 devient la formule 2.16.

$$X(z) = U(z)H(z) \tag{2.16}$$

Où H(z), qui représente la transmittance du conduit vocal, est un filtre *tous-pôles* d'ordre p, dont la formule est donnée par l'équation 2.17.

$$H(z) = \frac{\sigma}{1 + \sum_{i=1}^{p} a(i).z^{-i}} \tag{2.17}$$

En rajoutant un terme correspondant au gain de la source d'excitation, nous pouvons donc dire que la production de la parole peut être modélisée par l'équation 2.18.

$$X(z) = \frac{\sigma.U(z)}{1 + \sum_{i=1}^{p} a(i).z^{-i}} \tag{2.18}$$

Le terme de modélisation autorégressive vient du fait qu'un signal est dit autorégressif lorsqu'il est engendré par récurrence, c'est-à-dire que chaque échantillon peut être déterminé à partir des échantillons précédents. Ce qui est clairement le cas ici, étant donné que si l'on repasse la formule 2.18 dans le domaine temporel, nous obtenons l'expression 2.19.

$$x(n) + \sum_{i=1}^{p} a(i).x(n-i) = \sigma.u(n) \tag{2.19}$$

Comme nous l'avons expliqué précédemment, l'excitation u(n) est soit une suite d'impulsions périodiques, soit un bruit blanc, et les coefficients a(i) représentent les caractéristiques du conduit vocal. Ces coefficients sont d'ailleurs appelés « coefficients de prédiction », car en cas d'excitation nulle, chaque échantillon x(n) pourrait être prédit à partir des p échantillons précédents, associés aux coefficients de prédiction. La figure 2.7 représente le modèle autorégressif de production de la parole.

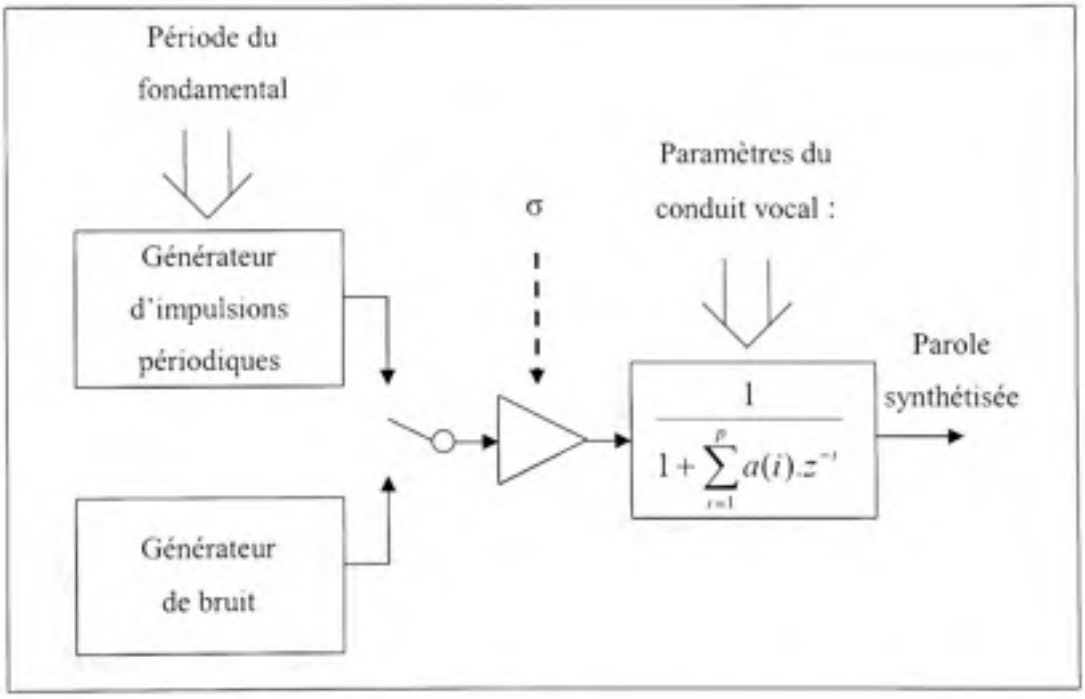

Figure 2.7 ***Modèle autorégressif de production de la parole.***

Pour mieux comprendre l'intérêt du modèle autorégressif, il est utile de visualiser son spectre. Le spectre d'un signal synthétique obtenu au moyen du système décrit par la figure 2.7 est en effet une approximation du spectre du signal qui aurait été généré au travers d'un conduit vocal dont on aurait déterminé les caractéristiques. Cette approximation étant d'autant meilleure que l'ordre du filtre de prédiction est élevé. Si l'on applique la formule

2.13 à la transformée de Fourier du signal synthétisé, on obtient donc la densité spectrale de ce signal, présentée par la formule 2.20.

$$S_X(e^{j\omega}) = S_U(e^{j\omega}).\left|\frac{\sigma}{1+\sum_{k=1}^{p} a(k).(e^{j\omega})^{-k}}\right|^2 \tag{2.20}$$

Où $S_U(e^{j\omega})$ est la densité spectrale de la source. On peut ainsi définir le spectre du modèle autorégressif par la formule 2.21.

$$S_M(e^{j\omega}) = \left|\frac{\sigma}{1+\sum_{k=1}^{p} a(k).(e^{j\omega})^{-k}}\right|^2 \tag{2.21}$$

Nous pouvons donc observer, sur la figure 2.8, la densité spectrale d'un modèle autorégressif d'ordre 20, superposé sur la densité spectrale du signal original, le son « a » déjà utilisé précédemment. La figure 2.9 présente quant à elle la densité spectrale du même modèle, mais pour un ordre de 100. L'objectif de cette comparaison étant de visualiser l'importance de l'ordre de prédiction, dans la précision de modélisation d'un signal.

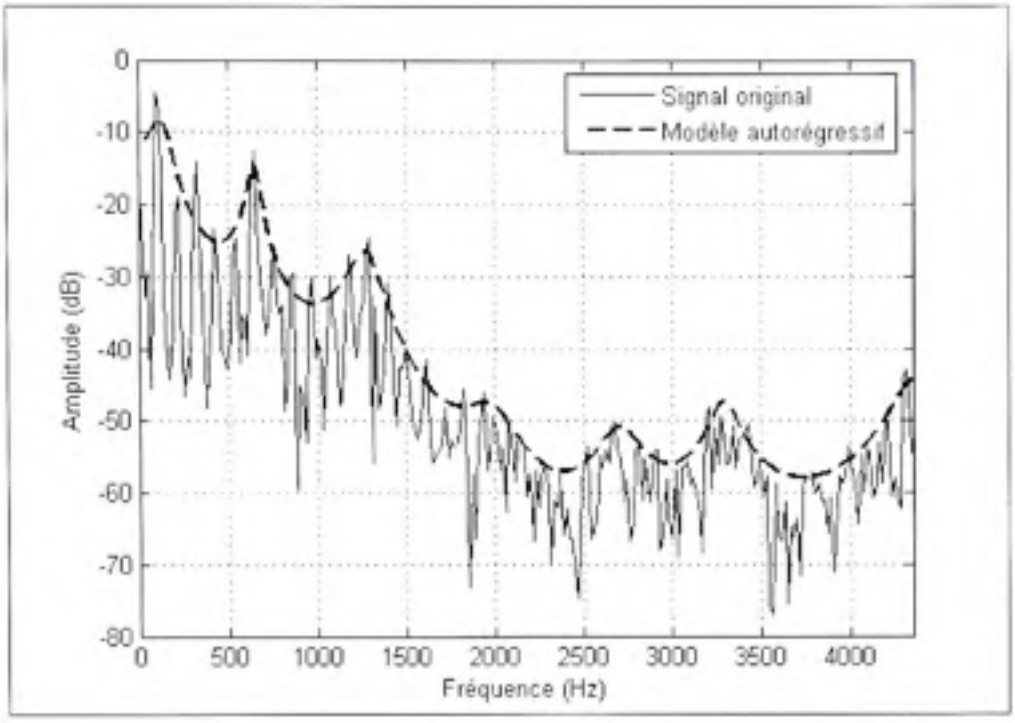

Figure 2.8 ***Densité spectrale d'un modèle autorégressif d'ordre 20.***

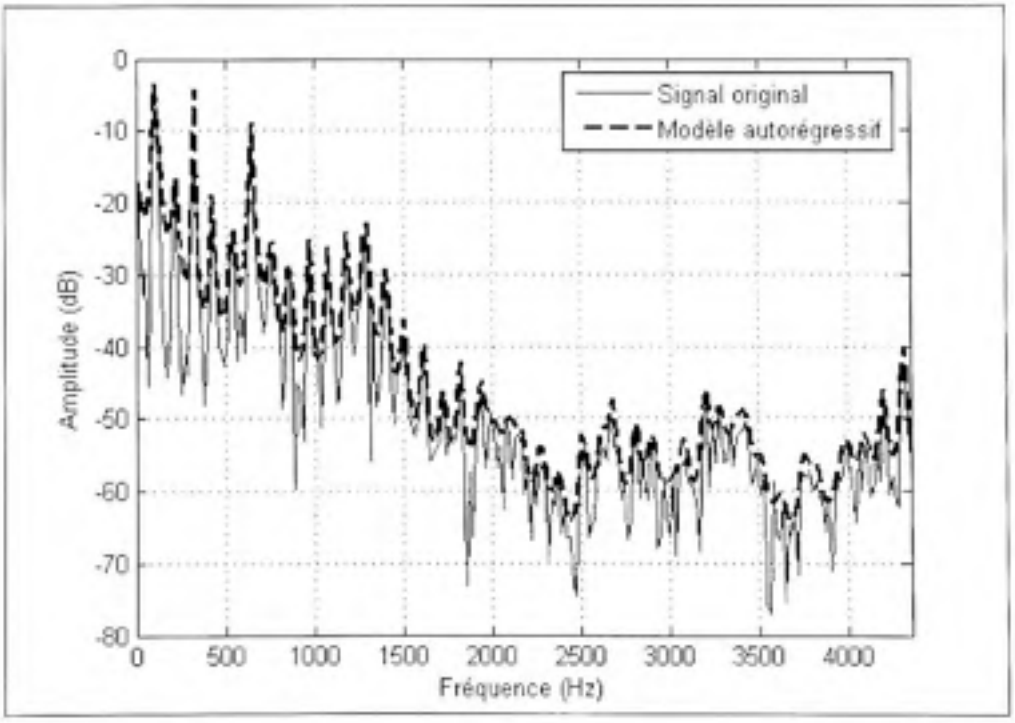

Figure 2.9 ***Densité spectrale d'un modèle autorégressif d'ordre 100.***

Nous observons ainsi que, pour une valeur de l'ordre de prédiction relativement faible, la fonction $S_M(e^{j\omega})$ reproduit fidèlement l'enveloppe de la densité spectrale réelle. Pour un ordre très élevé, la reproduction est quasi-parfaite. L'intérêt de cet outil réside donc dans le fait qu'il est possible de représente la majeure partie de l'information spectrale du signal à l'aide d'un nombre très restreint de paramètres : les coefficients de prédiction, le gain du modèle et la fréquence fondamentale de la source; là où la TFD nécessite une très grande quantité de données et de calculs pour arriver à peu près au même résultat. Nous allons donc décrire à présent l'algorithme permettant d'obtenir ces coefficients de prédiction linéaire.

2.4.2 Les coefficients LPC

Introduite par Wiener en 1966, la notion de prédiction linéaire a depuis été reprise dans de nombreuses applications. Utilisée pour la première fois dans le domaine du traitement et de la synthèse de la parole par Itakura et Saito en 1968, elle tire son intérêt du fait qu'elle utilise idéalement le caractère autorégressif du modèle de production de la parole, pour représenter efficacement l'information vocale au moyen d'une dizaine de coefficients, obtenus à l'aide d'une faible quantité de calculs. Rabiner et Schafer (1978) consacrent une grande partie de leur ouvrage à l'utilisation de la paramétrisation LPC dans le domaine de la reconnaissance vocale, et détaillent les étapes nécessaires au calcul des coefficients, la première étant l'établissement d'un système d'équations à résoudre, les équations de Yule-Walker.

2.4.2.1 Établissement des équations de Yule Walker

Nous avons vu que le signal vocal obéissait au modèle autorégressif d'ordre p défini par la formule 2.19, faisant de lui un signal autorégressif. Cette formule peut également s'écrire d'une autre façon, présentée par l'équation 2.22.

$$\tilde{x}(n) = -\sum_{k=1}^{p} a_k x(n-k) \qquad (2.22)$$

Où $\tilde{x}(n)$ est une estimation du signal original. La différence entre cette estimation et le signal réel est appelée erreur de prédiction linéaire, notée e(n), définie par la formule 2.23.

$$e(n) = x(n) - \tilde{x}(n) = x(n) + \sum_{k=1}^{p} a_k x(n-k) \tag{2.23}$$

Plus généralement, on estimera donc l'erreur de prédiction selon la formule présentée par l'équation 2.24.

$$e(n) = \sum_{k=0}^{p} a_k x(n-k) \qquad avec \quad a_0 = 1 \tag{2.24}$$

Il va de soi que lorsque les coefficients de prédiction sont exacts, l'erreur n'est pas nulle, mais correspond au signal d'excitation, et le signal vocal est ainsi parfaitement modélisé. Le critère usuellement retenu pour estimer les coefficients de prédiction est la minimisation de l'énergie de l'erreur de prédiction, appelée énergie résiduelle, dont la formule est donnée par l'équation 2.25.

$$E_p = \sum_{n=0}^{N+p-1} e^2(n) \tag{2.25}$$

Soit, en associant les formules 2.24 et 2.25 :

$$E_p = \sum_{n=0}^{N+p-1} \left[x(n) + \sum_{k=1}^{p} a_k x(n-k) \right]^2 \tag{2.26}$$

Les valeurs de a_k qui minimisent E_p sont donc les valeurs qui annulent sa dérivée. Il faut donc dériver E_p par rapport à chacun des coefficients a_k, comme présenté par la formule 2.27.

$$\frac{\partial E_p}{\partial a_k} = 0, \qquad k = 1,2,\ldots, p \tag{2.27}$$

Il reste ainsi à résoudre un système de p équations à p inconnues, dites équations de Yule-Walker, définies par la relation 2.28.

$$-\sum_{n=0}^{N+p-1} x(n-i).x(n) = \sum_{k=1}^{p} a_k \sum_{n=0}^{N+p-1} x(n-i).x(n-k), \qquad i = 1,\ldots, p \tag{2.28}$$

Pour simplifier ces équations, on peut faire intervenir la formule de l'autocovariance, donnée par l'équation 2.29, permettant ainsi d'exprimer les équations de Yule-Walker sous la forme donnée par la formule 2.30.

$$\phi(i,k) = \sum_{n} x(n-i).x(n-k) \tag{2.29}$$

$$\boxed{\phi(i,0) = \sum_{k=1}^{p} a_k.\phi_n(i,k)} \tag{2.30}$$

Pour résoudre ce système, les deux méthodes les plus couramment utilisées sont la méthode de l'autocorrélation, introduite pour la première fois par Itakura et Saito en 1968, et la méthode de la covariance, établie par Atal et Schroeder en 1968. Nous allons présenter la première méthode, car c'est celle que nous utiliserons au cours de notre étude. Son hypothèse de départ est la réduction de la fonction d'autocovariance, présentée en 2.30, à la fonction d'autocorrélation, comme présenté par l'équation 2.31.

$$\phi(i,k) = r(i-k) = \sum_{n=0}^{N-1-(i-k)} x(n).x(n+i-k) \tag{2.31}$$

On peut ainsi exprimer les équations de Yule-Walker sous la forme donnée par la formule 2.32, ou bien, pour plus de clarté, sous une forme matricielle, présentée par l'équation 2.33.

$$\sum_{k=1}^{p} a_k r(i-k) = -r(i), \qquad i = 1,\ldots, p \tag{2.32}$$

$$\begin{bmatrix} r(0) & r(1) & r(2) & . & . & r(p-1) \\ r(1) & r(0) & r(1) & . & . & r(p-2) \\ r(2) & r(1) & r(0) & . & . & r(p-3) \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ r(p-1) & r(p-2) & r(p-3) & . & . & r(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ . \\ . \\ a_p \end{bmatrix} = - \begin{bmatrix} r(1) \\ r(2) \\ r(3) \\ . \\ . \\ r(p) \end{bmatrix} \tag{2.33}$$

Ces équations jouent donc un rôle très important, car ce sont elles qui nous permettent de déterminer les coefficients de prédiction. Mais résoudre un tel système d'équations par des méthodes algébriques classiques demanderait un nombre faramineux de calculs. C'est pour cette raison que des algorithmes spéciaux de résolution ont été mis en place. Outre les algorithmes de Burg (1975) et Schlur (1917), la méthode la plus utilisée dans le domaine du codage de la parole est celle de Levinson-Durbin (1960), qui solutionne le système au moyen d'un algorithme récursif.

2.4.2.2 Algorithme de Levinson-Durbin

Cette technique de résolution utilise la structure particulière de la grande matrice d'autocorrélation présente dans la formule 2.33. En effet, cette matrice est Toeplitz symétrique, c'est-à-dire que la diagonale joue le rôle d'axe de symétrie. Cela permet de simplifier grandement les calculs, en autorisant une résolution par récursivité sur l'ordre de prédiction. Rowden (1991) présente cet algorithme de la façon suivante :

a) pour le premier tour, calculer le terme d'erreur initial E_0, le coefficient de réflexion k_1 et le coefficient prédicteur $\alpha_1(1)$;

$$E_1 = r(0) \tag{2.34}$$

$$k_1 = \frac{r(1)}{E_0} \tag{2.45}$$

$$\alpha_1(1) = k_1 \tag{2.36}$$

b) à partir du second tour, et jusqu'au p^{ième} tour, en notant j l'indice du tour, effectuer les opérations des équations 2.37, 2.38, 2.39 et 2.40;

$$k_j = \frac{r(j) - \sum_{n=1}^{j-1} \alpha_{j-1}(n).r(j-n)}{E_{j-1}} \tag{2.37}$$

$$\alpha_j(j) = k_j \tag{2.38}$$

$$\alpha_j(m) = \alpha_{j-1}(m) - k_j.\alpha_{j-1}(j-m) \qquad m = 1,\ldots, j-1 \tag{2.39}$$

$$E_j = E_{j-1}.(1 - k_j^{\,2}) \tag{2.40}$$

c) enfin, déterminer les coefficients de prédiction, solutions du système, à partir des coefficients α calculés au cours de l'algorithme.

$$\boxed{a_j = \alpha_p(m) \qquad m = 1,\ldots, p} \tag{2.41}$$

Largement utilisés dans la plupart des systèmes de traitement de la parole, les coefficients LPC ont également donné lieu à d'autres représentations paramétriques du signal, notamment dans le but d'améliorer l'efficacité de la quantification nécessaire à la transmission du signal, domaine auquel se prêtent mal les coefficients de prédiction. Ainsi,

les coefficients de réflexion, ou coefficients PARCOR (pour PARtial CORrelation), les coefficients Log Area Ratio (LAR) et les coefficients Line Spectrum Pair (LSP) apparaîtront souvent dans les systèmes de télécommunication. Néanmoins ils sont très peu utilisés dans le domaine de la reconnaissance vocale, et on leur préfère grandement d'autres paramètres, pouvant également être déterminés à partir de l'analyse LPC, les coefficients LPCC, appelés également coefficients cesptraux, auxquels nous consacrerons la dernière partie de ce chapitre, après avoir défini la notion de cepstre.

2.4.3 Étude cepstrale

L'analyse cepstrale est une autre façon d'étudier le signal de parole, le premier à utiliser cette technique en traitement du signal vocal étant Peter Noll en 1967. L'hypothèse de départ est la même que précédemment : le signal de parole résulte de la convolution d'une excitation et d'une réponse impulsionnelle. L'analyse cepstrale a pour objectif de séparer les contributions de la source et du conduit vocal, apportant ainsi certains avantages par rapport à la transformée de Fourier : d'une part l'observation d'un spectre lissé correspondant aux caractéristiques du conduit vocal, et d'autre part une détermination aisée de la fréquence fondamentale. De très bonnes explications sur l'analyse cepstrale sont fournies par Calliope (1989) ou Rabiner et Schafer (1978).

Le point de départ de l'analyse cepstrale est le traitement homomorphique qui consiste à effectuer une déconvolution du signal donné par la formule 2.15. Le fonctionnement de cette déconvolution est donné par la figure 2.10.

Figure 2.10 ***Traitement homomorphique pour le calcul du cepstre.***

Ainsi, la première étape est d'effectuer une TFD du signal vocal, pour obtenir le résultat présenté par la formule 2.42.

$$X(k) = U(k).H(k) \tag{2.42}$$

La TFD a donc transformé la convolution des deux signaux en une multiplication, rendant ainsi la séparation plus aisée. En effet, nous savons que le logarithme d'un produit de deux signaux donne une addition des logarithmes de ces signaux. L'étape suivante est ainsi donnée par la formule 2.43.

$$\hat{X}(\omega) = Log[|X(\omega)|] = Log[|U(\omega)|] + Log[|H(\omega)|] \tag{2.43}$$

Que nous noterons donc selon la formule 2.44.

$$\hat{X}(\omega) = \hat{U}(\omega) + \hat{H}(\omega) \tag{2.44}$$

Une transformée de Fourier inverse nous ramène alors dans un domaine pseudo-temporel appelé domaine quéfrentiel. Le signal ainsi obtenu est présenté par la formule 2.45.

$$\tilde{c}(n) = \tilde{u}(n) + \tilde{h}(n) \tag{2.45}$$

Les $\tilde{c}(n)$ sont les coefficients cepstraux approchés du signal x(n), alors que $\tilde{u}(n)$ et $\tilde{h}(n)$ sont les transposées dans le domaine quéfrentiel de $u(n)$ et $h(n)$. La figure 2.11 nous montre la représentation du cepstre sur l'échelle quéfrentielle (qui s'exprime en secondes).

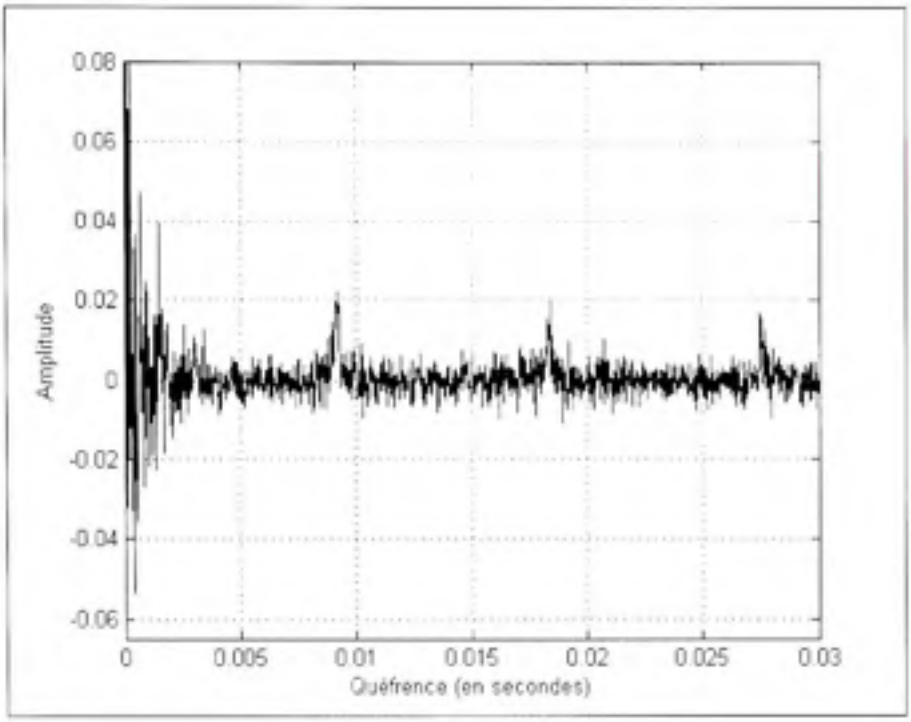

Figure 2.11 ***Cepstre du son voisé « a ».***

Cette figure, qui semble présenter une structure périodique à partir d'un certain point, peut s'expliquer par les hypothèses suivantes :

- $u(n)$ est la contribution de l'excitation. C'est donc une suite d'impulsions séparées de n_0 échantillons (où n_0 correspond à la fréquence fondamental F_0). Sur la figure 2.11, nous observons effectivement une répétition régulière de pics, qui est donc due à la présence de $\tilde{u}(n)$ dans le cepstre $\tilde{c}(n)$;

- $h(n)$ est la réponse impulsionnelle d'un filtre à réponse impulsionnelle finie (RIF). Ses valeurs sont donc élevées au début, mais décroissent au cours des échantillons, pour devenir rapidement négligeables lorsque l'on arrive à n_0. C'est effectivement le phénomène observé sur les premiers échantillons du cepstre.

En isolant les signaux $\tilde{u}(n)$ et $\tilde{h}(n)$, il est donc possible d'observer la réponse fréquentielle du conduit vocal, ainsi que la fréquence fondamentale de l'excitation. Étant donné que la

contribution de $\tilde{h}(n)$ dans $\tilde{c}(n)$ a lieu principalement sur les n_0 premiers échantillons, une simple fenêtre temporelle F de longueur n_0, comme présentée par la formule 2.46, nous permet d'isoler $\tilde{h}(n)$.

$$\begin{cases} F_n = 1 & \text{si } n < n_0 \\ F_n = 0 & \text{si } n \geq n_0 \end{cases} \tag{2.46}$$

On peut également isoler $\tilde{u}(n)$ en utilisant un filtre qui ne retiendrait que les échantillons après n_0. Pour retrouver les signaux originaux u(n) et h(n) exprimés dans un domaine temporel, il nous faut ensuite réaliser un traitement homomorphique inverse de celui réalisé auparavant, qui éliminerait ainsi la composante logarithmique. La figure 2.12 présente un tel traitement.

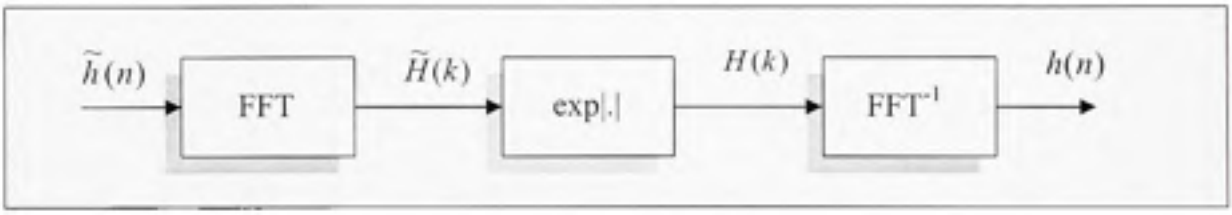

Figure 2.12 ***Traitement homomorphique inverse.***

Les figures suivantes présentent le résultat de l'analyse cepstrale. Nous rappelons tout d'abord, au moyen de la figure 2.13, la densité spectrale de puissance du signal étudié, puis présentons les densités spectrales de puissance de h(n) sur la figure 2.14 et u(n) sur la figure 2.15.

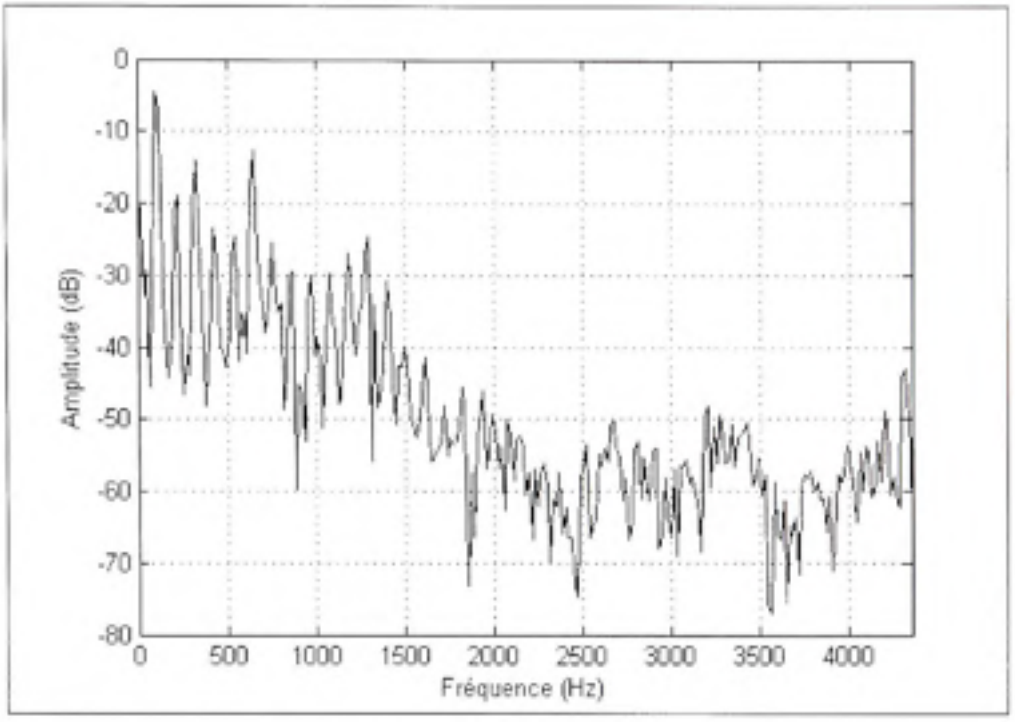

Figure 2.13 ***Densité spectrale du son voisé « a ».***

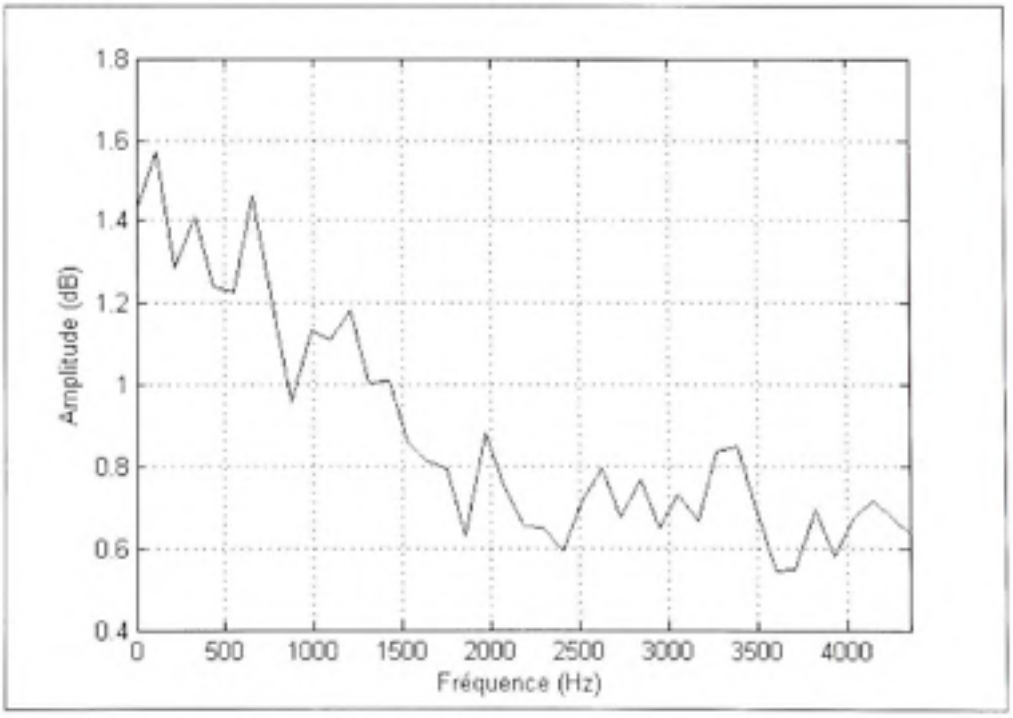

Figure 2.14 ***Densité spectrale du conduit vocal.***

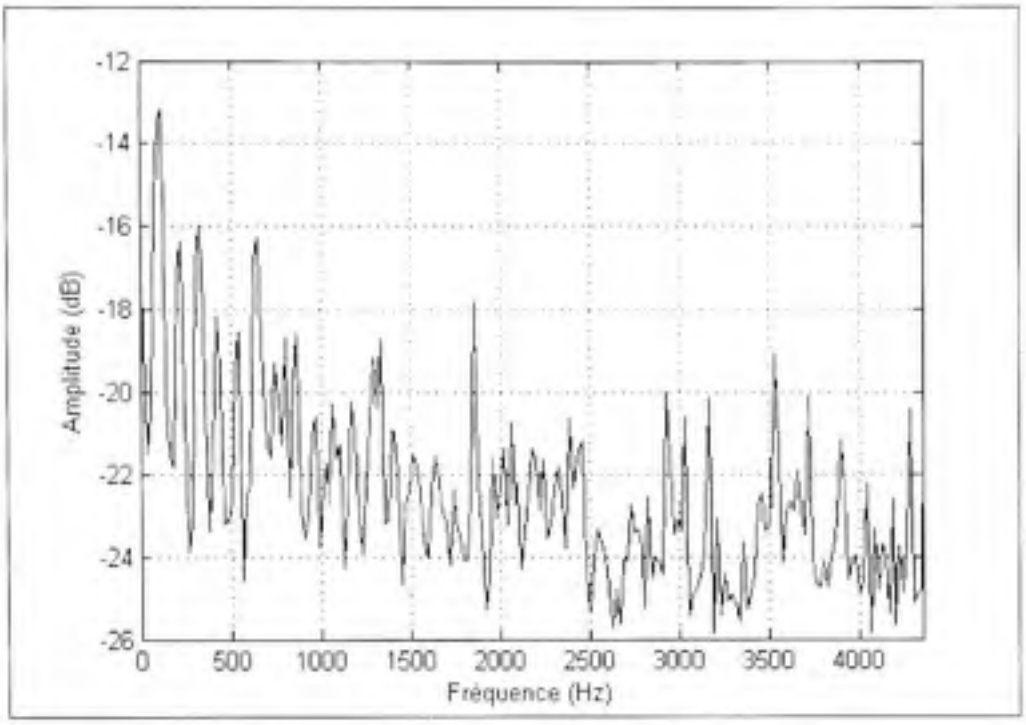

Figure 2.15 ***Densité spectrale de la source.***

Il est donc fortement possible, grâce à l'analyse cepstrale, de déterminer avec une bonne qualité l'enveloppe spectrale du conduit vocal, ainsi que le signal d'excitation. Cela permet, pour l'analyse du signal vocal, d'extraire certaines caractéristiques intéressantes sur le signal étudié. L'analyse cepstrale peut d'ailleurs être effectuée dans bien d'autres domaines que le traitement du signal vocal, dès lors que l'on souhaite séparer différents signaux convolués. Une bonne étude de ce que représente le cepstre était nécessaire, étant donné que c'est un outil très important pour extraire des informations caractérisant le signal vocal, malheureusement, la détermination du cepstre par l'intermédiaire de la TFD est assez couteux en calculs, et donc peu adapté à une utilisation sur un système à faibles capacités. Toutefois, il est possible de l'estimer au moyen des coefficients LPCC, obtenus à partir des coefficients de prédiction linéaire. Nous allons donc tout d'abord décrire les étapes de ce processus, pour ensuite mentionner les raisons pour lesquelles l'utilisation des coefficients cepstraux peut s'avérer très judicieuse pour la conception de notre système de reconnaissance vocale.

2.4.4 Les coefficients LPCC

2.4.4.1 Détermination des coefficients LPCC

La méthode d'obtention des coefficients LPCC à partir des coefficients LPC a été introduite pour la première fois par Markel et Gray en 1976. On obtient tout d'abord le premier coefficient au moyen de la formule 2.47.

$$c_0 = \ln(\sigma^2) \tag{2.47}$$

Où σ^2 est le gain du modèle LPC. Les coefficients restants sont obtenus à l'aide de la formule 2.48.

$$c_m = a_m + \sum_{k=1}^{m-1}\left(\frac{k}{m}\right)c_k\, a_{m-k} \qquad m = 1,\ldots, p \tag{2.48}$$

Généralement, l'ordre de la représentation cepstrale est plus grand que l'ordre de prédiction. Ainsi, les coefficients c_m, pour m supérieur à p, sont calculés selon la formule 2.49.

$$c_m = \sum_{k=1}^{m-1}\left(\frac{k}{m}\right)c_k\, a_{m-k} \qquad pour\ \ m > p \tag{2.49}$$

De nombreuses études, notamment celles de Juang et al en 1986, Itakura et Umezaki en 1987 ou Junqua et al en 1993, ont prouvé que, parmi les différents types de paramétrisation, les coefficients cepstraux étaient ceux qui fournissaient les meilleurs résultats en terme d'efficacité de la reconnaissance. Réputés pour être particulièrement robustes, et nécessitant une faible quantité de calculs, les coefficients cepstraux s'avèrent donc être la forme de paramétrisation idéale à la conception d'une méthode robuste de reconnaissance vocale destinée à fonctionner sur un processeur de traitement du signal. De plus, nous allons

maintenant présenter diverses caractéristiques de ces coefficients, observées au cours de notre étude, qui pourront s'avérer fort utiles par la suite.

2.4.4.2 Caractéristiques particulières des coefficients LPCC

La figure 2.16 présente les valeurs des coefficients LPCC du son « a » utilisé précédemment, pour un ordre de paramétrisation fixé à 20. Les valeurs des coefficients ont été reliées entre elles, de façon à mieux visualiser leur évolution, au fur et à mesure que l'on se rapproche du coefficient d'ordre maximal.

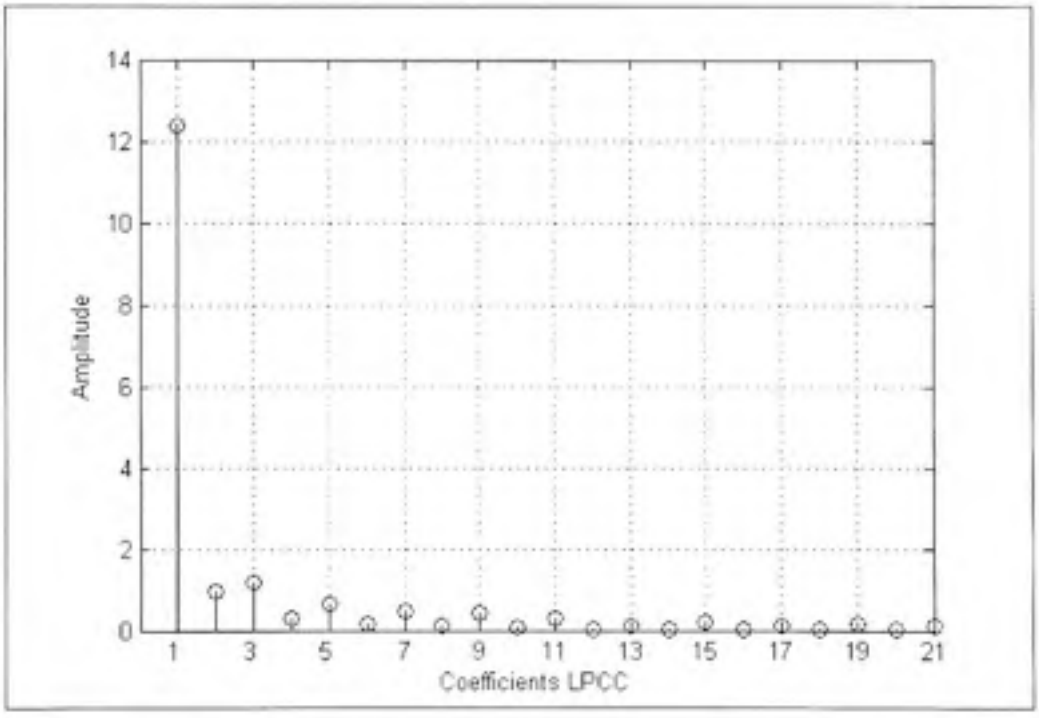

Figure 2.16 ***Coefficients LPCC du son « a », pour un ordre 20.***

Deux caractéristiques sont à retenir de cette figure. Tout d'abord, l'amplitude du premier coefficient est nettement plus élevée que celles des coefficients suivants. Deuxièmement, il semblerait que plus on s'éloigne du premier coefficient, plus l'amplitude diminue, selon une décroissance exponentielle, pour se stabiliser autour d'une valeur très faible, à partir du onzième coefficient dans le cas de la figure 2.16. La différence, nettement visible, entre le premier coefficient et les autres est expliquée par la formule 2.47. Il est en effet obtenu à

partir du gain du modèle LPC, là où les suivants sont obtenus par récursivité à partir des coefficients précédents, et des coefficients de prédiction (de faible amplitude). Si cette caractéristique nous intéresse, c'est tout simplement car elle signifie que le premier coefficient LPCC est porteur d'informations sur l'énergie du signal. La figure 2.17 va nous permettre de mieux observer cette propriété.

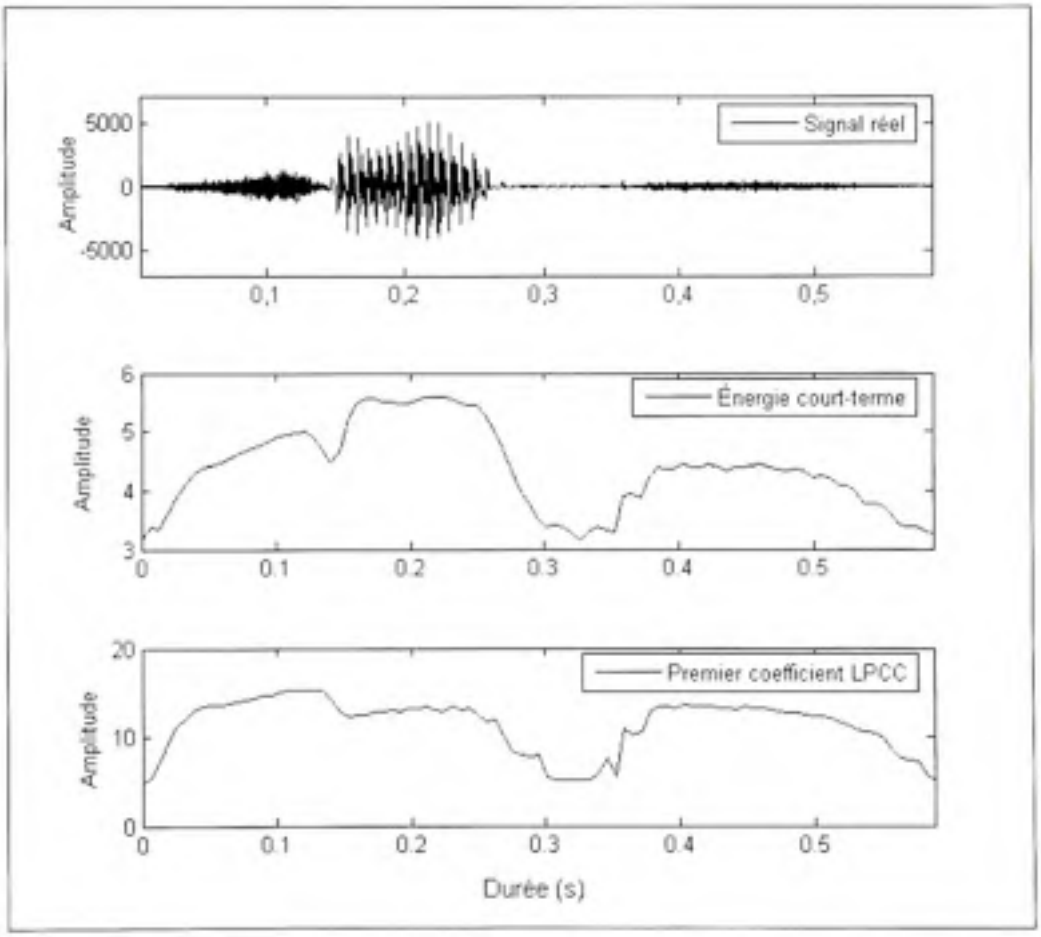

Figure 2.17 ***Évolution de l'énergie et du premier coefficient LPCC pour le chiffre 6.***

On observe aisément, au moyen de la figure 2.17, que les évolutions de l'énergie et du premier coefficient LPCC suivent globalement la même courbe. Très intéressante, cette caractéristique pourrait notamment nous permettre d'éviter d'utiliser du temps de calcul pour calculer l'énergie du signal, quand le coefficient déjà déterminé peut jouer ce rôle.

Nous avons pu effectuer une autre observation intéressante, concernant cette fois-ci les second et troisième coefficients : leur capacité à se comporter différemment selon que l'on est en présence de son voisé ou non voisé. Plus précisément, leur valeur est positive en présence de son voisé, et devient négative en présence de son non voisé. La figure 2.18 présente cette caractéristique, en affichant l'évolution des ces deux coefficients pour le chiffre 6.

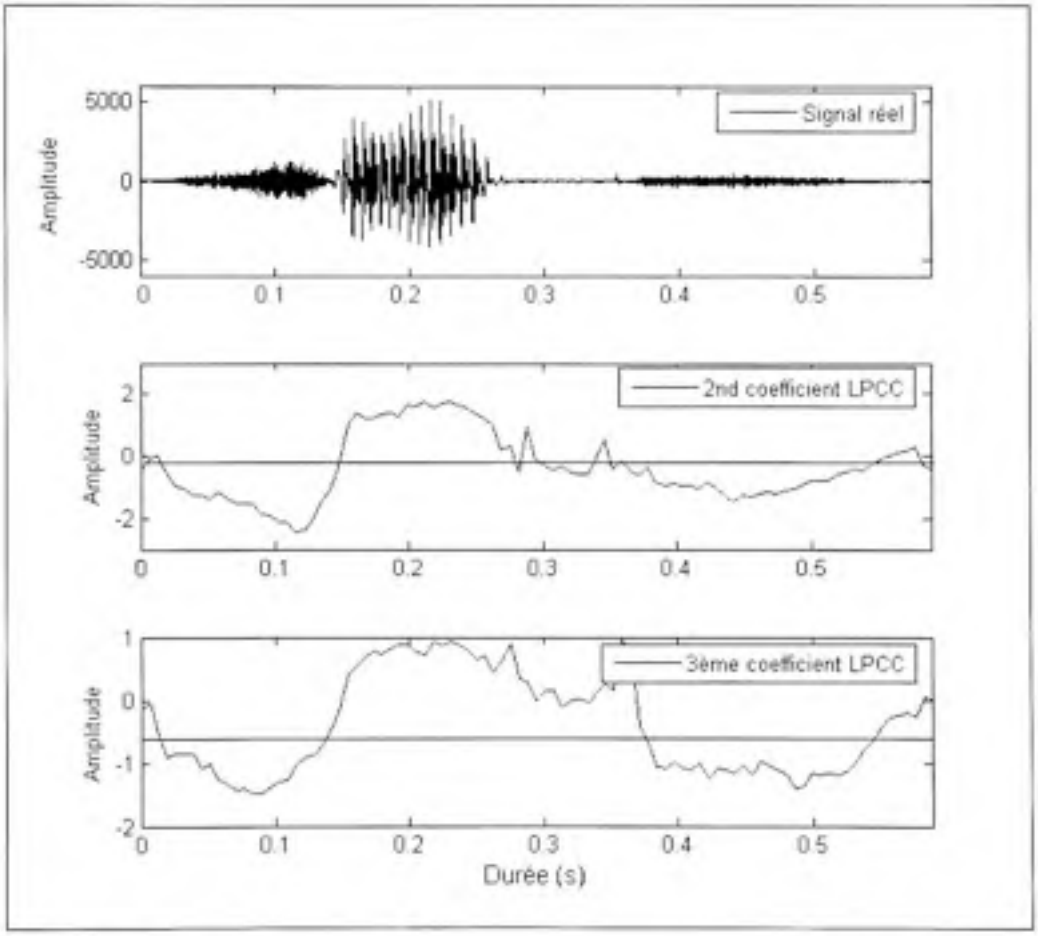

Figure 2.18 ***Évolution des second et troisième coefficients LPCC pour le chiffre 6.***

Parfaitement adapté pour étudier cette propriété, le chiffre 6 (en anglais) est constitué de parties non voisées à chaque extrémité, et d'une zone voisée suivie d'une zone de silence au centre. Sur la figure 2.18, les coefficients 2 et 3 présentent bien des valeurs négatives aux deux extrémités du chiffre, tandis que la zone voisée est également parfaitement repérée grâce aux valeurs positives. Quant à la zone de silence, si le coefficient 3 semble

parfaitement faire la distinction avec la zone non voisée, le coefficient 2 oscille entre valeurs positives et négatives.
Les coefficients d'ordre supérieur ne semblent pas présenter d'information particulière sur le signal, lorsqu'ils sont pris séparément. Utilisés ensemble, ils sont en revanche porteurs de renseignements sur les caractéristiques plus fines du signal.

Ces constatations effectuées sur le comportement des coefficients LPCC vont nous aider dans notre choix d'utiliser ce jeu de paramètres pour concevoir le système de reconnaissance vocale. En effet, contrairement aux coefficients LPC, qui représentent les coefficients d'un filtre, et n'ont donc aucun intérêt lorsqu'ils sont pris séparément les uns des autres, les coefficients LPCC portent des informations directes sur l'état du signal, comme l'énergie, ou la présence de voisement. En plus de disposer d'une représentation robuste de la parole, il nous sera donc possible d'utiliser ces paramètres, séparément ou ensemble, pour optimiser le plus possible notre méthode.

Ce chapitre nous aura donc permis de présenter tous les éléments indispensables à la réalisation d'un système de reconnaissance de la parole, en ce qui concerne la phase d'extraction de l'information contenue dans l'onde sonore. Diverses méthodes de traitement ont été étudiées, allant du simple calcul de l'énergie jusqu'aux procédés se basant sur la structure particulière du processus de traitement de la parole, et une méthode de paramétrisation adaptée à nos besoins a été retenue pour représenter le signal. En l'état, nous allons pouvoir présenter, au cours du prochain chapitre, les différents procédés généralement utilisés pour transcrire l'information extraite du signal en un jeu de caractères représentant la signification visuelle du son émis par le locuteur.

CHAPITRE 3

LA RECONNAISSANCE VOCALE

3.1 Introduction

Sujet de nombreuses recherches depuis plusieurs décennies, la reconnaissance automatique de la parole est considérée comme le Saint Graal des chercheurs en traitement du signal vocal. Ayant pour objectif de faciliter la relation homme-machine en annihilant le lien physique qui oblige souvent l'être humain à utiliser ses mains pour interagir avec une machine ou un ordinateur, cette technologie suscite bien des attentes dans de nombreux domaines. Si le milieu industriel prévoit l'utilisation de la reconnaissance vocale pour des commandes de machines ou des contrôles de processus, les applications grand public en feront plutôt un usage dédié à l'amélioration de la vie quotidienne, comme les logiciels de dictée automatique ou la composition vocale des numéros de téléphone. La grande diversité des tâches et des conditions d'utilisation est, en plus de la complexité du signal vocal, la raison principale pour laquelle plus de cinquante ans de recherche ont été nécessaires pour permettre la conception de systèmes fonctionnels.

Pour mieux situer le problème, il est intéressant de présenter les progrès effectués depuis les débuts de la recherche sur la reconnaissance vocale, à la fin des années 40. Si nous ne décrirons pas en détail toutes les avancées réalisées au cours des cinq dernières décennies, dont Furui (2005) fournit un très bon résumé, nous présenterons en revanche les grandes phases de recherche qui nous ont amenés à la situation actuelle.

3.1.1 Historique

Depuis la naissance de la reconnaissance automatique de la parole en 1949 et le sténo-sonographe de J. Dreyfus-Graf, on peut dire que chaque décennie aura apporté son lot de découvertes. Les années 50 auront ainsi vu naître des avancées significatives sur les

systèmes monolocuteurs, avec notamment le premier reconnaisseur de chiffres, en employant principalement les unités phonétiques du langage.

C'est au cours des années 60 que sont apparus les premiers systèmes utilisant l'électronique numérique, permettant ainsi de réaliser de nombreux progrès, principalement grâce à l'augmentation de la taille des vocabulaires utilisés, et à la numérisation des banques de filtres. C'est également au cours de cette décennie qu'ont été utilisés pour la première fois les coefficients de prédiction linéaire ainsi que les algorithmes de programmation dynamique, deux avancées importantes pour les recherches futures, ces derniers ne commençant à être réellement utilisés qu'au début des années 80.

La recherche dans le domaine de la reconnaissance vocale a pris toute son ampleur à partir des années 70, en grande partie grâce à l'apparition des méthodes statistiques basées sur les chaînes de Markov, qui utilisent une modélisation stochastique du langage. Nécessitant toutefois une puissance de calcul très élevée et un apprentissage très long, ces méthodes commenceront à être de plus en plus utilisées à partir de la décennie suivante, avec l'apparition des premiers processeurs de traitement du signal.

Les années 80 auront donc surtout servi à essayer de profiter des processeurs de plus en plus puissants pour appliquer les méthodes développées auparavant, les recherches se concentrant alors sur trois axes principaux :

- les méthodes basées sur la reconnaissance de forme, et la programmation dynamique;
- la reconnaissance se fondant sur approche probabiliste de la production de la parole;
- la reconnaissance utilisant les réseaux de neurones et les systèmes experts.

Depuis cette période, et jusqu'à nos jours, aucune avancée majeure n'a été effectuée, et la recherche est quelque peu en stagnation, tentant surtout d'améliorer les méthodes développées au cours des décennies précédentes, ou de travailler sur des façons

d'augmenter la qualité de la reconnaissance en conditions défavorables. Ces dernières années ont en effet vu apparaître de plus en plus d'applications intégrant une fonction de reconnaissance vocale, entraînant par-là même des utilisations dans des environnements variés, loin des conditions idéales des expériences réalisées en laboratoire. Ainsi, les applications de reconnaissance automatique de la parole embarquées dans les téléphones cellulaires, ou autres appareils transportables, amènent les chercheurs à essayer d'améliorer l'efficacité des méthodes en milieux bruités, tandis que les logiciels de dictée vocale, encore tous basés sur les modèles de Markov, nécessitent de meilleurs algorithmes d'apprentissage ou des vocabulaires de plus en plus grands.

Ces cinquante années de recherche ont donc permis d'aboutir à des méthodes efficaces, utilisées dans de plus en plus d'applications, mais encore loin d'être parfaites, et il faudra encore de longues études avant de trouver la méthode universelle permettant à l'être humain de communiquer oralement avec une machine dans toutes les circonstances.

3.1.2 Classification des méthodes de reconnaissance vocale

Si les recherches sur la reconnaissance vocale ont emprunté de nombreuses voies, c'est en partie à cause des différents types de problèmes à résoudre. Tout d'abord au niveau de la façon d'aborder la tâche à accomplir, et ensuite au niveau des conditions associées à la réalisation de cette tâche.

Pour le premier point, il existe effectivement différents types de tâches, selon la nature du système que l'on souhaite créer (Calliope, 1989) :

- reconnaissance ou compréhension : à savoir si l'on souhaite simplement reconnaître les mots, ou bien faire en sorte que la machine comprenne également leur signification, comme dans le cadre d'un dialogue;
- système « auto-organisateur » ou « fondé sur des connaissances », entre lesquels la différence vient principalement de la façon de traiter le signal : simplement extraire

les informations à l'aide d'outils mathématiques, ou bien se fonder directement sur les propriétés linguistiques pour effectuer un choix;

- reconnaissance globale par mots ou reconnaissance analytique, où il s'agit de déterminer si l'on souhaite identifier un mot en le considérant dans son ensemble, ou bien en effectuant une segmentation *à priori* pour appliquer ensuite un traitement spécifique selon la nature du segment.

Une fois le type de méthode sélectionné, viennent ensuite s'ajouter ce que l'on appelle les facteurs de complexité, qui dépendent du type d'application visé :

- reconnaissance mono-locuteur ou multi-locuteurs;
- reconnaissance de mots isolés ou de parole continue;
- reconnaissance d'un vocabulaire restreint ou large;
- reconnaissance en environnement calme ou bruyant.

Ces critères, qui viennent réduire ou augmenter la complexité de la reconnaissance, sont les principaux éléments à prendre en compte lors de la conception d'une méthode de reconnaissance de la parole, afin de s'adapter aux conditions d'utilisation du système. Ainsi, nous pouvons préciser la classification de la méthode développée au cours de ce mémoire : l'objectif sera de concevoir une méthode s'appuyant sur des paramètres suffisamment robustes pour garantir de bonnes performances en milieu bruité, en vue d'une simple identification des mots prononcés aussi bien de façon isolée que continue, par des locuteurs différents, et pour un vocabulaire restreint constitué des dix chiffres de 1 à 0. Quant à la façon d'accomplir cette tâche, le caractère continu de la prononciation des mots nécessitera une méthode basée sur la segmentation du signal, et une bonne connaissance du vocabulaire concerné.

3,1.3 Architecture d'une méthode de reconnaissance vocale

Bien que très variés et utilisant différentes approches, les systèmes de reconnaissance vocale se basent généralement sur la même architecture, constituée de plusieurs étapes, que l'on peut visualiser au moyen de la figure 3.1.

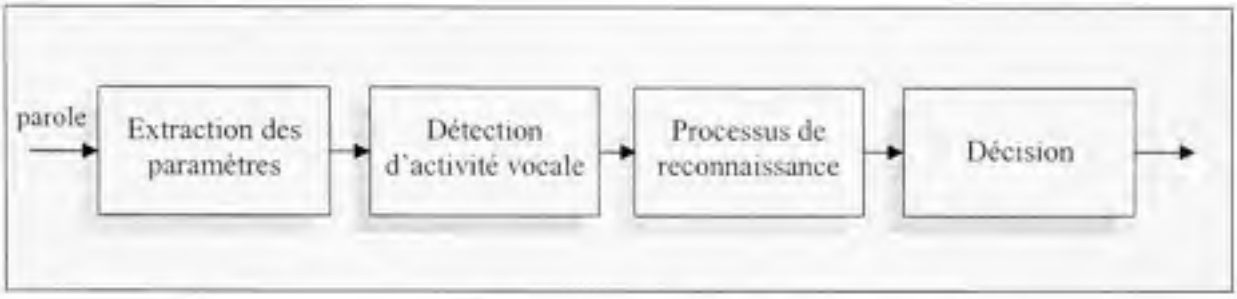

Figure 3.1 ***Structure d'une méthode de reconnaissance vocale.***

Ainsi, après la phase d'extraction des paramètres que nous avons présentée au chapitre 2, une détection d'activité vocale est habituellement utilisée afin de repérer les limites du mot ou de la chaîne de mots que l'on souhaite reconnaître. Cette étape peut d'ailleurs se situer avant la phase d'extraction des paramètres, mais les données nécessaires à la détection d'activité vocale sont généralement déterminées au cours de celle ci. La seconde grande partie de ce chapitre sera donc consacrée à la présentation des méthodes utilisées pour différencier la parole du bruit de fond.

Une fois le signal vocal isolé, vient le cœur du système : la phase de reconnaissance et la prise de décision. Comme mentionné lors de l'historique, différentes approches peuvent être utilisées pour décoder le signal de parole en se basant sur les paramètres extraits auparavant. Nous en retiendrons particulièrement deux : tout d'abord l'approche acoustico-phonétique, qui se fonde sur une très bonne connaissance des propriétés acoustiques de la parole pour associer une unité phonémique à chaque région du signal et ainsi déterminer quel est le mot ou la chaîne de mots qui auraient pu être produits à partir de la séquence de phonèmes observée. La seconde approche, de type « reconnaissance de forme », s'appuie sur la comparaison des paramètres du signal avec un ensemble de références, afin de déterminer

quelle est la référence la plus proche du mot à reconnaître. Principalement différentes dans la façon qu'elles ont de traiter l'information, nous consacrerons les parties 3 et 4 de ce chapitre à une étude plus approfondie de ces approches.

3.2 La détection d'activité vocale

L'objectif de la détection d'activité vocale est de séparer les zones de parole des autres zones qui sont sans intérêt pour le traitement de la parole. D'une grande utilité dans le domaine des télécommunications, lorsqu'il s'agit de libérer un canal de transmission si l'on n'y repère aucune activité de parole, cette étape est également très importante dans le domaine de la reconnaissance vocale afin de maximiser la qualité de la reconnaissance, en travaillant sur des formes représentant idéalement les mots du vocabulaire. Wilpon et *al* (1984) ont en effet démontré que d'un taux de réussite de 93%, pour un vocabulaire de 10 chiffres parfaitement isolés, on pouvait tomber à moins de 70% de réussite lorsque l'on déplaçait les limites de début et de fin jusqu'à 150ms. Plus précisément, une simple erreur de 4 trames de 15ms dans la délimitation des mots entrainait une chute de 3%.

Si, pour des mots prononcés dans des conditions idéales et correctement articulés, la détection des zones de parole est aisée, ce n'est malheureusement pas toujours le cas dans la pratique. Entre les problèmes attribués au locuteur, tels que les bruits liés à l'ouverture et la fermeture de la bouche ou à la respiration, les bruits de fond, ou la distorsion introduite par le canal de transmission, la dégradation du signal est parfois trop grande pour détecter idéalement les limites des zones de parole. Ce problème a été l'objet de nombreuses recherches, qui ont abouti à diverses approches. Nous pouvons ainsi mentionner l'algorithme de l'annexe G.729B de l'ITU (Benyassine et *al*, 1997), très utilisé dans les applications de téléphonie mobile, en raison de sa capacité à détecter les zones de parole en temps réel, à partir des coefficients LPC. Mais les algorithmes les plus couramment utilisés en reconnaissance de la parole utilisent plutôt des segments de parole entiers, dont ils éliminent les zones non désirables entourant la parole. C'est le cas de l'algorithme basé sur

l'énergie court-terme et le taux de passage par zéro, présenté par Rabiner et Sambur en 1975, et dont le déroulement général est présenté sur la figure 3.2.

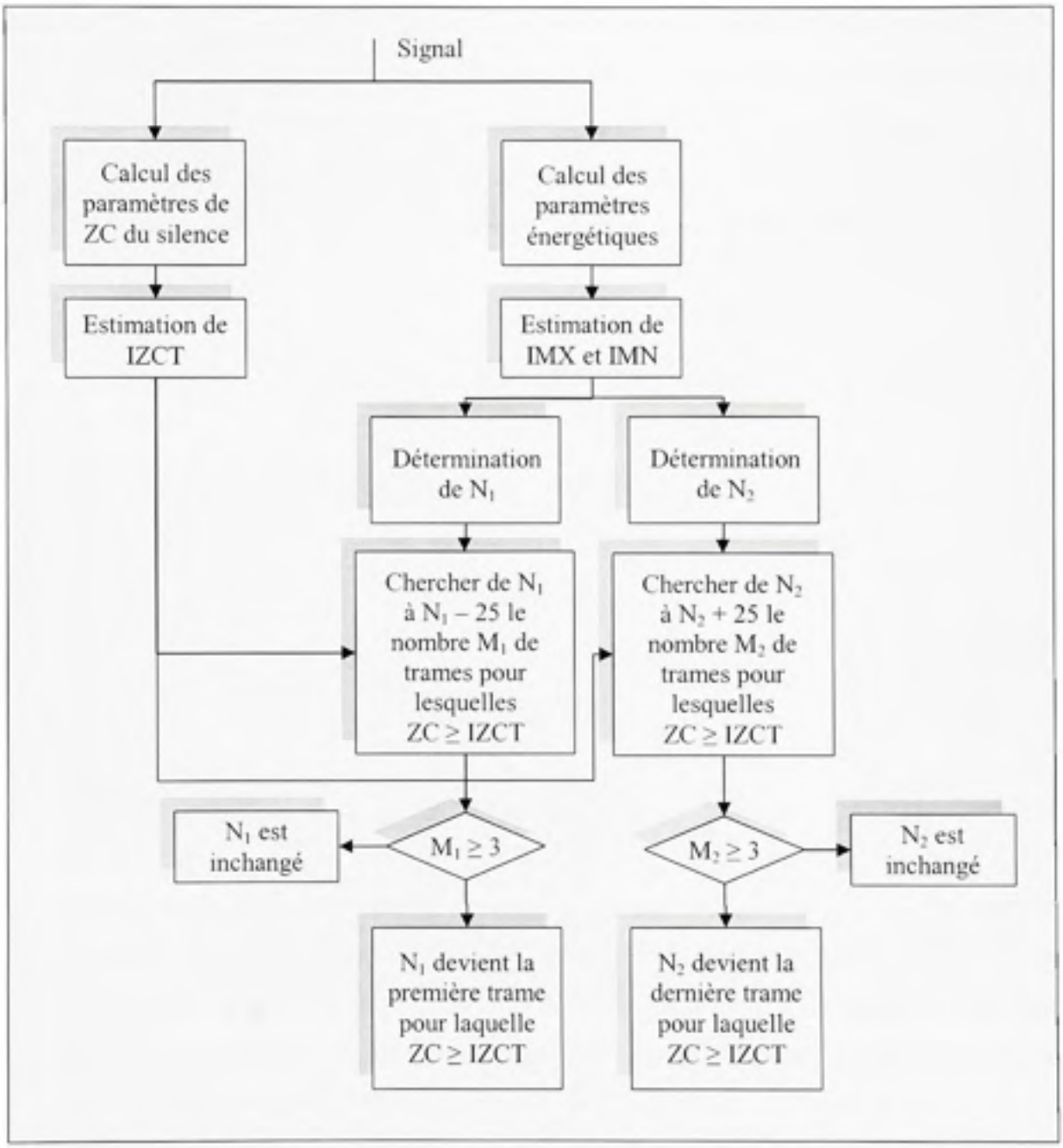

Figure 3.2 ***Déroulement de la détection d'activité selon Rabiner et Sambur.***
(Tiré de Rabiner et Sambur, 1975)

L'idée est ici de déterminer séparément les points de début et de fin de la zone de parole, en se basant tout d'abord sur l'énergie, puis en affinant la recherche à l'aide du taux de passage par zéro (ZC). Les 100 premières millisecondes du signal sont considérées comme étant une zone de silence et utilisées pour déterminer les paramètres du bruit de fond : la moyenne et l'écart-type du taux de passage par zéro, respectivement IZC et σ_{IZC}, ainsi que la moyenne de l'énergie court-terme, IMN. La seconde étape consiste à calculer les valeurs de seuils de taux de passage par zéro selon la formule 3.1, et d'énergie, selon les formules 3.2 à 3.5 (IF étant une valeur fixée par l'utilisateur, et IMX la valeur maximale de l'énergie sur toute la durée du signal).

$$\boxed{IZCT = \min(IF, IZC + 2.\sigma_{ZCT})} \tag{3.1}$$

$$I_1 = 0{,}03 \times (IMX - IMN) + IMN \tag{3.2}$$

$$I_2 = 4 \times IMN \tag{3.3}$$

$$\boxed{ITL = \min(I_1, I_2)} \tag{3.4}$$

$$\boxed{ITU = 5 \times ITL} \tag{3.5}$$

La détection de la trame de début de la zone d'activité vocale, notée N_1, se fait selon le schéma de la figure 3.3, la procédure pour déterminer la trame de fin de parole, N_2, se déroulant de la même façon, mais en partant de la fin du signal, et en opérant à reculons. L'idée est de rejoindre les régions pour lesquelles l'énergie dépasse le seuil ITL, puis le seuil ITU, sans être retombée sous ITL.

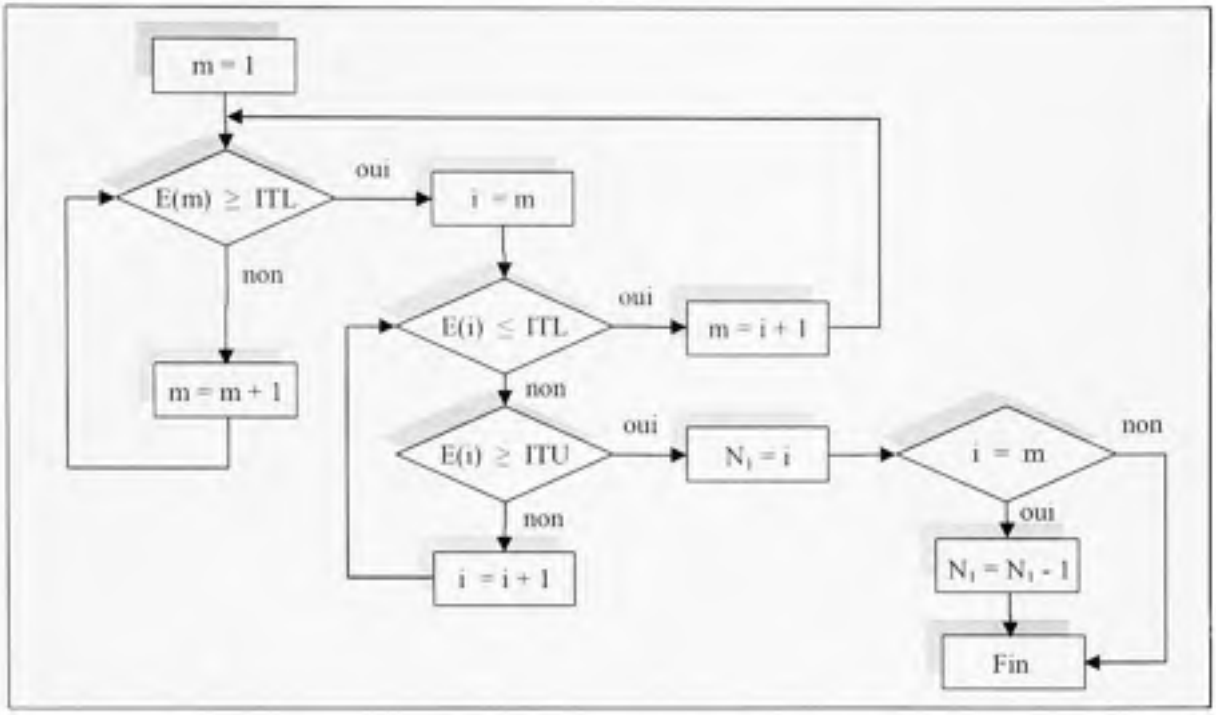

Figure 3.3 ***Détection énergétique de la trame de début de la zone d'activité vocale.***
(Tiré de Rabiner et Sambur, 1975)

L'affinage de la recherche se fait ensuite en utilisant le taux de passage par zéro, selon la dernière partie du schéma de la figure 3.2, en se déplaçant d'un certain nombre de trames, à partir du premier point déterminé, vers le début (ou la fin) du signal. Si le taux de passage par zéro dépasse plus de 3 fois le seuil IZCT, la point de limitation de l'activité vocale devient la dernière trame ayant dépassé IZCT. Cette phase permet d'inclure dans la zone de parole les éventuelles zones non voisées, donc l'énergie est plus faible que les zones voisées.

Pour illustrer le résultat obtenu grâce à cette méthode, nous présentons sur la figure 3.4 une détection d'activité vocale pour le chiffre 7 (prononcé en anglais) entouré de larges zones de silence. Très intéressante pour sa relative simplicité, cette méthode a également fourni de très bons résultats dans de nombreuses applications de reconnaissance de la parole, aussi bien en milieux calmes que bruités, si toutefois les paramètres du bruit de fond restent stables durant toute la durée du signal considéré.

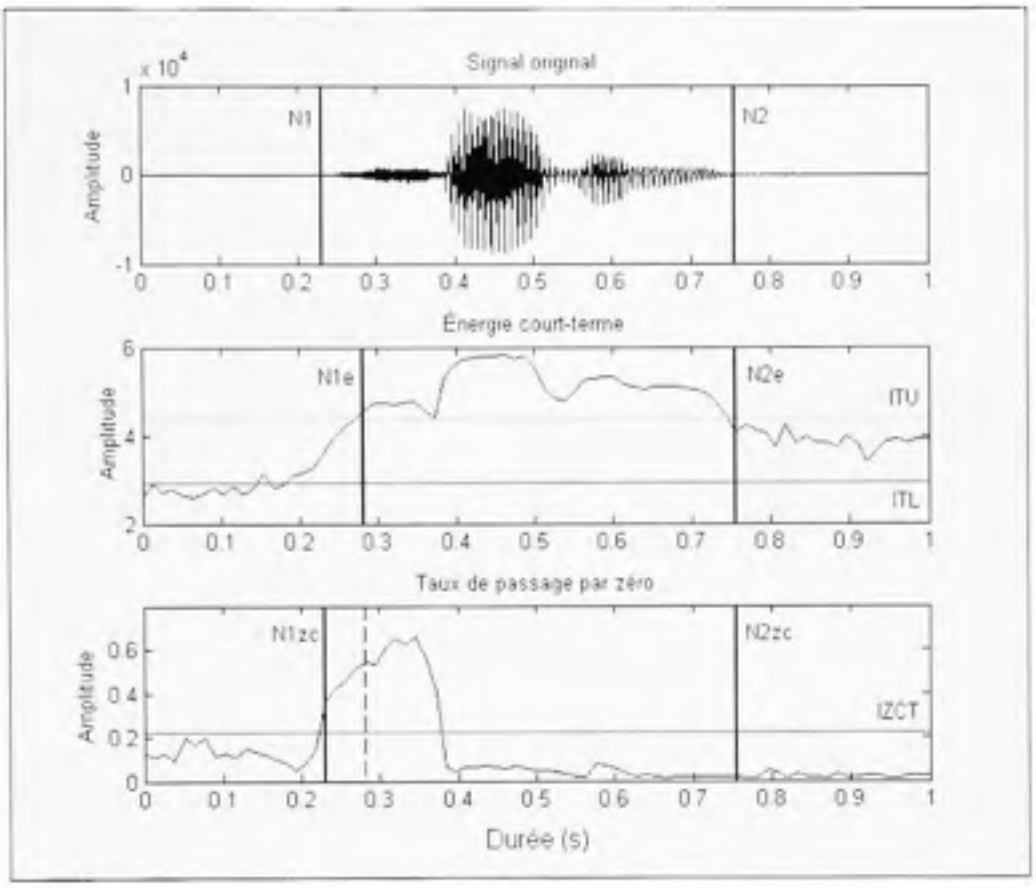

Figure 3.4 ***Détection des limites d'activité vocale pour le chiffre 7.***

3.3 La reconnaissance de type acoustico-phonétique

Cette approche se base sur l'hypothèse que le langage humain est constitué d'une quantité finie d'unités phonétiques, et que ces unités peuvent être identifiées au travers de certaines propriétés visibles du signal ou de son spectre (Rabiner, 1993).

L'objectif des méthodes basées sur cette approche est donc de segmenter le signal en régions dans lesquelles les propriétés acoustiques sont représentatives d'une classe d'unités phonétiques, et ensuite d'étiqueter ces zones en se basant sur une très bonne connaissance du langage concerné et des unités acoustiques qui le constituent. Bien évidemment, plus le vocabulaire est grand, plus le nombre de phonèmes considéré est important, et l'identification phonémique des régions du signal de parole n'en devient que plus

compliquée. La dernière phase consiste à déterminer quel est le mot constitué par cette chaîne de phonèmes. On peut visualiser le déroulement global du décodage acoustico-phonétique de la parole au moyen de la figure 3.5.

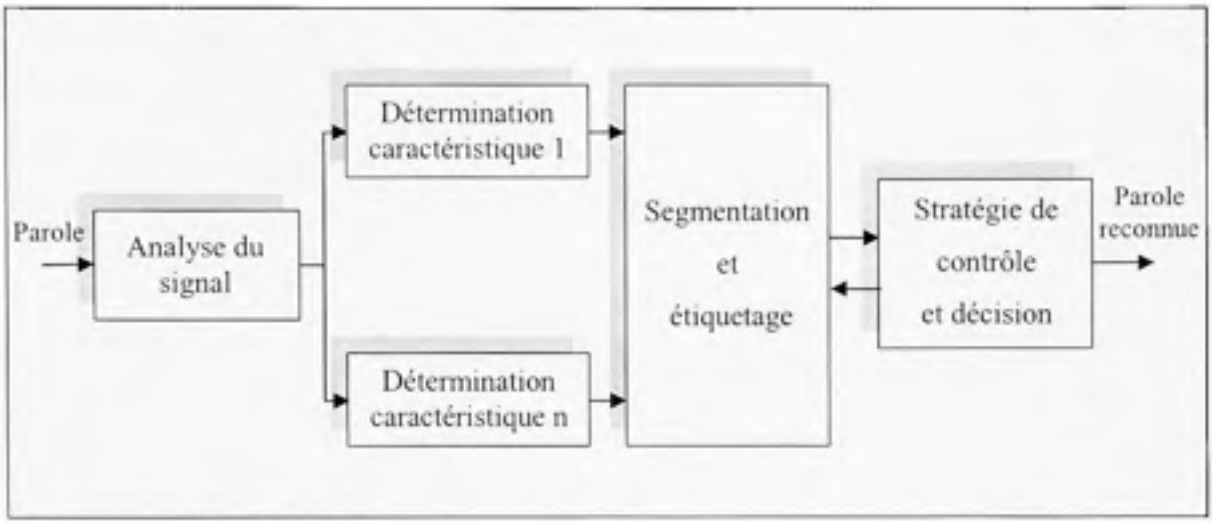

Figure 3.5 ***Déroulement d'un système de reconnaissance acoustico-phonétique.***

Ainsi, après l'habituelle phase d'analyse du signal destinée à fournir une représentation des caractéristiques spectrales du signal, un système de reconnaissance de la parole par décodage acoustico-phonétique se compose de trois étapes, que nous allons décrire sommairement.

3.3.1 Extraction des paramètres pertinents

L'objectif de cette étape est de convertir les caractéristiques spectrales du signal en un jeu d'indices acoustiques décrivant les propriétés physiques des sons. Ces indices peuvent être binaires, et donc notifiés par leur présence ou leur absence, comme :

- le voisement;
- la nasalité;
- la frication.

On peut également les considérer selon leur valeur, comme :

- la fréquence fondamentale;
- l'énergie (haute et basse fréquence).

Pour déterminer les valeurs de ces différents indices, on peut donc aussi bien utiliser des outils élémentaires tels que le taux de passage par zéro ou le calcul de l'énergie, que la représentation fréquentielle du signal, la prédiction linéaire, ou l'analyse cepstrale.

À la sortie de cet étage, chaque trame de signal est donc représentée par un ensemble d'indices simples fournissant des indications sur ses propriétés acoustiques. Il reste alors à combiner judicieusement ces indices pour distinguer les unités phonétiques du mot. C'est le rôle de la phase de segmentation et d'étiquetage.

3.3.2 Segmentation et étiquetage des sons de parole

La segmentation consiste à découper le signal de parole en unités clairement définies, en se basant sur l'évolution temporelle des indices déterminés à l'étape précédente. L'objectif est donc de trouver des régions stables, à l'intérieur desquelles les valeurs des indices ne varient pas, ou peu. Si à l'origine, cette tâche était réalisée manuellement par des phonéticiens entraînés, de nombreuses méthodes de segmentation automatique ont depuis fait leur apparition, qui utilisent principalement des fonctions mesurant les discontinuités locales du signal. Mais ces discontinuités n'étant pas toujours synchrones selon les locuteurs, les langages, ou les modèles phonétiques utilisés, l'objectif des chercheurs a été de déterminer une façon idéale d'interpréter les discontinuités acoustiques du signal en termes d'informations linguistiques globales. Ainsi, plutôt que de parler de phonèmes, la notion de « phones homogènes » introduite par Caelen et *al* (1983) considère le regroupement en un seul segment de tous les échantillons spectraux homogènes, sans se baser sur une connaissance du langage. Il appartient ensuite au système de rattraper la sursegmentation éventuellement engendrée par cette technique, en greffant un modèle interprétatif sur ces

segments pour former des unités phonétiques plus larges (Perennou et *al*, 1985). Dans cette approche, nous pouvons mentionner les récentes techniques de segmentation, basées sur l'algorithme de *Hull Convex* (Li et Liu, 1999), ou bien sur des techniques de programmation dynamique (Sharma et Mammone, 1996).

L'étiquetage phonétique d'un segment se fait ensuite en combinant les valeurs des indices déterminés auparavant, afin d'identifier le son parmi l'ensemble des phonèmes ou unités acoustiques du vocabulaire considéré. Si la grande diversité des phonèmes peut rendre cette tâche très ardue, comme l'a démontré Calliope (1989), l'utilisation d'un arbre de décision, permettant de procéder par élimination successive des différentes classes de sons, simplifie grandement le processus. Rabiner (1993) a ainsi présenté diverses méthodes de classification, et une façon d'accomplir aisément une telle procédure est présentée à la figure 3.6.

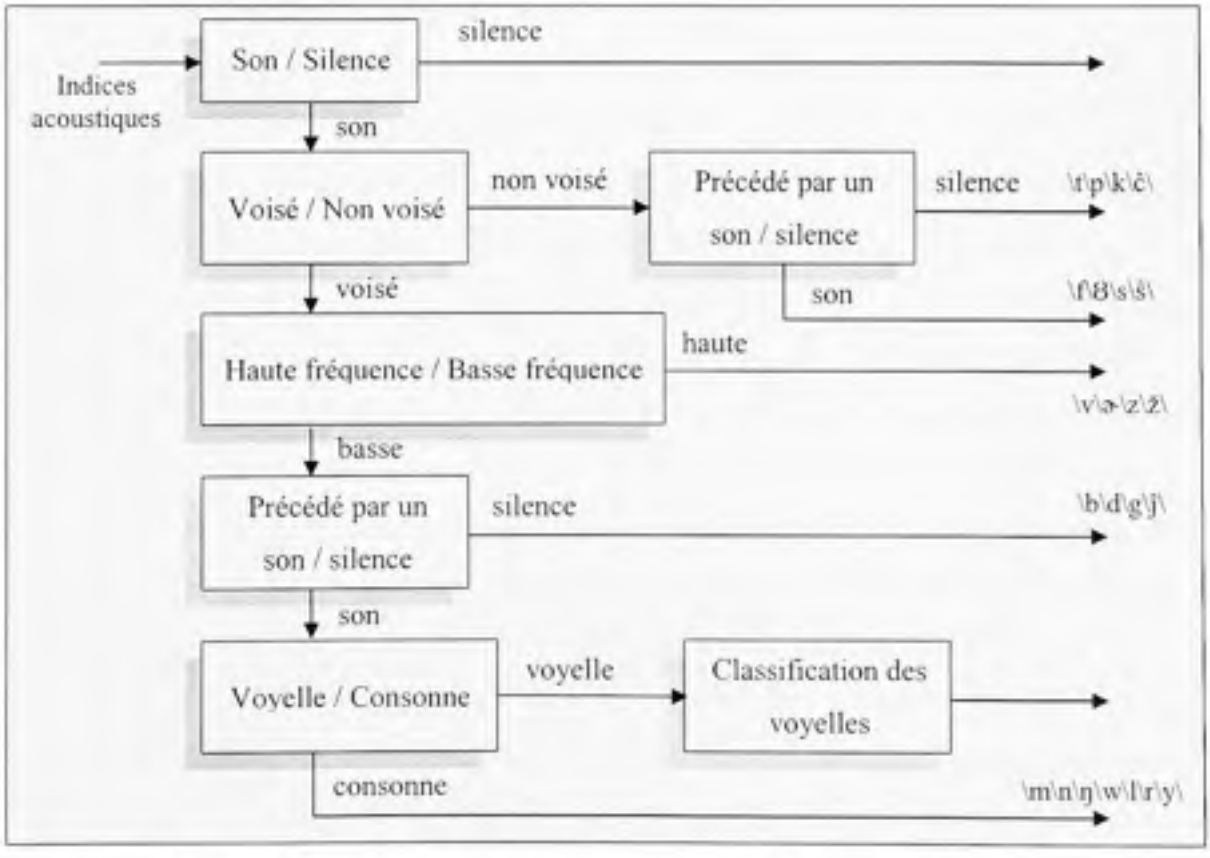

Figure 3.6 ***Arbre de classification binaire des sons de parole.***

Une telle procédure offre la possibilité de détecter rapidement les classes de sons plus « simples », tandis que les sons plus complexes sont identifiés après un processus plus long, et pouvant faire appel à d'autres arbres de décision, comme par exemple le classificateur de voyelles. Toutefois, ce procédé basé sur des critères de décision basiques ne prend pas entièrement en compte toute la complexité du langage humain, dont le comportement est loin d'être binaire. Cela nous permet malgré tout de mieux visualiser la tâche à accomplir pour identifier les différentes catégories de sons.

3.3.3 Décision et stratégie de contrôle

Les étapes décrites jusqu'ici donnent ainsi naissance à une chaîne de phonèmes, plus communément appelée « treillis de phonèmes » en raison de la possibilité d'identifier chaque segment par plusieurs choix de phonèmes correspondants aux meilleurs candidats issus de l'arbre de décision. Il appartient ensuite au système de se baser sur une très bonne connaissance du langage et des contraintes linguistiques pour déterminer quel est le mot correspondant le plus à cette séquence. L'intelligence artificielle et les systèmes experts sont souvent favorisés par les chercheurs pour effectuer cette tâche (Carbonnel et *al*, 1983), et l'on parle également de stratégie de contrôle qui consiste à exploiter les contraintes lexicales pour revenir éventuellement sur la phase de segmentation et réajuster le découpage.

Cette approche est donc fortement appréciée des chercheurs, en raison de sa propension à mimétiser les actions du cerveau humain, ainsi que la possibilité de fonctionner aussi bien sur des mots isolés que des chaînes de mots. Nous noterons plus particulièrement la notion très intéressante de classification des sons qui permet, grâce à une très bonne connaissance du vocabulaire considéré et une utilisation intelligente des propriétés acoustiques, de déterminer la catégorie phonétique des éléments de parole. Toutefois, bien que théoriquement très attirantes, ces méthodes rencontrent peu de succès une fois mises en pratique, notamment en raison de la trop grande complexité du langage humain, que cinquante années de recherche n'ont toujours pas permis de maîtriser parfaitement. Les

chercheurs préfèrent donc souvent appliquer des méthodes alternatives, comme celles basées sur la reconnaissance de formes.

3.4 Méthodes basées sur la reconnaissance de formes

Les méthodes basées sur la reconnaissance de formes sont apparues pour la première fois dans les années 60-70 (Junqua, 1996), après que des nombreuses années de recherche acoustico-phonétique n'aient abouti à aucun résultat probant. Ici, l'idée est de comparer la forme globale du mot à reconnaître avec l'ensemble des mots constituant le vocabulaire de référence. Le grand intérêt de ces méthodes réside donc dans le fait qu'elles ne requièrent pas une excellente connaissance théorique du langage et de ses unités phonétiques, offrant ainsi au système la possibilité de travailler sur des langages différents, en modifiant « simplement » le dictionnaire contenant les références. La figure 3.7 présente la structure générale d'un système basé sur la technique de reconnaissance globale.

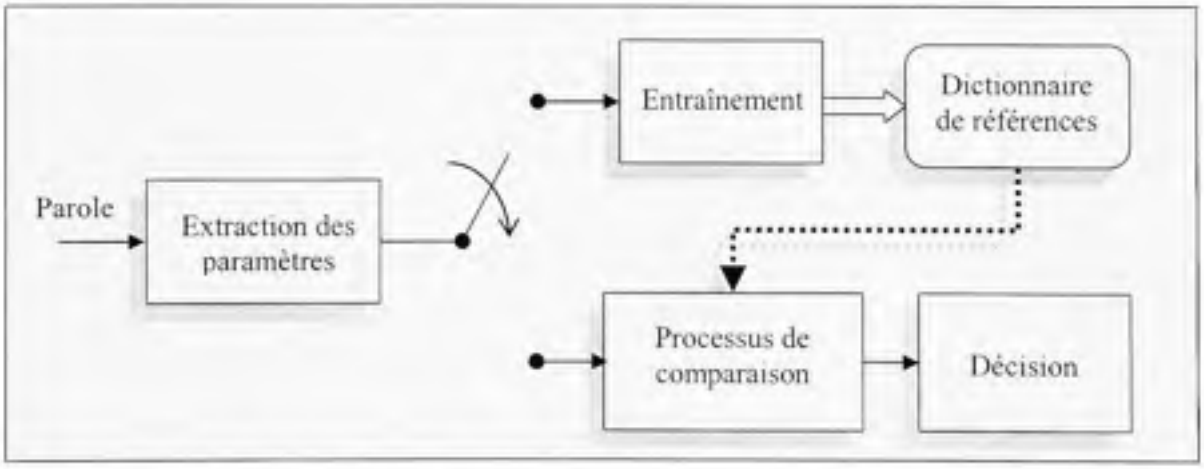

Figure 3.7 ***Déroulement général d'un système de reconnaissance de formes.***

Les trois phases principales sont donc :

- l'entraînement, qui consiste à déterminer la « forme » moyenne de chaque mot du vocabulaire, en le prononçant de façon répétée. Le résultat idéal est généralement constitué d'un grand nombre de versions du même mot, afin de bien prendre en compte

les variations possibles intra et inter locuteurs. Nous présenterons plus loin les méthodes utilisées pour créer un dictionnaire réduit à partir d'un très grand nombre de références;

- la comparaison entre les mots de référence et le mot à reconnaître, faisant appel à des mesures de dissemblance spectrale, que nous présenterons par la suite, et des méthodes d'alignement temporel entre les deux formes à comparer;
- la décision, qui va sélectionner la référence la plus proche de l'image acoustique du mot à identifier, à partir des résultats de l'étape précédente.

Le cœur du problème se situe dans le processus de comparaison. Deux méthodes sont généralement mises en avant : l'une basée sur un calcul de dissemblance générale, après alignement temporel des deux formes à comparer, et l'autre basée sur des modèles statistiques de production pour chaque mot du vocabulaire, l'identification d'un mot revenant alors à rechercher le modèle qui la plus forte probabilité de l'avoir produit. Nous allons donc décrire brièvement ces deux types de méthodes, après avoir présenté quelques mesures de dissemblance spectrale entre deux vecteurs acoustiques.

3.4.1 Mesure de dissemblance entre deux vecteurs acoustiques

Une des clés de la plupart de ces algorithmes de reconnaissance de formes réside dans la possibilité de mesurer la distance, ou plutôt la dissemblance, entre deux vecteurs de caractéristiques spectrales. Contrairement aux méthodes acoustico-phonétiques, dont les indices acoustiques sont essentiellement qualitatifs et par conséquent fixés de façon empirique, les mesures de dissemblance fournissent un véritable résultat numérique facilement interprétable. Considérant que x et y sont deux vecteurs spectraux de dimension p, d(x,y) représentant la distance entre ces deux vecteurs, Gray et Markel (1976) ont alors défini les distances associées aux normes L_n ($n \geq 1$) présentées à la formule 3.6.

$$d_n(x,y) = \left[\sum_{k=1}^{p} (x_k - y_k)^n \right]^{\frac{1}{n}} \qquad (3.6)$$

Où x_k et y_k sont les coefficients du spectre obtenus après analyse par TFD. Les plus utilisées sont les distances d_1 et d_2 (distance euclidienne). Dans le cas des coefficients de prédiction linéaire, la mesure appelée *Log Likelihood Ratio* (LLR), introduite par Itakura (1975), est généralement utilisée. Elle est définie par la formule 3.7 où, p étant l'ordre du modèle de prédiction, x et y sont deux vecteurs de p+1 coefficients, et R est la matrice des coefficients d'autocorrélation évalués sur le signal correspondant à y.

$$d_{LLR}(x,y) = \log\left(\frac{xRx^T}{yRy^T}\right) \tag{3.7}$$

Puisque les coefficients cepstraux se prêtent parfaitement à la représentation du signal de parole dans les systèmes de reconnaissance, plusieurs mesures de distance, basées sur ces coefficients, ont été proposées, la plus simple étant la distance cepstrale euclidienne présentée par la formule 3.8.

$$d^2_{cep}(x,y) = \sum_n (c_y(n) - c_x(n))^2 \tag{3.8}$$

Afin de normaliser la contribution de chaque coefficient, des fonctions de pondération peuvent être utilisées, comme présenté sur la formule 3.9.

$$d^2_{cepw}(x,y) = \sum_n (w(n) \times (c_y(n) - c_x(n)))^2 \tag{3.9}$$

Si la fonction w(n) = n a prouvé son efficacité à de nombreuses reprises puisqu'elle permet d'uniformiser l'amplitude des coefficients (Hanson et Wakita, 1986), on lui préfèrera la fonction de pondération introduite par Juang et *al* (1986), présentée à la formule 3.10.

$$w(n) = \begin{cases} 1 + h \times \sin(\frac{n\pi}{L}), & 1 \le n \le L \\ 0, & n \le 0, \quad n > L \end{cases} \tag{3.10}$$

L étant le nombre de coefficients cepstraux, et h généralement fixé à L/2. La séquence $w(n) \times c(n)$, correspondant à une forme plus adoucie du spectre, permet donc de réduire la sensibilité du spectre aux conditions d'analyse (bruit,...), sans altérer sa structure, et s'avère donc particulièrement efficace en milieu bruité.

Il existe de nombreuses autres méthodes pour mesurer la distance entre deux formes acoustiques, et l'ouvrage de Rabiner (1993) en présente une description très détaillée. Ici, nous avons présenté les mesures de dissemblance les plus utilisées en pratique, et ayant prouvé leur efficacité dans de nombreux systèmes.

3.4.2 Reconnaissance basée sur l'alignement temporel des formes

3.4.2.1 Présentation du problème

La tâche de reconnaissance consiste ici à identifier un mot prononcé par un locuteur en déterminant sa distance vis-à-vis de l'ensemble des mots du vocabulaire de référence. Étant donné qu'un mot est représenté par une succession de vecteurs acoustiques court-terme, l'opération consiste donc à mesurer les distances locales entre chacun des vecteurs spectraux, x_i et y_i, des deux formes à comparer, X et Y, de longueur T, la distance totale entre ces deux mots étant alors obtenue par accumulation de toutes ces distances locales, comme présenté par la formule 3.11.

$$D(X,Y) = \sum_{i=1}^{T} d(x_i, y_i) \tag{3.11}$$

Toutefois, et c'est bien là l'une des complexités majeures de la reconnaissance de la parole, un vocabulaire est généralement constitué de mots de tailles différentes, et un même mot peut être prononcé à des vitesses d'élocution variables, engendrant ainsi une transformation non linéaire de l'échelle temporelle. Afin de rendre la mesure de distance indépendante des

fluctuations du rythme de prononciation et de la vitesse d'élocution, les chercheurs ont mis en œuvre des algorithmes de comparaison dynamique.

3.4.2.2 La programmation dynamique

Également appelée DTW, pour *Dynamic Time Warping*, la notion de programmation dynamique a été introduite par Bellman en 1954, mais utilisée pour la première fois en reconnaissance vocale par Vintsyuk en 1968. L'algorithme consiste alors à calculer de façon récursive la distance minimale accumulée pour chaque point (i,j) en tenant compte de certaines contraintes locales sur la façon avec laquelle le chemin optimum atteint ce point. Pour plus de clarté, la figure 3.8 présente la comparaison de deux mots X et Y, de tailles respectives T_x et T_y (en terme de nombre de trames). Chaque point visible sur cette figure correspond à une distance locale $d(x_i,y_j)$.

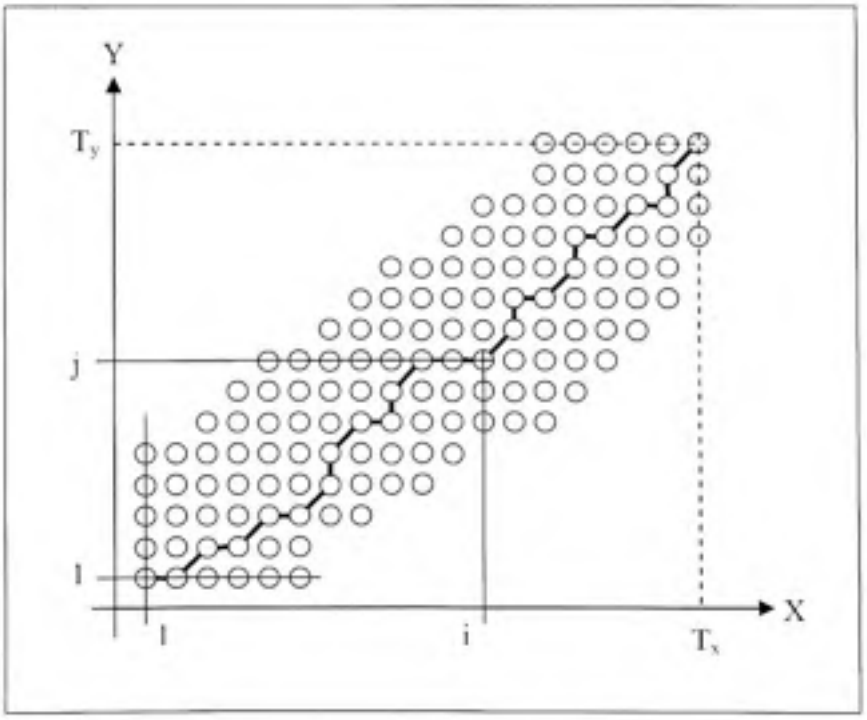

Figure 3.8 ***Comparaison des mots X et Y.***

La comparaison est donc basée sur la recherche du chemin optimum w de longueur K, défini par la formule 3.12, et dont la distance associée est donnée par la formule 3.13 (Kunt, 1984).

$$\begin{cases} w: \ \{i(k), j(k)\} & k = 1,2,\ldots,K \\ i(1) = 1 & j(1) = 1 \\ i(K) = T_x & j(K) = T_y \end{cases} \tag{3.12}$$

$$D_w(X,Y) = \frac{\sum_{k=1}^{K} d(x_{i(k)}, y_{j(k)}) \times g(k)}{N(g)} \tag{3.13}$$

Où g(k) est un coefficient de pondération, et N(g) un facteur de normalisation généralement fixé selon la formule 3.14.

$$N(g) = T_x + T_y \tag{3.14}$$

La recherche du chemin optimum, c'est-à-dire de distance accumulée minimale, se fait donc au moyen de contraintes locales destinées à réduire le nombre de calculs et à prendre en compte le caractère temporel, donc unidirectionnel, du chemin. Sakoe et Chiba ont proposé en 1978 plusieurs contraintes, dont les principales sont présentées sur la figure 3.9.

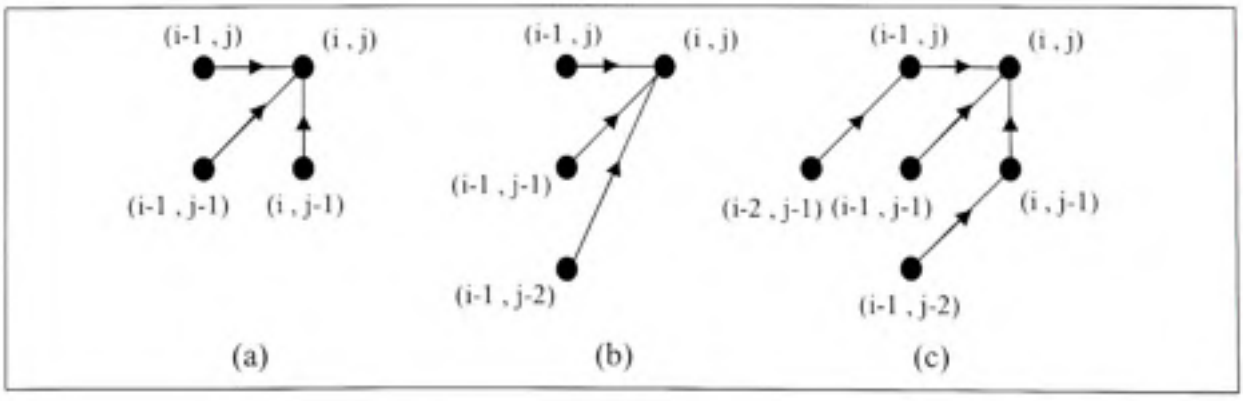

Figure 3.9 ***Contraintes locales sur le chemin optimal.***

Ainsi, si l'on souhaite déterminer le chemin optimal en se basant sur les contraintes de type a), la distance accumulée entre le point (1 , 1) et le point (i , j), notée D(i,j), sera déterminée par la formule 3.15.

$$D(i,j) = d(i,j) + \min\begin{cases} D(i-1,j) \\ D(i-1,j-1) + d(i,j) \\ D(i,j-1) \end{cases} \tag{3.15}$$

L'algorithme général de programmation dynamique peut donc être réalisé selon le schéma-bloc présenté à la figure 3.10 (r étant l'espace de recherche accordé autour de la diagonale, dans le but de réduire le nombre de calculs).

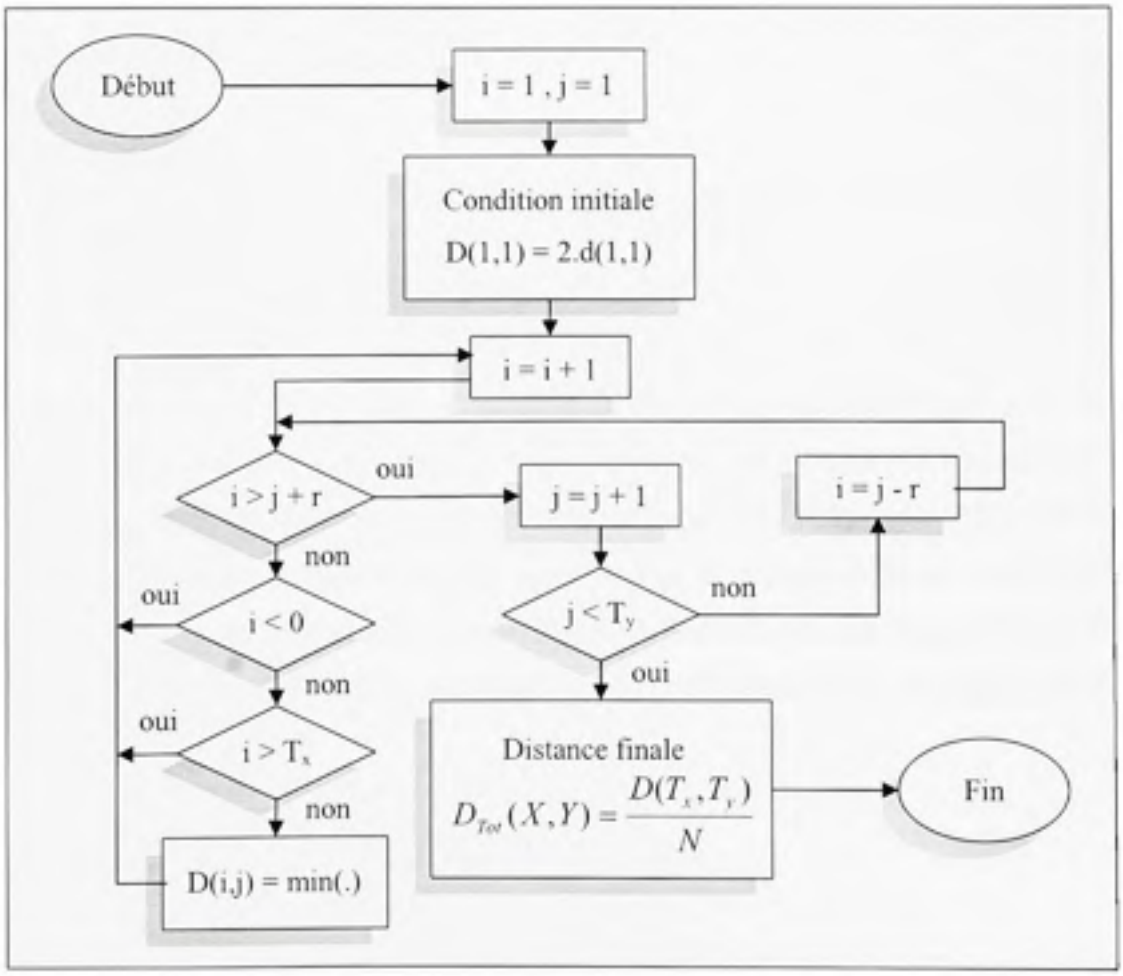

Figure 3.10 ***Algorithme de programmation dynamique.***

Les méthodes de reconnaissance par programmation dynamique ont déjà prouvé leur efficacité dans de nombreux systèmes, produisant des taux d'erreur inférieurs à 1% pour des petits vocabulaires monolocuteurs. Quant à l'augmentation de la taille du vocabulaire, elle ne pose que le problème de l'augmentation du volume de calcul et de place mémoire nécessaire. La programmation dynamique peut également être utilisée pour la reconnaissance de mots enchaînés, en effectuant une comparaison globale de la suite à reconnaître avec des suites de référence constituées par la concaténation des mots du vocabulaire. Cette tâche s'accomplit néanmoins au prix d'une très grande quantité d'opérations, et d'une capacité à détecter les limites exactes de chaque mot à l'intérieur de la phrase. Cette tendance à nécessiter de nombreux calculs, ainsi qu'un espace mémoire imposant, puisque chaque mot du vocabulaire doit être représenté par l'ensemble de ses vecteurs spectraux court-terme, est la raison pour laquelle les algorithmes de programmation dynamique sont particulièrement difficiles à implémenter sur des systèmes aux ressources limitées.

3.4.3 Reconnaissance basée sur une modélisation statistique des formes

3.4.3.1 Concept général

L'approche probabiliste de la reconnaissance de la parole consiste à modéliser la production d'une unité linguistique (phonème, syllabe, mot,...) en prenant en considération la statistique des diverses prononciations de cette même unité, le but étant d'identifier une unité inconnue en déterminant le modèle ayant la plus forte probabilité de l'avoir produite (Rabiner, 1993). Le modèle utilisé est une chaîne de Markov, une séquence composée d'un nombre fini d'états, correspondant chacun à un évènement observable, et de probabilités de transition entre chaque état.

3.4.3.2 Présentation des modèles de Markov cachés

En reconnaissance de la parole, on parlera de modèle de Markov caché (ou HMM pour Hidden Markov Model), étant donné que seuls les vecteurs acoustiques émis sont observés,

mais que leur observation est une fonction de probabilité de l'état. Autrement dit, à chaque état peuvent correspondre plusieurs observations. La figure 3.11 montre un exemple de modèle de Markov caché à 4 états représentant une unité de parole (nous considèrerons que c'est un mot).

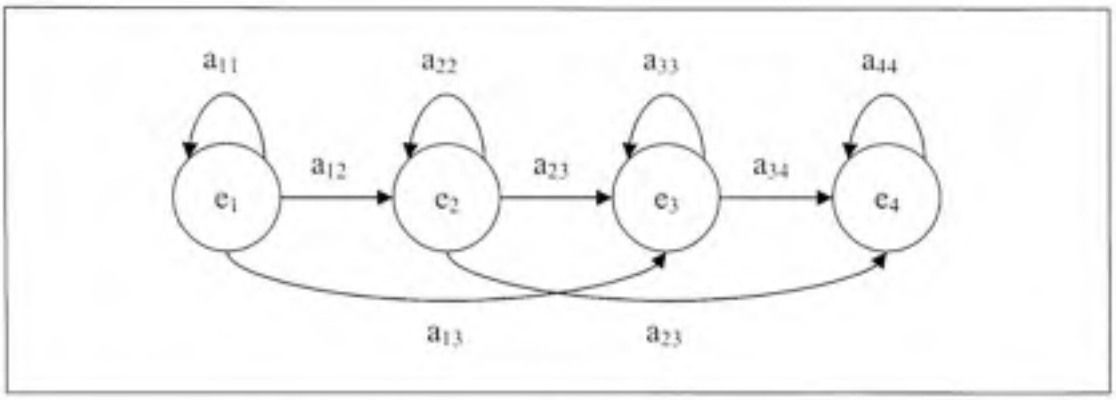

Figure 3.11 ***Modèle de Markov Caché à 4 états.***

Un modèle de Markov caché est donc défini par plusieurs éléments :

- un nombre d'états, noté N. Chaque état peut représenter un phonème, une unité acoustique, ou même un mot entier, selon la nature de ce que l'on souhaite modéliser;
- un nombre M de symboles distincts, $[o_1, o_2, \ldots, o_M]$, correspondant aux observations possibles à chaque état. Ainsi, si le modèle représente la production d'un mot, ces symboles pourront être un ensemble de vecteurs acoustiques représentant les différents phonèmes du vocabulaire, chacun ayant une probabilité plus ou moins forte d'être observé à un état du modèle. Chaque vecteur acoustique issu du mot à reconnaître est alors comparé à l'ensemble de ces symboles, pour être représenté dans le modèle par le symbole le plus proche;
- des probabilités de transition entre les états, la probabilité que le modèle passe de l'état s_i à s_j entre l'instant t et l'instant t+1 étant notée selon la formule 3.16;

$$a_{ij} = P(t+1 = e_j \mid t = e_i) \tag{3.16}$$

- des probabilités de distribution des M symboles pour chacun des N états, notées selon la formule 3.17;

$$b_j(k) = P(o_k \mid e_j) \quad 1 \le k \le M, \quad 1 \le j \le N \tag{3.16}$$

- un état initial, qui est la probabilité d'être à l'état e_i à l'instant t = 1.

Afin de prendre en compte les variations de prononciation des mots, chaque état peut donner lieu à trois transitions : retour sur lui-même, transition sur l'état suivant, ou saut de l'état suivant. Il est intéressant de remarquer que cela correspond aux contraintes utilisées dans les algorithmes de programmation dynamique, et la conséquence en est identique, puisque, peu importent les tailles des mots à reconnaître, les modèles s'aligneront toujours avec eux.

3.4.3.3 Apprentissage des modèles

Le nombre d'états d'un modèle est généralement choisi de façon empirique, après une longue étude du vocabulaire concerné, tandis que les paramètres a_{ij} et $b_j(k)$ sont obtenus au cours d'une phase d'apprentissage, à partir d'un nombre important d'énoncés du même mot. Cette phase d'entraînement, qui permet donc d'ajuster les paramètres du modèle en maximisant les probabilités d'observation jusqu'à stabilisation, est réalisée en utilisant une procédure itérative telle que l'algorithme de Baum-Welch (Baum, 1972).

3.4.3.4 Reconnaissance des modèles

Supposons qu'un mot inconnu $X = [x_1, x_2, \ldots, x_L]$ est émis par une suite d'états constituant un certain parcours S de longueur L :

$$S : \left[e_{s(1)}, e_{s(2)}, \ldots, e_{s(L)}\right] \tag{3.17}$$

avec n(1) = 1 et n(L) = N. Le mot peut donc être produit par tous les parcours issus du premier état, et aboutissant au dernier état, comme le montre la figure 3.12 pour un mot de 7 observations, sur un modèle à 4 états.

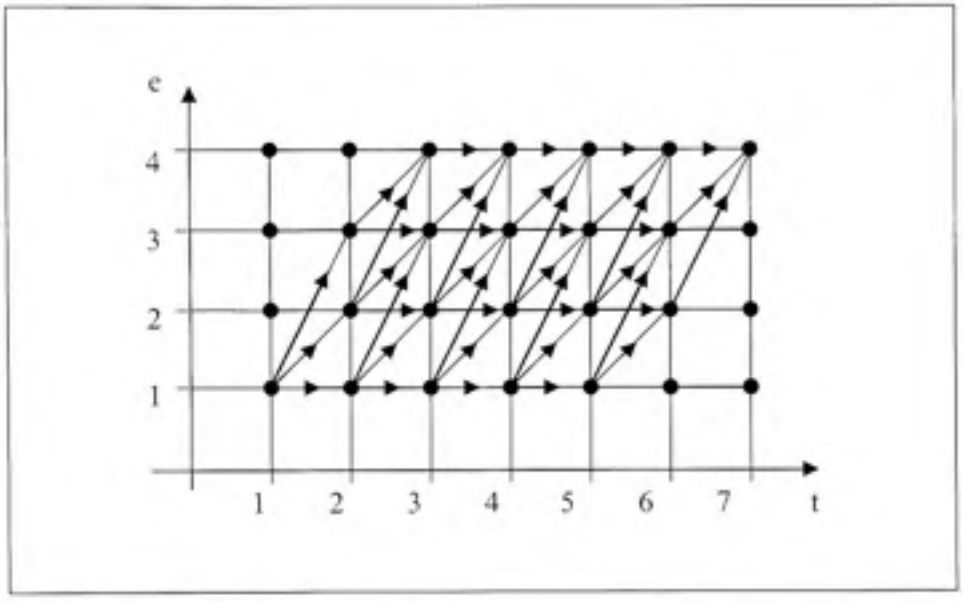

Figure 3.12 ***Parcours possibles dans un modèle de Markov.***

Le long d'un parcours, le mot est émis avec une probabilité $p_r(x)$ égale au produit des probabilités des transitions effectuées multiplié par le produit des probabilités des émissions successives. L'objectif est de déterminer la probabilité finale maximale, c'est-à-dire celle qui est associée au chemin optimum (Kunt, 1984) :

$$P_{\max}(X) = \max_{r} p_r(X) \tag{3.18}$$

Cette tâche est réalisée au moyen de l'algorithme de Viterbi (1967). S'apparentant quelque peu à l'algorithme de programmation dynamique, il procède en ne retenant, à chaque instant t et pour chaque état j, que les parcours auxquels est associée la plus grande probabilité d'avoir émis la séquence $[x_1, x_2, \ldots, x_t]$, selon la formule 3.19.

$$p_t(j) = b_j(x_t) \times \max \begin{cases} p_{t-1}(j) \times a_{jj} \\ p_{t-1}(j-1) \times a_{(j-1)j} \\ p_{t-1}(j-2) \times a_{(j-2)j} \end{cases} \quad (3.19)$$

Au final, la probabilité associée au chemin optimum est la probabilité à l'instant t = L et à l'état e = N :

$$P_{max}(X) = p_L(N) \quad (3.18)$$

L'identification du mot inconnu est alors effectuée en sélectionnant, parmi tous les modèles de référence, celui fournissant la plus forte probabilité d'émission.

Introduite pour la première fois dans les systèmes de reconnaissance vocale à la fin des années 70, cette méthode a depuis largement prouvé son efficacité et fourni de très bons résultats dans de nombreuses applications, notamment pour les tâches de reconnaissance de parole continue. Néanmoins, comme pour la DTW, la quantité de calculs requise à chaque reconnaissance est très élevée, et la réussite est également très dépendante de la phase d'entraînement. C'est pour cela que les résultats deviennent nettement moins bons en environnements bruités, et que ce procédé s'accorde peu avec une implémentation sur des systèmes aux ressources limitées. Des HMM, nous retiendrons donc le fait de modéliser un mot du vocabulaire en une succession d'états principaux, et de « reconstruire » le mot selon ce modèle, à partir des observations acoustiques.

3.4.4 Algorithmes de classification

Il est un dernier point important que l'on doit considérer lors de la conception d'une méthode de reconnaissance vocale, c'est la gestion du dictionnaire de références. Comme mentionné précédemment, pour être efficaces les méthodes de reconnaissance de forme travaillent à partir d'un large vocabulaire de référence, obtenu au moyen de prononciations répétées de chaque unité de parole. Que ces unités soient des mots entiers, comme pour la

programmation dynamique, ou bien des phonèmes, vecteurs spectraux ou autres petits évènements acoustiques, comme pour les HMM, plus le système contient de versions d'une même unité à comparer, et plus la reconnaissance a de chances de rencontrer un taux de réussite élevé, notamment dans le cas des systèmes multilocuteurs.

Malheureusement, les limitations techniques imposées par les systèmes, aussi bien en terme de rapidité de calcul que de place mémoire, nous empêchent souvent d'embarquer une trop grande quantité de références. Il est donc important de pouvoir optimiser le contenu du dictionnaire, afin de ne garder qu'un nombre restreint de versions de chaque mot, tout en conservant une grande diversité dans les représentations possibles de ce mot. Cette tâche peut être accomplie au moyen d'algorithmes de classification. Les premiers, apparus à la fin des années 70 (Levinson et *al*, 1979), nécessitaient une intervention humaine pour guider le processus. C'est pourquoi une nouvelle classe d'algorithmes de classification automatique a été développée, dont les plus utilisés encore de nos jours sont le « *Unsupervised clustering Without Averaging* » (UWA) et le « *Modified K-Means* » (MK-M) présentés par Wilpon et Rabiner en 1985. L'algorithme MK-M, dérivé des méthodes de quantification vectorielle, est présenté par le schéma bloc de la figure 3.13.

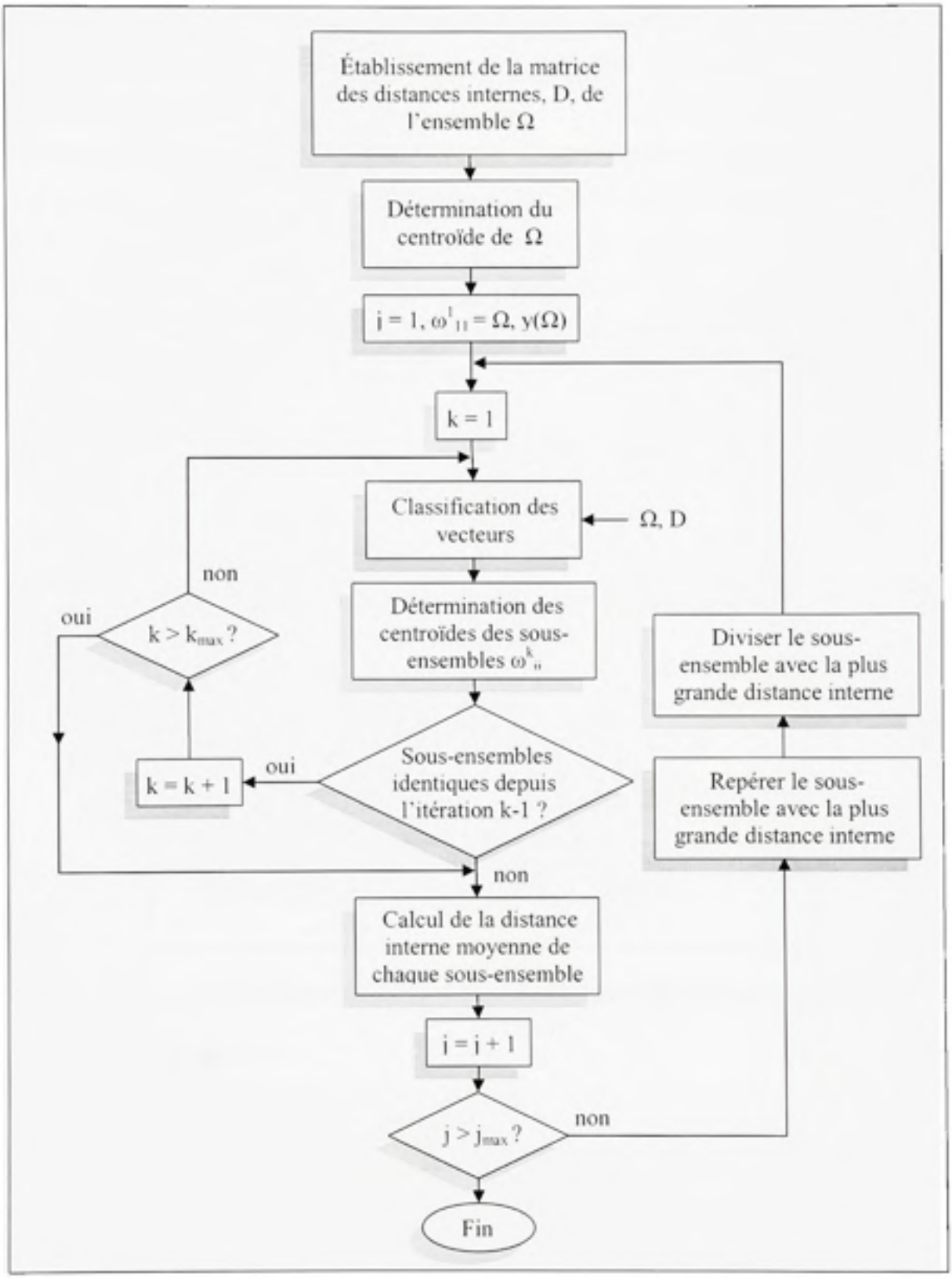

Figure 3.13 ***Schéma bloc de la procédure de classification M-KM.***

À partir d'un ensemble de plusieurs références, cet algorithme fournit un nombre de sous-ensembles J_{max} fixé par l'utilisateur, chacun étant représenté par son centroïde, c'est-à-dire l'élément ayant la distance la plus rapprochée de tous les autres éléments du sous-ensemble. Les distances entre les éléments peuvent être des distances locales comme celles présentées dans la partie 3.4.1, si l'on souhaite classer des petites unités acoustiques représentées par un simple vecteur de paramètres, ou bien une mesure de dissemblance globale obtenue, par exemple, par DTW si l'on traite avec des mots de parole complets. La figure 3.14 schématise le résultat d'une classification d'un ensemble de 38 éléments en 4 sous-ensembles, les centroïdes étant repérés par un petit cercle.

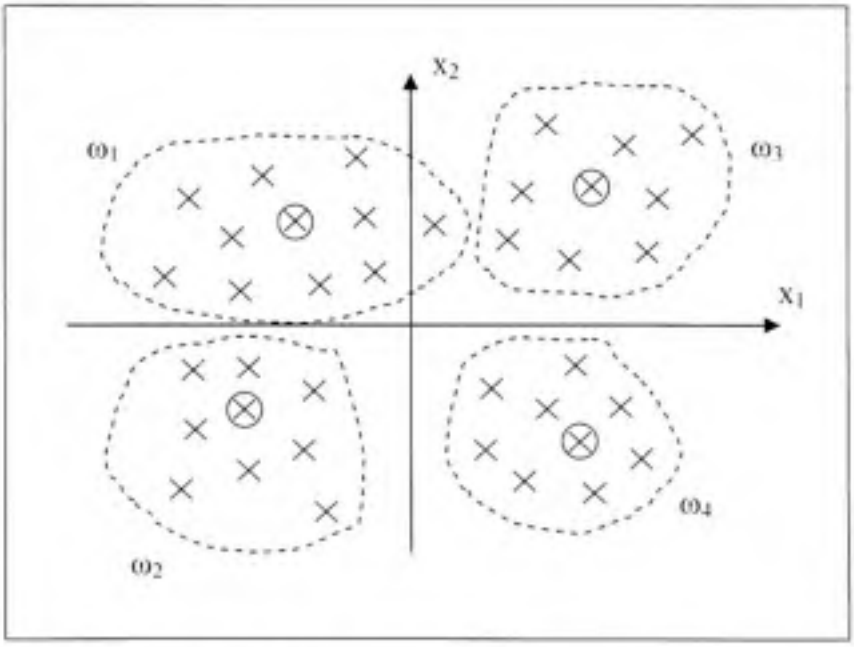

Figure 3.14 ***Classification d'un ensemble de 38 éléments en 4 sous-ensembles.***

3.5 Conclusion

Ce chapitre nous a permis de présenter la structure globale d'une méthode de reconnaissance de la parole. De la détection d'activité vocale à la création d'un dictionnaire de références, les tâches à effectuer sont nombreuses pour arriver à une reconnaissance optimale. Mais le choix le plus important teste celui de la procédure à utiliser pour

reconnaître le mot de parole, et les trois principales méthodes décrites au cours de ce chapitre abordent toutes le sujet selon des aspects différents, et avec plus ou moins de réussite, dépendamment des conditions d'application.

Pour les applications fonctionnant sur des systèmes aux ressources limitées, comme les processeurs de traitement du signal utilisés pour la téléphonie mobile, ce choix se porte régulièrement sur les algorithmes de programmation dynamique ou les modèles de Markov cachés, des sacrifices devant toutefois être réalisés pour s'adapter aux conditions imposées par le système, notamment en ce qui concerne la taille du vocabulaire. Les résultats sont donc généralement très bons, tournant autour de 5% d'erreur, pour les systèmes monolocuteurs utilisant des vocabulaires de quelques mots. Ainsi, Suyay *et al* (2004) obtiennent un taux de réussite de 94,85% lors d'une reconnaissance de 5 chiffres isolés utilisant la DTW, sur le processeur TMS320LF2407. Le passage en reconnaissance indépendante du locuteur réduit toutefois grandement la qualité de la reconnaissance si le nombre de références n'est pas suffisamment large. Lévy *et al* (2004) ont effectivement démontré que, d'un pourcentage d'erreur de 4,8%, on pouvait descendre à 42,54% en passant en multilocuteur, là où les expériences en laboratoires fournissent des résultats proches du 95% de réussite.

De ce point de vue, les méthodes utilisant les modèles de Markov cachés rencontrent plus de succès. Pour un temps de calcul à peu près semblable et une taille mémoire raisonnable, Levy et al (2004) sont donc passés de 42% d'erreur obtenues par DTW à 6% d'erreur en utilisant les HMM. Plus généralement, cette méthode fournit des taux de réussite monolocuteur aux alentours de 99% pour des petits vocabulaires de mots isolés, comme des chiffres ou des mots de commande, mais avec des tailles mémoire de l'ordre de plusieurs Mo. En applications multilocuteurs, Levy et *al* (2005) sont passés d'un taux d'erreur de 11% pour un vocabulaire de 47Ko à 3,7% pour un vocabulaire de 8 737 Ko. Le dernier point problématique concerne le temps de calcul, et peut être visualisé grâce à Hui et *al* (1998), pour qui la phase de reconnaissance en DTW représentait plus de 70% du temps de calcul total (le reste étant occupé par la phase de traitement du signal).

Toutes ces méthodes sont donc très séduisantes, mais souffrent du fait qu'elles ont initialement été conçues pour des systèmes assez généraux disposant de ressources élevées, et seule une augmentation constante de ces ressources a permis jusqu'ici d'améliorer les performances. Il serait donc intéressant d'essayer d'utiliser chacun des aspects attractifs de ces différentes méthodes, en les regroupant au sein d'une même méthode, afin d'obtenir de bons résultats pour une application précise. La description de notre méthode, basée sur cette considération, sera présentée au cours du chapitre suivant.

CHAPITRE 4

CONCEPTION D'UNE MÉTHODE DE RECONNAISSANCE DE LA PAROLE POUR UN SYSTÈME EMBARQUÉ

4.1 Introduction

Notre objectif est donc de développer une méthode pouvant reconnaître des chiffres, isolés ou connectés, prononcés par un seul ou plusieurs locuteurs différents. La grande difficulté provient du fait que cette méthode doit pouvoir être utilisée sur un processeur de traitement du signal, et donc produire de bons résultats tout en satisfaisant aux contraintes de mémoire et de rapidité imposées par un tel système. En empruntant certaines caractéristiques des algorithmes présentés au chapitre précédent, et en nous basant sur une bonne connaissance du vocabulaire concerné, nous tenterons donc de concevoir une méthode la plus optimisée possible pour l'utilisation que nous souhaitons en faire, tout en obtenant des résultats proches des systèmes plus classiques.

Le déroulement général de notre méthode est présenté sur la figure 4.1, et se rapproche donc fortement de ce qui a été présenté au chapitre précédent.

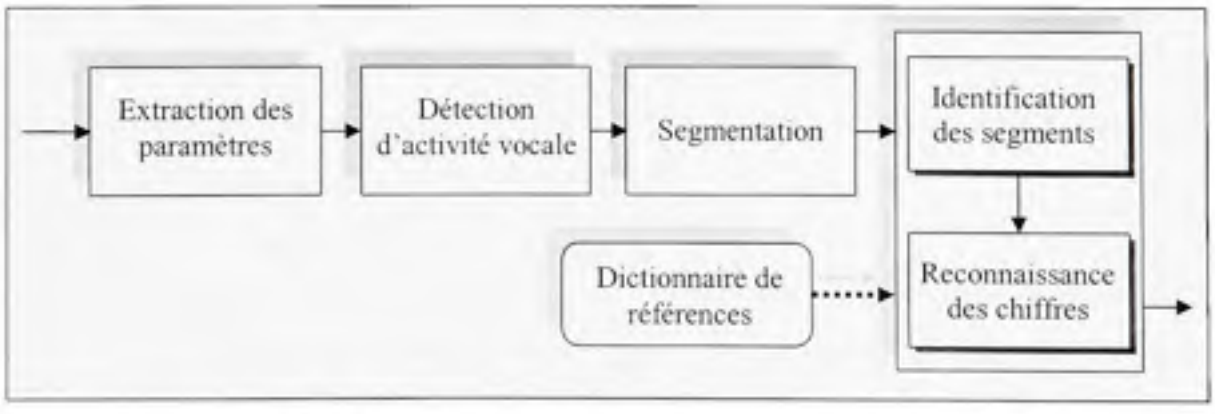

Figure 4.1 ***Déroulement de notre méthode de reconnaissance.***

La première étape, qui consiste à extraire les informations nécessaires au décodage de la parole, comporte les tâches décrites au chapitre 2 : à partir d'un signal échantillonné, que nous découpons en trames d'environ 13ms superposées à 50%, sur lesquelles on applique une fenêtre de Hamming, une analyse par prédiction linéaire est effectuée, de laquelle on en déduit les coefficients cepstraux. Comme mentionné auparavant, le choix se porte sur ces coefficients en raison de leur robustesse et de leur capacité à transporter explicitement des informations sur le signal. Un nombre de 20 coefficients sera sélectionné pour représenter une trame de parole, afin de disposer d'une information très précise sur le signal.

La seconde étape concerne la détection d'activité vocale, nécessaire pour éliminer les zones de silence entourant les chiffres ou les chaînes de chiffres. Nous utiliserons la méthode travaillant à partir de l'énergie et du taux de passage par zéro, présentée au chapitre 2, en y apportant toutefois quelques modifications destinées à l'adapter au système et au vocabulaire sur lequel nous travaillons. Nous présenterons ces arrangements au cours de la seconde partie de ce chapitre.

L'étape suivante se trouve être le découpage du signal vocal en segments acoustiquement uniformes. Si la reconnaissance de chiffres isolés peut s'effectuer en prenant la forme globale du mot, dans le cas de chiffres connectés il faut être capable de déterminer les frontières entre ces mots. Ici, nous tenterons plutôt de repérer les frontières entre les zones de parole à l'intérieur desquelles les caractéristiques spectrales varient peu. Ce procédé, faisant appel à une mesure des variations spectrales au cours des trames successives, sera décrit lors de la troisième partie de ce chapitre.

La quatrième et dernière étape se décompose en deux parties différentes. L'objectif sera d'utiliser la succession de segments obtenus grâce à la phase précédente afin de déterminer quels sont le ou les chiffres engendrés par cette séquence. La première partie du processus consistera donc à identifier tout d'abord les segments en les comparant avec l'ensemble des références contenues dans le dictionnaire. Une fois chaque segment étiqueté, on procèdera alors à une « reconstruction » du chiffre, un peu à la manière des modèles de Markov

cachés, mais en s'adaptant plutôt aux spécificités et caractéristiques du vocabulaire utilisé ici. Ces deux tâches seront décrites au cours des parties 5 et 6 de ce chapitre, la quatrième étant consacrée à la description de la constitution du dictionnaire de référence.

4.2 La détection d'activité vocale

Comme mentionné au chapitre 3, la détection d'activité vocale est une étape très importante en vue d'obtenir les résultats les plus satisfaisants possibles, et de réduire la quantité de calculs et la place mémoire utilisés pour la procédure de reconnaissance. La figure 4.2 présente bien ce problème, le chiffre 8 (prononcé en anglais), tiré de la base de données TI-DIGITS, étant entouré de larges zones de silence non nécessaires à la reconnaissance, dont la taille est égale, sinon supérieure, à la taille du chiffre lui-même.

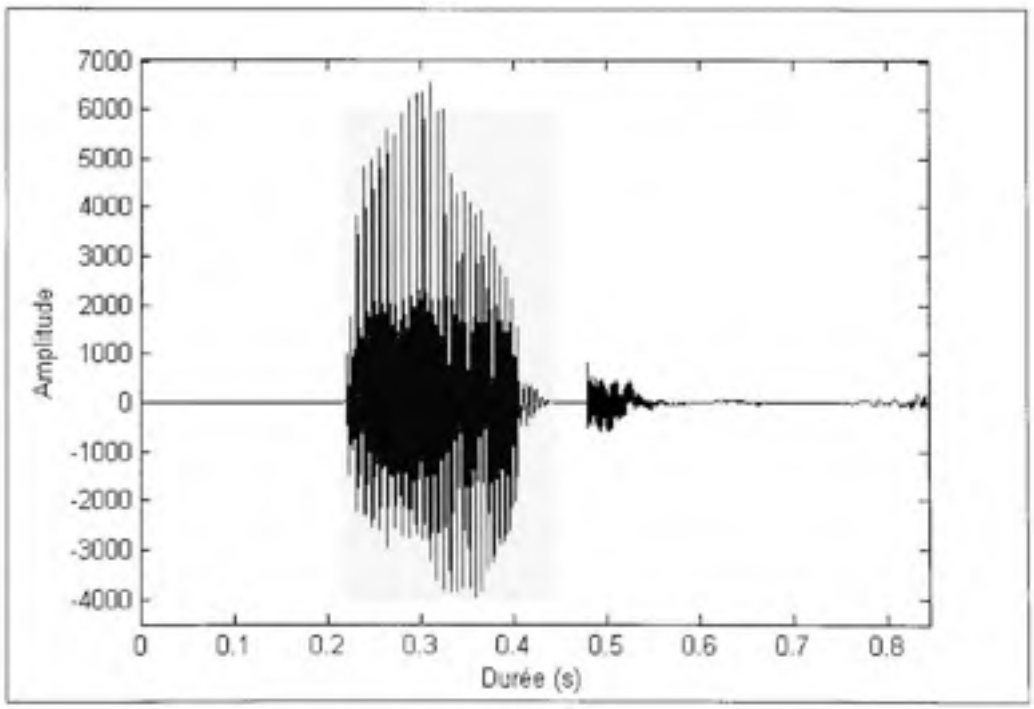

Figure 4.2 ***Chiffre 8 non isolé des zones de silence.***

Mais cette figure présente également une autre information intéressante : tout comme le chiffre 6, le chiffre 8 possède une zone de silence « intérieure » qu'il est nécessaire de

conserver afin de ne pas couper le chiffre en deux. La méthode de détection d'activité vocale (DAV) décrite dans le chapitre 3 apparaît donc comme étant idéale pour notre vocabulaire, puisqu'elle effectue la détection à partir des limites extrêmes du signal, en s'arrêtant à la première zone de parole rencontrée, sans se soucier des éventuelles zones de silence à l'intérieur des frontières obtenues. Néanmoins, cette caractéristiques des chiffres 6 et 8 présente également un inconvénient pour cette méthode : les zones « isolées » sont des sons non-voisés, autrement dit leur énergie peut parfois être faible, et donc non détectée par l'algorithme de Rabiner. Nous allons donc présenter les légères modifications à apporter à l'algorithme afin de le rendre fonctionnel pour notre vocabulaire, puis nous introduirons ensuite une seconde modification portant sur l'utilisation du paramètre de l'énergie, dont nous pouvons simplifier le calcul en utilisant les coefficients LPCC.

4.2.1 Modification de l'algorithme de Rabiner

Si l'on observe le schéma-bloc de la figure 3.2, une fois la première limite fixée grâce aux seuils d'énergie, l'algorithme affinait sa détection en parcourant le signal sur une dizaine de trames à partir de ce point, et modifiait sa décision si le seuil du taux de passage par zéro était dépassé sur plus de trois trames. Cette technique n'est donc pas adaptée si une zone de parole non détectée au premier abord se trouve trop éloignée du reste du mot, comme c'est souvent le cas pour les chiffres 6 et 8. De plus, la zone d'affinage par le taux de passage par zéro apparaît parfois trop grande, et il arrive que la présence de pics isolés, due à une mauvaise qualité du signal, entraîne un déplacement indésirable de la limite d'activité vocale. Nous utiliserons donc une méthode d'affinage personnalisée faisant appel à trois tests consécutifs, comme présenté sur la figure 4.3.

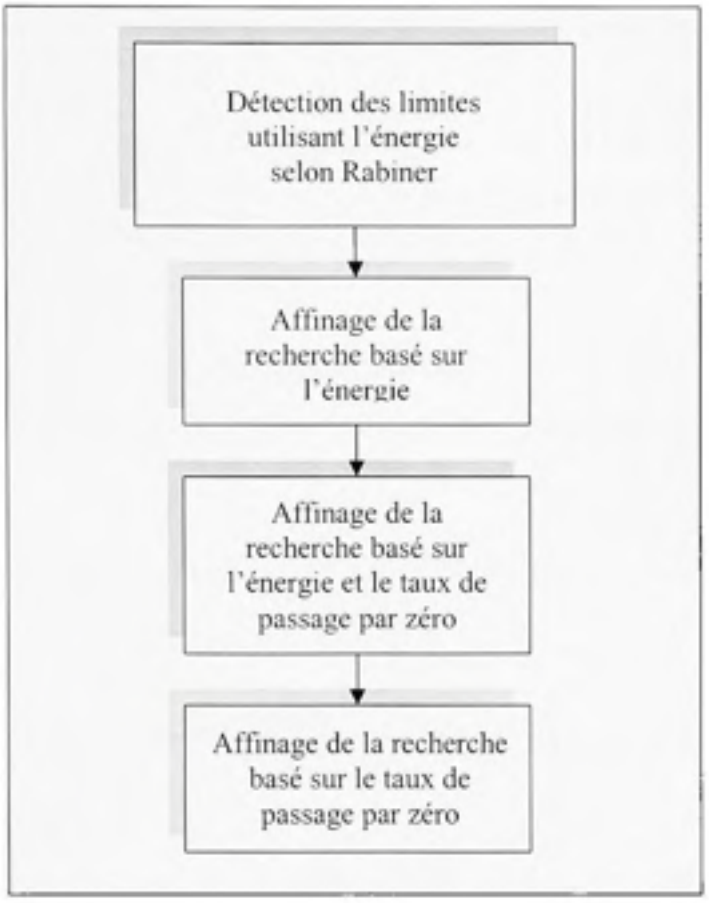

Figure 4.3 ***Déroulement de la modification apportée à l'affinage de détection.***

Les opérations à effectuer sont donc les suivantes :

- décalage du premier point de limite tant que la valeur de l'énergie reste supérieure au seuil S_1 défini par la formule 4.1. Cela permet de prendre en compte une pente énergétique qui serait plutôt douce;

$$S_1 = \frac{2.(ITU + ITL)}{5} \tag{4.1}$$

- décalage basé cette fois-ci sur deux paramètres : on déplace le point de limite aussi longtemps que l'énergie reste supérieure au seuil ITL, et que dans le même temps le taux de passage par zéro reste supérieur au seuil IZCT. Cette opération a pour but de

mieux prendre en compte les parties non voisées ayant une faible amplitude énergétique;

- décalage effectué tant que le taux de passage par zéro est supérieur au seuil IZCT sur plus de trois trames consécutives. Cela permet donc d'éviter de prendre en compte une valeur qui serait un simple pic d'amplitude isolé.

Cette succession de tests nous a fourni des résultats satisfaisants, dont on peut observer un exemple sur la figure 4.4, représentant la détection des limites du chiffre 6 (prononcé en anglais).

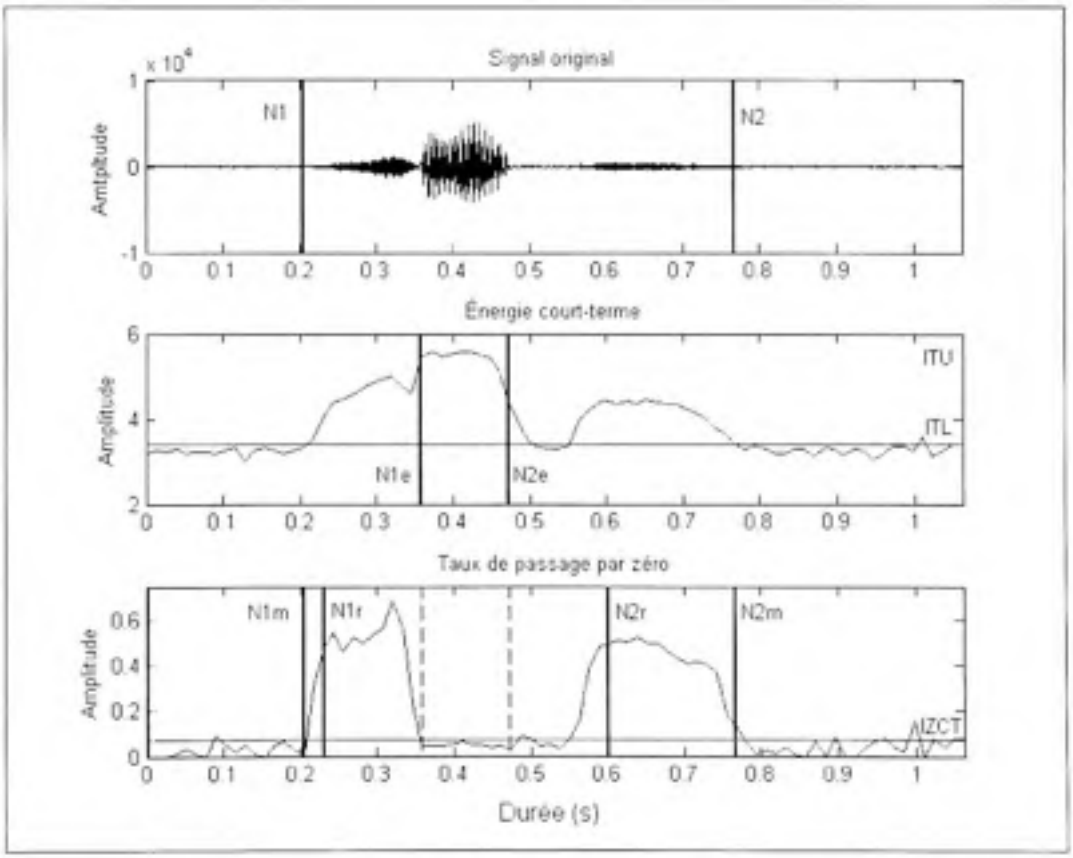

Figure 4.4 ***Exemple de détection d'activité vocale modifiée pour le chiffre 6.***

On peut y observer les limites N_{1r} et N_{2r}, telles qu'obtenues selon l'algorithme initial, après affinage à partir des points N_{1e} et N_{2e}. Les limites N_{1m} et N_{2m} représentent quant à elles les frontières obtenues grâce à la série de vérifications présentées précédemment.

4.2.2 Apport des coefficients LPCC

La méthode présentée jusqu'à présent fait appel à l'énergie court-terme, calculée selon la formule 2.5. Cela implique donc de parcourir une première fois le signal entier en calculant l'énergie pour chaque trame, avant de le reparcourir une seconde fois pour déterminer les frontières des zones de parole. Cette procédure consomme donc du temps de calcul et de la mémoire, mais peut être simplifiée en utilisant la caractéristique des coefficients LPCC présentée au chapitre 2.

En effet, le premier coefficient correspond à l'énergie du signal, comme nous avions pu le constater au moyen de la figure 2.17. Si l'échelle d'amplitude est différente, entre l'énergie et ce coefficient, la forme globale de leur évolution est identique. Il serait donc judicieux, puisque les coefficients LPCC sont déjà à notre disposition, de remplacer l'énergie par ce premier coefficient lors de la procédure de recherche des zones de parole. Cela impliquerait donc simplement d'effectuer le calcul du taux de passage par zéro en même temps que l'extraction des paramètres, afin que le module de détection d'activité vocale n'ait qu'à utiliser des données déjà calculées auparavant, sans avoir à reparcourir entièrement le signal. Cette modification est donc très importante en vue d'une optimisation maximale de notre méthode de reconnaissance vocale, et la figure 4.5 présente l'absence de différences entre une détection utilisant l'énergie et une détection utilisant le premier coefficient cepstral.

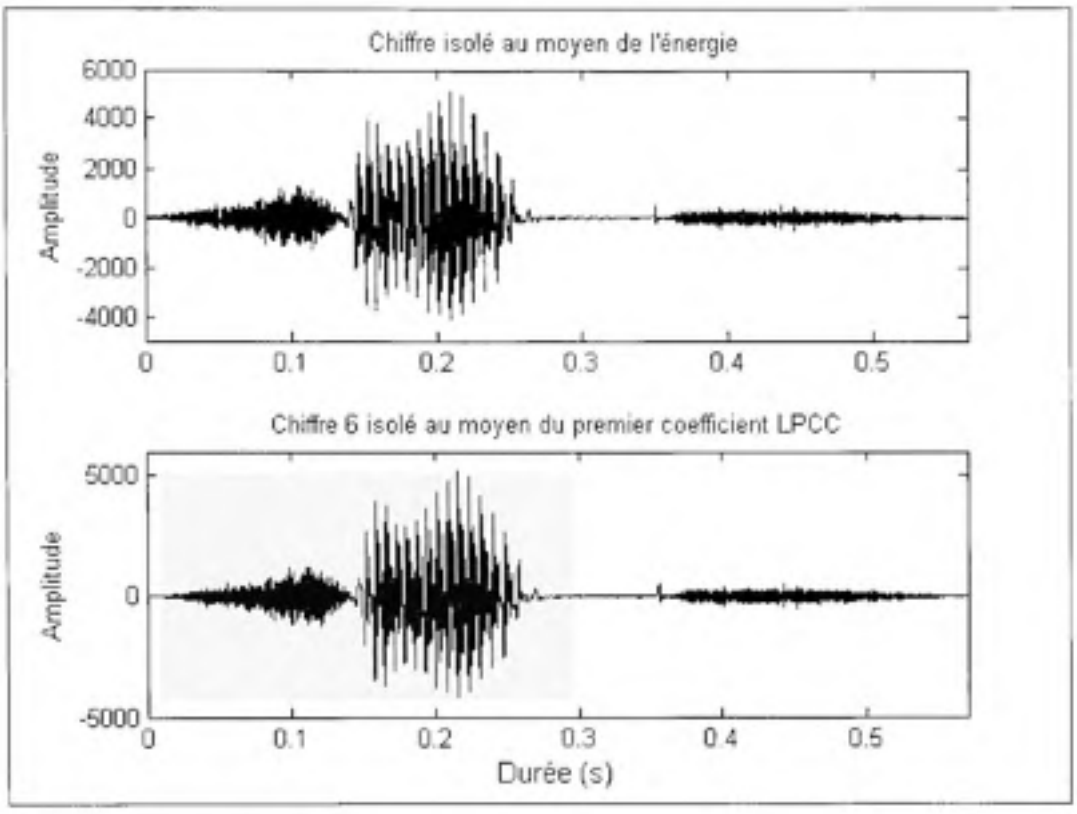

Figure 4.5 ***Comparaison de la DAV utilisant l'énergie et le premier coefficient LPCC.***

4.3 La segmentation

Maintenant que nous possédons la forme exacte du mot ou de la chaîne de mots à reconnaître, il est possible de commencer le processus de reconnaissance. Mais étant donné que notre application a pour objectif de fonctionner sur des chaînes de chiffres connectés, il est impératif de pouvoir séparer les chiffres afin de les étudier isolément les uns des autres. La segmentation en chiffres entiers est rendue d'autant plus difficile que ces derniers sont généralement prononcés sans interruption entre chaque mot, comme le montre la figure 4.6 qui présente la séquence « two zero nine zero four ».

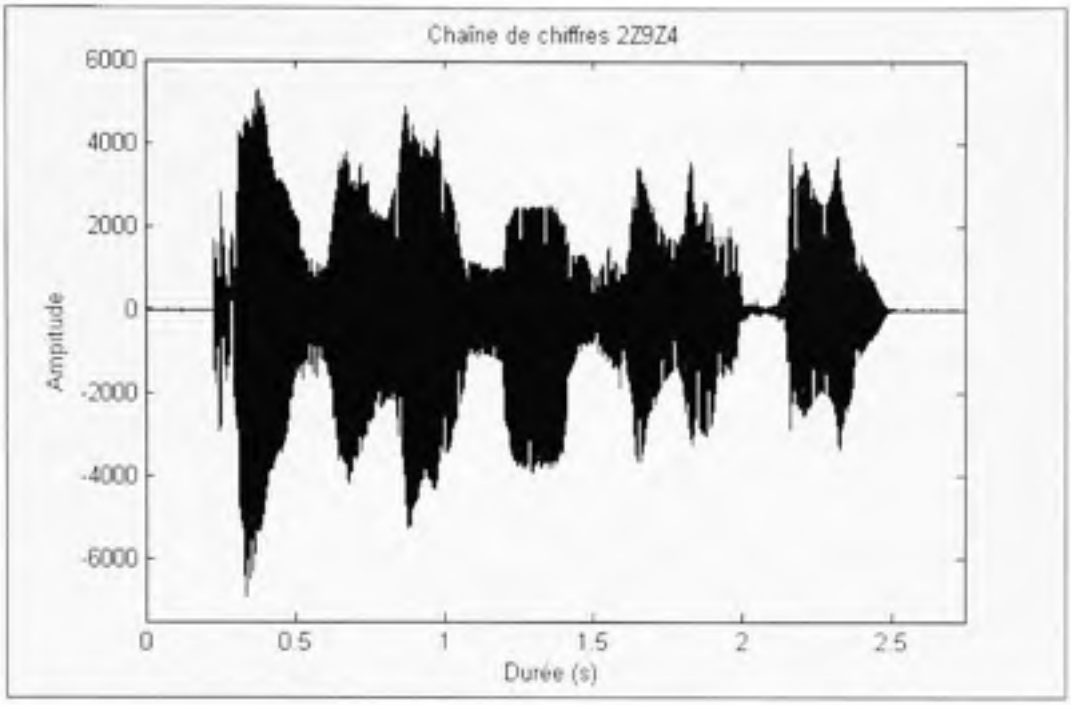

Figure 4.6 ***Exemple d'une séquence de chiffres prononcés successivement.***

De nombreuses méthodes tentent de repérer les frontières entre chaque chiffre, comme Zaabi (2004) qui utilise la méthode de *Hull Convex* associée à l'énergie du signal pour repérer les zones de creux énergétique, considérant qu'elles représentent généralement une transition entre deux chiffres. Dans le cadre de notre étude, nous considèrerons plutôt un découpage en segments acoustiquement homogènes. Pour ce faire, nous ferons appel à la méthode de segmentation basée sur les filtres hybrides multi-niveaux, également appelés MHF pour *Multi Hybrid Filter*, introduits part Faundez et Vallverdu (1996). Cette méthode, qui s'appuie sur les variations des caractéristiques spectrales du signal au cours du temps est très intéressante car elle permet de travailler à partir des coefficients de prédiction linéaire. Nous allons donc présenter le déroulement général d'une segmentation par MHF, avant de décrire la façon dont nous procèderons pour l'appliquer à notre système.

4.3.1 Algorithme de segmentation par filtre hybrides multi-niveaux

Cette méthode s'appuie donc sur la mesure des discontinuités locales des caractéristiques spectrales de la parole, en travaillant trame par trame sur toute la durée du signal. Pour mesurer ces discontinuités, l'algorithme sélectionne à chaque pas un groupe de neuf trames, au centre duquel se trouve la trame en cours, comme présenté sur la figure 4.7. Chacune de ces trames est donc représentée par un vecteur pouvant contenir les coefficients spectraux du signal obtenus après une analyse par TFD, ou même les coefficients de prédiction linéaire.

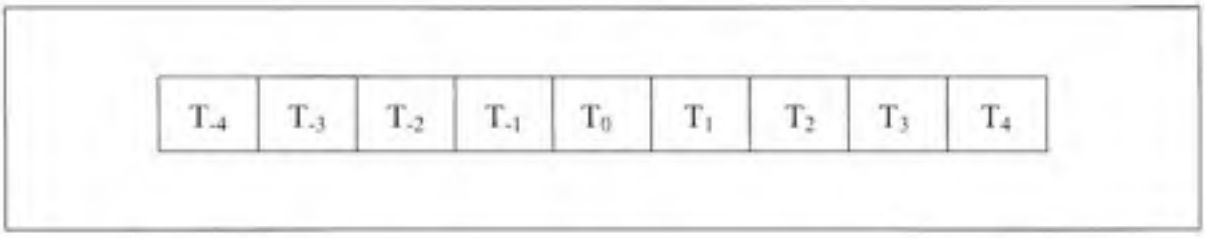

Figure 4.7 ***Ensemble de trames utilisées pour un tour d'algorithme.***

L'algorithme travaille ensuite sous la forme de filtres hybrides multi-niveaux : multi-niveaux car la sortie est le résultat de deux étapes successives de filtrage, et hybrides car cela combine du filtrage linéaire et non linéaire. Au premier étage, on obtiendra huit ensembles de paramètres spectraux, M_i, qui sont en fait les moyennes des paramètres de huit groupes de trames. Une différence est alors effectuée entre chacun de ces ensembles au moyen d'une mesure de dissemblance comme la distance spectrale euclidienne ou la distance LLR présentées au chapitre 2. Nous obtenons donc un ensemble de sept valeurs D_i qui représentent les différences spectrales entre les parties situées à gauche de la trame, et celles situées à sa droite. La figure 4.8 schématise les opérations réalisées durant cette étape.

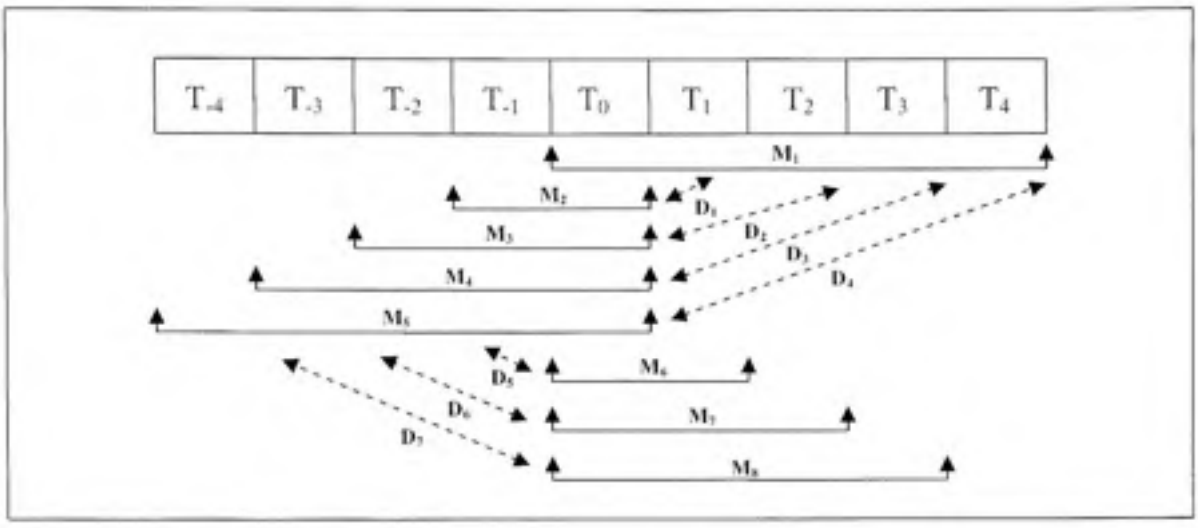

Figure 4.8 ***Premier étage du MHF.***

Les sept différences spectrales D = [D_1 D_2 D_3 D_4 D_5 D_6 D_7] sont alors utilisées dans le second étage, qui consiste à appliquer le filtre non linéaire présenté par la formule 4.2.

$$S = \min\{D\} \tag{4.2}$$

Si on utilisait le filtre *max{}*, il y aurait une tendance à détecter les transitions avec anticipation, et cela influera sur les trames suivantes. Le filtre *min{}*, en revanche, prendra en compte la plus petite valeur de la variation. Si la plus petite différence est déjà haute, cela signifiera donc d'autant plus que la variation spectrale entre deux trames est réellement élevée.

Répétées pour chaque trame du signal, ces opérations fournissent une variable représentant l'évolution des caractéristiques spectrales au cours du signal. Ainsi, lorsqu'on perçoit un pic d'amplitude très élevée de cette variable, cela signifie qu'à cet instant précis on se trouve entre deux zones aux caractéristiques spectrales différentes. La figure 4.9 présente cela pour une chaîne de plusieurs chiffres, avec un algorithme se basant sur les paramètres LPC.

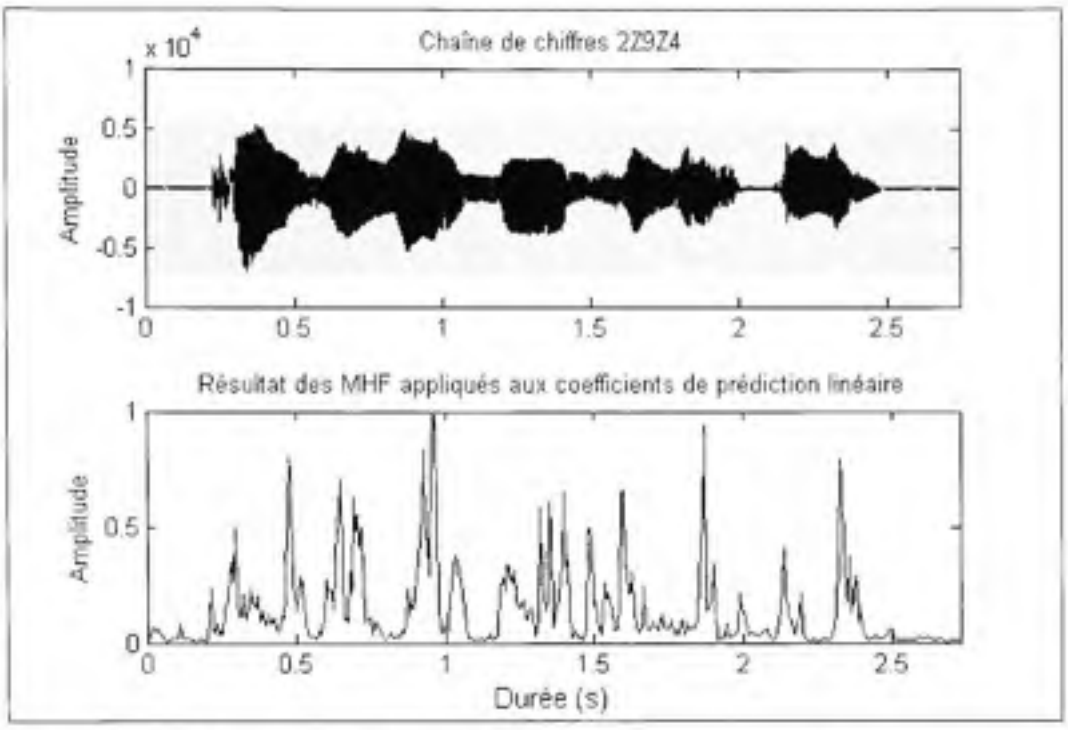

Figure 4.9 ***Résultat des MHF appliqués sur les coefficients LPC.***

Il nous reste donc à déterminer une façon judicieuse d'interpréter des variations spectrales afin d'en déduire les frontières des différents segments acoustiques. La figure 4.9 nous montre bien à quel point cette tâche s'avère compliquée, en raison du caractère très instable du signal. Un filtrage pourrait permettre de lisser la courbe, facilitant ainsi la détection des pics majeurs, mais serait très coûteux en calculs. Nous ferons donc appel aux coefficients LPCC, tout d'abord car il est plus intéressant de les utiliser étant donné qu'ils sont déjà en notre possession, et ensuite car nous pouvons tirer avantage de leurs caractéristiques afin de simplifier la détection des pics. Nous allons maintenant présenter le déroulement de notre algorithme complet de segmentation.

4.3.2 Déroulement de la méthode utilisée

Les coefficients cepstraux représentant les caractéristiques spectrales du signal, leur utilisation dans les filtres hybrides multi-niveaux est donc parfaitement justifiée. Mais, fait encore plus intéressant, contrairement aux coefficients LPC, les coefficients cepstraux

peuvent être utilisés séparément, comme nous l'avons déjà mentionné auparavant. Notre algorithme de segmentation tirera donc profit de ces caractéristiques, en effectuant trois segmentations MHF, chacune travaillant sur des groupes distincts de coefficients. L'objectif sera ensuite d'arriver à combiner intelligemment les trois résultats afin d'obtenir la segmentation désirée. La figure 4.10 présente le déroulement global de la procédure.

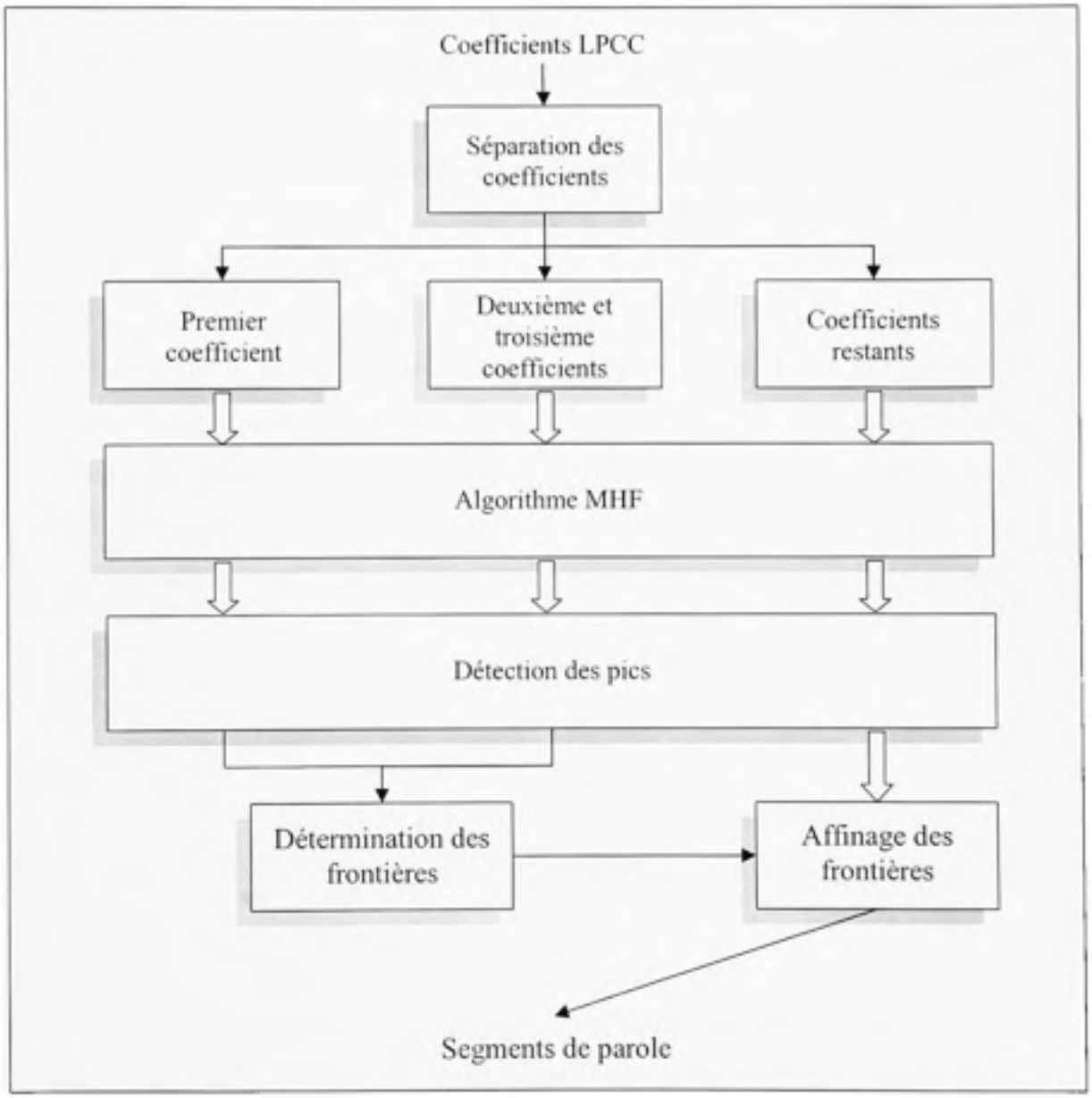

Figure 4.10 ***Déroulement de la segmentation utilisant les coefficients LPCC.***

4.3.2.1 Séparation des coefficients LPCC

Le premier étage de la procédure consistera à isoler les coefficients en trois groupes distincts :

- le premier groupe sera constitué du seul premier coefficient LPCC, utilisé tel quel sous forme d'un vecteur contenant toutes ses valeurs prises au cours du signal de parole. L'intérêt de ce « groupe » sera de permettre une segmentation énergétique du signal;
- le second groupe contiendra les deuxième et troisième coefficients cepstraux. L'objectif sera ici de permettre notamment un découpage entre les zones voisées et non voisées du signal;
- quant au troisième groupe, il rassemblera l'ensemble des coefficients cepstraux restants, afin de permettre une segmentation du signal selon les variations de ses caractéristiques spectrales plus fines, à la manière de ce qui était fait avec l'ensemble des coefficients LPC sur la figure 4.9. Ces derniers coefficients n'ont aucune signification quand ils sont pris isolément, mais leurs variations très rapides font qu'ils représentent une source d'information importante lorsqu'ils sont utilisés en groupe.

4.3.2.2 Détection des pics de variations spectrales

Le second étage de l'algorithme consiste à appliquer les filtres hybrides multi-niveaux à chaque groupe de coefficients, et à en déterminer les pics de variations majeures. Pour ce faire, on fixe trois valeurs importantes :

- un seuil d'amplitude minimale requise pour sélectionner un pic : *seuil_pic*;
- un seuil permettant de déterminer si la variable de variation spectrale est repassée sous une certaine amplitude minimum entre la sélection de deux pics successifs : *seuil_retour*;

- une durée minimale requise entre la sélection de deux pics successifs : *écart_min*.

L'algorithme de sélection des pics à partir du résultat fourni par le passage dans les filtres hybrides multi-niveaux, présenté par le schéma de la figure 4.11, vérifie la valeur de la variation spectrale pour chaque trame, et effectue les tests suivants :

- si les trames adjacentes sont toutes les deux d'amplitude inférieure, on est en présence d'un pic. On vérifie donc si son amplitude est suffisamment élevée pour que l'on puisse la valider;
- si c'est le cas, on teste la valeur « retour », qui nous indique si l'amplitude est retombée sous le seuil *seuil_retour* depuis la sélection du dernier pic, ou non. En fonction du résultat, on fixe la valeur d'écart minimum autorisé entre deux pics consécutifs : soit *écart_min*, soit le double lorsque le retour n'a pas eu lieu;
- on vérifie ensuite si cet écart a été respecté entre le dernier pic validé et celui que l'on est en train de tester : si c'est le cas, on valide le pic; sinon cela signifie que les deux pics sont considérés comme étant trop proches pour être validés tous les deux, on garde donc celui ayant l'amplitude la plus élevée;
- si l'on n'est pas arrivé à la fin du signal, on passe à la trame suivante, en prenant soin de réactualiser l'indice de retour.

Cette méthode fournit un résultat correct pour les trois groupes de coefficients, résultat que l'on peut visualiser au moyen des figures 4,12, 4.13 et 4.14. La figure 4,12 nous montre notamment que lorsque les chiffres sont prononcés de façon rapprochée, il est très difficile de réaliser un découpage énergétique, tandis que la figure 4.13 présente clairement l'efficacité d'un découpage en zones de catégories différentes. Enfin, la figure 4.14 permet de visualiser qu'un découpage basé sur les variations spectrales fines est nettement plus difficile à réaliser.

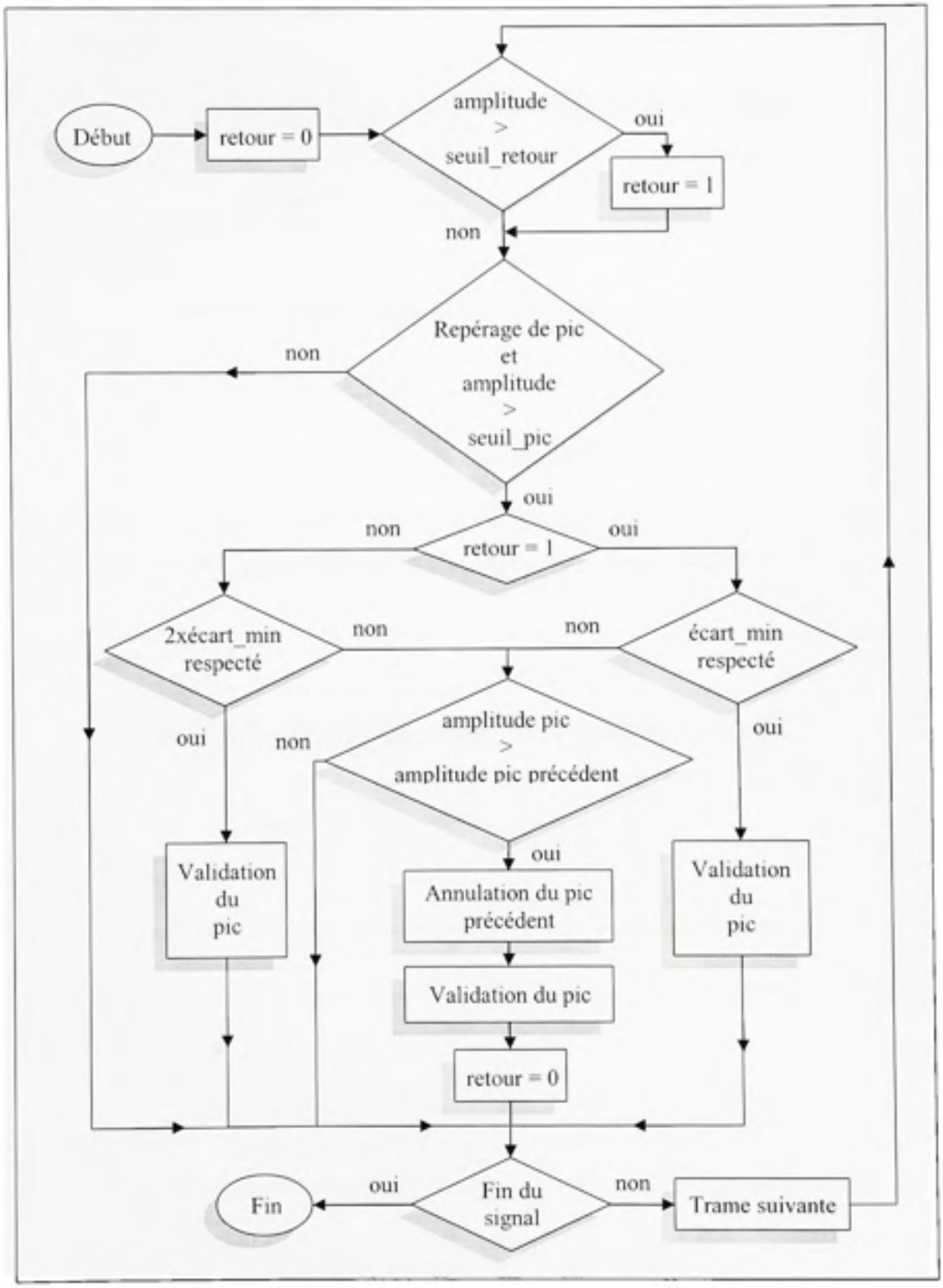

Figure 4.11 ***Algorithme de sélection des pics de variations spectrales.***

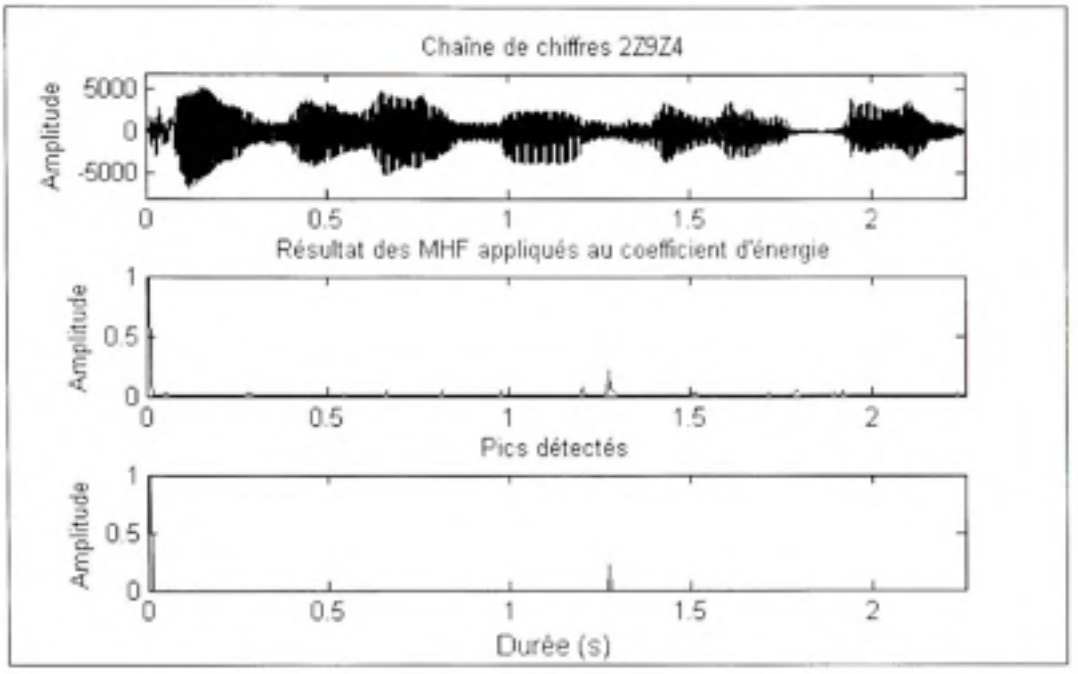

Figure 4.12 ***Segmentation MHF basée sur le premier coefficient LPCC.***

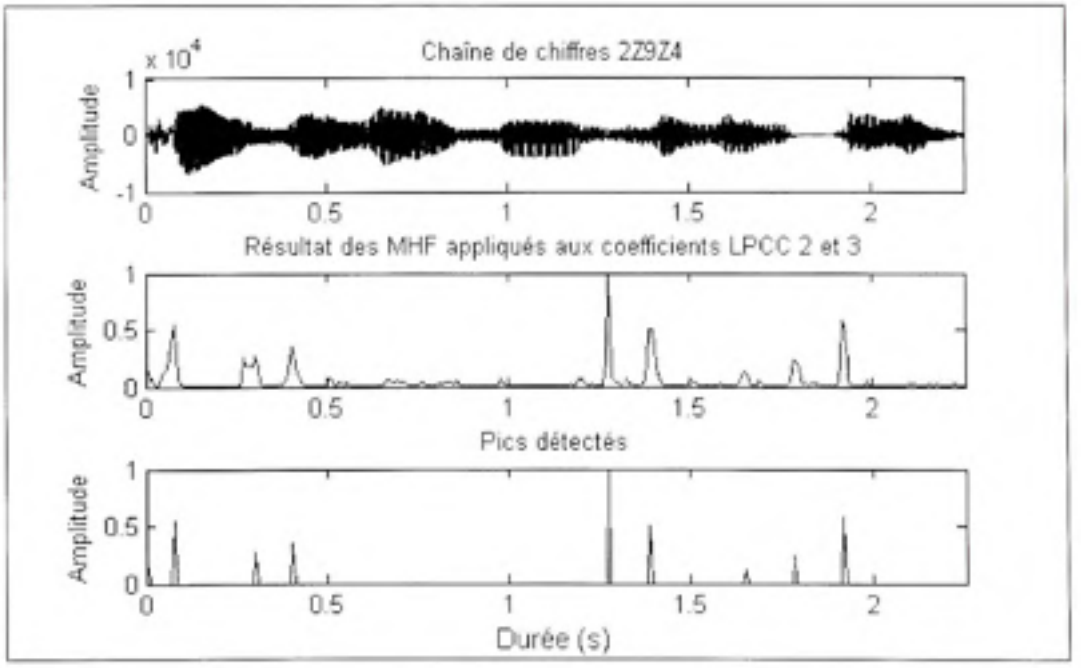

Figure 4.13 ***Segmentation MHF basée sur les 2nd et 3ème coefficients LPCC.***

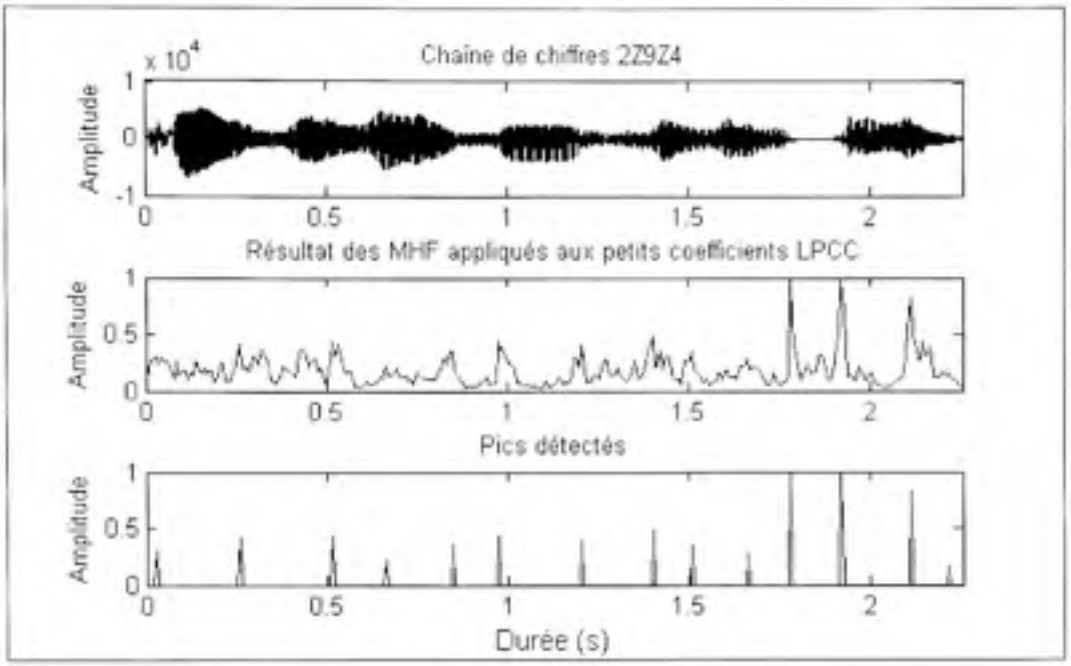

Figure 4.14 ***Segmentation MHF basée sur les petits coefficients LPCC.***

4.3.2.3 Détermination des frontières finales

Chaque groupe de coefficients semble donc présenter des propriétés de segmentation intéressantes; mais pris isolément ils sont soit incomplets soit trop précis pour permettre une bonne segmentation. La dernière étape de l'algorithme de segmentation consiste donc à combiner judicieusement ces trois groupes pour obtenir un résultat satisfaisant. Cela sera réalisé en deux phases :

- la première phase consistera à regrouper la segmentation basée sur le premier coefficient avec celle basée sur les second et troisième coefficients. Les deux jeux de pics seront en fait superposés en un seul et même ensemble : lorsque deux pics venant de deux groupes différents seront jugés trop proches, on ne retiendra que celui ayant la plus forte amplitude. Le résultat sera donc un ensemble de frontières nettement plus complet que ceux obtenus auparavant;
- la seconde phase consistera à compléter encore la segmentation, en ajoutant, dans l'ensemble des deux premiers groupes, le pic de plus forte amplitude venant du

troisième groupe, si toutefois il ne coïncide pas avec une frontière déjà retenue. Un processus d'affinage de la segmentation sera ensuite effectué, en insérant les pics du troisième groupe dans les intervalles entre deux frontières consécutives jugés trop grands. Nous cherchons effectivement à découper le signal en phones homogènes, c'est-à-dire en unités relativement courtes; un segment trop long pourrait donc être du à la non-détection d'une transition entre deux zones acoustiquement différentes.

Le résultat final présente une bonne segmentation entre les différents évènements sonores, même si une sur-segmentation est parfois obtenue. Mais le principal objectif de ce module est de manquer le moins de frontières possibles, puisque cette éventuelle sur-segmentation pourra être rattrapée par la suite de la méthode, qui consistera à reconstruire le signal à partir des segments obtenus. La figure 4.15 présente un exemple de segmentation finale pour un chiffre isolé, tandis que la figure 4.16 concerne une chaîne de chiffres connectés. Notons que cette segmentation nous ramène à la notion de phone homogène, évoquée au cours du chapitre 2. Il reste donc à déterminer comment identifier et utiliser cette succession d'unités, et la partie qui suit nous aidera dans cet exercice, en détaillant la constitution du dictionnaire contenant les références qui seront utilisées pour réaliser cette tâche.

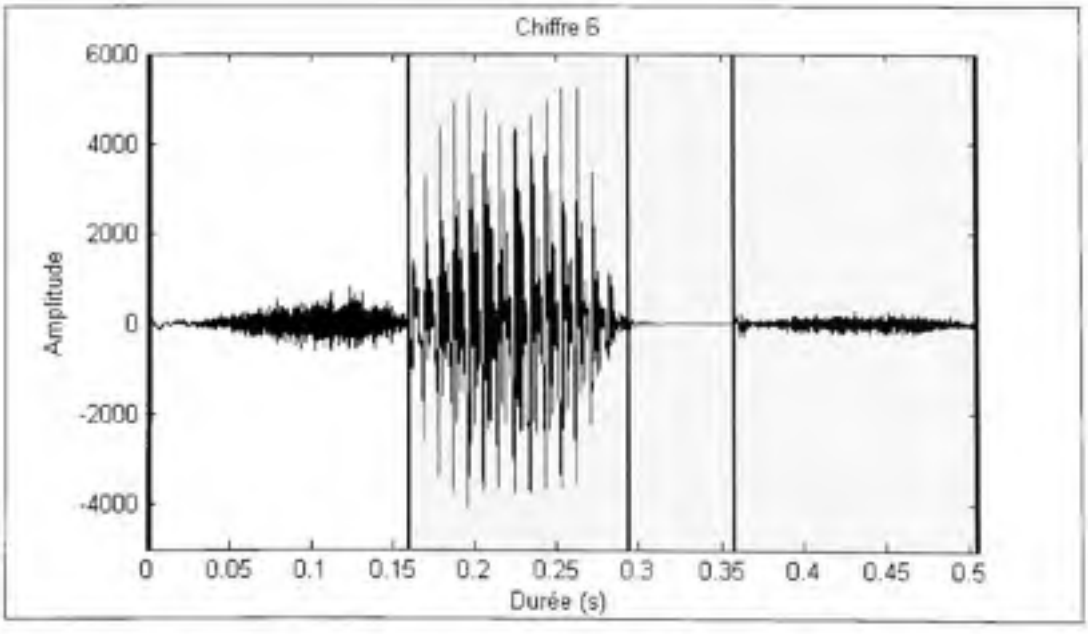

Figure 4.15 ***Segmentation complète d'un chiffre isolé.***

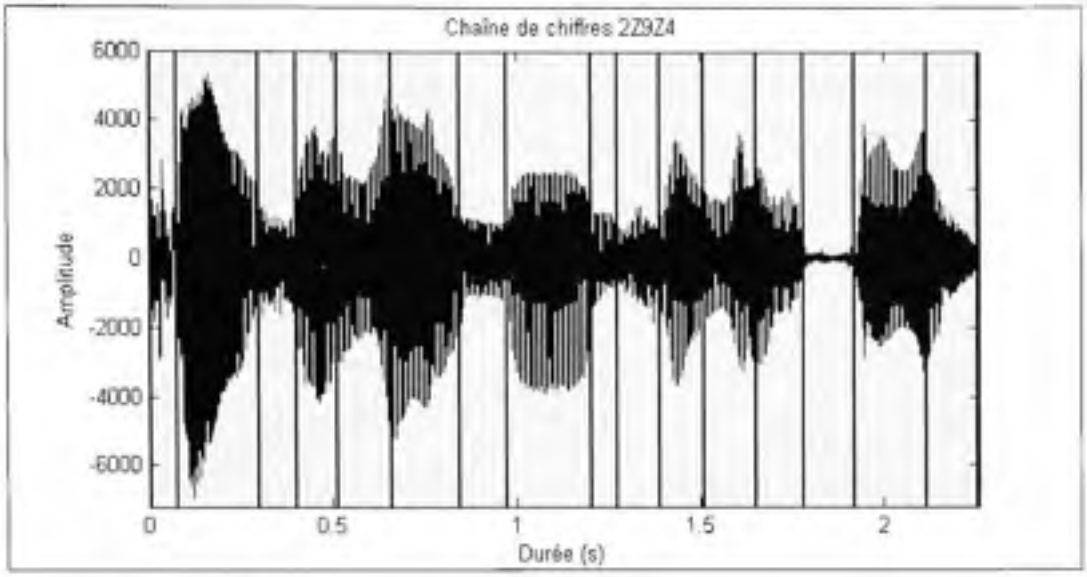

Figure 4.16 ***Segmentation complète d'une chaîne de chiffres.***

4.4 Le dictionnaire de références

Rappelons-le, notre objectif est d'effectuer la reconnaissance de chiffres prononcés en anglais, allant de 0 à 9, le zéro ayant deux prononciations possibles. Afin d'optimiser au maximum notre méthode de reconnaissance vocale, il est important de s'appuyer sur les caractéristiques spécifiques du vocabulaire. Parmi ces spécificités, on peut notamment citer sa taille, et donc par conséquent la faible quantité de sons différents qui le composent. Ce sont justement ces sons qui nous intéressent ici, puisque, la segmentation nous ayant permis de les isoler, il nous faudra à présent être capable de les identifier. Cela sera fait en partie à l'aide d'un dictionnaire contenant l'ensemble des sons de référence, et dont nous allons décrire la constitution, au moyen d'une rapide étude du vocabulaire.

4.4.1 Étude du vocabulaire

De nombreuses observations sur le comportement de la segmentation pour les chiffres de ce vocabulaire nous ont permis de constater que chaque chiffre se découpait généralement de la même façon sur ses différentes versions. Autrement dit, pour chacun des onze chiffres, il est possible d'extraire un modèle segmenté global, pouvant être constitué de trois à six

segments selon les chiffres. Sur l'ensemble du vocabulaire, nous observons donc un total de 40 segments différents, et l'**annexe 1** présente le modèle découpé de chaque chiffre, ainsi que la répartition des 40 segments sur l'ensemble du vocabulaire.

Mais « 40 segments » ne signifie pas « 40 sons différents ». Parmi les onze chiffres, on retrouve en effet une redondance de certains sons, suffisamment importante pour classer les différents sons en un petit nombre de catégories, comme présenté sur la figure 4.17. Cette catégorisation des unités acoustiques du vocabulaire nous permettra d'étiqueter chaque segment selon sa catégorie, simplifiant ainsi la procédure de reconnaissance en étant capable de détecter la classe de l'élément à identifier. L'avantage de s'adapter aux caractéristiques d'un vocabulaire donné est que le nombre de catégories obtenues est bien plus réduit que celui généralement retenu pour classer l'ensemble des phonèmes du langage. Notons également que, du fait de cette classification, un segment peut être associé à plusieurs chiffres, et pas seulement à celui dont il est tiré.

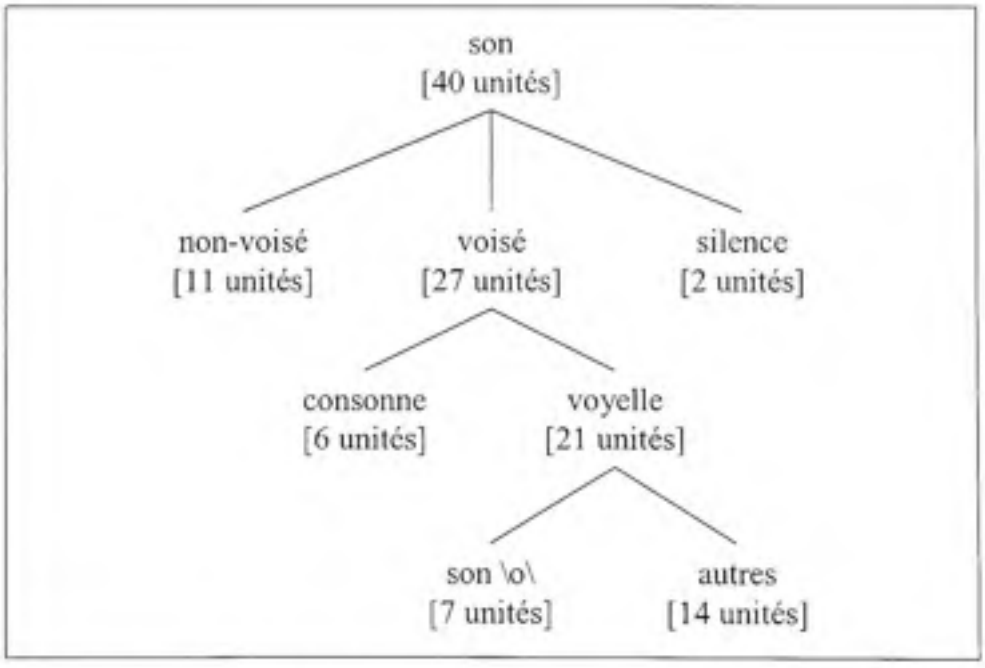

Figure 4.17 ***Classification des sons du vocabulaire.***

Il reste enfin une dernière caractéristique intéressante concernant ce vocabulaire. La segmentation nous permet d'observer que les onze chiffres sont généralement constitués d'états, similaires à ceux présents dans un modèle de Markov cachés. Plus précisément, nos observations nous amènent à considérer que le modèle de conception des chiffres est formé de quatre états : un début, un milieu, un intérieur, et une fin.

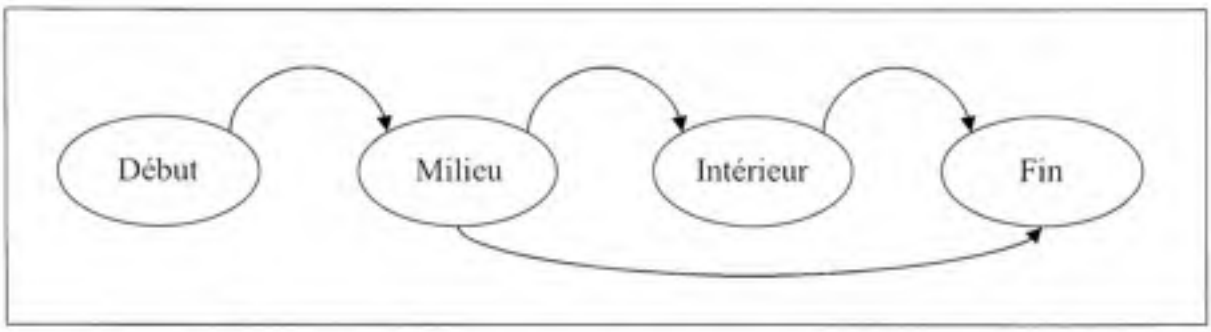

Figure 4.18 ***Modélisation d'un chiffre en quatre états.***

Le modèle présenté à la figure 4.18 représente toutefois un aspect général des chiffres, tous n'étant effectivement pas constitués de ces quatre parties. Ainsi, si le chiffre 6 est construit avec les quatre états différents, le chiffre 1 n'en comporte que deux : un milieu et une fin. Un modèle ne commence donc pas forcément à l'état de début et ne s'arrête pas obligatoirement à l'état de fin, et l'état d'intérieur peut être sauté puisqu'il n'est pas présent dans tous les chiffres. L'état de milieu est en revanche l'élément clé de tous les chiffres, puisqu'il est présent pour chacun d'entre eux. Nous pouvons retrouver en **annexe 1** les états dont sont constitués chacun des onze chiffres du vocabulaire. Un détail intéressant est à relever afin de compléter cette analyse : les états de début et de fin sont systématiquement constitués de sons non-voisés (pour les chiffres 2, 3, 4, 5, 6, 7, 8 et zero) ou de consonnes (pour les chiffres 1 et 9), tandis que les états de milieu toujours constitués de voyelles, que ce soit des sons \o\ ou autres.

4.4.2 Constitution du dictionnaire de références

Pour le processus de reconnaissance, un dictionnaire est nécessaire afin de pouvoir comparer une forme inconnue avec l'ensemble des références. Dans notre cas, ce dictionnaire sera constitué des 40 segments mentionnés précédemment, chaque segment étant représenté par un vecteur de coefficients LPCC. En vue de couvrir une vaste gamme de prononciations différentes, tous les segments de référence seront représentés par plusieurs versions, ces versions étant sélectionnées parmi un ensemble bien plus large, grâce à l'algorithme de classification présenté au chapitre 3. Dans l'optique de prendre en compte les caractéristiques décrites à la partie précédente, chaque segment de référence sera identifié selon :

- le chiffre auquel il est associé;
- sa catégorie phonétique;
- l'état auquel il appartient.

Pour ce faire, nous utiliserons un tableau à trois dimensions, nommé « tableau d'association », constitué de 40 lignes et 16 colonnes, chacune contenant deux informations, comme cela est schématisé sur la figure 4.19.

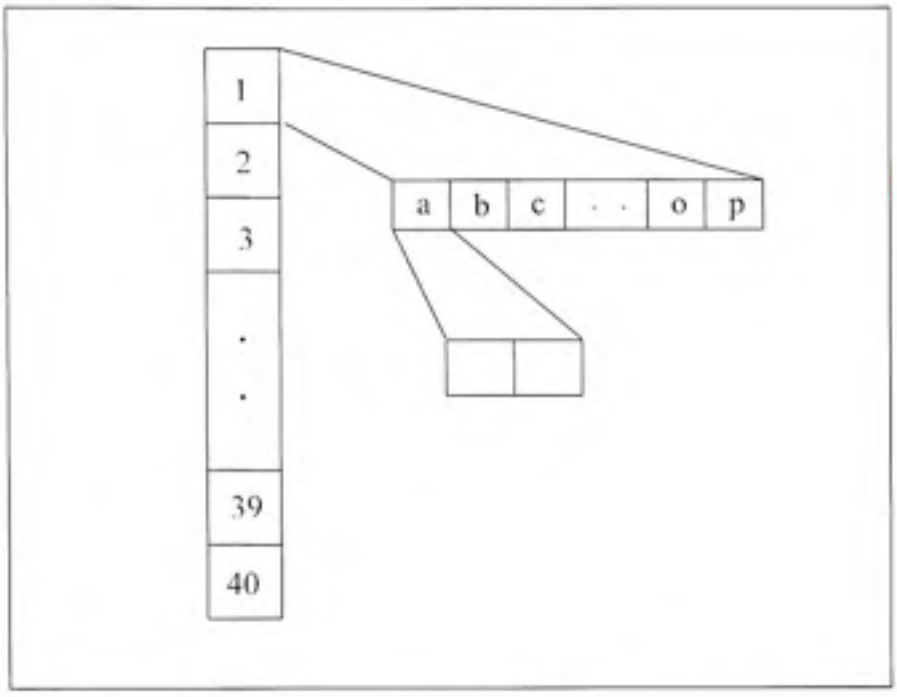

Figure 4.19 ***Schématisation du tableau d'association des segments.***

Ce tableau d'association se présente de la façon suivante :

- les 40 lignes représentent les 40 segments acoustiques du vocabulaire;
- chaque ligne se voit associer 16 colonnes : les neuf premières représentent les chiffres de 1 à 9; les deux suivantes représentent les deux prononciations du zéro; et les colonnes suivantes correspondent aux catégories auxquelles sont susceptibles d'appartenir les segments : silence, non-voisé, voyelle, son \o\, et superposition. Cette dernière catégorie indique en fait si le segment en question peut éventuellement, dans le cas de chiffres connectés, représenter à la fois la fin d'un chiffre et le début du suivant;
- chacune des onze premières colonnes contient donc deux informations : l'état auquel le segment est associé dans le chiffre, indiqué sous la forme d'une valeur allant de 1 à 5 (la valeur 5 indiquant que le segment en question peut aussi bien être associé à un état de début qu'à un état de fin dans le chiffre concerné); et le poids que l'on accorde au segment dans le chiffre concerné, indiqué sous la forme d'une valeur

comprise entre 0 et 3 (0 si l'on considère que le segment n'a rien à voir avec le chiffre en question, et 3 s'il a une très forte probabilité de lui appartenir). Ces valeurs peuvent être assimilées aux probabilités d'émission présentes dans les modèles de Markov cachés.

- Quant aux cinq dernières colonnes, elles ne contiendront qu'un simple indice qui sera activé en fonction de l'appartenance du segment à la catégorie.

Ce tableau d'association regroupe donc toutes les informations nécessaires à la construction d'un chiffre à partir d'une séquence de segments, un peu comme les modèles de Markov contiennent des informations de probabilités de distribution des vecteurs acoustiques pour chaque état, et pour chaque modèle, excepté que la phase d'entraînement des modèles a été remplacée par une étude minutieuse du vocabulaire concerné. L'**annexe 2** présente le contenu du tableau d'association pour le vocabulaire constitué des chiffres 0 à 9.

4.5 L'identification des segments

La procédure de segmentation nous fournit donc une succession de segments acoustiques qu'il faut maintenant être capable d'identifier afin de pouvoir compléter le processus de reconnaissance. L'objectif ici est donc d'arriver à étiqueter chaque segment inconnu selon le numéro d'une ou plusieurs références contenues dans le dictionnaire, cette procédure se réalisant en deux étapes : la première consistant à représenter le segment inconnu par un seul vecteur spectral significatif de l'information globale contenue dans le segment, et la seconde utilisant ce vecteur pour le comparer avec l'ensemble des références. Nous décrivons maintenant ces deux opérations.

4.5.1 Représentation d'un segment

Un segment de parole, tel qu'obtenu grâce à la phase précédente, est donc défini par deux frontières entourant une succession de vecteurs acoustiques. Notre objectif est de représenter l'information contenue entre ces deux frontières au moyen d'un seul vecteur.

Étant donné que la procédure de segmentation décrite au cours de la partie précédente est censée avoir délimité des zones aux caractéristiques spectrales constantes, une moyenne de l'ensemble des vecteurs spectraux contenus par chaque zone devrait théoriquement convenir. On obtient cela avec la formule 4.3, N étant la taille du segment en nombre de trames et C le vecteur de coefficients $p+1$ cepstraux.

$$C(k)\frac{\sum_{i=1}^{N} C_i(k)}{N}, \qquad k = 1,\ldots, p+1 \tag{4.3}$$

Néanmoins, en pratique cette uniformité acoustique des segments n'est pas toujours respectée. En effet, si le découpage est basé sur la sélection des pics de variations spectrales majeures, cela ne signifie pas pour autant que les paramètres du signal ne varient pas entre deux pics sélectionnés, en atteste la difficulté que l'on éprouvait pour extraire les frontières désirées. Comme nous l'avions mentionné, un signal n'est stable que sur des petites durées, de l'ordre de la dizaine de millisecondes, et c'est pour cette raison que nous travaillons par trames. Sur des zones de plusieurs centaines de millisecondes, comme c'est souvent le cas pour nos segments, le signal n'est donc que très rarement stationnaire, mais il transporte tout de même une information globale, et c'est cette information que nous souhaitons représenter. Effectuer une moyenne des caractéristiques spectrales de l'ensemble d'un segment risquerait donc d'altérer cette information, et ce notamment en raison des zones d'extrémités qui sont bien souvent des zones de transition entre deux segments.

Pour « éliminer » ces parties indésirables, nous effectuerons une mini-segmentation à l'intérieur du segment, en séparant les zones lorsqu'apparaît une grande différence spectrale, mais en se basant cette fois-ci simplement sur un calcul progressif de la moyenne des caractéristiques spectrales, en partant du début du segment. Ainsi, lorsque l'on rencontre une trame dont la distance par rapport à cette moyenne est supérieure à un certain seuil, on recommence le processus en prenant comme nouveau point de départ la trame en question, tout en mémorisant les informations du « mini-segment » précédent. Au final, si le segment

n'était effectivement pas stationnaire dans sa totalité, on se retrouve avec plusieurs segments, chacun étant représenté par un vecteur, parmi lesquels il ne faudra en retenir qu'un seul. Cette sélection s'opère selon la procédure présentée par la figure 4.20.

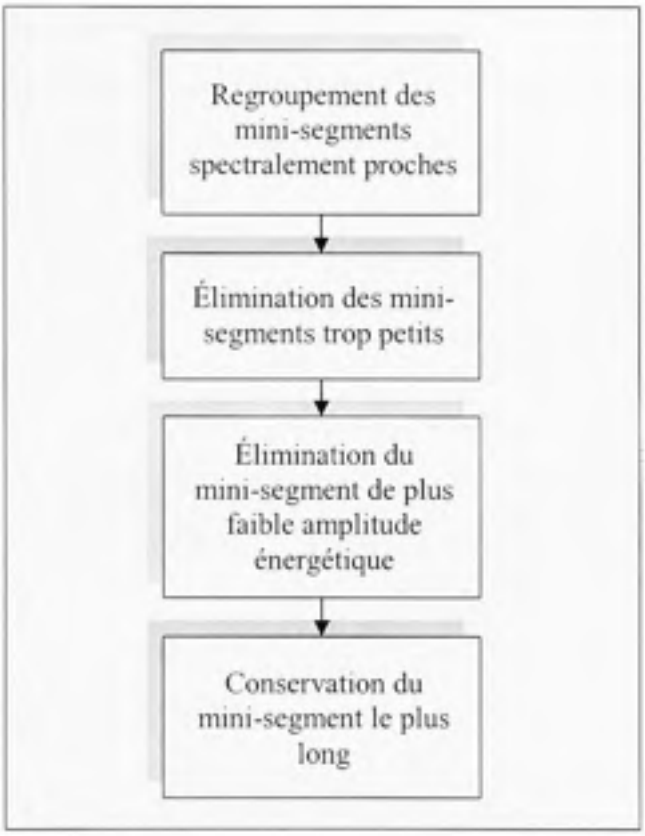

Figure 4.20 ***Processus d'uniformisation de segment.***

On peut décrire les quatre principales étapes comme suit :

- on regroupe tout d'abord les segments dont la différence spectrale est inférieure à un certain seuil : il est effectivement possible que deux mini-segments acoustiquement identiques ou proches aient été séparés par une zone de signal « indésirable ». On effectue cette opération à l'aide d'une matrice contenant les distances entre tous les segments, réactualisant cette matrice à chaque regroupement éventuel de mini-segments, cette procédure étant réalisée jusqu'à stabilisation;

- la seconde opération consiste à éliminer les mini-segments que l'on juge trop petits pour être représentatifs de la totalité du segment. Cette sélection se fait par rapport à un pourcentage de la taille de ce dernier, ce pourcentage étant plus ou moins élevé selon l'amplitude énergétique moyenne du segment (ceci afin de mieux éliminer les zones indésirables aux extrémités);
- le segment d'amplitude énergétique la plus faible est ensuite éliminé;
- au final, on ne retient que le mini-segment le plus long.

Ces opérations permettent d'uniformiser les segments sur lesquels on sera amenés à travailler, tout en essayant le plus possible de ne pas alourdir inutilement les calculs : lorsque le signal est idéal, le calcul progressif de la moyenne des caractéristiques ne rencontre aucun obstacle. Ce n'est que lorsque la segmentation ne s'est pas faite dans les conditions idéales que la procédure d'uniformisation est enclenchée et, là encore, si toutes les zones indésirables ont été éliminées avant la fin du processus, les tests restants ne seront pas sollicités.

4.5.2 Identification d'un segment

L'étiquetage d'un segment peut se faire de façon très « simple » : à l'aide d'une mesure de dissemblance, on compare le segment en question avec l'ensemble des références contenues dans le dictionnaire, en ne retenant que la référence ayant présenté la distance la plus proche. Toutefois, deux problèmes surviennent en procédant de la sorte : le processus de comparaison peut être lent, dépendamment du nombre de références à comparer; et si cela fonctionne très bien lorsque les conditions de reconnaissance sont idéales, il en est autrement dans un contexte multilocuteurs ou bruité. Afin de répondre à ces deux problèmes, notre méthode d'identification sera basée sur une reconnaissance *a priori* de la catégorie à laquelle appartient le segment, en nous basant sur deux des distinctions sonores présentées sur la figure 4.17 : la détection de voisement/non-voisement, puis, pour les sons voisés, la distinction consonne/voyelle, le déroulement général de ce processus d'identification étant présenté sur la figure 4.21, et détaillé par la suite.

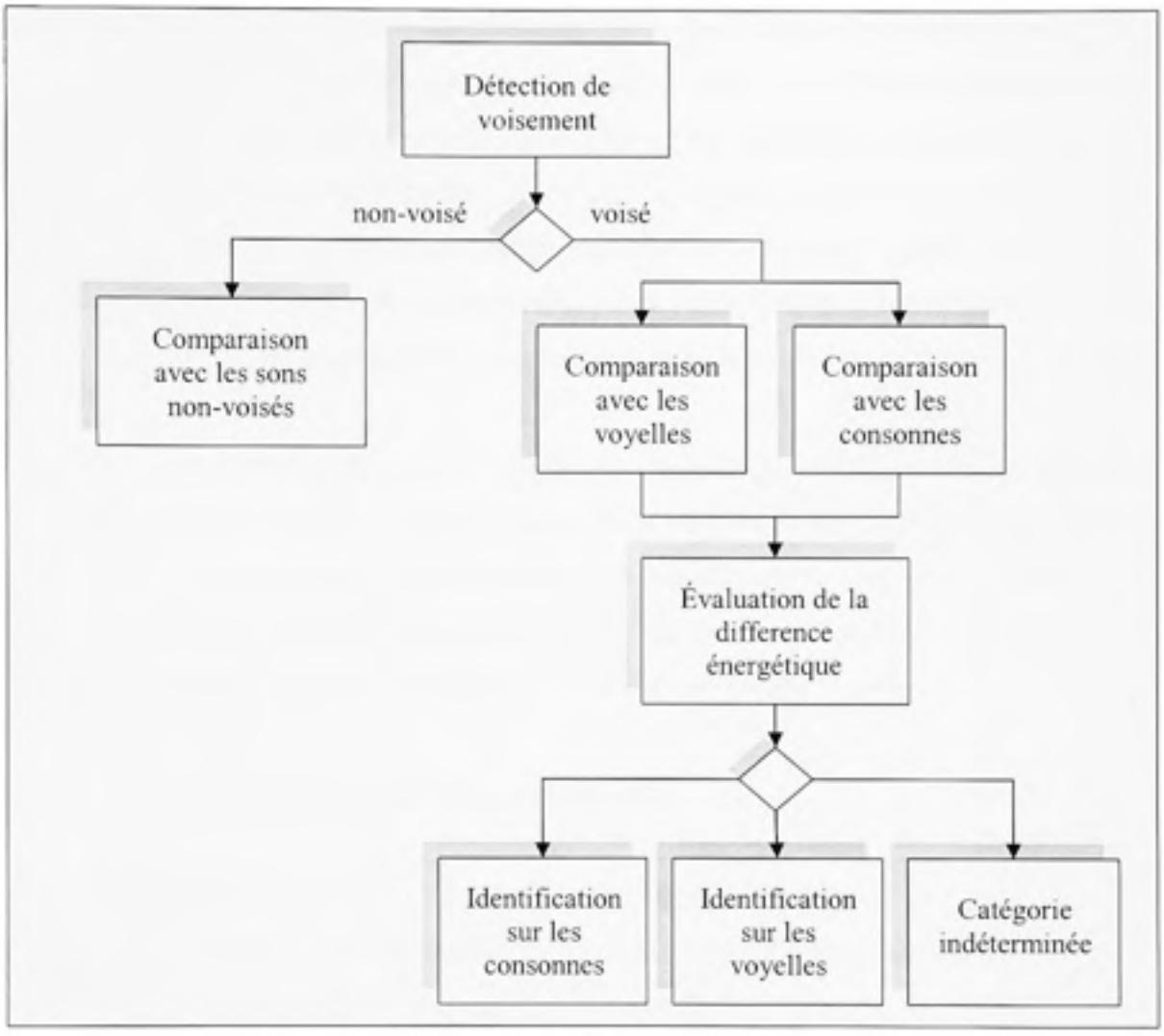

Figure 4.21 ***Processus d'identification de segment.***

4.5.2.1 Distinction de catégorie voisée/non-voisée

Ce sont les deux grandes catégories de sons constituant notre vocabulaire. Comme mentionné précédemment, ce sont les second et troisième coefficients cepstraux qui nous renseignent sur l'état de voisement du signal; plus précisément, la détection d'un son non-voisé s'effectue selon deux tests :

- un premier test est effectué sur le second coefficient LPCC : si sa valeur est inférieure à zéro, on déclarera le segment comme étant non-voisé;

- si le premier test n'a pas été concluant, un second est opéré, portant sur le deuxième coefficient LPCC et l'énergie du signal : une valeur du coefficient inférieure à zéro, associée à une faible amplitude énergétique sera considérée comme une preuve de non-voisement. L'introduction du critère énergétique est due au fait que cette caractéristique du troisième coefficient cepstral est moins « stable » que pour le second, mais les sons non-voisés ayant généralement une faible amplitude énergétique, la réunion de ces deux critères fournit une bonne assurance de réussite.

Lorsqu'aucun des deux tests n'est concluant, le segment est alors considéré comme étant voisé. Cette technique de classification fournit de très bons résultats : sur un ensemble de 30 versions de chacun des 40 segments du vocabulaire, 100% des sons voisés ont été correctement détectés, tandis que les sons non-voisés l'ont été à 94.33%. S'opère ensuite la seconde classification, selon que le son est une consonne ou une voyelle.

4.5.2.2 Distinction de catégorie consonne/voyelle

Cette distinction se fait selon l'énergie du segment, plus précisément en mesurant la différence entre l'énergie du segment inconnu et la moyenne énergétique des références de chaque catégorie. En effet, dans notre vocabulaire de onze chiffres, les consonnes correspondent au son \n\. Si aucune caractéristique des coefficients cepstraux ne nous permet d'effectuer la différenciation entre ces deux catégories, nous avons en revanche remarqué que ces sons \n\ étaient généralement d'une amplitude énergétique bien plus faible que les autres sons voisés. Ce critère n'étant toutefois pas d'une précision redoutable, et donc porteur d'éventuelles erreurs si l'on base toute notre décision finale sur la simple observation de son comportement, nous en déduirons trois conclusions possibles présentées ci-après :

- le son est effectivement une consonne si la différence entre l'énergie du segment et l'énergie de l'ensemble des voyelles de référence est supérieure à un certain seuil;

- le son est effectivement une voyelle si la différence entre l'énergie du segment et l'énergie de l'ensemble des consonnes de référence est supérieure à un certain seuil;
- la différence est soit trop faible soit trop grande dans les deux cas pour que l'on puisse effectuer un choix, on se déclare alors en catégorie indéterminée.

La classification d'un segment n'est donc pas une finalité, elle sert essentiellement à diriger le processus d'identification afin de réduire l'espace de recherche. Néanmoins, cette catégorisation étant effectuée sur des critères certes satisfaisants, mais relativement simples, à la manière de ce qui est réalisé dans les approches acoustico-phonétiques décrites au chapitre 3, il est important d'incorporer au processus d'identification des mesures de sécurité permettant d'élargir l'espace de recherche lorsque les conditions idéales de reconnaissance ne semblent pas rencontrées.

4.5.2.3 Processus d'identification finale

La finalité du processus d'identification est de retenir un ou plusieurs candidats choisis parmi les références appartenant à la catégorie du segment inconnu, ces candidats étant en fait les références ayant fourni les mesures de dissemblance les plus basses, selon les formules présentées à la partie 3.4.1. Ici, nous retiendrons les quatre références les plus proches, ceci dans le but de s'assurer une certaine marge de correction, et nous conserverons également les valeurs des distances associées à ces quatre « étiquettes ». Ces distances seront en effet des indices de « sécurité » sur lesquelles nous nous baserons pour éventuellement revenir sur le choix de la catégorie puisque de leur comportement, nous pouvons déduire plusieurs hypothèses présentées ci-dessous :

- une distance très faible augure de façon quasi certaine une reconnaissance réussie;
- une distance moyenne laisse planer le doute sur la réussite de la reconnaissance, et certaines mesures de correction devraient être mises en œuvre afin de s'assurer de l'exactitude du résultat;

- une distance trop élevée ne laisse que très peu de doutes quant à l'échec de la reconnaissance, échec probablement du à une erreur dans la sélection de la catégorie.

Les procédures de correction à entamer dans les second et troisième cas dépendent alors de la catégorie du segment :

- dans le cas d'un son non-voisé, étant donné que le critère de sélection fournissait une réussite de l'ordre de 94%, la reconnaissance est généralement réussie, et les distances relativement faible, compte tenu du fait que cette catégorie comporte peu d'éléments, tous spectralement assez proches les uns des autres. Une distance trop élevée entraîne donc immédiatement le passage du segment dans la catégorie des sons voisés;
- dans le cas des consonnes, le constat est le même que pour les sons non-voisés : cette catégorie comporte peu d'éléments, tous très proches puisque de sonorité \n\, et leur faible amplitude permet de les repérer avec une très bonne précision. Néanmoins, l'utilisation de l'énergie comme critère de classification est loin de s'avérer aussi efficace que les second et troisième coefficients cepstraux, et c'est pour cela que l'on préférera se rediriger vers la catégorie indéterminée en cas de distance moyennement élevée;
- dans le cas des voyelles, la classification selon l'énergie s'avère plus efficace, et une distance trop élevée sera plutôt synonyme de mauvaise reconnaissance à l'intérieur même de la catégorie, étant donné la grande diversité des sons qui la composent. On privilégiera alors une procédure de correction se basant sur un nouveau calcul des distances ne tenant compte que des coefficients cepstraux d'ordre élevé (afin de prendre en compte les caractéristiques plus fines du signal);
- dans le cas où la catégorie est indéterminée, la procédure de correction fera surtout appel à une plus longue succession de tests basés sur différents seuils de distances, permettant de diriger le choix vers la catégorie la plus probable.

L'objectif de tout ce processus est donc de s'assurer le plus possible de la réussite de l'identification du segment, et cela de façon intelligente : lorsque toutes les conditions de reconnaissance semblent correctes (catégorie repérée, distance faible,...), le résultat est sélectionné très rapidement, tandis que plus le choix est indécis, et plus les mécanismes de vérification et de correction sont enclenchés. Si une telle procédure est inutile dans le cas d'une reconnaissance monolocuteur, elle est en revanche fort utile en présence de locuteurs différents, où la diversité des sons prononcés est bien plus élevée. La sortie de ce module de notre système de reconnaissance est schématisée sur la figure 4.22, qui présente bien l'intérêt de sélectionner plusieurs candidats pour chaque segment, puisque pour le deuxième segment notamment, la référence qui lui est théoriquement associée n'est placée qu'en troisième position.

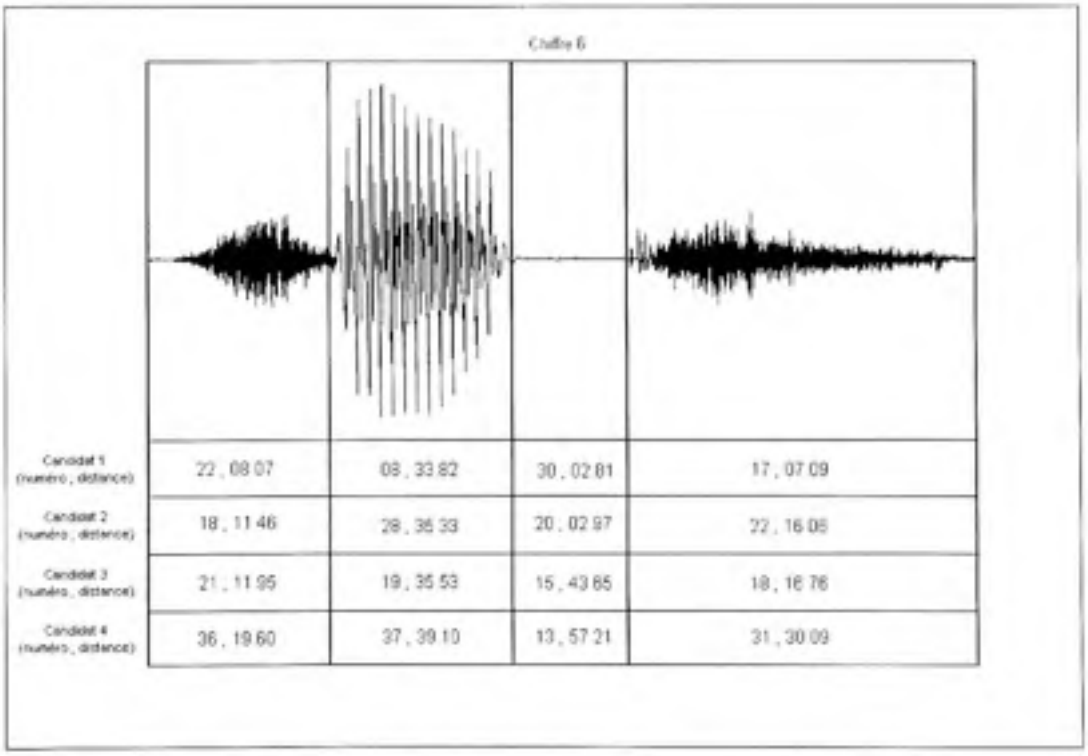

Figure 4.22 ***Sortie du module d'identifications de segments.***

4.6 Reconnaissance des chiffres

4.6.1 Présentation

La dernière phase de notre système de reconnaissance vocale consiste à reconstruire les chiffres du vocabulaire à partir du treillis d'unités acoustiques obtenu grâce aux opérations décrites jusqu'ici. Le chapitre 2 nous avait permis de présenter diverses méthodes remplissant ce rôle : les modèles de Markov cachés, notamment, qui utilisent un modèle constitué d'états pour reconstruire les mots au moyen de fonctions probabilistes; ou même la DTW, dont le déroulement s'apparente quelque peu à une reconstitution continue (*i.e.* trame par trame) du chiffre suivant les formes des références, une telle reconstitution pouvant éventuellement être réalisée de façon moins continue, à partir d'unités acoustiques plus larges. Ces deux méthodes ont pour objectif de repérer la forme ou le modèle se rapprochant le plus du mot à reconnaître, mais les algorithmes pour y arriver sont très lourds, en raison de la nécessité de déterminer le « chemin » le plus court parmi un très grand nombre. Ici, nous nous appuierons sur certaines caractéristiques intéressantes de ces deux méthodes afin de les associer dans un algorithme plus adapté à nos besoins.

Cet algorithme prendra la forme d'une reconstruction des chiffres selon un modèle acoustique prédéfini, en utilisant, pour chaque segment, l'ensemble des informations obtenues à l'étage précédent, associées à un modèle de connaissance spécifique du vocabulaire concerné. En partant d'un état initial, l'objectif sera alors de calculer, au fur et à mesure que les segments se succèdent, une mesure de ressemblance avec les onze chiffres de référence, jusqu'à sélectionner celui dont le choix semble le plus cohérent. Les outils nécessaires à ce procédé, ainsi que son déroulement complet, seront présentés au cours des parties suivantes.

4.6.2 Le tableau de construction

Alors que, pour reconnaître un chiffre de taille K_I, les HMM à N états et les algorithmes de programmation dynamique travaillaient sur des espaces de recherche respectifs de K_I*N et

K_I*K_R « cases », K_R étant la taille de la référence à comparer, et ce pour chacun des chiffres de référence, notre méthode travaillera sur un espace de dimension 11*6, que l'on peut modéliser par un tableau, présenté sur la figure 4.23, dont les lignes représentent les 11 chiffres du vocabulaire, et les colonnes se définissent de la façon suivante :

- les quatre premières colonnes correspondent aux quatre états du modèle présenté précédemment par la figure 4.18, et contiendront chacune un poids de construction, que nous définirons par la suite;
- la cinquième colonne contient la somme des quatre premières;
- la sixième colonne indique le nombre d'états activés (*i.e.* dont le poids de construction n'est pas nul) sur la ligne.

	Début	Milieu	Intérieur	Fin	Poids	États actifs
One						
Two						
Three						
Four						
Five						
Six						
Seven						
Eight						
Nine						
Zero						
Oh						

Figure 4.23 ***Représentation du tableau de construction.***

Pour chaque segment de parole, ce tableau sera amené à être rempli selon un processus présenté sur la figure 4.24, et décrit plus en détail par la suite.

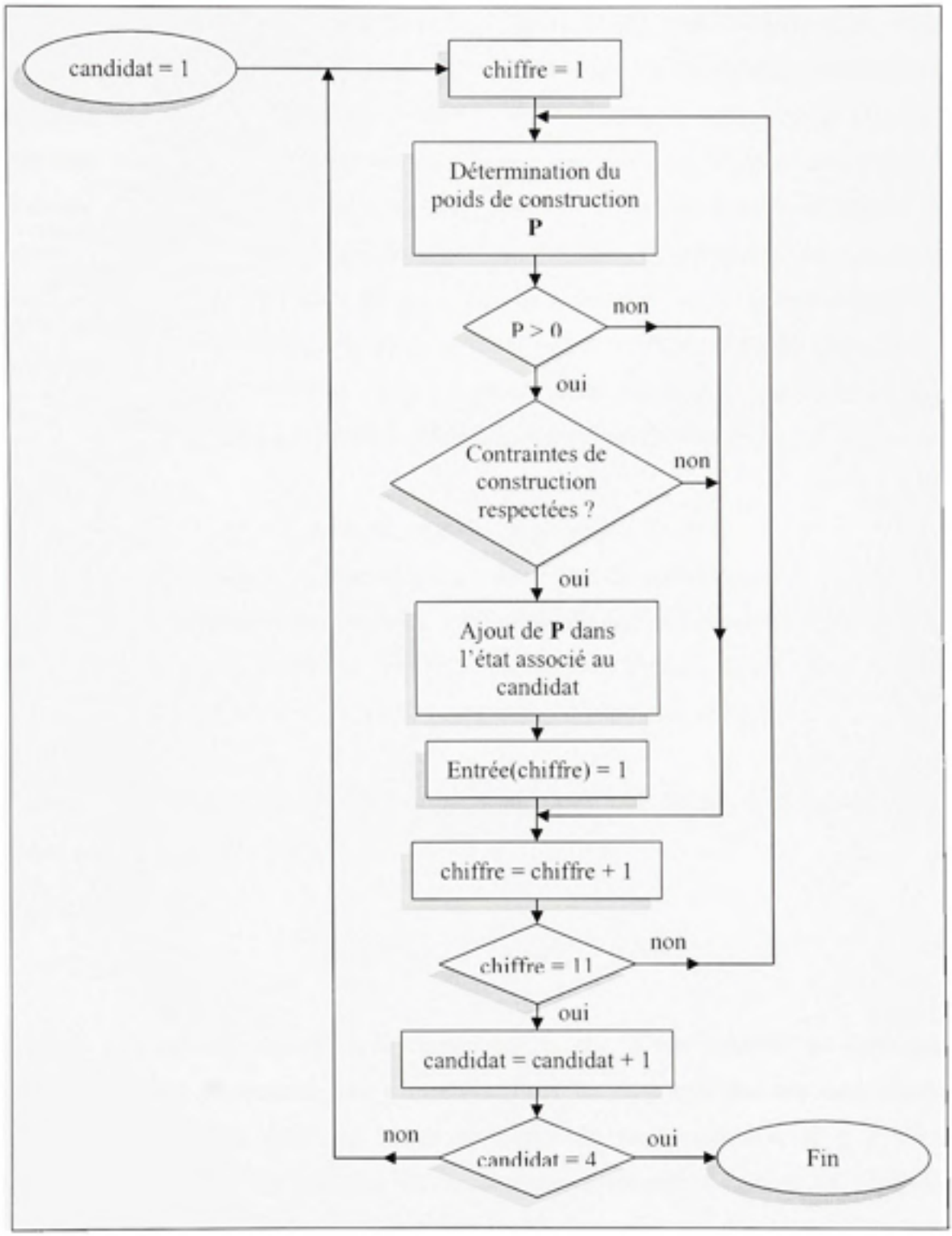

Figure 4.24 ***Schématisation du processus de remplissage du tableau de construction.***

Le poids de construction, tel que mentionné ci-dessus, est une valeur s'apparentant quelque peu aux probabilités de distribution des modèles de Markov cachés (plus la probabilité que le segment soit associé à un état est grande, et plus la donnée se rapproche de 1), excepté qu'ici cette valeur statistique prend la forme d'une mesure de dissemblance inverse entre un état donné et le segment acoustique considéré, c'est-à-dire que plus le poids est élevé et plus le segment est susceptible d'appartenir à l'état en question. Au début de la reconnaissance, toutes les cases du tableau sont donc initialisées à zéro, et seront ensuite remplies en fonction de l'appartenance des segments aux chiffres (lignes) et aux états (colonnes), au moyen du poids de construction. Chaque segment étant constitué de quatre candidats, le poids à insérer dans une case d'état est déterminé en fonction de trois critères :

- la position du candidat, indiquée par une valeur allant de 1 à 4;
- la distance du candidat, représentée par une valeur comprise entre 0 et 4, suivant que cette mesure se situe dans certains intervalles de distance fixés empiriquement;
- le poids d'association du candidat au chiffre (ligne), fourni par le tableau d'association présenté en **annexe 2**, indiqué par une valeur comprise entre 0 et 3.

Le poids d'un candidat est alors donné par la multiplication de ces trois valeurs, comme indiqué par la formule 4.4.

$$P_{candidat} = P_{position} \times P_{distance} \times P_{association} \tag{4.4}$$

Ce poids sera ensuite additionné au contenu de la case d'état associée au candidat, si toutefois certaines contraintes sont respectées. En effet, à la manière des contraintes de cheminement imposées dans les autres méthodes de reconnaissance, il y a ici des contraintes, basées sur la structure du vocabulaire à reconstruire, dont la vérification détermine si le poids d'un candidat peut-être inséré dans un état selon le statut des autres états du chiffre considéré. On peut définir ces contraintes de la façon suivante :

a) on ne peut pas insérer de début si le milieu est déjà activé;

b) on ne peut pas insérer de milieu si l'intérieur et/ou la fin sont déjà activés;
c) on ne peut pas insérer d'intérieur si le milieu n'est pas activé;
d) on ne peut pas insérer de fin si le milieu n'est pas activé;
e) les chiffres 2, 4, 6, 7, 9 et zero doivent impérativement avoir un début activé pour pouvoir insérer un milieu;
f) on ne peut pas insérer de début si le début est déjà activé;
g) on ne peut pas insérer de fin si la fin est déjà activée.

Ces deux dernières contraintes ont été fixées afin de faciliter la reconnaissance de chiffres connectés, pour permettre la réinitialisation du tableau lorsque l'on repère que la construction d'un chiffre est terminée et qu'il faut alors passer au suivant. Si tous ces critères sont respectés, la case est donc mise à jour en y ajoutant le poids déterminé auparavant, et ces opérations sont répétées pour chacune des 11 lignes du tableau, en indiquant à chaque fois au système quelles sont les lignes ayant connu une entrée sur ce segment.

À la fin du remplissage du tableau par l'ensemble des candidats d'un segment, il reste une opération importante à effectuer afin de compléter cette phase, concernant la prise en compte de certaines catégories acoustiques : les sons non voisés et les voyelles (autres que sons \o\). Ces deux catégories sont en effet constituées de nombreux éléments, tous relativement proches les uns des autres, spectralement parlant. Le fait qu'ils apparaissent donc assez souvent dans le processus de reconnaissance, et que leur étiquetage précis soit parfois plus difficile à obtenir que pour les autres catégories engendre la nécessité de fournir au système la preuve qu'ils étaient présents, si jamais une erreur d'étiquetage n'a pas permis l'activation des cases adéquates. Il s'agit donc ici de compenser cette éventuelle erreur en activant les états de poids nul correspondant à la catégorie détectée sur au moins un des candidats du segment grâce au tableau d'association. L'objectif est alors de pouvoir continuer la reconstruction avec les segments suivants, en prenant en compte le fait que certains états auraient peut-être du être activés, sans toutefois en être certain. Cette

incertitude est notifiée au moyen de poids spéciaux, dépendamment des catégories considérées :

- pour un segment de catégorie non-voisée, un poids de 0,5 est choisi;
- pour un segment de catégorie voyelle, un poids de 0,25 est choisi.

À la fin de l'analyse de chaque segment, le tableau de reconstruction indique donc l'état d'avancement de la reconstruction de tous les chiffres potentiellement candidats à la reconnaissance. Un exemple est fourni sur la figure 4.25, qui présente le tableau de reconstruction après l'analyse du premier segment du chiffre 6 de la figure 4.22.

	Début	Milieu	Intérieur	Fin	Poids	États actifs
One						
Two	0.5				0.5	1
Three	0.5				0.5	1
Four	0.5				0.5	1
Five	0.5				0.5	1
Six	60				60	1
Seven	60				60	1
Eight						
Nine						
Zero	9				9	1
Oh						

Figure 4.25 ***Exemple de tableau de reconstruction rempli pour un segment.***

Sachant comment remplir le tableau pour un segment, il nous reste alors à répéter successivement ces actions sur toute la séquence de segments jusqu'à repérer que la reconstruction du chiffre en cours de reconnaissance semble complétée, auquel cas on se

basera sur les deux dernières colonnes de chaque ligne (poids total et nombre d'états activés) pour déterminer quel est le chiffre le plus susceptible de coller au chiffre inconnu. Un élément important de ce processus consiste donc à détecter les transitions entre les chiffres afin de permettre la réinitialisation du tableau de construction en vue de poursuivre la reconnaissance après la validation d'un chiffre. L'algorithme utilisé pour réaliser ces tâches sera donc présenté ci-après.

4.6.3 Déroulement de l'algorithme complet de reconnaissance

L'algorithme complet utilisera simultanément deux tableaux de construction, un troisième étant également tenu à disposition pour être utilisé dans certains cas de figure. Le premier tableau sert en fait à effectuer la reconstruction du chiffre en cours de reconnaissance. Ce qui signifie qu'une fois le premier segment passé, son remplissage s'effectue de la même manière que présentée précédemment, mais uniquement sur les lignes déjà activées, l'objectif étant de pouvoir continuer la reconstruction en tenant compte des états activés au cours des segments précédents. Le second tableau, quant à lui, est utilisé pour permettre le démarrage d'une nouvelle construction, dans le cas des chiffres connectés, lorsque l'on détecte, grâce au tableau 1, que la construction en cours est complétée. Remis à zéro à chaque nouveau segment, ce tableau 2 est alors rempli de façon régulière sur l'ensemble de ses lignes, comme si l'on était au tout début d'un chiffre. Cette méthode permet donc aussi bien la reconnaissance de chiffres isolés que connectés, sans nécessité de connaître à l'avance le nombre de chiffres à reconnaître. Mais le processus complet ne consiste pas seulement à remplir des tableaux, il y a un nombre important d'opérations à effectuer pour chaque segment, afin notamment de repérer les fins de chiffres, d'assurer les transitions entre les chiffres successifs, et de valider des constructions complètes.

L'algorithme entier est donc schématisé sur la figure 4.26, et des explications plus précises concernant le déroulement des parties importantes de cet algorithme seront données par la suite.

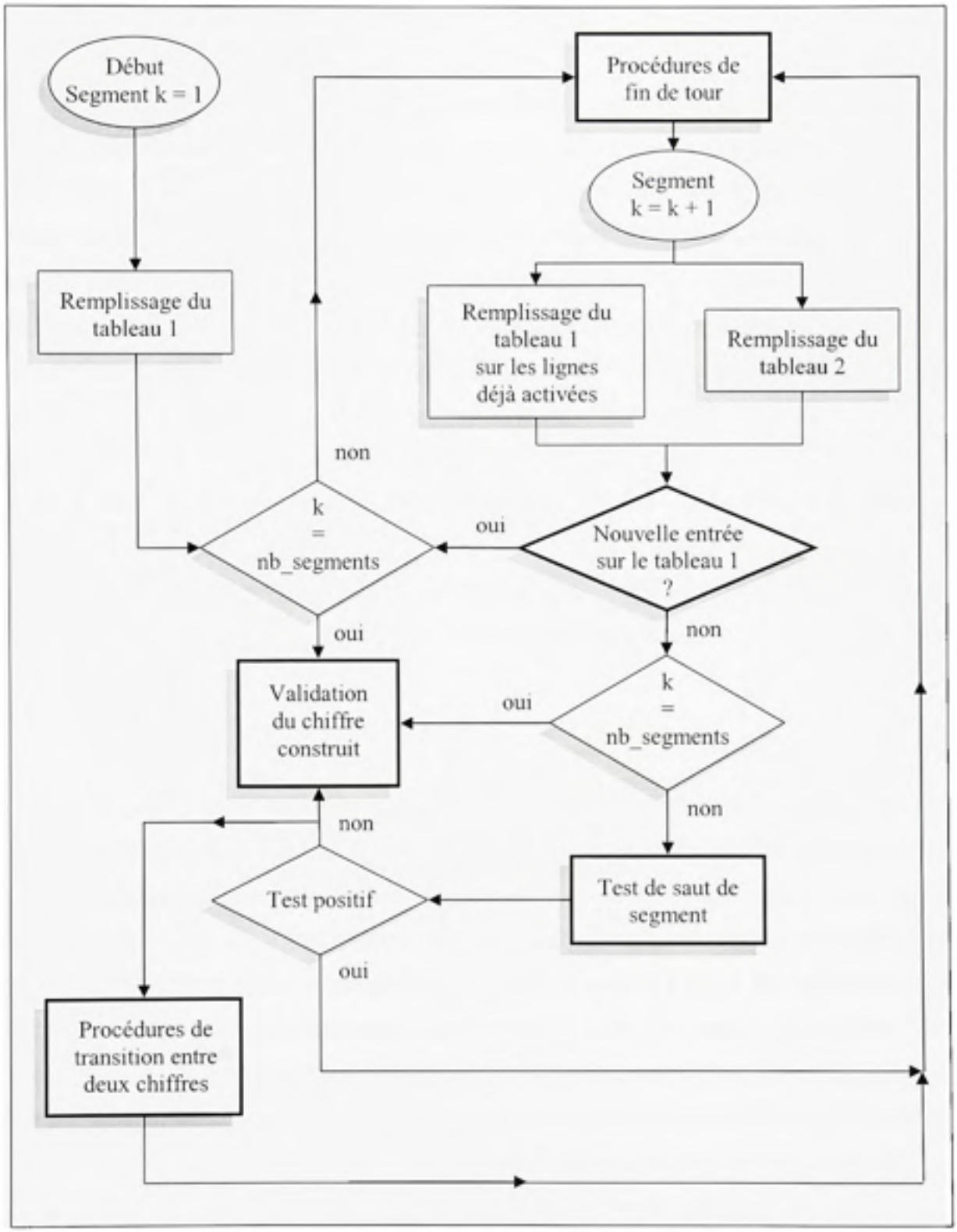

Figure 4.26 ***Schématisation de l'algorithme de reconnaissance de chiffres.***

4.6.3.1 Vérifications post-remplissage

Suite au remplissage du tableau de construction, certaines vérifications sont à effectuer afin d'orienter la suite du processus, dépendamment de la position du segment venant d'être traité. Pour le premier segment, il faut s'assurer que ce n'est pas également le dernier, auquel cas on lance immédiatement la procédure de validation pour stopper le processus, au lieu de passer au tour suivant. Pour les segments ultérieurs, les actions post-remplissage dépendent principalement des entrées qu'a connues le tableau 1. Si le segment en cours a effectivement fait avancer la construction, on peut alors passer au tour suivant, si tour suivant il y a. En revanche, si le tableau de construction n'a connu aucune évolution par rapport à son état précédent, cela peut signifier que la construction en cours est complétée, et qu'il faut alors passer à la reconnaissance du chiffre suivant. Mais cette étape de transition s'accompagne d'une vérification importante, dont l'objectif est de s'assurer que la stagnation de la construction en cours est bien due à l'apparition d'un nouveau chiffre, et non à un segment mal identifié. Cette procédure est décrite ci-après.

4.6.3.2 Test de saut de segment

L'objectif est de simuler la continuation de la reconstruction comme si le segment en cours n'avait jamais existé. C'est donc ici que l'on fait appel au troisième tableau de construction, dans lequel on copie le contenu du premier tableau, et ce afin que la simulation n'en modifie pas la teneur au cas où ce test ne serait pas concluant. On effectue donc une extrapolation sur les évènements à venir, en remplissant ce nouveau tableau à partir des informations du segment suivant, mais cela uniquement sur la ou les lignes de poids le plus élevé. Il est effectivement important que cette vérification n'ait valeur que d'éventuelle correction de dernier recours, et qu'elle n'entraîne pas la continuation d'une construction qui n'aurait pas lieu d'être. Le résultat retenu est uniquement la présence ou non d'une entrée dans ce troisième tableau. Si c'est positif, on jugera alors qu'il est judicieux de continuer la construction en cours sur le tableau 1. Sinon, on peut alors lancer le processus de validation du chiffre, et passer à la construction du chiffre suivant.

Cette procédure n'est toutefois enclenchée que si certaines conditions sont rencontrées :

- le segment sauté ne doit pas dépasser une certaine taille, auquel cas on jugerait qu'il est inconscient de négliger entièrement une partie du signal pouvant avoir une certaine importance dans la signification du chiffre;
- le segment sauté ne doit pas être un silence, puisqu'un tel son ne peut apparaître que lors des transitions entre deux chiffres, ou bien en présence des chiffres 6 et 8, mais pour lesquels il aurait alors engendré une entrée dans le tableau de construction;
- le segment sauté ne doit pas appartenir à la catégorie des non-voisés, étant donné que ces sons n'apparaissent qu'en début ou fin de chiffre. Le fait qu'ils n'engendrent aucune entrée dans le tableau signifie donc qu'ils n'ont pas répondu aux contraintes de construction mentionnées précédemment et qu'ils représentent plutôt le commencement d'un nouveau chiffre.

Autrement dit, ce test n'est effectué qu'en présence de voyelles de petites tailles, catégorie la plus représentée, et la plus propice aux sursegmentations et autres erreurs d'identification.

4.6.3.3 Procédures de fin de tour

Lorsque la construction est amenée à continuer, certaines opérations doivent être effectuées avant de passer au segment suivant, afin d'assurer une bonne poursuite de la reconstruction en tenant compte des évènements produits au cours de ce tour. On cherchera notamment à :

- apporter un soutien aux chiffres dont la construction semble avancer;
- vérifier que la reconstruction ne s'étend pas au-delà de durées raisonnables.

Concernant le premier point, on souhaite donc favoriser les lignes du tableau qui évoluent, par rapport à celles qui stagnent. Plus concrètement, en fonction de la détection ou non d'une entrée au cours d'un remplissage, on va récompenser/sanctionner chaque ligne activée du tableau, de la façon suivante :

- si la ligne connaît une entrée au cours du traitement d'un segment, son poids est majoré d'un tiers de sa valeur;
- si la ligne ne connaît aucune entrée au cours du traitement d'un segment, son poids est minoré d'un tiers de sa valeur, et son indice d'inactivité est incrémenté. Précisons également que lorsque qu'une ligne est inactive durant deux segments consécutifs, elle est automatiquement désactivée, et son poids est remis à zéro.

Le but de ce procédé est d'assurer une véritable évolution de la construction au fur et à mesure que les segments se succèdent, afin que le tableau de construction semble progresser en même temps que l'on avance dans la chaîne de segments, sans que les actions isolées réalisées plusieurs tours auparavant n'aient trop d'incidence dans le résultat.

Concernant la durée de la reconstruction, il faut garder à l'esprit que la durée moyenne des mots du vocabulaire que nous souhaitons reconnaître n'excède que très rarement la demie seconde. Sachant cela, il est important de mettre régulièrement à jour la taille du chiffre en construction, en fonction de la taille des segments rencontrés, afin de pouvoir mettre un terme à cette construction lorsque la durée du chiffre dépasse un seuil limite. De la même manière, on fixera une taille limite aux sons \o\ : présents dans quatre mots du vocabulaire (1, 4, 0 et zero), ces sons peuvent aussi bien débuter les chiffres que les conclure, et donc nuire à la reconnaissance lorsque les chiffres en question sont connectés. Le dépassement de cette taille limite entraînera alors le commencement de la construction d'un nouveau chiffre.

Ces procédures, tout comme le test de saut de segment, sont rendues nécessaires par le caractère complexe et inconstant du signal vocal, qui force régulièrement le système à mettre en œuvre des mesures de correction pour obtenir une bonne qualité de reconnaissance. Dans des conditions de segmentation et d'identification idéales, de telles actions ne seraient effectivement pas nécessaires, et la reconstitution des chiffres pourrait se dérouler avec un simple remplissage des tableaux de construction.

4.6.3.4 Procédures de transition entre deux chiffres

Lorsqu'un chiffre est complété, il convient de lancer, si besoin est, la construction du chiffre suivant. Comme nous l'avons mentionné précédemment, dans la majeure partie des cas les informations nécessaires à ce nouveau départ se trouvent dans le tableau 2, mais certaines situations requièrent un traitement particulier. Les actions à réaliser dépendront donc des conditions rencontrées :

- lorsque le segment en cours est un silence, on initialise toutes les cases du tableau 1 à zéro;
- lorsque le tableau 1 s'avère entièrement vide, ce qui arrive quand le segment précédent était un silence, un simple transfert du contenu du tableau 2 vers le tableau 1 permet de débuter la construction d'un nouveau chiffre;
- lorsque l'on est en présence d'un son non-voisé, on copie également le contenu du tableau 2 dans le tableau 1;
- lorsqu'aucune des conditions mentionnées ci-dessus n'est rencontrée, la procédure est plus complexe, le nouveau contenu du tableau 1 dépendant alors du résultat d'un test de superposition destiné à déterminer si le dernier segment du chiffre venant de se conclure se confond avec le premier segment du chiffre à venir.

Ce test de superposition est rendu nécessaire en raison de la présence, dans le vocabulaire, de sons pouvant être situés aussi bien en début qu'en fin de chiffre. C'est notamment le cas du son \n\, présent à la fois en fin des chiffres 1, 7 et 9, et en début de chiffre 9, mais également des sons \o\ et \s\, voire \t\, dont l'appartenance à cette catégorie de superposition est notifiée dans le tableau des associations. Cette particularité ne permet donc pas une détection idéale des frontières lorsque les chiffres concernés sont prononcés de façon continue les uns à la suite des autres, comme le montre la figure 4.27. Sur cette figure, présentant les prononciations successives des chiffres 1 et 9, le troisième segment, qui représente à la fois le son \n\ final du one et le son \n\ initial du nine, est découpé en un seul

bloc, sans distinction entre la partie appartenant au premier chiffre, et celle appartenant à son successeur.

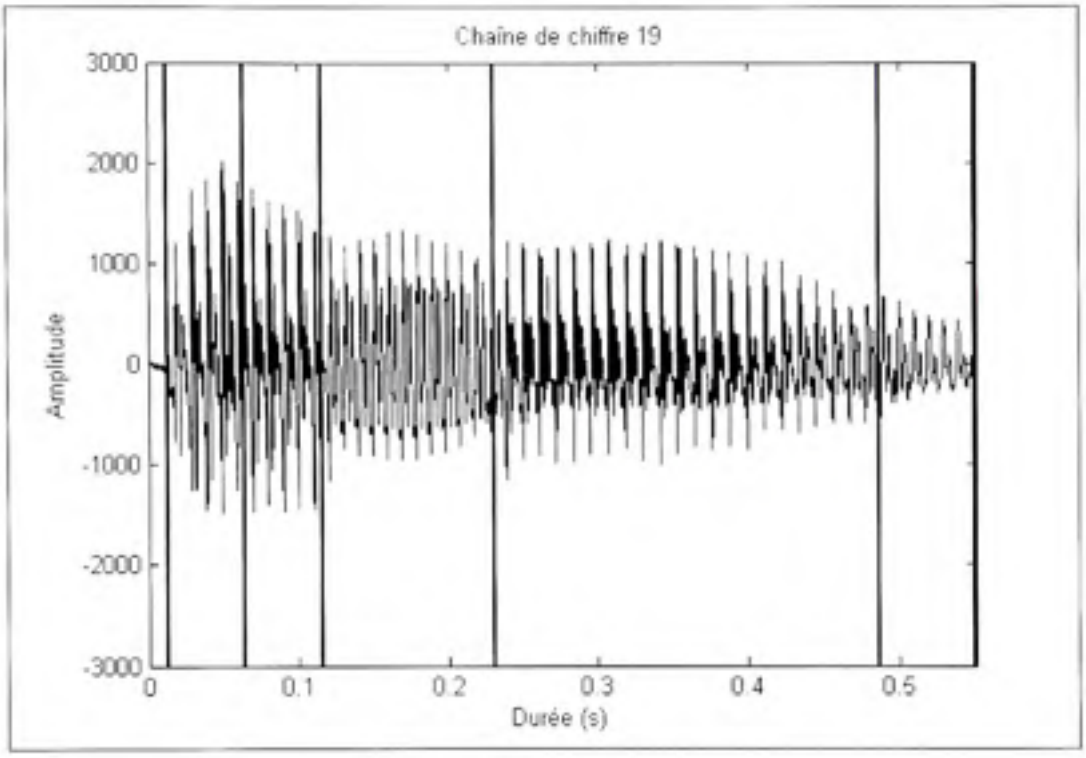

Figure 4.27 ***Exemple de chiffres consécutifs superposés.***

Lorsque le segment précédent a donc été repéré comme appartenant à une catégorie de superposition, la procédure de vérification est lancée, faisant une nouvelle fois appel au troisième tableau de construction. Ce test se déroule alors en deux étapes :

- le tableau 3 est tout d'abord rempli avec les quatre candidats du segment précédent;
- on effectue ensuite second un remplissage uniquement sur les lignes activées, avec les informations du segment en cours, en réduisant toutefois les valeurs des seuils de distances tolérées, afin que ce test ne soit pris en compte que lorsque les conditions de reconnaissance sont jugées satisfaisantes.

Si une entrée est repérée à l'issue de la seconde étape, on estimera alors qu'il y a effectivement une superposition des deux chiffres consécutifs, et on place alors le contenu du tableau 3 dans le tableau 1. Si le test n'est en revanche pas concluant, ou que le segment précédent n'appartenait pas à une catégorie de superposition, on copie simplement le contenu du tableau 2 dans le tableau 1.

Toutes ces opérations permettent donc d'assurer la continuité du processus de reconnaissance au cours des différents chiffres, mais la réussite de cette continuité dépend principalement de la capacité du système à repérer avec exactitude la fin de construction d'un chiffre, capacité d'autant plus élevée que l'étiquetage des segments est correct. Concernant les chiffres dont la reconstitution est complétée, il reste d'ailleurs à les identifier en analysant le contenu final du tableau de reconstruction.

4.6.3.5 Validation de la construction

Cette procédure d'identification sert en fait à valider toute la construction effectuée auparavant, puisqu'il ne s'agit pas simplement de sélectionner le chiffre ayant le poids le plus élevé, mais également de s'assurer que ce chiffre répond à certains critères de modélisation, comme :

- une taille minimale;
- un nombre minimal d'états activés, selon les chiffres considérés;
- l'activation obligatoire de certains états clés, selon les chiffres considérés.

Concrètement, une fois l'assurance acquise que la taille du chiffre construit est suffisamment grande pour autoriser la validation, la ligne de poids maximal est sélectionnée, et son contenu est vérifié pour garantir la cohérence du choix de validation, compte tenu de la structure des mots du vocabulaire :

- les chiffres 3, 8 et 0 peuvent ne posséder qu'un seul état activé;

- les chiffres 1, 2, 4, 5 et zero doivent posséder au moins deux états actifs;
- les chiffres 6 et 7 doivent impérativement présenter un intérieur actif lorsque leur milieu n'est activé que par un poids de 0,25;
- le chiffre 9 doit posséder une fin active, si le poids de son milieu est égal à 0,25.

Ces exigences assurent une certaine logique dans la validation des chiffres, puisque si elles ne sont pas respectées par la ligne de poids maximal, c'est alors la ligne venant en second qui est sélectionnée pour procéder à la validation. Il serait effectivement illogique de valider, par exemple, un chiffre 7 ne présentant qu'un début, quand le chiffre 5, de poids moins fort, semble bien plus complet. Ces procédures de correction sont donc présentes pour rattraper certaines erreurs, mais dans la majorité des cas, c'est bien la ligne de poids maximal qui sera retenue, les erreurs provenant alors plutôt d'étiquetages incorrects. Pour conclure cette partie, et mieux visualiser la tâche de validation, la figure 4.28 présente le contenu final du tableau de construction pour le chiffre 6 présenté sur la figure 4.22.

	Début	Milieu	Intérieur	Fin	Poids	États actifs
One						
Two	0,5	32			32,5	2
Three						
Fou						
Five						
Six	60	24	60	60	204	4
Seven	60	0,5			60,5	2
Eight						
Nine						
Zero	12	28			40	2
Oh						

Figure 4.28 ***Exemple de construction complétée pour le chiffre 6.***

4.7 Conclusion

De la segmentation en unités phoniques, au processus de reconnaissance basé sur la modélisation acoustique du vocabulaire, on retrouve dans cette méthode plusieurs aspects mentionnés au cours du chapitre précédent, et utilisés de façon différente dans de nombreux systèmes. Ici, nous nous sommes principalement focalisés sur le caractère restreint du vocabulaire étudié, et la possibilité d'en classer les sons selon plusieurs catégories afin de réduire l'espace de recherche et de faciliter l'identification des unités acoustiques inconnues. Découlant de cette analyse lexicale, une technique se présentait alors, consistant à simuler la reconstruction des chiffres, à la manière des méthodes de modélisation probabiliste, mais selon une approche plus adaptée à nos besoins, tirant parti d'une bonne connaissance du vocabulaire pour permettre la reconnaissance d'une façon peu coûteuse en calculs et en mémoire.

Cette dernière caractéristique est d'ailleurs l'un des atouts majeurs de la méthode développée pour ce mémoire, et ce grâce à aux particularités apportées à chacune des étapes importantes du processus. Au-delà du choix de représenter l'information vocale au moyen des paramètres robustes que sont les coefficients cepstraux, c'est donc surtout leur utilisation intelligente pour les tâches de détection d'activité vocale et de segmentation qui rend leur présence intéressante dans l'optique d'optimiser au maximum le processus pour une application sur un système aux ressources limitées. La structure de la tâche de reconnaissance présente quant à elle un très grand intérêt, de par son espace de recherche réduit et l'absence de longs calculs destinés à trouver le chemin optimal. Le second avantage introduit par cette méthode concerne la taille du vocabulaire de référence, puisque ce dernier n'est constitué que de 40 unités différentes, chacune représentée par un simple vecteur spectral. Cette volonté d'optimisation ne doit néanmoins pas nuire à la qualité de la reconnaissance, et le chapitre 5 nous permettra de présenter les résultats des divers tests effectués pour s'assurer de l'efficacité de notre méthode de reconnaissance vocale.

CHAPITRE 5

PRÉSENTATION DES RÉSULTATS

5.1 Introduction

La méthode présentée au chapitre précédent semble donc théoriquement très séduisante, mais il faut maintenant évaluer ses performances pratiques, afin de pouvoir tirer des conclusions quant à son efficacité. Ce chapitre aura donc pour but de présenter les résultats des tests de reconnaissance des dix chiffres constituant notre vocabulaire, effectués sur un grand ensemble de locuteurs différents. Les paramètres utilisés pour représenter le signal sont ceux décrits au début du chapitre 4 : un jeu de 20 coefficients obtenus à partir de trames de 13ms superposées à 50%. Nous présenterons donc tout d'abord la base de données utilisée pour constituer notre dictionnaire de référence ainsi que l'ensemble des échantillons de test. Viendra ensuite la présentation des résultats axée sur trois éléments majeurs :

- le gain de place mémoire obtenu grâce à la création d'un dictionnaire spécialement adapté aux caractéristiques du vocabulaire étudié;
- la rapidité du processus de reconnaissance;
- la réussite de la reconnaissance en milieux mono et multi-locuteurs, pour des chiffres tant isolés que connectés.

Concernant ces deux derniers points, il sera intéressant d'observer l'apport de certains aspects du processus d'identification de segments. Cette phase est en effet une des clés de la reconnaissance, puisque tout le processus de reconstruction est basé sur les résultats de l'étiquetage, et des erreurs à ce niveau peuvent donc réduire considérablement l'efficacité de la méthode. C'est notamment pour cela que l'on a mis en œuvre tout un système d'identification des segments par catégorisation acoustique, censé réduire l'espace de comparaison et ainsi apporter une plus grande liberté de traitement, selon la catégorie dans laquelle se trouve le son. Ainsi, alors que l'on devrait initialement comparer l'unité

inconnue avec chacune des 40 références, la détection de voisement permet de diviser l'espace de comparaison en deux parties, avant que la sélection énergétique ne divise elle-même la partie des sons voisés. Le processus d'identification final vient ensuite compléter le tout, en appliquant tout un système de vérifications et corrections pour revenir en arrière en cas de doute. Il serait donc intéressant de mesurer l'évolution de la qualité de la reconnaissance au fur et à mesure que l'on développe la catégorisation, et la figure 5.1 présente le cheminement des tests que nous effectuerons, incorporant également l'apport de l'uniformisation acoustique des segments, décrite au paragraphe 5.1 du chapitre 4.

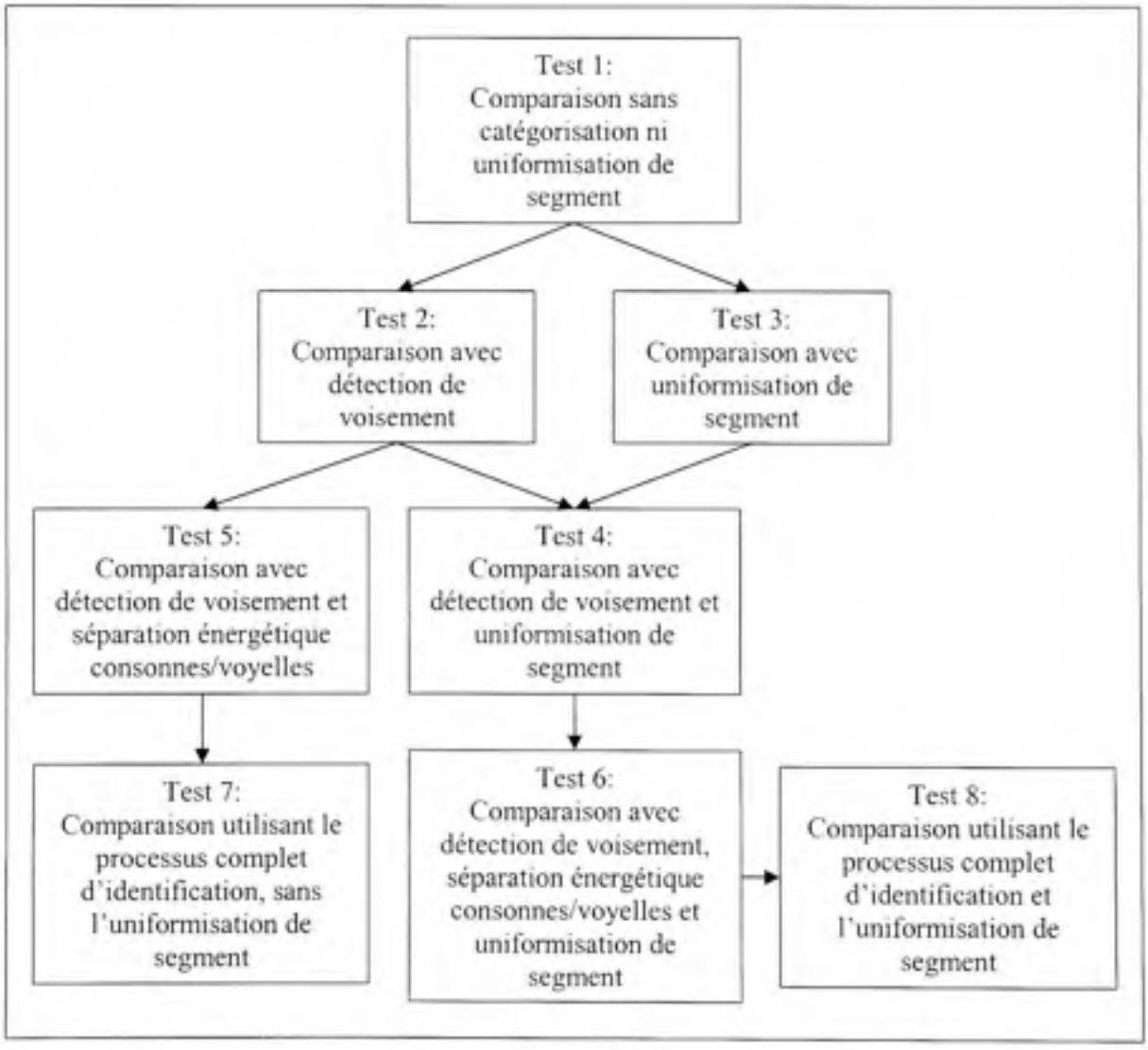

Figure 5.1 ***Répartition des tests effectués pour évaluer l'efficacité de la reconnaissance.***

5.2 La base de données TI-DIGITS

Les données utilisées pour constituer nos références et les occurrences de test proviennent du corpus TI-DIGITS. Cette base de données, qui regroupe plus de 25 000 séquences de chiffres prononcées par plus de 300 hommes, femmes et enfants différents dans un environnement calme, et échantillonnées à 20 000 Hz, a été construite au Texas en 1982, pour les besoins des laboratoires de Texas Instruments Incorporated (Leonard, 1984). Plus précisément, pour chacun des locuteurs, 77 séquences de chiffres sont mémorisées, comprenant aussi bien des chiffres isolés (mais entourés de larges zones de silence), que des chaînes de deux à sept chiffres connectés. Il est à préciser que les locuteurs ont été sélectionnés parmi 21 régions dialectiques des États-Unis, afin de fournir une bonne représentation de l'ensemble des prononciations possibles à travers les différents accents. L'objectif d'une telle base de données est de fournir suffisamment d'échantillons de parole pour pouvoir évaluer les performances des systèmes de reconnaissance vocale multilocuteurs.

Ici, notre intérêt sera de clairement séparer l'ensemble de référence et l'ensemble de test, afin de pouvoir opérer les tests de façon entièrement indépendante des locuteurs d'entraînement. Sur les 55 locuteurs masculins en notre possession, nous en utiliserons donc 25 pour constituer notre dictionnaire de référence, tandis que les tests de reconnaissance seront réalisés sur les chiffres prononcés par les 30 locuteurs restants. Chacun des 40 segments acoustiques du vocabulaire sera donc représenté par 25 versions différentes, à partir desquelles nous procèderons à une classification pour n'en conserver qu'un nombre réduit, tout en gardant un ensemble suffisamment représentatif de la grande variabilité des prononciations. Quant à l'ensemble de test, ces 30 locuteurs possédant chacun deux versions de chaque chiffre isolé, cela nous fournit un assortiment de 660 mots à identifier, ce qui représente une quantité suffisamment importante pour juger de l'efficacité de la méthode.

5.3 Taille du dictionnaire de références

Comme nous l'avons mentionné au chapitre 3, la taille du dictionnaire de référence joue un rôle très important dans bien des applications de reconnaissance de la parole. Plus celle-ci est grande, et plus la réussite est assurée, cette affirmation étant d'autant plus vraie pour les méthodes de reconnaissance de formes comme la programmation dynamique ou les modèles de Markov cachés. Un des objectifs principaux de la méthode développée au cours de ce mémoire était donc d'arriver à travailler à partir d'un nombre réduit de références, pour pouvoir être applicable sur des systèmes aux ressources limitées. Le choix d'un découpage des chiffres en segments acoustiques prédéterminés allait dans ce sens, puisque l'ensemble des chiffres du vocabulaire n'est alors représenté qu'au moyen de 40 vecteurs spectraux. Chaque segment étant représenté par plusieurs versions (nous choisirons un nombre de 16), toutes constituées d'un jeu de 21 coefficients cepstraux, nous pouvons détailler le contenu du dictionnaire de la façon suivante :

- 40 segments acoustiques représentant les 11 chiffres;
- 16 versions de chaque segment;
- 21 coefficients LPCC pour représenter un segment acoustique;
- 16 bits pour coder un coefficient.

Le dictionnaire ainsi constitué nécessitera donc un espace mémoire de 27 Ko. Pour bénéficier d'un point de comparaison, il peut être intéressant de présenter un tel calcul pour une méthode de reconnaissance par programmation dynamique. Sachant que dans ce cas là, chacun des chiffres doit être représenté par l'ensemble des vecteurs court-terme qui le composent et que, pour le vocabulaire présentement étudié, on dénombre une moyenne de 74 trames par chiffre, on peut donc détailler le contenu du dictionnaire de la façon suivante :

- 11 chiffres entiers;
- 74 trames, en moyenne, pour constituer un chiffre;
- 16 versions de chaque chiffre;

- 21 coefficients LPCC pour représenter une trame;
- 16 bits pour coder un coefficient.

Ce qui nous donne un dictionnaire dont la taille est égale à 547 Ko. Le tableau 5.1 nous permet de comparer les deux valeurs obtenues, et rend bien compte du gain de place réalisé grâce à la création d'un dictionnaire spécialement adapté à la structure du vocabulaire étudié. On peut également comparer cela avec les 47 Ko et 8737 Ko requis par Lévy pour obtenir des taux de réussite respectifs de 89% et 96,3%, en utilisant les modèles de Markov cachés (Lévy *et al*, 2004).

Tableau 5.1

Comparaison de la taille des dictionnaires de référence

	Méthode développée pour ce mémoire	**Méthode par programmation dynamique**
Taille du dictionnaire de référence	27 Ko	547 Ko

5.4 Rapidité de la reconnaissance

La taille réduite du dictionnaire de référence n'offre pas des avantages uniquement en ce qui concerne l'espace mémoire, elle est également très importante pour réduire le temps de calcul nécessaire à la reconnaissance des chiffres. Le processus d'identification est effectivement basé sur une comparaison avec l'ensemble des références, et moins il y en a, plus cette phase est rapide. C'est donc là tout l'intérêt de la classification des sons, qui permet de réduire d'avantage l'espace de comparaison. Le tableau 5.2 présente le temps de calcul, en secondes, nécessaire à la phase d'étiquetage des segments pour la reconnaissance de chacun des 11 chiffres du vocabulaire, effectuée au moyen des huit tests présentés sur la figure 5.1.

Tableau 5.2

Durée du processus d'étiquetage des segments (en secondes)

Chiffre	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8
One	0,0646	0,0413	0,0635	0,0443	0,0473	0,0494	0,0542	0,0563
Two	0,0438	0,0234	0,0442	0,0229	0,0228	0,0255	0,0281	0,0308
Three	0,0619	0,0361	0,061	0,0349	0,0323	0,0366	0,044	0,0483
Four	0,0579	0,0323	0,0603	0,0354	0,0322	0,0416	0,0353	0,0447
Five	0,0589	0,0290	0,0609	0,0319	0,0296	0,0369	0,0354	0,0427
Six	0,0604	0,0225	0,0609	0,0255	0,0207	0,0241	0,0299	0,0333
Seven	0,0599	0,0355	0,0615	0,037	0,035	0,037	0,0485	0,0505
Eight	0,0403	0,0198	0,0406	0,0216	0,0208	0,0249	0,0282	0,0323
Nine	0,0681	0,0423	0,0693	0,0465	0,0447	0,0453	0,0655	0,0661
Zero	0,068	0,0395	0,0692	0,0418	0,0395	0,0447	0,0499	0,0551
Oh	0,0391	0,0257	0,0401	0,0276	0,0249	0,0265	0,0359	0,0375
moyenne	**0,0566**	**0,0316**	**0,0574**	**0,0336**	**0,0318**	**0,0357**	**0,0413**	**0,0486**

L'analyse de ce tableau nous montre l'intérêt d'une séparation des sons voisés/non-voisés pour effectuer la reconnaissance, puisque le test 2 est 1,8 fois plus rapide que le test 1. L'uniformisation des segments, introduite dans les tests 3 et 4, ne rallonge quant à elle que très peu le temps de calcul, de même que l'ajout du procédé de sélection consonne/voyelle par différenciation énergétique, présent dans les tests 5 et 6. Le processus complet d'étiquetage, incluant les opérations de vérification et de correction, rallonge en revanche nettement plus le temps de traitement, mais ce dernier reste tout de même largement inférieur à celui obtenu au moyen du test de base. Pour juger de son intérêt, il restera alors à déterminer son influence sur la qualité de la reconnaissance.

Il est également important de mesurer le temps de traitement des autres modules de notre méthode de reconnaissance, à savoir :

- le module 1 constitué des étapes d'extraction des paramètres et de détection d'activité vocale;
- le module 2 constitué de la phase de segmentation;
- le module 4 constitué de la phase de reconstruction et de reconnaissance des chiffres.

Le tableau 5.3 présente donc ces mesures pour la reconnaissance de chacun des 11 chiffres du vocabulaire, incluant également la durée de traitement du module 3, qui correspond à l'étiquetage des segments effectué au moyen du test 8.

Tableau 5.3

Temps de traitement des différentes parties de la méthode de reconnaissance (en secondes)

Chiffre	Durée du chiffre	Durée du module 1	Durée du module 2	Durée du module 3	Durée du module 4	Durée totale de la reconnaissance
One	0,4399	0,2107	0,0281	0,0563	0,0155	**0,3129**
Two	0,3621	0,2037	0,0228	0,0308	0,0161	**0,2734**
Three	0,4489	0,2021	0,0274	0,0483	0,0162	**0,2940**
Four	0,4468	0,2213	0,0283	0,0447	0,0162	**0,3105**
Five	0,5536	0,2313	0,0332	0,0427	0,0156	**0,3228**
Six	0,6146	0,265	0,0338	0,0333	0,0157	**0,3478**
Seven	0,5351	0,2386	0,0312	0,0505	0,016	**0,3363**
Eight	0,3351	0,201	0,0214	0,0323	0,0156	**0,2703**
Nine	0,5255	0,2253	0,0335	0,0661	0,016	**0,3409**
Zero	0,5342	0,2323	0,031	0,0551	0,0172	**0,3356**
Oh	0,3555	0,201	0,0244	0,0375	0,0156	**0,2785**
moyenne	**0,4684**	**0,2211**	**0,0286**	**0,0452**	**0,0159**	**0,3108**

Ce tableau nous permet donc de constater que la durée de la phase de reconstruction est négligeable par rapport à la durée totale de traitement. C'est une grande réussite, étant donné que ce processus de recherche du chemin optimum était l'un des éléments les plus complexes des autres méthodes de reconnaissance. Mais plus globalement, nous pouvons noter la rapidité de l'ensemble du processus de reconnaissance par rapport à l'étape de traitement du signal (module 1), comme cela est clairement présenté dans le tableau 5.4.

Tableau 5.4

Répartition du temps de calcul

	Phase de traitement du signal	**Phase de segmentation**	**Phase de reconnaissance**
Pourcentage de la durée totale	71%	9,2%	19,8%

Si l'on y inclut la phase de segmentation, le processus de reconnaissance ne représente donc que 29% du temps total de traitement. Ces chiffres sont à comparer avec les résultats obtenus par Hui *et al* (1998), lesquels présentaient une reconnaissance par programmation dynamique occupant plus de 70% de la durée totale de traitement. Pour plus de précision, nous avons donc également effectué une reconnaissance par programmation dynamique des 11 chiffres de notre vocabulaire, dans des conditions semblables à celles utilisées pour les tests présentés plus haut : même ensemble de référence, même ensemble de test, même nombre de coefficient cesptraux…Le tableau 5.5 présente donc une comparaison entre les durées de traitement de la reconnaissance par notre méthode, et par la méthode de programmation dynamique.

Tableau 5.5

Comparaison des durées de reconnaissance par deux méthodes différentes (en secondes)

Chiffre	Méthode développée pour ce mémoire	Méthode de programmation dynamique
One	0,3129	3,9626
Two	0,2734	4,1098
Three	0,294	4,0574
Four	0,3105	4,1589
Five	0,3228	4,4630
Six	0,03478	4,6713
Seven	0,3363	4,3194
Eight	0,2703	3,6743
Nine	0,3409	4,3969
Zero	0,3356	4,4261
Oh	0,2785	3,7354
Moyenne	**0,3108**	**4,1796**

On peut ainsi observer que la méthode présentée au cours de ce mémoire affiche un temps de calcul 13 fois inférieur au temps requis pour effectuer la reconnaissance en utilisant un algorithme de programmation dynamique. Cette différence importante provient du fait que l'on a été capable d'adapter la méthode de reconnaissance à la structure du vocabulaire concerné, réduisant alors considérablement l'espace de comparaison ainsi que la complexité du processus de recherche de la référence la plus proche du mot inconnu. Précisons tout de même que ces tests ont été réalisés au moyen du logiciel Matlab, sur un ordinateur doté d'un processeur relativement récent. Une utilisation sur un processeur de traitement du signal moins puissant ne permettrait donc pas d'obtenir une telle rapidité de calcul, mais l'intérêt

ici est de pouvoir observer le gain de temps obtenu grâce aux spécificités de notre méthode par rapport à d'autres plus exigeantes. Il reste maintenant à déterminer si tous ces gains d'espace mémoire et de rapidité n'ont pas engendré trop de sacrifices au niveau de la qualité de la reconnaissance.

5.5 Réussite de la reconnaissance

L'efficacité d'une méthode de reconnaissance vocale se mesure principalement sur sa capacité à assurer un taux de réussite élevé. La méthode développée pour ce mémoire est conçue pour fonctionner aussi bien sur des chiffres isolés que connectés, mais nous nous focaliserons tout d'abord sur les résultats des reconnaissances de chiffres isolés, puisque c'est de là que découle majoritairement la réussite de la reconnaissance de chiffres connectés. En premier lieu, il est intéressant de s'assurer du succès de la reconnaissance monolocuteur, c'est-à-dire utilisant comme locuteurs de tests ceux employés pour constituer le dictionnaire de références. Le résultat global obtenu est de 94,55% de chiffres reconnus correctement, et bien que légèrement inférieur aux résultats obtenus par les méthodes plus complexe, il en reste malgré tout acceptable.

Les résultats de la reconnaissance multilocuteurs fournissent quant à eux plus d'indications quant à la robustesse de la méthode. Le succès y est effectivement plus difficile à obtenir, étant donné les grandes variations de prononciation entre chaque locuteur différent. Le processus d'identification de segments joue donc ici un rôle très important, et il est intéressant de présenter les résultats en fonction de l'évolution de sa complexité. Le tableau 5.6 regroupe donc l'ensemble des résultats de reconnaissance pour chacun des huit tests présentés sur la figure 5.1.

Tableau 5.6

Taux de réussite global de la reconnaissance indépendante du locuteur, pour des chiffres isolés

	Test 1	**Test 2**	**Test 3**	**Test 4**	**Test 5**	**Test 6**	**Test 7**	**Test 8**
Pourcentage de chiffres correctement reconnus	76,36	76,67	80	78,18	77,88	80,91	84,24	90,15

On peut ainsi observer l'apport des différents éléments du processus d'étiquetage de segments. En premier lieu, c'est l'uniformisation des segments qui semble présenter une grande efficacité par rapport à un simple calcul de moyenne spectrale de l'ensemble des trames d'un segment. Son association avec le processus complet d'étiquetage nous permet ensuite d'obtenir un résultat largement supérieur aux autres tests. En retenant le meilleur test, on peut alors effectuer une comparaison avec la reconnaissance par programmation dynamique, comme présenté par le tableau 5.7.

Tableau 5.7

Comparaison des taux de réussite de la reconnaissance par deux méthodes différentes (en secondes)

	Méthode développée pour ce mémoire	**Méthode de programmation dynamique**
Pourcentage de chiffres correctement reconnus	90,15	93,94

La différence de réussite est certes à l'avantage de la méthode de reconnaissance par programmation dynamique, mais l'écart est suffisamment faible pour que l'on puisse considérer le résultat de notre méthode comme étant satisfaisant, compte tenu des différences de complexité entre ces deux algorithmes.

Le dernier point à vérifier concerne la réussite de la reconnaissance sur des chaînes de chiffres connectés, puisque c'est une des caractéristiques importantes de notre méthode. Nous effectuons donc des tests sur des séquences de deux, trois et quatre chiffres, les résultats étant présentés dans le tableau 5.8.

Tableau 5.8

Résultats de la reconnaissance de chiffres connectés

	Séquence de 2 chiffres	**Séquence de 3 chiffres**	**Séquence de 4 chiffres**
Pourcentage de chiffres correctement reconnus	85,56	81,98	77,08

Si le résultat est correct pour de courtes séquences de chiffres, nous pouvons observer que la réussite diminue en même temps que le nombre de chiffres d'une séquence augmente. La segmentation est notamment en cause, étant donné que la détection des pics de variations spectrales est d'autant plus difficile que la zone de parole traitée est grande. Un second élément d'explication peut être apporté par la séparation consonnes/voyelles par sélection énergétique, qui est moins efficace lorsque l'énergie moyenne de la zone de parole est élevée, ce qui est le cas quand plusieurs chiffres sont prononcés successivement en un temps restreint.

5.6 Conclusion

Le tableau 5.9 présente un récapitulatif des résultats principaux de la méthode de reconnaissance vocale présentée au cours de ce mémoire.

Tableau 5.9

Récapitulatif des résultats principaux de la reconnaissance multilocuteurs sur des chiffres isolés

Chiffre	Réussite (%)	Durée du chiffre (s)	Durée totale de reconnaissance (s)
One	95	0,4399	0,3129
Two	96,67	0,3621	0,2734
Three	86,67	0,4489	0,2940
Four	93,33	0,4468	0,3105
Five	86,67	0,5536	0,3228
Six	98,34	0,6146	0,3478
Seven	95	0,5351	0,3363
Eight	88,34	0,3351	0,2703
Nine	85	0,5255	0,3409
Zero	75	0,5342	0,3356
Oh	93,33	0,3555	0,2785
moyenne	**90,30**	**0,4684**	**0,3112**

On en retient donc que le taux de reconnaissance est satisfaisant pour la taille du dictionnaire à partir duquel nous travaillons. Pour rappel, la plupart des systèmes de reconnaissance travaillent avec des dictionnaires de plusieurs centaines, voire milliers, de

Ko. La rapidité observée est également très appréciable, grâce à un algorithme de reconnaissance spécialement adapté au vocabulaire concerné. Notre méthode devrait donc pouvoir fonctionner sur des systèmes aux ressources limitées, et fournir des résultats convenables. Il reste toutefois à améliorer la reconnaissance des chiffres connectés, en se focalisant principalement sur les phases de segmentation et d'étiquetage, puisque c'est de là que proviennent les principales erreurs.

CONCLUSION

L'objectif de ce mémoire était de développer une méthode de reconnaissance robuste de la parole pouvant travailler sur un vocabulaire de dix chiffres, à partir d'un système possédant des ressources limitées. La plupart des méthodes utilisées habituellement sont en effet très lourdes en calculs et en place mémoire, et voient leurs résultats chuter dès lors que l'on tente de les alléger. Nous avons donc jugé judicieux d'analyser le vocabulaire à identifier, afin d'en extraire certaines caractéristiques qui pourraient nous aider à élaborer un moyen. La méthode ainsi enfantée se base sur un modèle général de construction des chiffres, à partir duquel nous pouvons simuler leur reconstruction. Les éléments utilisés pour cette reconstruction sont des segments acoustiques obtenus après un découpage basé sur les variations spectrales du signal, sur toute la durée du chiffre à identifier. L'algorithme de reconstitution prend également en compte de nombreuses propriétés propres aux éléments du vocabulaire étudié, qu'elles soient d'ordre acoustique ou lexical, afin de réduire l'espace de recherche, et de favoriser la sélection intelligente du chiffre le plus similaire à celui que nous souhaitons identifier.

Les résultats obtenus sur un grand nombre de locuteurs différents se montrent très satisfaisants, surtout si l'on prend en compte les larges gains d'espace mémoire et de temps d'exécution obtenus par rapport aux autres méthodes plus classiques. On peut ainsi noter que ces gains sont si imposants qu'il nous serait possible de rajouter plus de références dans le dictionnaire, ou d'améliorer le processus d'identification de segments, afin de rehausser les performances, tout en restant encore dans les limites de ressources fixées par le système. L'intérêt de cette méthode se situe également dans la possibilité de reconnaître des chaînes de chiffres connectés sans ajout de complexité par rapport à la reconnaissance de chiffres isolés. Cette tâche fonctionne d'ailleurs correctement, mais est tout de même très dépendante de la qualité des phases de segmentation et d'identification. Une amélioration de ces deux éléments serait effectivement une des clés pour garantir de meilleurs résultats de reconnaissance.

Enfin, il pourrait être intéressant d'adapter cette méthode à d'autres vocabulaires, en réajustant simplement le modèle et les contraintes de construction aux nouvelles caractéristiques de l'espace de travail. Si cela a fonctionné pour notre vocabulaire de dix chiffres, il n'y a effectivement pas de raison que ce ne puisse pas être le cas pour n'importe quel autre vocabulaire restreint, si tant est que l'on est capable de l'analyser suffisamment minutieusement pour en déterminer les propriétés de modélisation.

ANNEXE I

RÉPARTITION DES SEGMENTS DE RÉFÉRENCE PARMI LES ONZE CHIFFRES DU VOCABULAIRE

La méthode de reconnaissance développée au cours de ce mémoire s'appuie sur un dictionnaire de 40 unités acoustiques, pour reconnaître les chiffres allant de 0 à 9 prononcés en anglais (le zéro ayant alors deux prononciations différentes). Cette annexe présente donc, au moyen d'une série de figures, la répartition des ces 40 segments sur l'ensemble des 11 chiffres. Ces segments ont effectivement été sélectionnés selon les résultats les plus courants du découpage effectué au moyen de la méthode de segmentation présentée au chapitre 4. Il est donc important de préciser que, en pratique, le découpage obtenu n'est pas toujours si parfait, dépendamment de la qualité du signal et du locuteur, mais ce document a surtout pour objectif de permettre la visualisation des parties importantes constituant chacun des chiffres du vocabulaire.

Chaque figure de cette annexe se présentera donc de la façon suivante :

- la partie centrale contient la forme temporelle du chiffre, les frontières entre les segments successifs étant clairement affichées;
- la partie supérieure affiche les états auxquels sont associés les segments en question : début, milieu, intérieur ou fin;
- la partie inférieure affiche la numérotation des segments, telle qu'elle sera utilisée pour étiqueter les segments inconnus lors de la phase de reconnaissance.

Il est enfin important de préciser que les voyelles de certains chiffres ont été découpées en trois segments afin d'assurer une meilleur représentation de ces sons, connaissant la difficulté avec laquelle ils sont reconnus. Les dites voyelles sont ainsi découpées en deux segments, là où une variation spectrale est régulièrement observée, tandis qu'un troisième segment représente la totalité du son.

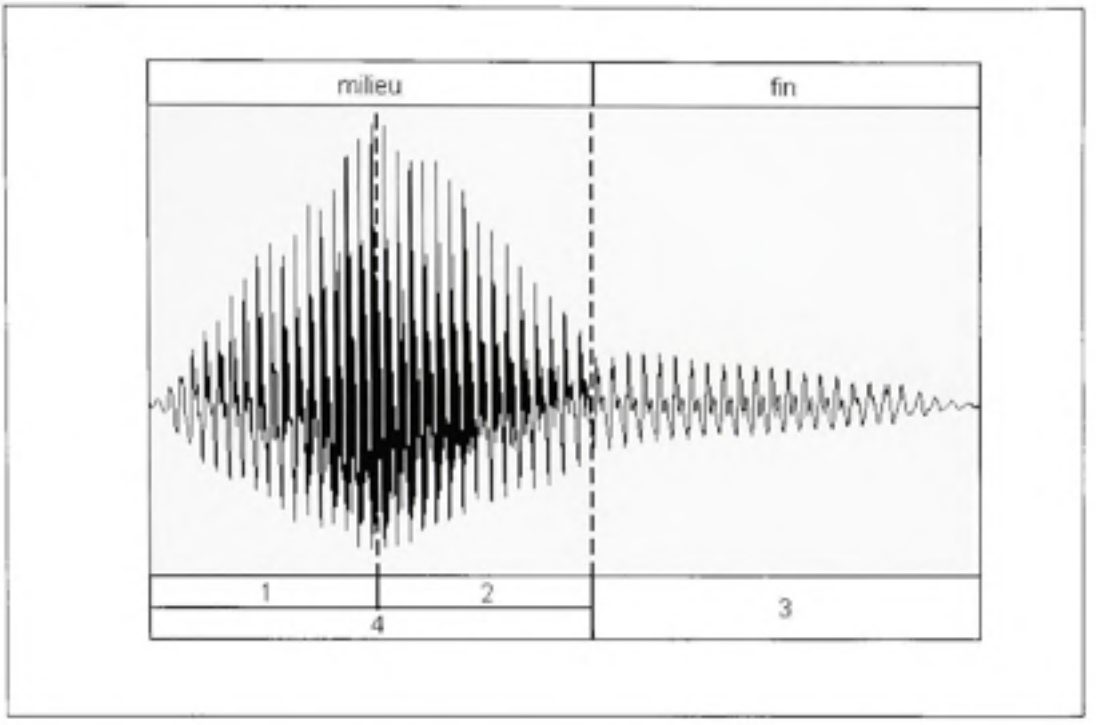

Figure A1.1 ***Segmentation du chiffre « one ».***

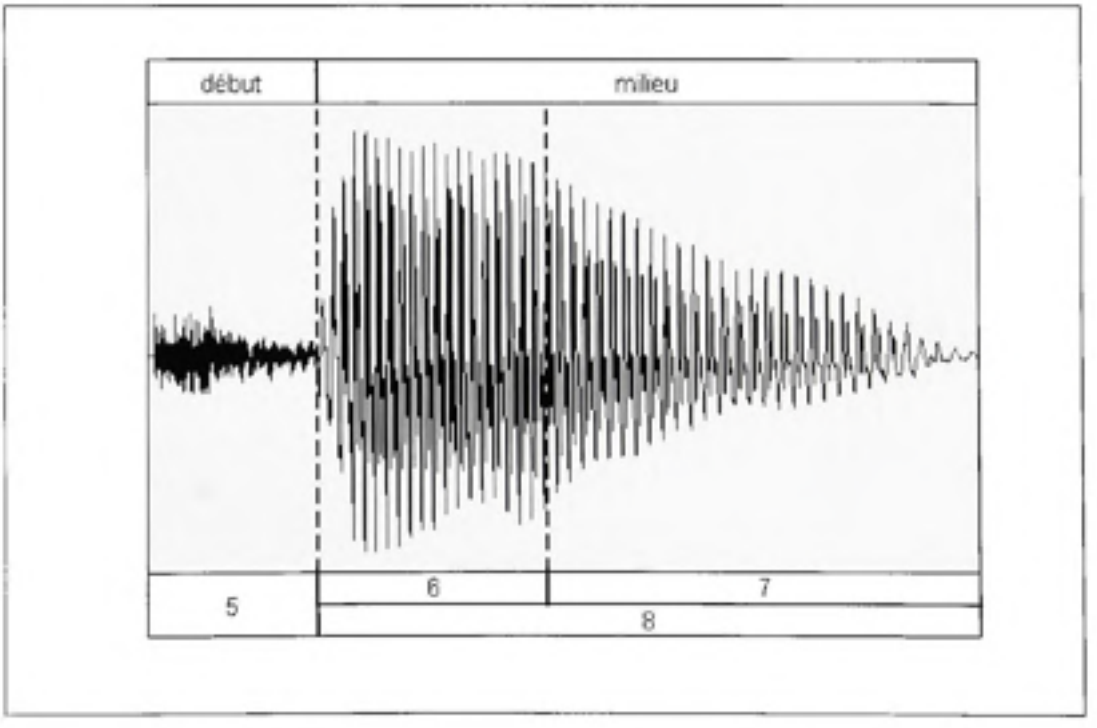

Figure A1.2 ***Segmentation du chiffre « two ».***

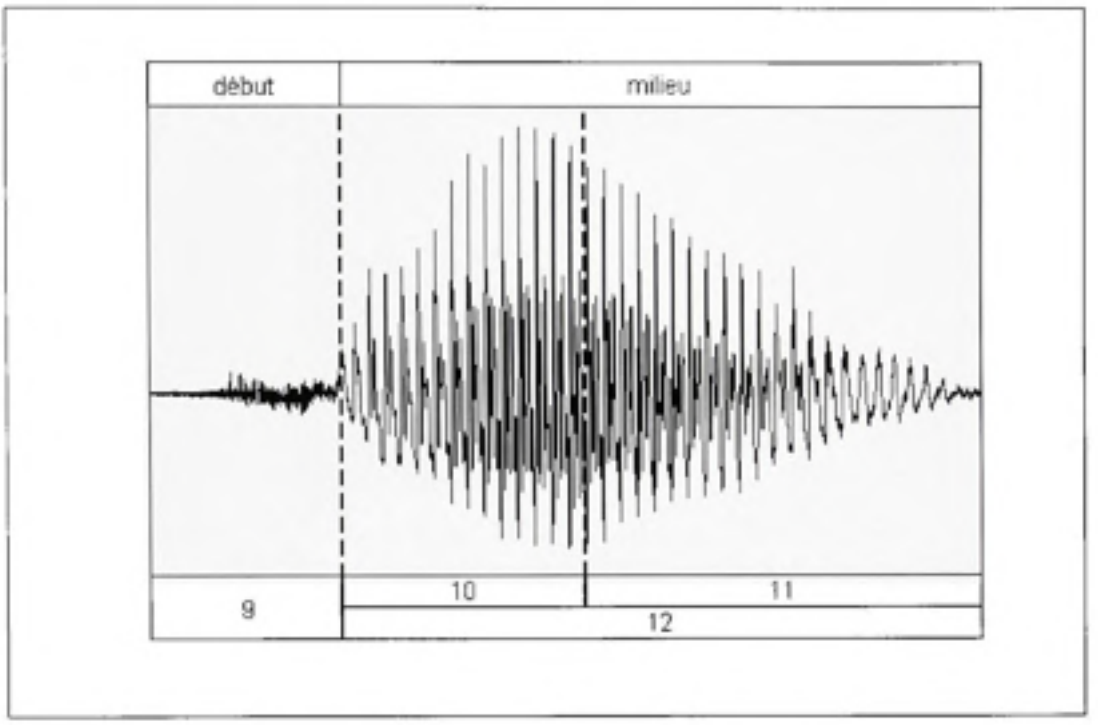

Figure A1.3 ***Segmentation du chiffre « three ».***

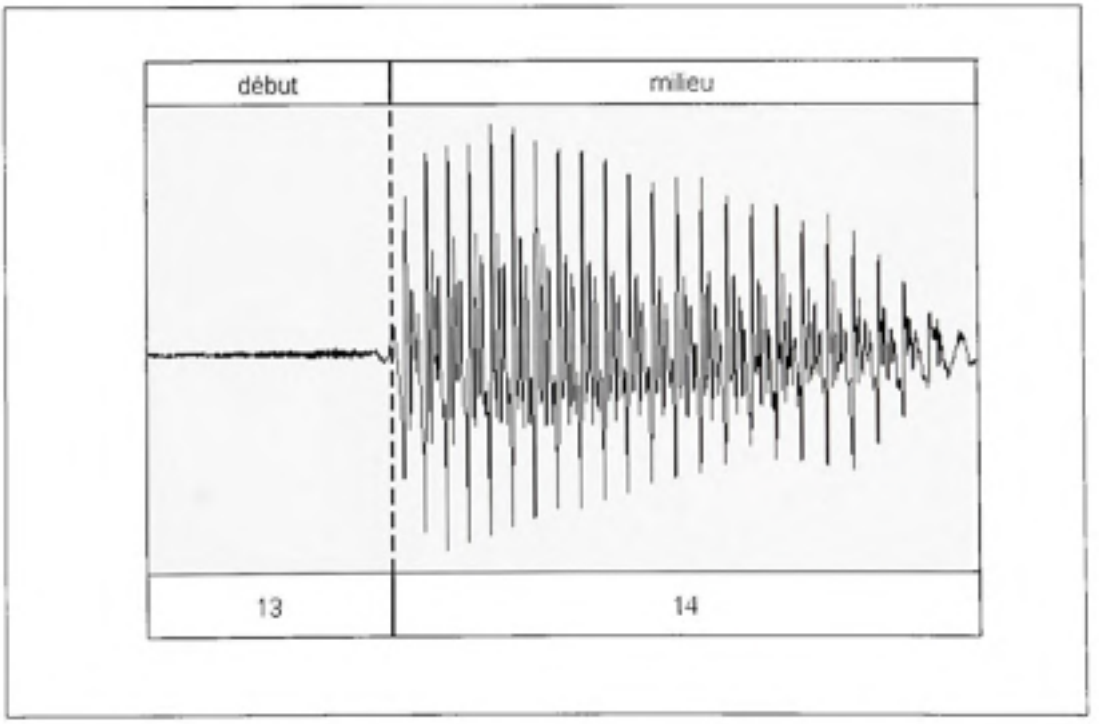

Figure A1.4 ***Segmentation du chiffre « four ».***

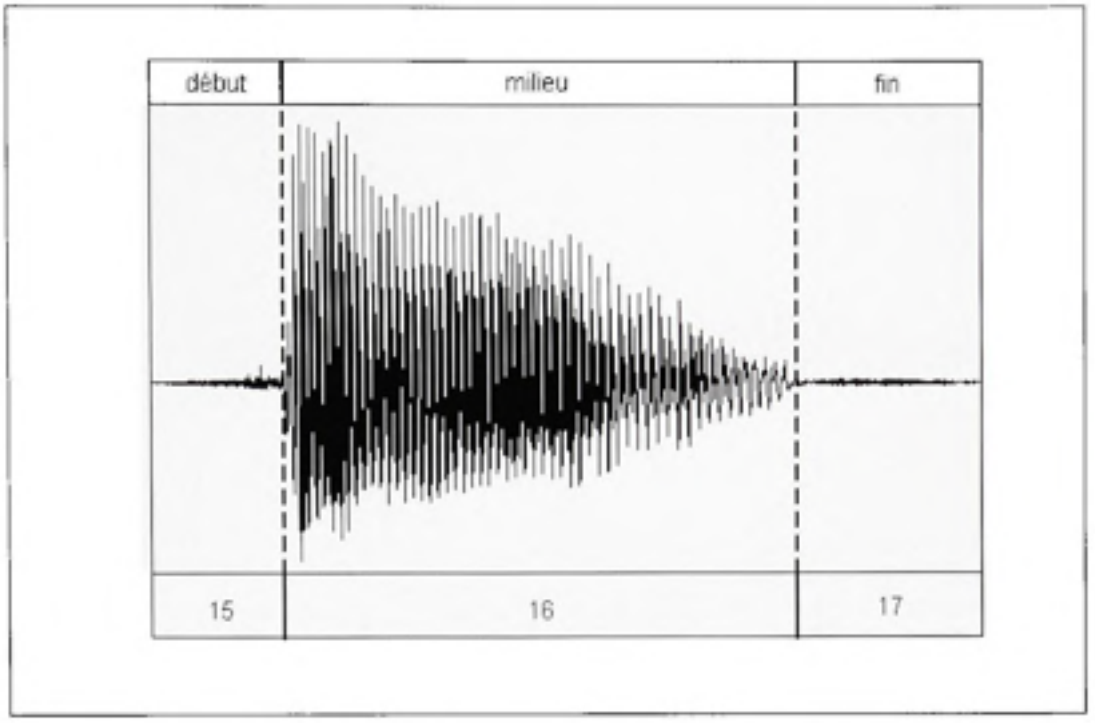

Figure A1.5 ***Segmentation du chiffre « five ».***

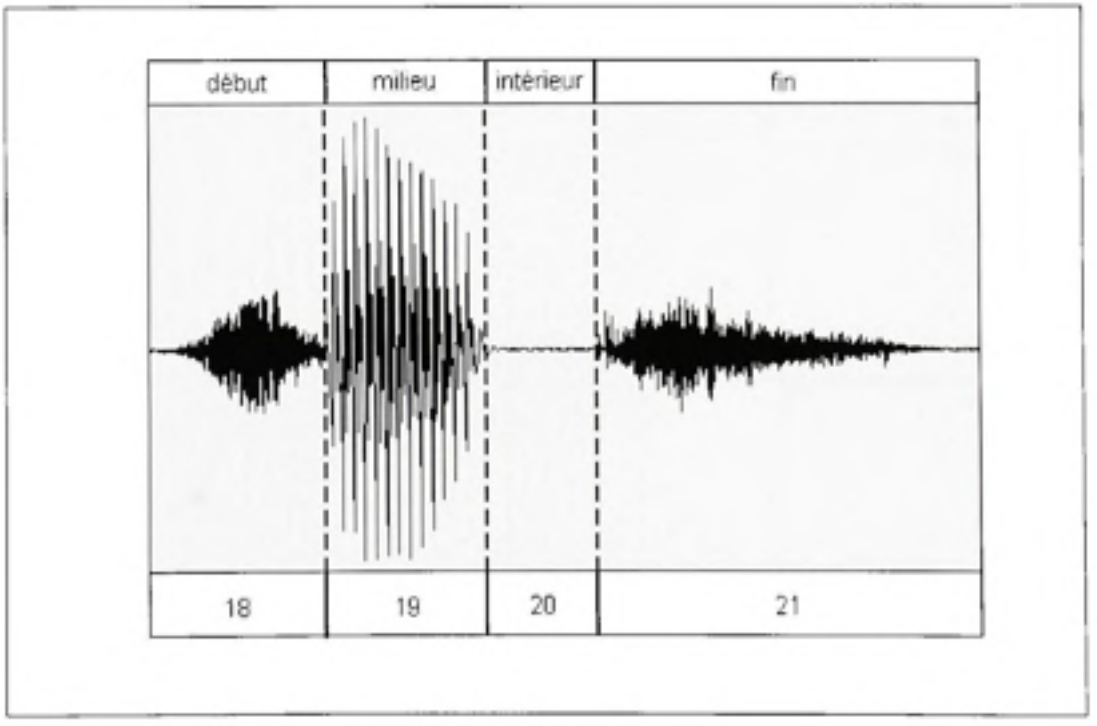

Figure A1.6 ***Segmentation du chiffre « six ».***

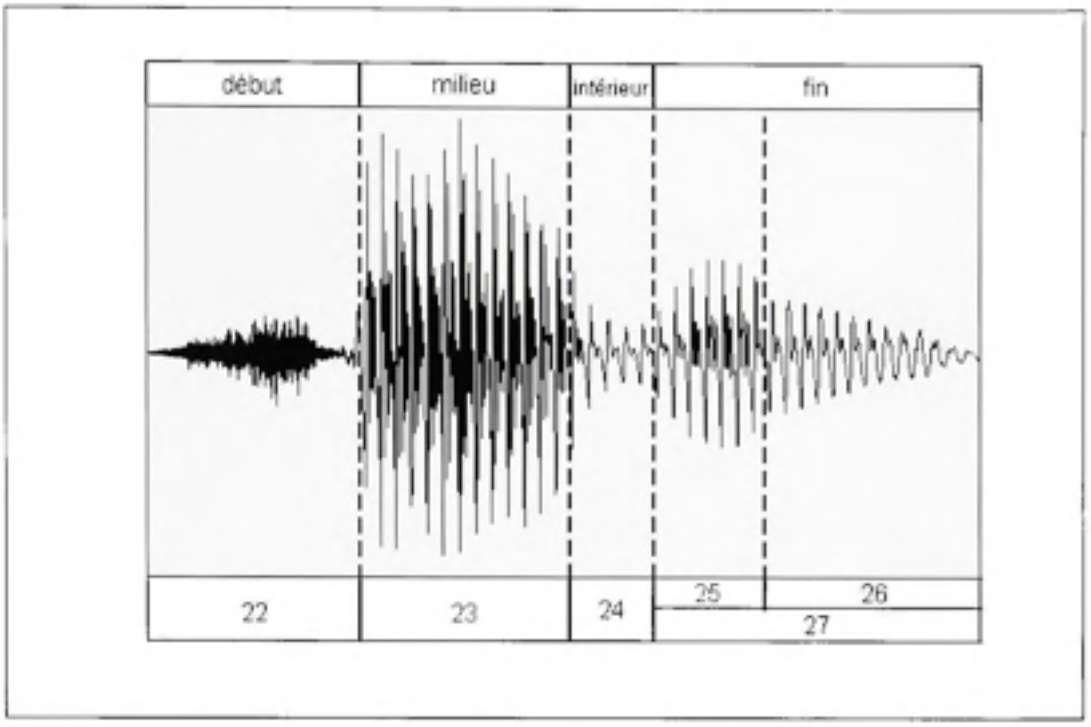

Figure A1.7 ***Segmentation du chiffre « seven ».***

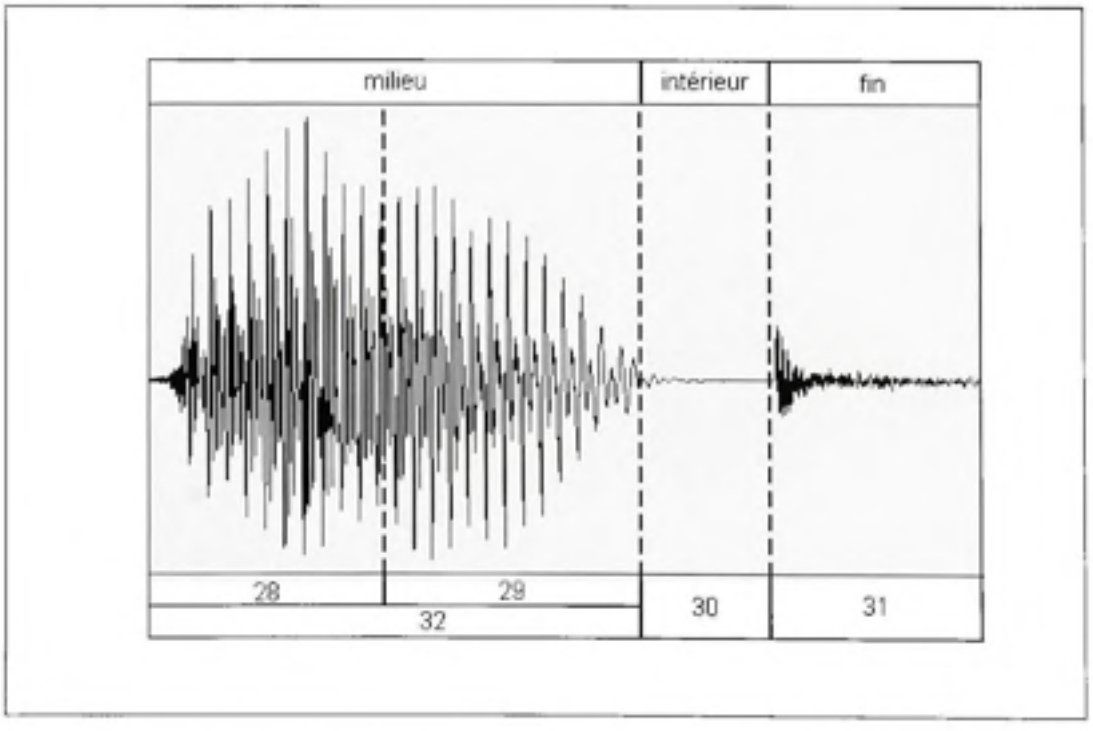

Figure A1.8 ***Segmentation du chiffre « eight ».***

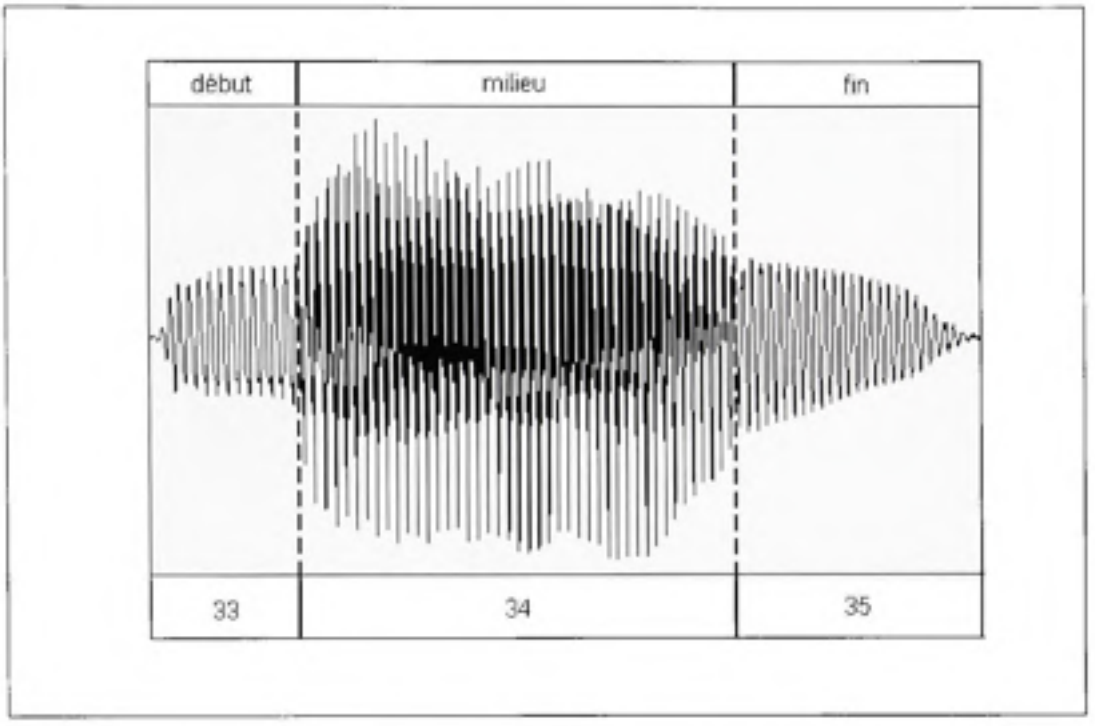

Figure A1.9 ***Segmentation du chiffre « nine ».***

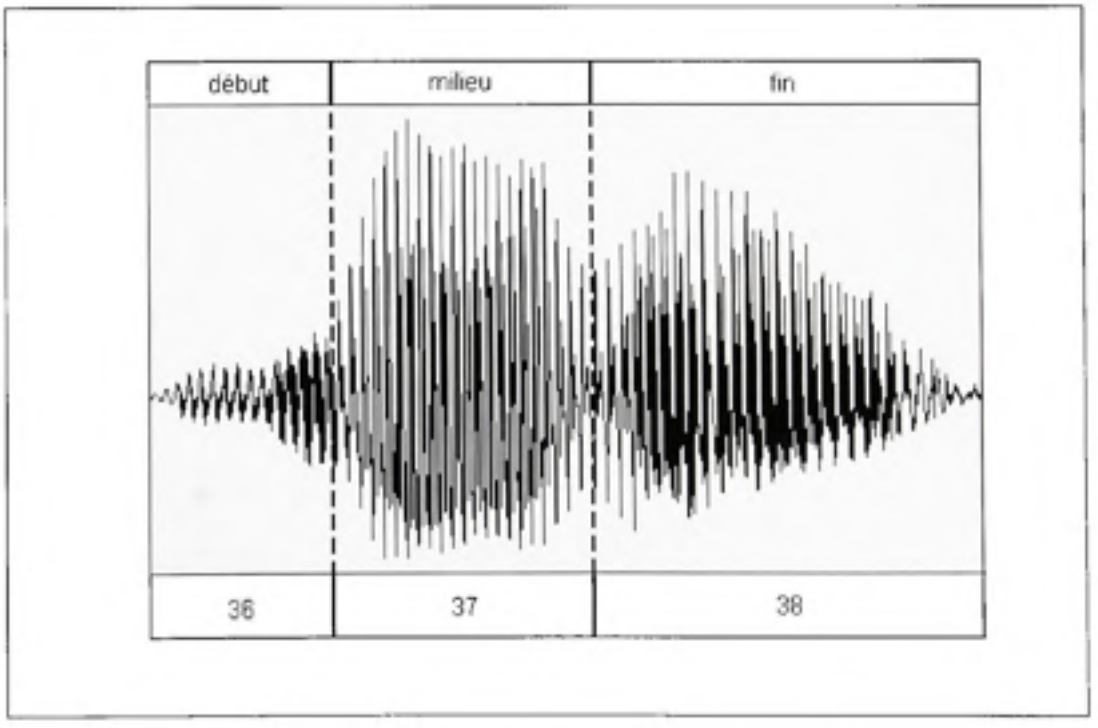

Figure A1.10 ***Segmentation du chiffre « zero ».***

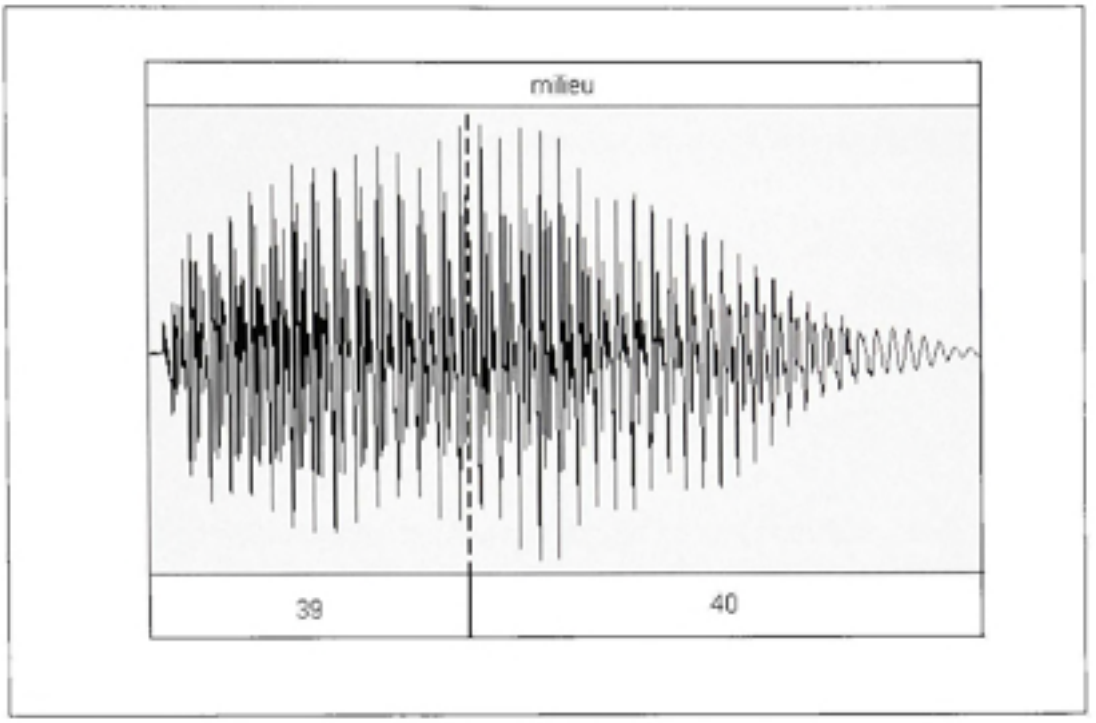

Figure A1.11 ***Segmentation du chiffre « oh ».***

ANNEXE II

TABLEAU D'ASSOCIATION DES SEGMENTS DE RÉFÉRENCE

Lors de la phase de reconstruction, les 40 segments présentés en annexe 1 sont utilisés par le biais d'un tableau indiquant leurs associations avec les différents chiffres du vocabulaire, ainsi que leur appartenance à certaines catégories. Ce tableau est donc constitué de 40 lignes, chacune voyant ses informations réparties de la façon suivante :

- les onze premières colonnes indiquent lorsqu'un son peut être retrouvé dans un ou plusieurs chiffres du vocabulaire. Quand c'est le cas, la colonne correspondante présente deux informations : une première valeur indiquant l'état du son dans le modèle du chiffre, comprise entre 1 et 5 (la valeur 5 signifiant que le segment en question peut aussi bien être associé à un état de début que de fin). La seconde valeur, comprise entre 1 et 3, représente la force d'association du segment avec le chiffre. Lorsque le segment n'est pas associé à un chiffre, la colonne contient simplement la valeur 0;
- les cinq dernières colonnes, activées par la valeur 1 ou désactivées par la valeur 0, indiquent la catégorie à laquelle appartient le segment : silence, non-voisé, voyelle, son \o\ et superposition. Il est important de noter qu'un segment peut n'avoir aucune de ces catégories activées. La catégorie « voyelle », notamment, n'est en effet activée que pour les segments de voyelles les plus difficiles à identifier avec précision.

Le programme de reconnaissance stocke toutes ces informations dans un seul tableau à trois dimensions. Pour des raisons de commodité, nous le présenterons ici sous la forme de quatre paires de tableaux, chacune étant constituée d'un premier tableau indiquant les associations des segments avec les chiffres, tandis que le second indiquera les catégories des segments. Ces quatre paires représentent les quatre catégories de sons : non-voisé (incluant les silences), consonnes, voyelles \o\ et autre voyelles.

Tableau A2.1

Tableau d'association des segments de sons non-voisés

	One	Two	Three	Four	Five	Six	Seven	Eight	Nine	Zero	Oh
5	0	1 3	0	0	0	0	0	4 3	0	0	0
9	0	0	1 2	0	0	0	0	0	0	0	0
13	0	0	0	1 3	1 3	0	0	0	0	0	0
15	0	0	0	1 3	1 3	0	0	0	0	0	0
17	0	0	0	0	4 3	0	3 2	0	0	0	0
18	0	0	0	0	0	5 3	1 3	0	0	0	0
20	0	0	0	0	0	3 3	0	3 3	0	0	0
21	0	0	0	0	0	5 3	1 3	0	0	0	0
22	0	0	0	0	0	5 3	1 3	0	0	0	0
24	0	0	0	0	4 1	0	3 3	0	0	0	0
30	0	0	0	0	0	3 3	0	3 3	0	0	0
31	0	1 3	0	0	0	0	0	4 3	0	0	0
36	0	0	0	0	0	0	0	0	0	1 3	0

Tableau A2.2

Tableau de catégorisation des segments de sons non-voisés

	Silence	**Non-voisé**	**Voyelle**	**Son \o**	**superposition**
5	0	1	0	0	1
9	0	1	0	0	0
13	0	1	0	0	0
15	0	1	0	0	0
17	0	1	0	0	0
18	0	1	0	0	1
20	1	0	0	0	0
21	0	1	0	0	1
22	0	1	0	0	1
24	0	0	0	0	0
30	1	0	0	0	0
31	0	1	0	0	1
36	0	1	0	0	0

Tableau A2.3

Tableau d'association des segments de consonnes

	One		Two	Three	Four	Five	Six	Seven		Eight	Nine		Zero	Oh
3	4	3	0	0	0	0	0	4	2	0	5	3	0	0
25	0		0	0	0	0	0	4	3	0	0		0	0
26	4	3	0	0	0	0	0	4	3	0	5	3	0	0
27	0		0	0	0	0	0	4	3	0	0		0	0
33	4	3	0	0	0	0	0	4	2	0	5	3	0	0
35	4	3	0	0	0	0	0	4	2	0	5	3	0	0

Tableau A2.4

Tableau de catégorisation des segments de consonnes

	Silence	Non-voisé	Voyelle	Son \o\	superposition
3	0	0	0	0	1
25	0	0	0	0	0
26	0	0	0	0	1
27	0	0	0	0	0
33	0	0	0	0	1
35	0	0	0	0	1

Tableau A2.5

Tableau d'association des segments de voyelles

	One	Two	Three	Four	Five	Six	Seven	Eight	Nine	Zero	Oh
6	0	2 | 3	0	0	0	0	0	0	0	0	0
7	0	2 | 3	0	0	0	0	0	0	0	0	0
8	0	2 | 3	0	0	0	0	0	0	0	0	0
10	0	0	2 | 3	0	0	0	0	0	0	0	0
11	0	0	2 | 3	0	0	0	0	0	0	0	0
12	0	0	2 | 3	0	0	0	0	0	0	0	0
16	0	0	0	0	2 | 3	0	0	0	2 | 2	0	0
19	0	0	0	0	0	2 | 3	0	0	0	2 | 2	0
23	0	0	0	0	0	0	2 | 3	0	0	0	0
28	0	0	0	0	0	0	0	2 | 3	0	0	0
29	0	0	0	0	0	0	0	2 | 3	0	0	0
32	0	0	0	0	0	0	0	2 | 3	0	0	0
34	0	0	0	0	2 | 2	0	0	0	2 | 3	0	0
37	0	0	0	0	0	0	0	0	0	2 | 3	0

Tableau A2.6

Tableau de catégorisation des segments de voyelles

	Silence	**Non-voisé**	**Voyelle**	**Son \o**	**superposition**
6	0	0	1	0	0
7	0	0	1	0	0
8	0	0	1	0	0
10	0	0	1	0	0
11	0	0	1	0	0
12	0	0	1	0	0
16	0	0	1	0	0
19	0	0	1	0	0
23	0	0	1	0	0
28	0	0	1	0	0
29	0	0	1	0	0
32	0	0	1	0	0
34	0	0	1	0	0
37	0	0	1	0	0

Tableau A2.7

Tableau de catégorisation des segments de voyelles \o\

	One		Two	Three	Four		Five	Six	Seven	Eight	Nine	Zero		Oh	
1	2	3	0	0	2	2	0	0	0	0	0	3	2	2	1
2	2	3	0	0	2	1	0	0	0	0	0	0		2	3
4	2	3	0	0	2	2	0	0	0	0	0	3	2	2	1
14	2	1	0	0	2	3	0	0	0	0	0	3	1	2	1
38	2	2	0	0	2	2	0	0	0	0	0	3	3	2	2
39	2	1	0	0	2	1	0	0	0	0	0	0		2	3
40	2	1	0	0	2	1	0	0	0	0	0	0		2	3

Tableau A2.8

Tableau de catégorisation des segments de voyelles \o\

	Silence	Non-voisé	Voyelle	Son \o\	superposition
1	0	0	0	1	1
2	0	0	0	1	0
4	0	0	0	1	1
14	0	0	0	1	1
38	0	0	0	1	1
39	0	0	0	1	1
40	0	0	0	1	1

BIBLIOGRAPHIE

Baum, L. 1972. « An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov process ». *Inequalities*, Vol. 3, p. 1-8.

Bellman, R. 1954. *Dynamic Programming*. Princeton : Princeton University press.

Benyassine, A., E. Sholomot, H. Su, D. Massoloux, C. Lambline et J.P. Petit. 1997. « ITU-T Recommandation G.729 Annexe B : A silence compression schem for use with G.729 optimized for V.70 digital simultaneous voice and data application ». *IEEE Communication magazine*, Vol. 35, n° 9, p. 64-73.

Boite, René. 1987. *Traitement de la parole*. 1ère édition. Lausanne : Presses polytechniques romandes. 280p.

Bose, Tamal. 2004. *Digital Signal and Image Processing*. 1ère éd. Hoboken, N.J. : J. Wiley, 706 p.

Caelen, J., N. Vigouroux et G. Perennou. 1983. « Structuration des informations acoustiques dans le projet ARIAL ». *Speech Com.*, Vol. 2, n° 2-3, p. 219-222.

Calliope. 1989. *La parole et son traitement automatique*. Paris : Masson, 717 p.

Carbonnel, N., J.P. Haton, J.M. Pierrel, et F. Longchamp. 1983. « Élaboration d'un système expert pour le décodage phonétique automatique ». *Speech Com.*, Vol. 2, n° 2-3, p. 231-233.

Church, Kenneth. 1987. *Phonological Parsing in Speech recognition*. 1ère éd.. Norwell : Kluvers Academic Publisher, 261p.

Durbin, J. 1960. *The fitting of time series models*. Rev. Inst. Int. Stat, v28. p 233-243.

Flanagan, J.L. 1972. *Speech Analysis, Synthesis, and Perception*. 2ème éd.. New York : Springer-Verlag.

Furui, Sadaoki. 1989. *Digital Speech Processing, Synthesis, and Recognition*. 1ère éd. Marcel Dekker Inc, 390p.

Furui, Sadaoki. 2005. « *50 years of progresse in speech and speaker recognition* ». In *Proc. SPECOM*, Patras, Greece, p 1-9 .

Gargour, Christian S. 2001. *Traitement numérique des signaux.* 1ère éd. Montréal : Université du Québec, École de Technologie Supérieure, 336p.

Gray, A. et J. markel. 1976. « Distance measure for speech processing ». *IEEE trans. ASSP*, Vol. 5, n° 24, p. 380-391.

Hanson, B. et H. Wakita. 1986. « Spectral slope based distorsion measures for all-pole models of speech ». *ICASSP*, p. 757-780.

Harris, Frederic J. 1978. « On the use of windows for harmonic analysis with the discrete Fourier transform ». *Proceedings of the IEEE*, Vol. 66, n° 1, p. 51-83.

Hataoka, N., H. Kokubo, Y. Obuchi et A. Amano. 2002. « Compact and robust speech recognition for embedded use on microprocessors ». *IEEE Multimedia Signal Processing*, p. 288-291.

Hui G., Ho K.C. et Goh, Z. 1998. « A robust speaker-independent speech recognizer on ADSP2181fixed-point DSP ». In *ICSP '98 Fourth International Conference on.* Vol. 1, pp. 694-697. Pékin, Chine : ICSP.

Itakura, F. et S. Saito. 1968. « Analysis synthesis telephony based on the maximum likelihood method ». In *Intern. Congr. Acoust.* Tokyo.

Itakura, F. 1975. « Minimum prediction residual principle applied to speech recognition ». *IEEE Trans. ASSP*, Vol. 23, p. 67-72.

Itakura, F. et T. Umezaki. 1987. « Distance measure for speech recognition based on the smoothed group delay spectrum ». *Proc. of ICASSP*, Vol. 3, p. 1257-1260 .

Junqua, J.C. et J.P. Haton. 1996. *Robustness in Automatic Speech Recognition.* 1ère éd. Norwell : Kluwer Academic Publisher, 440p.

Junqua, J.C., H. Wakita, et H. Hermansky. 1993. « Evaluation and optimization of perceptually-based ASR front-end ». *Speech and Audio Processing*, Vol. 1, n° 1, p. 39-48.

Kunt, Murat. 1984. *Traitement numérique des signaux.* 2ème éd. Lausanne : Presses polytechniques romandes, 402p.

Leonard, R. 1984. « A database for speaker-independent digit recognition ». In *IEEE International Conference on ICASSP '84*, Vol. 9, p. 328-331. Dallas.

Levinson, S., L. Rabiner, A. Rosenberg, J. Wilpon. 1979. « Interactive clustering techniques for selecting speaker-independant reference templates for isolated word recognition ». *IEEE Trans. ASSP*, Vol. 27, p. 134-141.

Lévy C., G. Linares, P. Nocera et J.F. Bonastre. 2004. « Reconnaissance de chiffres isolés embarquée dans un téléphone portable ». In *JEP'04*. Fes, Maroc.

Lévy C., G. Linarès et J.F. Bonastre. 2005. « Mobile phone embedded digit-recognition ». In *Workshop on DSP in Mobile and Vehicular Systems*. Sesimbra, Portugal : ACTi communications internationales.

Li, B.N.L. et J.N.K Liu. 1999. « A comparative study of speech segmentation and feature extraction on the recognition of different dialects ». In *Systems, Man, and Cybernetics IEEE International conference on*, Vol. 1, p. 538-542.

Markel J. et A. Gray. 1976. *Linear Prediction of Speech*. Berlin : Springer-Verlag. 305 p.

Noll, P. 1972. « *Cepstrum picth determination* ». In *J. Acoust. Soc. Amer.*, p 293-309.

Phadke S., R. Limaye, S. Verma et K. Subramanian. 2004. « *On design and implementation of an embedded automatic speech recognition* ». In *Proceedings of the 17th International Conference on VLSI Design*, p. 127. IEEE Computer Society.

Parsons, Thomas. 1986. *Voice and Speech Processing*. 1ère éd. New York : McGraw-Hill, 402 p.

Pikles, J.O. 1988. *An Introduction to the Physiology of Hearing*. 2ème éd. Londres : Academic Press.

Vergin, R. et D. O'Shaughnessy. 1995. « Pre-Emphasis and Speech Recognition ». In *Electrical and Computer Engineering, Canadian Conference*, Vol. 2, (Montreal, Sep. 5-8 1995) , p. 1062-1065.

Rabiner, L. et B.H Juang. 1993. *Fundamentals of speech recognition*. Englewood Cliffs : PTR Prentice Hall, 507p.

Rabiner, L. et M.R. Sambur. 1975. « An algorithm for determining the endpoints of isolated utterances ». *Bell Sys. Tech. J.*, Vol. 54, n° 2, p. 297-315.

Rabiner, L., B. Juang et J. Wilpon. 1987. « On the use of bandpass filtering in speech recognition ». *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. 35, n° 7, p. 947-954.

Rabiner, L. et R. Schafer. 1978. *Digital Processing of Speech Signals*. Englewood Cliffs, N.J. : Prentice Hall, 512p.

Rowden, Chris. 1992. *Speech Processing*. 1ère édition. Londres : McGraw-Hill, 405 p.

Sakoe, H. et S. Chiba. 1978. « Dynamic programming algorithm optimization for spoken word recognition ». *IEEE-ASSP*, Vol. 26, p. 43-49.

Sharma, M. et R. Mammon. 1996. « "Blind" speech segmentation : automatic segmentation of speech without linguistic knowledge ». *In Fourth International Conference on Spoken Language*, Vol. 2, p. 1237-1240. ICSLP.

Perennou, G. et M. De Calmes. 1985. « Segmentation en évènements phonétiques et en unités syllabiques ». *XIV JEP*, p. 142-146. Paris : GALF.

Vallverdu, F. et M. Faundez. 1996. « *Speech Segmentation Using Multilevel Hybrid Filters* ». Barcelone : ETSE Telecomunicacio.

Vintsyuk, T.K. 1968. *Speech discrimination by dynamic programming*. Vol. 4, p. 81-88.

Viterbi, A. 1967. « Error bounds for convolutionnal codes and an asymptotically optimal decoding algorithm ». *IEEE Trans. on Information Theory*, p. 260-269.

Wilpon, J. et L. Rabiner. 1985. « A modified clustering algorithm for use in isolated word recognition ». *IEEE Trans. ASSP*, Vol. 3, n° 33, p. 587-594.

Wilpon, J.G., L. Rabiner, et T.B. Martin. 1984. « An improved word-detection algorithme for telephone-quality speech incorporating both syntactic and semantic constraints ». *AT&T Tech. J.*, Vol. 63, n° 3, p. 479-498.

Zaabi, K. 2004. « Implémentation d'une méthode de reconnaissance de la parole sur le processuer de traitement numérique du signal TMS320C6711 ».Mémoire de maîtrise en génie électrique, Montreal, École de technologie supérieure, 136p.