UNIVERSITÉ DE SHERBROOKE
Faculté de génie
Département de génie électrique et de génie informatique

TRAITEMENT BIO-INSPIRÉ DE LA PAROLE POUR SYSTÈME DE RECONNAISSANCE VOCALE

Thèse de doctorat
Spécialité génie électrique

Stéphane LOISELLE

Jury Jean ROUAT (directeur)
François DUVAL
Douglas O'SHAUGHNESSY
Ramin PICHEVAR

Sherbrooke (Québec) Canada Été 2010

Your file Votre référence
ISBN 978-0-494-75063-6
Our file Notre référence
ISBN 978-0-494-75063-6

Exploration really is the essence
of the human spirit, and to pause,
to falter, to turn our back on the
quest for knowledge, is to perish

–Frank Borman

RÉSUMÉ

Cette these presente un traitement inspire du fonctionnement du systeme auditif pour ameliorer la reconnaissance vocale Pour y parvenir, le signal de la parole est filtre par un banc de filtres et compresse pour en produire une representation auditive

L'innovation de l'approche proposee se situe dans l'extraction des éléments acoustiques (formants, transitions et *onsets*) à partir de la representation obtenue En effet, une combinaison de détecteurs composes de neurones à décharges permet de reveler la presence de ces éléments et génere ainsi une séquence d'évenements pour caractériser le contenu du signal

Dans le but d'évaluer la performance du traitement présente, la sequence d'evenements est adaptee a un systeme de reconnaissance vocale conventionnel, pour une tâche de reconnaissance de chiffres isoles prononces en anglais Pour ces tests, la sequence d'evenements agit alors comme une selection de trames automatique pour la génération des observations (coefficients cepstraux) En comparant les resultats de la reconnaissance du prototype et du systeme de reconnaissance original, on remarque que les deux systemes reconnaissent tres bien les chiffres prononces dans des conditions optimales et que le systeme original est legerement plus performant Par contre, la différence observee au niveau des taux de reconnaissance diminue lorsqu'une reverberation vient affecter les donnees à reconnaître et les performances de l'approche proposee parviennent à dépasser celles du systeme de reference De plus, la selection de trames automatique offre de meilleures performances dans des conditions bruitees

Enfin, l'approche proposee se base sur des caracteristiques dans le temps en fonction de la nature du signal, permet une selection plus intelligente des donnees qui se traduit en une parcimonie temporelle, presente un potentiel fort interessant pour la reconnaissance vocale sous conditions adverses et utilise une detection des caracteristiques qui peut être utilisee comme sequence d'impulsions compatible avec les réseaux de neurones à decharges

Mots-clés Intelligence artificielle, reconnaissance vocale, extraction de caractéristiques phonetiques, système auditif biologique, approche connectioniste, modèles de Markov caches

REMERCIEMENTS

Je tiens d'abord à remercier Jean Rouat, sans qui l'opportunité de ce doctorat n'aurait pas existe, et le personnel de l'Universite de Sherbrooke qui ont rendu ce projet possible

Pour leur aide et leur soutien, je souhaite remercier mes collègues et amis du groupe NECOTIS ainsi que les etudiants gradués du departement de génie électrique et de genie informatique que j'ai eu la chance de côtoyer

Enfin, je remercie ma famille et mes proches pour leur soutien inconditionnel et leurs encouragements

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ACRONYMES

Acronyme	Définition
DCT	Discrete cosine transform
HMM	Hidden Markov model
IaF	Integrate-and-fire
LSM	Liquid state machine
MFCC	Mel-frequency cepstral coefficients
ROC	Rank order coding
SNR	Signal-to-noise ratio

CHAPITRE 1

INTRODUCTION

1 1 Mise en contexte et problématique

Plusieurs penseurs s'accordent pour dire que depuis l'avènement de l'ordinateur personnel et l'émergence d'Internet, notre société serait entree dans l'Âge de l'information Cet âge se caracteriserait par une nouvelle capacite individuelle, d'une envergure surprenante, d'obtenir et de partager de l'information pratiquement instantanément sans être reellement affectee par les distances ou les lieux geographiques

Évidemment, l'Internet joue un rôle majeur dans cette transformation qui menace maintenant d'extinction les autres medias plus traditionnels De nos jours, ce réseau informatique global est omnipresent dans les domaines de la communication, l'education, la surveillance, la protection, la publicite, le divertissement et bien d'autres

Par contre, on a souvent de la difficulte à naviguer et filtrer l'immense quantite d'information qui se trouve à notre portée pour obtenir ce que l'on cherche En effet, l'interface homme machine a peu changé depuis plusieurs années et encore aujourd'hui, le clavier et la souris restent les outils les plus utilises pour transmettre nos commandes à un systeme informatique Des améliorations pour rendre ces interactions plus conviviales ont eté apportees et quelques alternatives sont proposees Cependant, il est difficile d'imaginer un moyen de communication plus efficace pour nous que celui dont l'être humain se sert pour communiquer depuis plusieurs milliers d'annees, c'est-à-dire la parole

Malgré les importants developpements en informatique des dernieres années, les performances des systemes de reconnaissance vocale conventionnels restent inferieures a celles du système auditif humain surtout dans des conditions adverses et ces outils restent fortement dependants aux donnees d'apprentissage [Sroka et Braida, 2005, Lippman, 1997, Picone, 1993] La reconnaissance vocale automatique n'est donc pas très attirante et elle ne se limite habituellement qu'à des applications simples et limitees

1 2 Définition du projet de recherche

Puisque l'être humain se sert de la parole depuis des milliers d'annees comme moyen de communications, il n'est pas surprenant de constater à quel point il peut accomplir efficacement la reconnaissance vocale En effet, notre système auditif peut extraire avec facilité la parole d'un individu dans un environnement bruité, déterminer les caractéristiques du locuteur (tel que son sexe, son âge et son accent) et reconnaître ce qui est prononcé

D'ailleurs certains travaux proposent de s'inspirer des systemes auditifs biologiques pour combler les faiblesses des systemes conventionnels telles que les problèmes lies à l'utilisation de fenêtres fixes des approches conventionnelles [Lewicki et Sejnowski, 1999] et diminuer la dependance aux donnees d'apprentissage en utilisant des caractéristiques qui definissent mieux la nature du signal [Ali *et al*, 1998, 2001]

À travers ce projet de recherche, on propose donc d'ameliorer la reconnaissance vocale en s'inspirant de ces eléments Le traitement permettra alors à un prototype d'obtenir un meilleur taux de reconnaissance dans des conditions adverses et pourra s'adapter aux approches connexionnistes

1 3 Objectifs du projet de recherche

Les objectifs de ce projet de recherche sont

- d'utiliser une representation auditive du signal inspiree de la biologie, qui permet une plus grande robustesse au système, pour une tâche de reconnaissance vocale, en exploitant le plus possible l'information utile et discriminante, qui se trouve dans les transitions des signaux acoustiques,
- d'extraire de cette representation des caracteristiques à l'aide de bases qui representent des eléments acoustiques afin de se departir de la contrainte de la fenêtre fixe glissante, que l'on trouve dans les systemes conventionnels,
- de generer une sequence d'evenements qui serait compatible avec les classificateurs des approches connexionnistes

Les performances du traitement proposé seront evaluees par la realisation d'un systeme de reconnaissance automatique de la parole et elles sont comparées à celles d'un système conventionnel Markovien sur la base de donnees *TI 46-Word* [Doddington et Schalk, 1981] (chiffres isoles prononces en anglais et enregistres par un seul microphone) Ces evaluations se feront sur des donnees propres, bruitees et avec reverberation

1 4 Contributions originales

La contribution originale du projet de recherche se situe au niveau du traitement du signal En effet, l'approche proposee

- se base sur des caracteristiques dans le temps en fonction de la nature du signal
  - débuts et changements importants,
  - formants,
  - transitions,
- presente une robustesse au niveau du bruit et de la reverbération,
- utilise une detection des caracteristiques qui peut être utilisee comme sequence d'impulsions et adaptee a un reseau de neurones,
- permet une selection plus intelligente des données qui se traduit en une parcimonie temporelle

1 5 Plan du document

D'abord, un bref survol de travaux interessants portant sur la reconnaissance vocale est effectue pour presenter l'approche conventionnelle et quelques élements de l'approche biologique

Par la suite, les etapes du traitement propose sont présentees La représentation auditive du signal utilisee est sélectionnee, on explore les éléments acoustiques qui serviront de bases au systeme propose et un mecanisme de generation d'impulsion sera implémente

Le chapitre suivant comparera les performances (données propres, bruit blanc et réverberation) du traitement complet face à un systeme de reférence conventionnel grâce à un prototype

Enfin, une courte synthèse du projet de recherche est faite et de nouvelles perspectives de recherche sont proposees

CHAPITRE 2

ÉTAT DE L'ART

2 1 Reconnaissance vocale

La reconnaissance vocale a pour but de permettre à un systeme informatique de reconnaître les mots prononces par un locuteur La parole peut ensuite être affichée comme texte, ce qui necessite aux systèmes d'avoir un vocabulaire etendu ainsi qu'une connaissance en grammaire On peut aussi limiter la reconnaissance a quelques commandes pour une application spécifique, un service a la clientele automatique par exemple Il est evident que ce moyen de communication est beaucoup plus naturel que l'utilisation du clavier et permet a l'utilisateur de garder l'usage de ses mains pour d'autres tâches

2 2 Approche conventionnelle

La grande majorite des systemes de reconnaissance vocale actuels se sert des coefficients cepstraux comme modele du signal et des chaînes de Markov cachees comme modeles de mots ou d'unites phonétiques [Jurafsky et Martin, 2008, Picone, 1993] Le signal de la parole est alors converti en vecteurs d'observations qui representent des événements dans un espace de probabilite et on se sert des chaînes de Markov cachees pour trouver la sequence la plus probable d'avoir produit ces evenements et ainsi reconnaître ce qui est prononce Cette approche est maintenant bien connue et peut offrir d'excellentes performances (aux alentours de 99% pour des chiffres isoles) dans des conditions optimales

Par contre, l'utilisation de systemes de reconnaissance uniquement dans des conditions optimales est particulierement limitée En effet, adapter le système au locuteur n'est pas toujours possible, les caracteristiques des signaux ne correspondent pas necessairement à celles utilises lors de l'entraînement du système et la presence de facteurs adverses selon l'environnement (bruits, reverberation,) vont dégrader rapidement la performance du système

2.2.1 *MFCC*

Les traitements homomorphiques, comme l'analyse cepstrale, se basent sur le modele elementaire source-filtre de la production de la parole Dans ce modèle, les vibrations des cordes vocales representent la source et le conduit vocal agit comme un filtre On tente alors d'eliminer l'influence de la source et on ne conserve que des caracteristiques liees au filtre pour obtenir une certaine indépendance au niveau de la voix du locuteur Pour calculer ces coefficients

- on effectue la transformée de Fourier à l'aide d'une fenêtre courte (10 à 20 millisecondes) qu'on glisse sur le signal selon un intervalle specifie,
- on projette la transformee sur une echelle frequentielle non lineaire inspiree du systeme auditif humain (l'échelle *Mel* [Stevens *et al*, 1937]),
- on calcule le logarithme pour chacune des energies de ces frequences,
- on calcule la transformee en cosinus discrete sur les energies des frequences

Comme les termes d'ordre inferieur correspondent à la correlation à court terme du signal (forme du conduit vocal), alors que les maximums locaux des termes d'ordre superieur demontrent la correlation à long terme, ou la périodicite de la forme d'onde (information sur l'excitation), on ne conserve generalement qu'un nombre limite de ces coefficients De plus, augmenter la precision de la representation du signal en augmentant le nombre de coefficients demande aussi une plus grande quantité de données necessaire à l'apprentissage des modèles

Les systèmes de reconnaissance vocale modernes utilisent les premiers (d'une dizaine à plus d'une vingtaine) coefficients cepstraux (*Mel-frequency cepstral coefficients* ou *MFCC*) pour representer le spectre d'amplitude à court terme de la parole Le logarithme de l'énergie de la trame est aussi conserve ainsi que les dérivees premieres et secondes en temps de ces coefficients pour obtenir les coefficients Delta et Delta-Delta [Jurafsky et Martin, 2008], dans le but de mieux caractériser la variation temporelle du signal

À chaque deplacement de la fenêtre, un vecteur de coefficients (aux dimensions assez bien decorrelees) qui representent le cepstre du signal est alors génere et ces vecteurs peuvent ensuite être utilises par une chaîne de Markov cachée

Par contre, ces coefficients ne sont pas très robustes aux bruits additifs De plus, l'utilisation des fenêtres est susceptible de lisser ou d'effacer des changements rapides, caracté-

ristiques importantes de certains elements acoustiques [Smith et Lewicki, 2005, Lewicki, 2002b]

2 2 2 *HMM*

Une chaîne de Markov cachée (*hidden Markov model* ou *HMM*) consiste en un automate à etats finis qui se déplace d'un etat à un autre à chaque unite de temps selon une densité de probabilités À chaque instant, un etat actif lui est donc associé et des observations sont génerees, une fois de plus, selon une densité de probabilite Le processus stochastique sous-jacent n'est donc observable que par la séquence d'observations qui est produite

Une chaîne de Markov cachee est caracterisee par

- N, le nombre d'etats du modele,
- M, la dimension du vecteur d'observations o_t genere par état,
- la densite de probabilite de transition de l'etat i à l'etat j, $A = \{a_{ij}\}$,
- la densite de probabilite d'observer o_t à l'etat j, $B = \{b_j(k)\}$,
- la probabilite de l'etat initial π, $\pi = \{\pi_i\}$

Dans le cadre de la reconnaissance vocale, on modélise un mot ou un phoneme par un modele HMM et les MFCC servent d'observations [Rabiner, 1989] Pour la reconnaissance de mots isoles, le mot (w) est representé par la séquence d'observations O, definit alors comme

$$O = o_1, o_2, \quad , o_T \tag{2 1}$$

et la tâche revient à trouver

$$argmax_i\{P(w_i|O)\} \tag{2 2}$$

Pour y arriver, il s'agit d'estimer les paramètres des modèles de Markov caches L'apprentissage de ces parametres pour les modeles peut s'effectuer par un algorithme de maximum de vraisemblance (*Expectation-maximisation algorithm*) et l'algorithme de Viterbi est souvent utilisé pour calculer la probabilité qu'un modèle génère une sequence donnée d'observations

Évidemment, pour qu'un tel modele fonctionne, il faut suffisamment de donnees pour que l'apprentissage soit efficace et ces données doivent representer le plus fidèlement possible celles a reconnaître

2.3 Améliorer la reconnaissance vocale

De plus en plus, on tente d'amelıorer la performance des systemes actuels en s'ınspırant du système audıtıf bıologıque Certaıns travaux proposent un traıtement qui permet une meılleure reconnaıssance dans des condıtıons adverses [Messıng *et al* , 2009, Dımıtrıadıs *et al* , 2005, Ravındran *et al* , 2004, Sandhu et Ghıtza, 1995] alors que d'autres presentent des approches dıfferentes basées sur des reseaux de neurones [Smıt et Barnard, 2009, Verstraeten *et al* , 2005, Loıselle *et al* , 2005, Rouat et Garcıa, 1998]

2 3 1 Traıtements pour une reconnaissance vocale plus robuste

Dans l'artıcle de D Messıng [Messıng *et al* , 2009], on presente un système qui produıt des confusıons sımılaires à celles de l'être humaın pour une tâche de reconnaıssance de consonnes dont le sıgnal est dégrade par du bruıt Le système propose est compose d'un modèle de l'oreılle moyenne, d'un banc de filtres qui peut adapter ses gaıns et ses bandes passantes pour ımıter les mecanısmes d'adaptatıon non lıneaires de la cochlee et d'une étape qui représente les cellules cılıees ınternes et le nerf audıtıf Cette archıtecture permet d'amplifier les elements de faıble energıe du sıgnal dependamment du bruıt et produıt des résultats ıntéressants

Par contre, la performance est évaluee sur des sıgnaux synthetıques De plus, une etape du traıtement consıste à lısser le sıgnal par un recouvrement de fenêtres pour trouver un taux de decharges moyen, ce qui peut reduıre ou élımıner des changements rapıdes ımportants a la caracterısatıon de certaıns sıgnaux

Dans un travaıl antérieur du groupe de O Ghıtza [Sandhu et Ghıtza, 1995], les auteurs proposent une autre alternatıve ınspırée par les propriétés du systeme audıtıf pour re-presenter le sıgnal et effectuer la reconnaıssance vocale contınue Cette foıs, les cellules cılıées ınternes sont sımulees par des detecteurs de nıveaux croıssants Chaque foıs qu'un nıveau d'energıe est depasse dans un canal, on accumule dans un hıstogramme l'ıntervalle de temps passe entre ces ınstants Par la suıte, les trames du sıgnal represente par ces hıstogrammes sont utılısees dans un systeme à base de *HMMs* pour effectuer la reconnaıs-sance

Comparee à un système conventıonnel (*HMMs* et *MFCCs*), leur approche offre une meılleure performance lorsque les donnees tests sont modifiées pour sımuler un sıgnal télephonıque, maıs elle n'offre aucun avantage en presence de reverberatıon ou lorsque les donnees ne

sont pas modifiées De plus, l'obtention des histogrammes demande un traitement par trames qui risque toujours de lisser des éléments acoustiques importants

Un banc de filtres qui imite le système auditif est aussi utilisé par D Dimitriadis [Dimitriadis *et al*, 2005] En estimant le logarithme de la moyenne à court-terme de l'operateur de Teager-Kaiser [Kaiser, 1990] pour chaque canal de ce banc de filtres, les auteurs montrent qu'il est possible d'obtenir une meilleure performance face aux bruits additifs Pour ce faire, ils remplacent les *MFCCs* du système de reconnaissance conventionnel par les coefficients qu'ils extraient et ils comparent leurs resultats avec ceux d'un systeme conventionnel non modifie

Les coefficients proposés incorporent l'amplitude et l'information fréquentielle, mais l'operateur Teager-Kaiser est plutôt sensible aux bruits De plus, cette approche risque aussi de souffrir du probleme lié a l'utilisation de trames

Un probleme similaire se retrouve dans l'approche proposee par S Ravindran [Ravindran *et al*, 2004] ou un modèle du systeme auditif est aussi utilise pour obtenir des coefficients qui remplacent les *MFCCs* Cette fois, la tâche se limite à la classification du type de signal (bruit, musique, parole et vocalisation d'animaux), mais dans un contexte de circuits integres limites en ressources

2 3 2 Classificateurs basés sur des réseaux de neurones

Du côte des classificateurs, il existe aussi certaines alternatives inspirees de la biologie et on s'interesse de plus en plus aux reseaux de neurones à impulsions L'article de W J Smit [Smit et Barnard, 2009] presente un systeme de reconnaissance vocale continue qui utilise des sequences d'impulsions pour modeliser les chiffres prononces Des bases non negatives sont obtenues des trames du spectrogramme du signal par une technique d'optimisation [1] et le résultat de la projection du spectrogramme du signal à reconnaître sur ces bases est consideré comme des séquences d'impulsions Pour générer une impulsion, la valeur de la projection pour une base doit simplement depasser un seuil defini et elle indique que cette base est utilisée pour reconstruire une partie du spectrogramme La valeur de cette impulsion définit alors sa contribution Un algorithme d'apprentissage supervise est ensuite utilise pour obtenir des modèles des mots à l'aide de treillis de Viterbi

[1] L'optimisation est effectuee par la methode de la descente de gradient presentee par B Olshausen [Olshausen et Field, 1997], mais en remplaçant le parametre de contrôle pour determiner la taille du pas par une recherche lineaire

Ce type de representation produit une certaine parcimonie puisque les bases ne sont pas toutes présentes à chaque instant et la distorsion temporelle est geree en créant des versions à echelles de temps differentes de chaque base

Par contre, l'apprentissage des modèles proposés constitue un problème laborieux à resoudre et demande des ressources considerables De plus, la performance du systeme n'est évaluee qu'avec des donnees propres et n'atteint pas le niveau offert par un systeme conventionnel Les auteurs sont aussi conscients que le spectrogramme n'est pas le meilleur choix pour representer le signal

Une equipe de Gand en Belgique a aussi exploré l'idee d'effectuer la reconnaissance vocale avec un reseau de neurones [Verstraeten *et al*, 2005] Ils se servent d'une *Liquid State Machine* (*LSM*) dans le contexte de la reconnaissance de chiffres isoles Dans le travail qu'ils presentent, le reservoir (ou liquide) de la *LSM* est composé de neurones a decharges qui sont reliés entre eux pour former un réseau recurrent [Maass *et al*, 2002] Puisque chaque neurone possede un etat, ce réseau presente une dynamique non-lineaire complexe dans un espace d'etats internes de grandes dimensions L'etat de ce reseau est modifie par ses entrees et les auteurs s'attendent a y retrouver l'information pertinente pour réaliser la reconnaissance Deux types d'entrees sont appliqués au réseau des *MFCCs* et les sorties du modèle auditif de Lyon [Lyon, 1982] Les fonctions de sorties de la *LSM* utilisees pour la classification sont des classificateurs linéaires et elles permettent de projeter les etats du reseau aux classes de sorties (une couche de neurones de classification)

Combine à l'entree inspiree par le système auditif, leur prototype offre une performance legerement inférieure à celle d'un systeme conventionnel sur donnees propres, mais plus robuste au bruit De plus, d'autres travaux du groupe présentent une méthode efficace de l'implementation materielle de cette approche [Schrauwen et Verstraeten, 2007]

Cependant, les *MFCCs* ne sont pas vraiment compatibles avec ce prototype et la taille et les caractéristiques des réseaux produisent des variations souvent importantes au niveau des performances ce qui a demandé d'evaluer cette approche sur un ensemble de reseaux Aussi, l'apprentissage des fonctions de sorties peut presenter des difficultes, mais des travaux récents proposent des moyens pour l'ameliorer [Ghani *et al*, 2008, Oliveri *et al*, 2007]

Une autre utilisation de sequences d'impulsions est presentee dans un travail precedent [Loiselle *et al*, 2005] Dans ce travail, on utilise les sequences d'impulsions générées par un modele simple du systeme auditif pour évaluer la performance d'un code par ordre de rang (*Rank Order Coding* ou *ROC*) [Thorpe *et al*, 2001] dans un contexte de reconnaissance de chiffres isoles

Dans le cas de données d'apprentissage très limitées, l'approche basée sur le *ROC* est plus performant Par contre, un systeme conventionnel peut rapidement s'ameliorer en augmentant la quantite de données disponible, ce qui n'est pas possible pour le prototype qui utilise le *ROC* De plus, le traitement de ce prototype favorise de façon importante le debut des prononciations, ce qui produit d'importantes confusions entre les chiffres cinq, six et sept (en français)

Combiner un reseau de neurones et une représentation inspiree du système auditif a aussi ete exploree par J Rouat [Rouat et Garcia, 1998] Cette fois, on effectue la reconnaissance de voyelles sur un reseau de neurones *Dystal* [Alkon *et al*, 1990] On extrait d'abord l'information qui se situe à la sortie de filtres cochléaires dans l'enveloppe du signal module et le reseau apprend les correlations et anti-corrélations à travers l'excitabilité de ses zones dendritiques Enfin, une deuxième couche sert à la classification sous forme de vecteurs de similarites

La reconnaissance des voyelles isolees avec l'approche proposee est efficace et utiliserait des caracteristiques plus robustes au bruit

Par contre, il s'agit d'une tâche simple sur des donnees propres et une etude plus poussee est nécessaire pour une evaluation plus serieuse des performances de ce dernier

Comme on peut le constater, s'inspirer du système auditif biologique permet d'améliorer les performances en reconnaissance vocale Par contre, ces gains en performances restent souvent modestes ou limites à des situations particulières Néanmoins, on ne doit pas oublier que les traitements et classificateurs inspires par la biologie sont bien recents compares aux approches conventionnels De plus, on souligne dans plusieurs cas que ces gains pourraient être plus importants, mais que la representation et le classificateur ne sont pas bien adaptes

On propose donc d'utiliser une représentation auditive du signal, qui nous permettra de sélectionner des élements acoustiques sur lesquels sera basée la reconnaissance vocale Ensuite, ce traitement devra presenter l'information sous un format compatible avec les classificateurs des approches connexionnistes

2 4 Représentation auditive du signal

Une representation populaire inspirée du système auditif biologique consiste à filtrer le signal par un banc de filtres cochleaires Par la suite, on peut appliquer une rectification, une compression et divers types d'inhibitions ou de gains pour représenter certains meca-

nismes d'adaptation que l'on retrouve à différents niveaux du système auditif [Lyon, 1982] La representation alors obtenue permet de mieux conserver les détails importants autant en temps qu'en frequence

La forme caracteristique des filtres cochleaires de type gammatone [Patterson, 1976] serait d'ailleurs le fruit d'une adaptation des mécanismes de la cochlée aux signaux naturels qui nous entourent (vocalisations, bruits du vent, craquements de brindilles,) puisqu'ils permettent d'en faire un codage efficace [Lewicki, 2002a]

Plus haut dans le traitement auditif des signaux, on retrouverait des zones sensibles a différentes échelles de modulations [Mesgarani *et al*, 2008, Pichevar et Rouat, 2003] qui pourraient servir à la segmentation et la classification Par contre, les mécanismes exacts du fonctionnement de ces niveaux superieurs du systeme auditif restent encore partiellement mystérieux

Bien que le fonctionnement du systeme auditif ne soit pas entièrement compris, on sait qu'il se base sur des élements acoustiques comme les formants, ses transitions et coarticulations ainsi que les debuts des phonemes et les changements brusques d'énergie (*onsets*), pour effectuer la reconnaissance vocale

2.5 Éléments acoustiques

Les sommets d'energie stationnaires qui apparaissent dans la representation du signal de la parole se nomment formants et ils definissent les résonances du conduit vocal Habituellement, les deux premiers formants (F_1 et F_2) sont suffisants pour caracteriser les voyelles a, e, i, o et u [Geisler, 1998]

Tableau 2 1 Liste des centres des formants F_1 et F_2 des voyelles a, e, i, o et u [Suga, 1990]

Voyelle	F_1 (Hz)	F_2 (Hz)
a	800	1100
e	550	2100
i	310	2500
o	590	900
u	340	850

Comme on peut le constater, l'identification de ces voyelles isolees est une tâche facilement realisée et ne presente pas de difficulte pour les systèmes conventionnels Par contre, ces voyelles ne representent qu'un composant du signal de la parole et les sommets d'énergie stationnaires ne permettent pas de bien detecter un grand nombre de phonèmes Plusieurs

consonnes ne sont pas voisées et les caractéristiques des voyelles sont modifiees par l'articulation, dependamment de ce qui les precède et de ce qui va les suivre Enfin, il faut noter que le tableau 2 1 presente la moyenne des positions des formants et qu'il existe une variabilite significative de ces positions, même dans le cas d'un seul locuteur

Des chercheurs ont aussi trouve de fortes evidences que des groupes de neurones dans les aires auditives détectent le debut des signaux [Moore, 1997, Popper et Fay, 1992, Pickles, 1988, Brugge *et al*, 1969] En effet, on a remarque que certains groupes de neurones vont décharger plus fortement au début du stimulus Ce changement important dans le signal est évidemment une caractéristique forte utile à la reconnaissance De plus, le début d'un stimulus est habituellement peu affecté par la reverberation

En combinant ce mecanisme avec la detection de formants et leurs transitions, il est alors possible de classer certains types de consonnes (occlusives [Bandyopadhyay et Young, 2004, Ali *et al*, 2001] et fricatives [Ali *et al*, 1998]) en plus des composants voises de la parole

2 6 Réseaux de neurones artificiels et séquences d'impulsions

La détection des élements acoustiques peut être realisée à l'aide de neurones artificiels à impulsions [Maass et Bishop, 1999] de type *Integrate and Fire* (*IaF*) À son plus simple, ce modele de neurone artificiel de troisieme generation [Maass, 1997] suppose que le neurone genere une impulsion lorsque son potentiel atteint un seuil Ce potentiel, qui represente alors le potentiel de la membrane d'un neurone biologique, evolue en fonction du temps selon la somme des entrées du neurone La sortie du neurone peut aussi être simplifiée comme information binaire (presence ou absence d'impulsion) et permet de coder l'information sous forme d'une sequence temporelle d'impulsions Suite à la génération d'une impulsion, le neurone retourne à son etat initial ou dans une période refractaire pour les modeles plus complexes Pour certains modeles, il n'est pas permis au neurone de generer une impulsion lorsqu'il est dans une periode refractaire Pour d'autres, il est simplement plus difficile pour le potentiel du neurone d'atteindre le seuil de decharge

Un neurone *IaF* peut être alors adapte pour être sensible à un certain motif d'entree Dans le cas présent, ces motifs correspondraient à la representation des caractéristiques acoustiques Plus les entrees du neurone ressemblent à ce motif, plus le potentiel interne de ce neurone augmente et s'il atteint un seuil determiné, une impulsion est genérée Cette impulsion indique alors la présence, à ce moment, d'entrées similaires au motif à detecter

L'utilisation de détecteurs de caractéristiques qui imitent le fonctionnement d'un neurone (ou d'un groupe de neurones) permettrait en plus d'obtenir une parcimonie en ne conservant que les instants ou les zones interessantes [Feldbauer *et al*, 2005, Lewicki, 2002b] Il est certainement plus efficace d'avoir un groupe de detecteurs d'eléments acoustiques qui n'indiquent que les instants où une caracteristique est présente, que d'avoir à traiter continuellement le signal sans se soucier de son contenu De plus, une séquence d'evenements eparse est plus facilement classifiée dans un espace à grandes dimensions, qu'une sequence dense dans un espace aux dimensions limitées

Par contre, une sequence d'impulsions eparse n'est évidemment pas un type de données qui est adapte aux modeles statistiques conventionnels tels que les *HMMs* Certains auteurs contournent ce problème en modifiant ces sequences pour les utiliser avec un systeme conventionnel, mais une partie de l'information temporelle est alors perdue [Holmberg *et al*, 2005]

Il est aussi possible d'utiliser des algorithmes de regroupement de sequences comme *CLUSS* [Kelil *et al*, 2007] ou *MUSCLE* [Edgar, 2004] En effet, ces algorithmes recherchent des similarités dans les sequences pour former des groupes Ces similarites sont basees sur la taille et la fréquence des segments identiques entre deux sequences Cependant, ces deux algorithmes ont ete conçus pour le séquençage des protéines et à notre connaissance, ils n'ont pas ete appliqués à la reconnaissance vocale ou aux sequences d'impulsions

Deduire la structure des connexions d'un réseau qui genère une telle sequence d'impulsions en analysant la frequence des épisodes qui s'y trouvent peut aussi produire des bases qui permettraient d'en faire la classification Dans ce sens, l'algorithme presenté par D Patnaik [Patnaik *et al*, 2008] offre un potentiel intéressant En effet, cet algorithme permet de deduire differents types de patrons de connexions dans un reseau de neurones, en cherchant la fréquence des segments identiques dans les séquences produites par ce dernier Par contre, cet algorithme devient lourd à simuler pour des séquences complexes

Une approche plus naturelle pour classer une séquence d'impulsions est évidemment l'utilisation des reseaux de neurones à impulsions Pour gerer la distorsion temporelle du signal de la parole, des mecanismes de delais [Unnikrishnan *et al*, 1992] ou de recurrences [Izhikevich, 2006, Graves et Schmidhuber, 2005, Verstraeten *et al*, 2005] remplacent les etats des modèles conventionnels

À la suite de ces travaux, on peut voir qu'une séquence d'impulsions genérée par la detection combinee des *onsets* et des changements au niveau des formants permettrait de conserver l'information temporelle necessaire à la reconnaissance vocale, d'extraire des ca-

racteristiques robustes pour une meilleure performance dans des conditions adverses et elle serait compatible aux reseaux de neurones Au chapitre suivant, les etapes du traitement propose basees sur ces elements sont expliquees

CHAPITRE 3

VUE D'ENSEMBLE DU TRAITEMENT PROPOSÉ

À l'interieur de ce chapitre, un traitement est propose pour coder sous forme de séquence d'evenements, la detection de certains eléments caracteristiques de la parole, afin d'en faire la reconnaissance

3 1 Représentation du signal de la parole

La representation du signal acoustique selectionnee est basee sur un banc de filtres L'utilisation de ces filtres permet d'eviter les contraintes liees aux fenêtres fixes glissantes et represente plus fidelement, qu'une représentation basee sur la transformée de Fourier (figure 3 3), les signaux qui evoluent rapidement

3 1 1 Banc de filtres

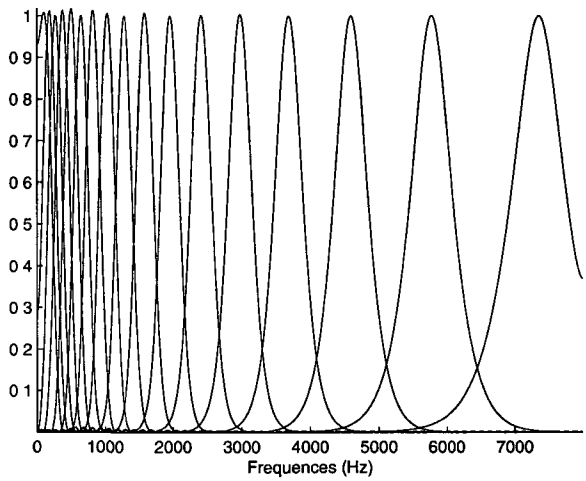

Figure 3 1 Un sur huit des cent vingt-huit filtres distribués pour couvrir 8 Khz

D'abord, le signal est filtré par un banc de filtres compose de 128 filtres à reponse impulsionnelle finie (une partie des filtres est illustree à la figure 3 1 et l'ensemble des frequences

centrales se trouve à l'annexe A) Ces filtres symetriques de 129 coefficients [1] sont distribues selon l'échelle *Bark* [Zwicker, 1961] pour couvrir les bandes critiques de l'audition

Le nombre de filtres est choisi afin de bien couvrir le spectre en fréquences du signal et il doit offrir suffisamment de diversité à l'étape de codage, que l'on retrouve plus loin dans le traitement Un nombre trop éleve de filtres devient très redondant en plus d'augmenter inutilement le temps de calcul et si ce nombre est trop limité, il sera difficile de bien caractériser les frequences

L'implementation en filtres a reponse impulsionnelle finie donne une phase lineaire aux sorties et permet une meilleure detection de coincidences pour les changements du signal qui s'etalent sur plus d'un canal

La transformee d'Hilbert de la sortie de ces filtres donne leur enveloppe et ainsi représente grossierement les deplacements de la membrane basilaire Ces enveloppes sont ensuite compressees par le calcul du logarithme pour reproduire l'effet de compression qui est observe au niveau de la cellule ciliée et du nerf auditif (figure 3 2)

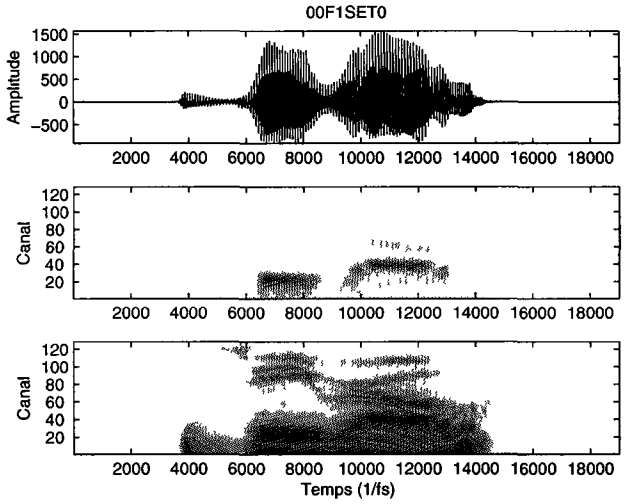

Figure 3 2 Signal de parole en fonction du temps d'une prononciation d'un *zero* en anglais (haut), enveloppe de la sortie des filtres (milieu) et résultat de la compression logarithmique (bas) La fréquence d'échantillonnage (f_s) est de 16 KHz et les canaux sont tries en ordre croissant selon leur fréquence centrale

[1] implementation de S Gagne, Y Liu et J Rouat, inspiree des travaux de R Patterson [Patterson, 1976]

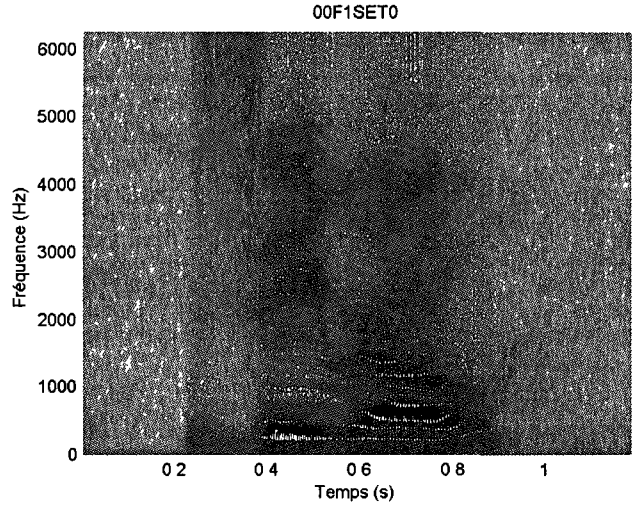

Figure 3 3 Spectrogramme de la prononciation d'un *zero* en anglais, calcule sur 1024 points, avec des intervalles de 128 echantillons et recouvrement de moitie, dont l'amplitude est compressee par un logarithme à base 10

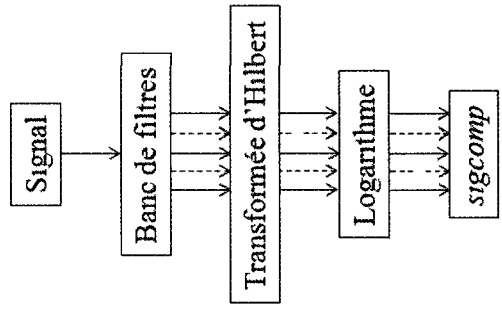

Figure 3 4 Étapes du traitement pour obtenir la representation temps/frequences du signal

3 1 2 Bases temporelles

Pour identifier les zones plus stables, periodes de temps où l'on n'observe peu ou pas de changement, du signal de parole des zones où il y a une rapide variation de l'amplitude, la representation du signal (*sigcomp* de la figure 3 4) est projetee sur des bases temporelles (figure 3 5)

Cette etape procede du même principe que celui que l'on retrouve dans l'article de D Zotkin [Zotkin *et al* , 2005] où l'on se sert de bases temporelles pour identifier les modulations temporelles Pour limiter la fréquence de variation possible et pour accélérer le traitement, la representation est d'abord sous echantillonnée à 1 KHz

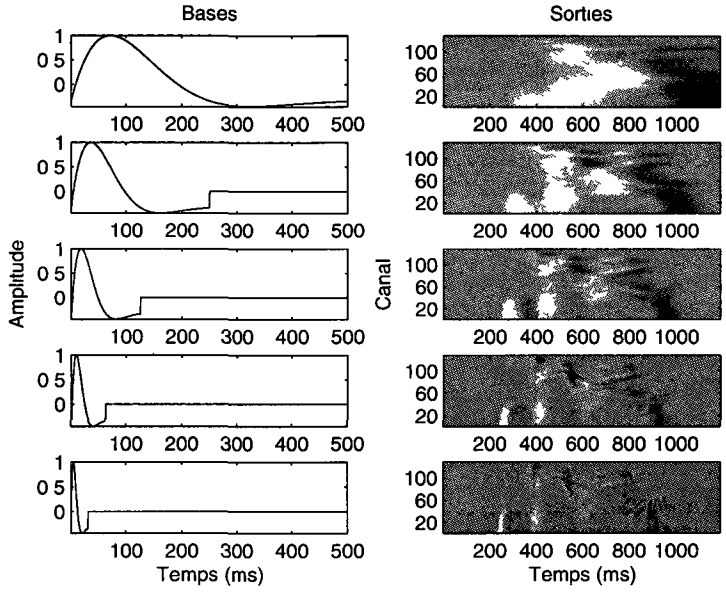

Figure 3 5 Les bases temporelles (gauche) filtrent chaque canal de la representation de la prononciation (bas de la figure 3 2) sous echantillonnee a 1 KHz et les résultats sont presentes (droite) Les bases sont triees selon leur frequence, de la plus petite (haut) à la plus elevee (bas)

Ces bases temporelles (b_{f_b} de l'equation 3 1) sont construites avec des sinusoides amorties de fréquences differentes (2, 4, 8, 16 et 32 Hz) Ces fréquences couvrent la zone qui correspond aux champs récepteurs spatio-temporels des neurones que l'on retrouve au niveau du cerveau des primates Les moyennes en amplitude de ces bases sont ensuite soustraites pour obtenir une somme en temps de 0 Chaque canal de la representation est filtre par ces cinq bases pour generer cinq nouvelles representations (sig_b)

$$\begin{aligned} tmp_{f_b} &= e^{-t} sin(2\pi f_b t), f_b = (2,4,8,16,32)Hz \\ b_{f_b} &= tmp_{f_b} - moyenne(tmp_{f_b}) \end{aligned} \tag{3 1}$$

Pour faire ressortir l'importance des changements rapides, l'amplitude des représentations est multipliée par un facteur La valeur de cette constante est choisie en fonction de la frequence de la base qui a genere la representation et elle augmente selon cette fréquence de telle sorte que les bases courtes sont légerement favorisees

Par la suite, on sélectionne le type d'information (évolution rapide ou lente du signal) le plus adequat pour caracteriser chaque instant (t) pour chaque canal (ch) en prenant la valeur la plus grande parmi les cinq sorties des filtres (figure 3 6), comme l'illustre

l'equation 3 2 et la figure 3 7 La sortie maximale des cinq filtres est alors conservée dans la matrice *maxsig*

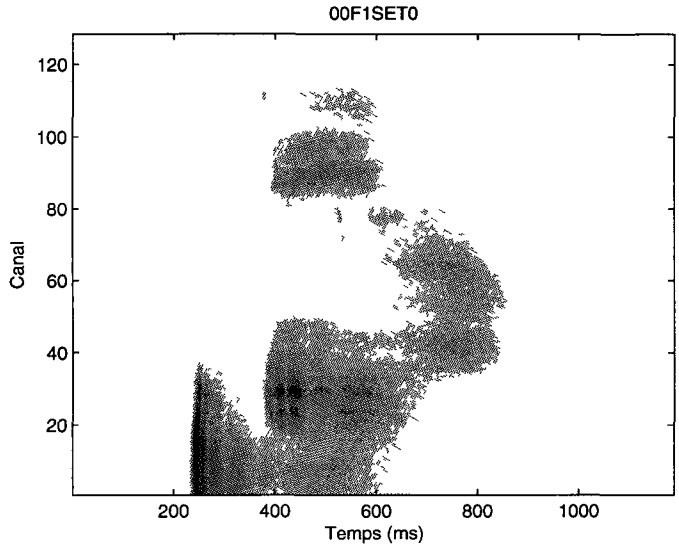

Figure 3 6 Maximum pour chaque instant et chaque frequence des sorties des bases temporelles (colonne de droite de la figure 3 5)

$$maxsig(ch,t) = \max_{b=2,4,8,16,32}(sig_b(ch,t)), ch = 1,2,3, \quad ,128 \tag{3 2}$$

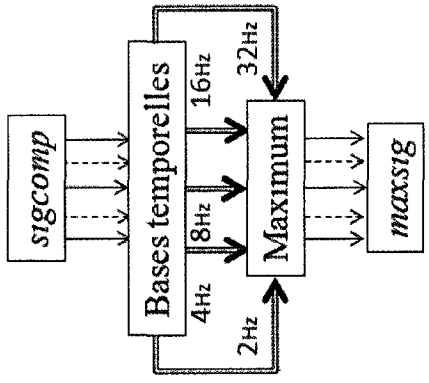

Figure 3 7 Étapes du traitement pour obtenir le maximum pour chaque instant et chaque frequence des sorties des bases temporelles

La representation obtenue fait bien ressortir les changements d'energies rapides, ainsi que les zones d'energies importantes du signal, sur une échelle en fréquences adaptée à la parole

3 2 Caractéristiques acoustiques recherchées

Maintenant que la representation est generee, il faut en extraire les caracteristiques (presentees à la section 2 5) qui vont servir à la reconnaissance Pour commencer, on trouve une approximation de la position des deux premiers formants ainsi que leur trajectoire dans le temps Ensuite, les changements importants du signal seront détectés

3 2 1 Localisation des formants

Detecter les pics d'energie importants en frequence permet d'obtenir de l'information sur la position possible de formants Pour y arriver, une base gaussienne etroite (figure 3 8) est deplacee en frequence sur les canaux de la representation a la figure 3 6 (*maxsig*) pour faire ressortir les maximums importants à chaque instant Si à un temps t, la partie gaussienne de la base couvre les canaux les plus importants de la représentation, le produit scalaire de cette base avec le vecteur des canaux de la représentation à cet instant donnera la valeur la plus importante La base gaussienne utilisée doit être relativement etroite pour conserver une bonne resolution en fréquences, mais elle doit aussi être suffisamment large pour eviter qu'un même pic d'energie soit detecte plus d'une fois Pour le traitement propose, des bases gaussiennes de moyenne 0 et d'ecart-type 2 seront utilisées

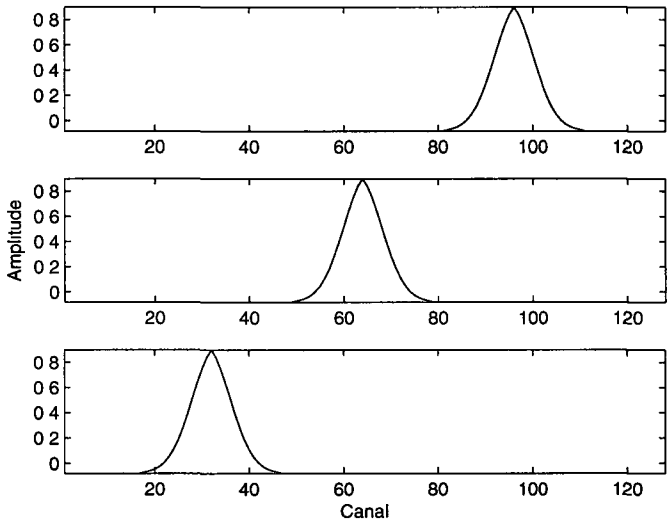

Figure 3 8 Base gaussienne de moyenne 0 et d'écart-type 2, utilisée pour la detection de pics d'energies importants en fréquence Pour ces exemples, elle est centree sur les canaux 96 (haut), 64 (milieu) et 32 (bas)

Dans un premier temps, la position du pic d'énergie le plus important à chaque instant est trouvee L'energie de la zone couverte par la base à cet endroit ainsi que l'energie des frequences inferieures sont ensuite soustraites à la représentation à cet instant, pour permettre la detection d'un second pic d'energie L'énergie de la zone soustraite à la représentation est cependant conservee comme valeur à la position du premier pic d'energie par la somme des énergies éliminées de la representation Par exemple, la zone des bases présentées à la figure 3 8 couvrent 31 canaux Dans le cas où la position du pic d'energie le plus important se situe au canal 32 à l'instant t, position qui correspond à la base de l'image du bas, la somme des énergies des canaux 1 a 42 est d'abord calculee, puis l'energie de ces canaux est mise à zero à l'exception du canal 32 qui prend la valeur de la somme calculee La position ainsi detectee est alors consideree comme une approximation du premier formant F_1 (milieu de la figure 3 9)

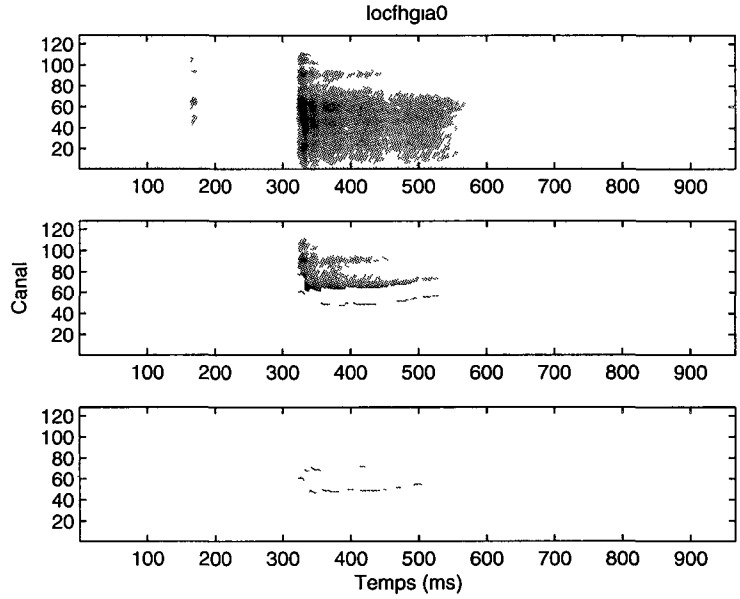

Figure 3 9 Représentation temps/frequences de la voyelle /a/ prononcee par la locutrice *hgi* (haut), résultat de la detection d'un premier pic d'énergie qui représente F_1 (milieu) et approximation de la position de F_2' (bas) La trajectoire en basses fréquences dans l'image du bas represente l'estimation de l'évolution en temps de F_1 et la trajectoire dans les frequences superieures represente celle de F_2'

Pour definir la position d'un second pic d'energie, on calcule le centre de masse de l'energie restante de la representation à chaque instant et on pondere encore cette deuxieme position par l'energie qu'elle represente (bas de la figure 3 9) Ce centre de masse (R de l'equation 3 3) est trouve en effectuant la somme des positions des canaux (P_{ch}), qui ne sont pas

affectes par le calcul du premier pic d'energie, ponderees par leur energie (e_{ch}) et divisee par la somme de ces energies L'energie des canaux est ensuite mise à zero a l'exception du canal R qui prend la valeur e_{ch} Cette information extraite est assimilee à la notion de F_2' introduite en psychoacoustique par des auteurs tels que L A Chistovich [Chistovich *et al*, 1978] et G Fant [Bladon et Fant, 1978, Carlson *et al*, 1970]

$$R = \frac{\sum e_{ch} P_{ch}}{\sum e_{ch}} \tag{3 3}$$

Dans le cas ou une vallee importante est detectee pres de la position du premier pic d'energie (troisieme image de la figure 3 10), la zone de detection du second pic sera decalée vers les hautes frequences au-dessus de cette vallee (quatrieme image de la figure 3 10)

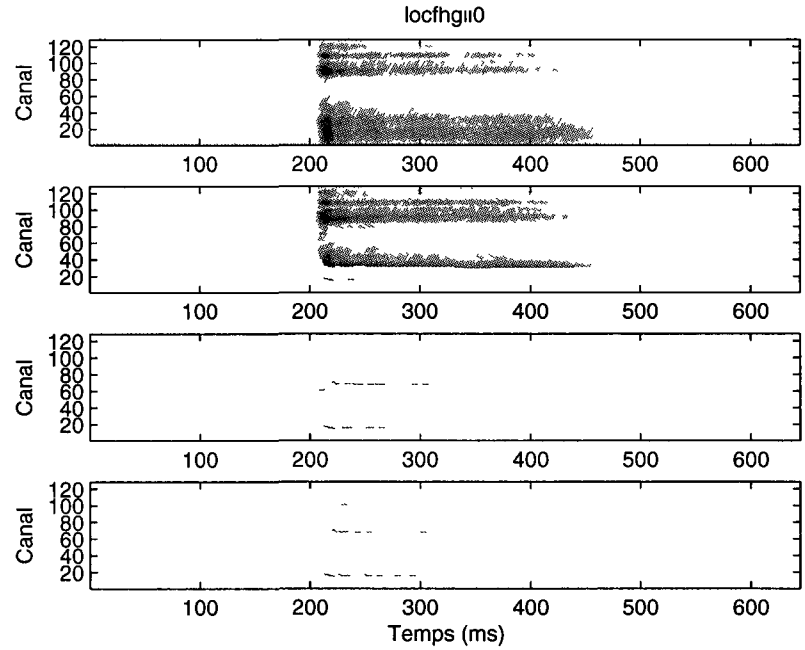

Figure 3 10 Représentation temps/fréquences de la voyelle /i/ prononcée par la locutrice *hgi* (haut), résultat de la détection d'un premier pic d'énergie qui represente F_1 (deuxieme image), detection d'une vallée importante (troisieme image) et approximation de la position de F_2' (bas) La trajectoire en basses fréquences dans l'image du bas représente l'estimation de l'évolution en temps de F_1 et la trajectoire dans les fréquences les plus hautes représente celle de F_2' Lorsque presente, la vallee importante apparaît comme un troisième trajectoire entre celles de F_1 et F_2'

Cette methode pour localiser des formants (figure 3 11) reste simple, mais comme on peut le voir a la sous-section suivante, elle fonctionne suffisamment bien pour classifier des voyelles isolees En effet, un prototype simple a éte conçu pour cette tâche et il peut parfaitement reconnaître les voyelles prononcées par une locutrice (*hgi*) lorsque les modeles sont génerés avec les prononciations de neuf autres locuteurs

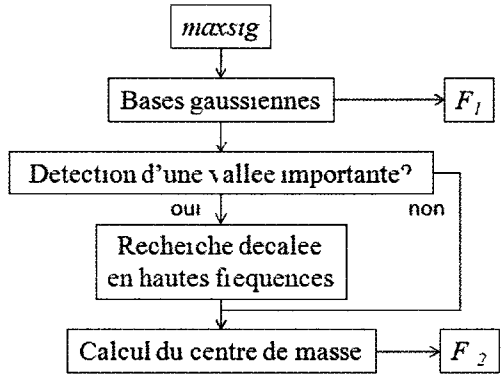

Figure 3 11 Étapes du traitement pour obtenir une estimation de F_1 et F_2'

3 2 2 Reconnaissance de voyelles isolées

Les prononciations des voyelles /a/, /e/, /i/, /o/ et /u/ de cinq locuteurs et cinq locutrices proviennent d'une petite base de donnees en français utilisee dans des travaux antérieurs [Loiselle, 2004] Ces dix prononciations isolees de chaque voyelle prononcees par chaque locuteur sont enregistrees sous format *wav* [IBM Corporation and Microsoft Corporation, 1991] et échantillonnées à 16 KHz Aucun bruit n'est ajoute à ces prononciations Par contre, il est a noter que ces enregistrements ont eté recueillis dans un bureau et à l'aide d'un equipement ordinaire, ce qui donne une qualite du signal moins bonne que celle habituellement trouvée dans les bases de données commerciales

Pour generer les cinq modeles (un par voyelle à reconnaître), on effectue la détection des pics d'énergie des prononciations de chaque voyelle pour les neuf locuteurs de l'apprentissage (exemple pour les prononciations precedentes du /a/ et du /i/ à la colonne de gauche de la figure 3 12) Ensuite, un vecteur est produit pour chaque représentation en effectuant la somme en temps des valeurs de chaque canal sur la duree complete de la prononciation (colonne de droite de la figure 3 12) Enfin, tous les vecteurs pour une même voyelle sont additionnes ensemble points par points et normalises, en divisant chaque valeur du vecteur resultant par la somme de ces valeurs La figure 3 13 presente les cinq vecteurs modèles utilisés pour la reconnaissance des voyelles prononcees par un dixième locuteur (la locutrice *hgi* dans cet exemple)

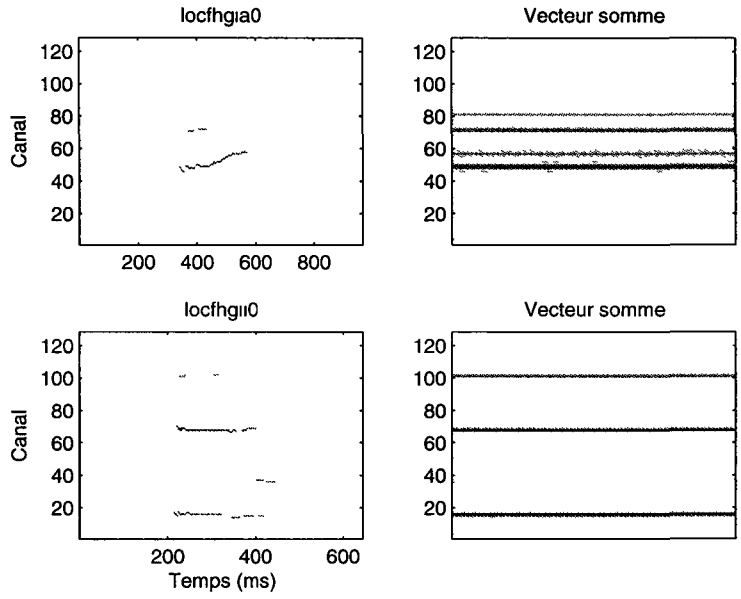

Figure 3 12 Detection de pics d'energie pour une prononciation du /a/ (en haut a gauche) et le vecteur produit par sa somme en temps (en haut à droite) Detection de pics d'energie pour une prononciation du /i/ (en bas à gauche) et le vecteur produit par sa somme en temps (en bas à droite)

La reconnaissance s'effectue en generant d'abord les vecteurs pour chaque prononciation à reconnaître Par la suite, on calcule la correlation de chaque vecteur produit avec les vecteurs modeles Si la plus haute valeur de ces corrélations provient bien du modele de la voyelle prononcee, celle-ci est consideree comme reconnue Comme indique precedemment, ce prototype simple reussit parfaitement a classer les voyelles /a/, /e/, /i/, /o/ et /u/ prononcees par la locutrice *hgi* lorsque l'apprentissage est basé sur les neuf autres locuteurs

L'estimation proposée des positions des deux premiers formants fonctionne pour réaliser une tâche simple de classification de voyelles isolées Par contre, la parole n'est pas composee que de voyelles isolées et la reconnaissance vocale est un probleme plus complexe (même lorsqu'elle se limite aux mots isolés comme dans ce travail de recherche) L'evolution des trajectoires reste cependant un bon indicateur quant à la nature du signal et permet d'identifier differents phonemes De plus, les transitions observees permettent de diviser le signal de parole en segments (partie voisée, non voisée ou transitoire) selon son contenu[Mariani et Lienard, 1977] L'étape suivante du traitement proposé consiste alors à caractériser ces trajectoires

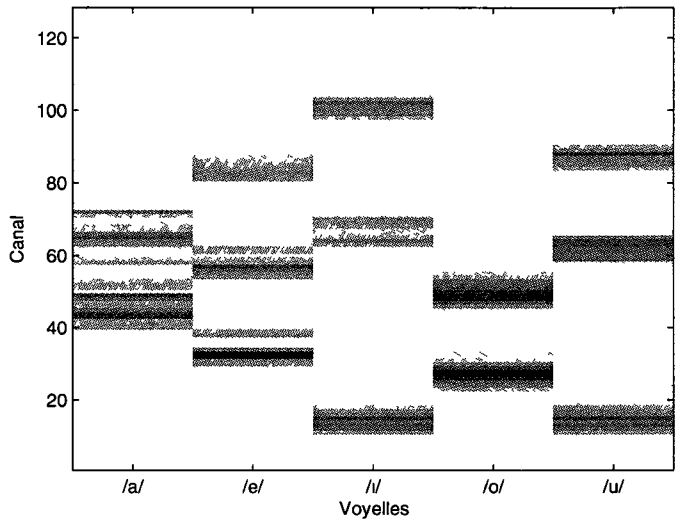

Figure 3 13 Vecteurs modeles des voyelles pour la reconnaissance des prononciations de la locutrice *hgi*

3 2 3 Identification des tendances des trajectoires

Pour caractériser davantage les trajectoires des pics d'énergie, les tendances ascendantes et descendantes sont aussi identifiées On qualifie de tendance ascendante, une augmentation en frequence de la position du pic d'énergie au cours du temps À l'inverse, une diminution en frequence de cette position est qualifiee de descendante Dans le cas où les changements en frequences d'une trajectoire sont minimes, cette trajectoire est considérée comme stable

En vision, on imite souvent le traitement du systeme visuel periphérique par l'application de bases de type filtres de Gabor (figure 3 14) à une image pour faire ressortir des contrastes importants ou des bordures [Daugman, 1980] On donne à ces filtres des orientations differentes et le filtre qui correspond le mieux à l'orientation la plus importante d'une partie de l'image, est celui qui produit la plus grande sortie à cet endroit Il n'est pas inconcevable qu'un mécanisme similaire se retrouve au niveau du systeme auditif

On fait donc ressortir les tendances stables, ascendantes et descendantes en filtrant la representation des trajectoires des pics d'energie avec des bases de type filtres de Gabor à deux dimensions Voici les étapes pour réaliser cette tâche

- on filtre la representation des trajectoires (les images au bas des figures 3 9 et 3 10 presentent un exemple de trajectoires pour les prononciations d'un /a/ et d'un /i/ et

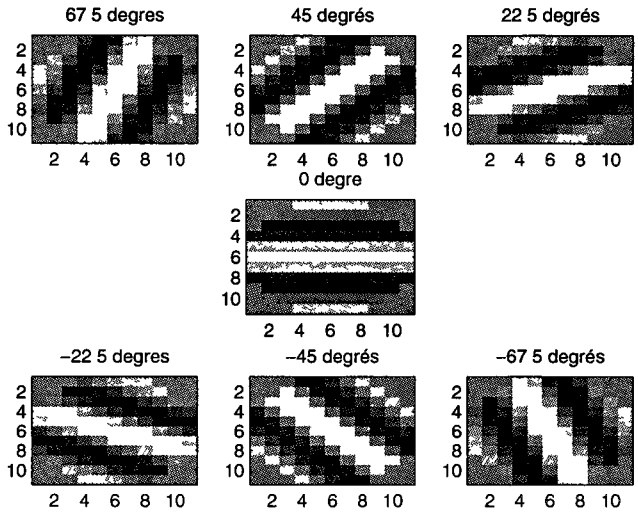

Figure 3 14 Les sept filtres a orientation de dimensions 11x11 utilises pour identifier les tendances des trajectoires Les filtres de Gabor orientés à 67 5 °, 45 ° et 22 5 ° (ligne du haut) répondent plus fortement à une trajectoire ascendante Le filtre de Gabor horizontal (ligne du milieu) repond plus fortement à une trajectoire stable Les filtres de Gabor orientes à −22 5 °, −45 ° et −67 5 ° (ligne du bas) répondent plus fortement à une trajectoire descendante Les zones pâles de ces figures sont positives et les zones foncées sont négatives

les images de la ligne du haut de la figure 3 15 illustre un exemple du filtrage d'une trajectoire simple par les bases ascendantes),

- pour chaque image à la sortie des filtres, on ne conserve que les valeurs qui se situent sur la position des trajectoires (ligne du milieu de la figure 3 15 pour les filtres ascendants),

- on regroupe les images des filtres de pente similaire en trois groupes (angle positif, angle negatif et filtre horizontal),

- pour chaque groupe, on ne conserve que la valeur maximale des trajectoires a chaque instant (ligne du bas de la figure 3 15 pour les filtres ascendants)

On obtient finalement trois versions des mêmes trajectoires, mais dont les valeurs sont modifiées par les filtres pour faire ressortir les tendances ascendantes pour une (image de gauche de la figure 3 16), les tendances descendantes pour l'autre (image de droite de la figure 3 16) et les tendances stables pour la dernière (image du milieu de la figure 3 16)

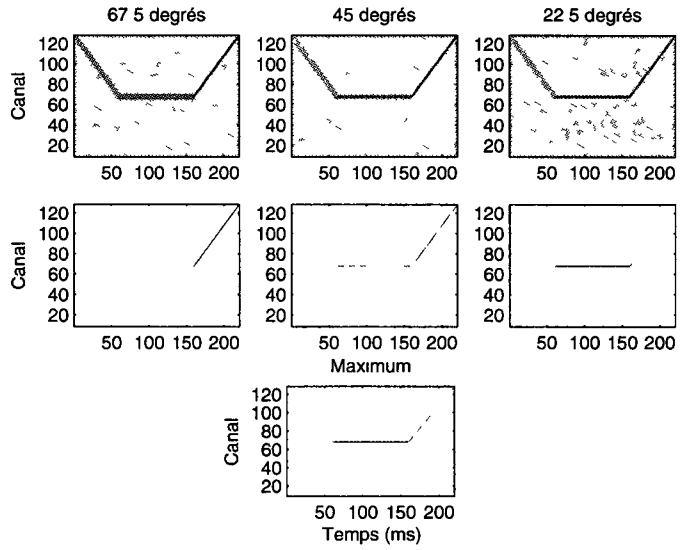

Figure 3 15 Exemple du filtrage d'une trajectoire simple (ligne du haut) par les bases ascendantes (les filtres de Gabor orientés à 67 5 °, 45 ° et 22 5 ° Seules les valeurs qui se situent sur la position des trajectoires sont conservées (ligne du milieu) et uniquement la valeur maximale des trajectoires à chaque instant pour le groupe des filtres ascendants est conserve (ligne du bas)

Ces tendances vont servir plus loin (sous-section 3 3 2) comme détection de transitoires Cependant, elles ne sont pas tres efficaces pour coder une caracteristique dont l'energie se trouve sur une courte periode de temps et qui couvre une large region en frequences, comme les fricatives et les plosives Pour cette raison, un autre type de détection sera utilise pour identifier les augmentations importantes d'energies sur une courte période de temps Cette etape permettra aussi de détecter le debut du signal de la parole, qui s'observe dans la représentation utilisee, comme une augmentation d'énergie importante simultanée de plusieurs canaux sur une courte période de temps

3 2 4 Détection des augmentations importantes d'énergies sur une courte période de temps

Des indices utiles à la reconnaissance vocale proviennent des changements importants ou rapides au niveau de la représentation spectro-temporelle Une augmentation d'energie importante peut indiquer le debut d'un phonème ou la transition rapide d'un formant

Grâce aux bases temporelles utilisées (b_{f_b}) dans le traitement proposé, ces changements importants de la representation ressortent bien En effet, lorsque le maximum d'amplitude

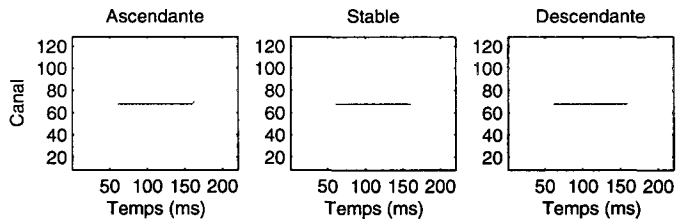

Figure 3 16 Resultat du filtrage de la trajectoire simple par le groupe des filtres ascendants (gauche), le filtre stable (milieu) et le groupe des filtres descendants (droite)

provient de la base la plus courte (b_{32Hz}), qui repond le mieux aux augmentations rapides d'energies, il y a de fortes chances qu'un evenement important (debut du mot ou d'un phoneme) ait eu lieu

Comme on peut le voir avec la prononciation du /a/ presentee précedemment (figure 3 17), conserver uniquement la base la plus courte, lorsque son amplitude dépasse celle des autres bases, nous permet d'en extraire son début

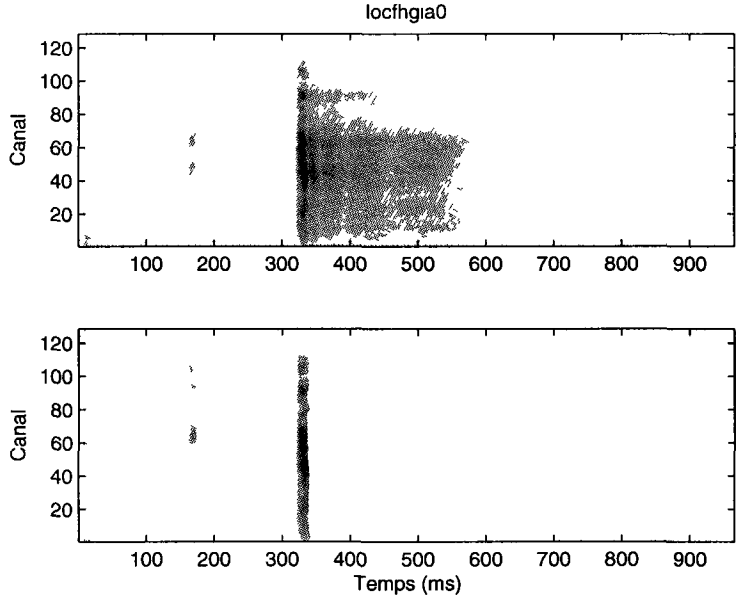

Figure 3 17 Representation temps/frequences de la voyelle /a/ prononcee par la locutrice *hgi* (haut) et zone du signal qui est le mieux caractérisée par la base b_{32Hz} (bas) Le debut de la voyelle correspond à la large bande en fréquences foncée qui se situe aux alentours de 330 millisecondes

Suite a des tests sur quelques prononciations, on remarque que cette approche n'isole pas parfaitement le debut de la prononciation et qu'une partie des bruits de fond dans le

silence est conservee Un mécanisme supplementaire est donc nécessaire pour améliorer la detection des changements importants

Une methode de detection des debuts des sons qui utilise des neurones à impulsions est proposee dans les travaux de L Smith [Smith et Fraser, 2004] et elle presente un potentiel intéressant pour améliorer cette etape de notre traitement

3.3 Neurone *Integrate and Fire* pour coder la détection des éléments acoustiques

À la section precedente, les etapes du traitement proposé pour estimer les trajectoires des formants et detecter les augmentations rapides d'energies importantes sont presentees Dans cette section, un mécanisme est presente pour coder l'information recueillie precedemment, en séquences d'evénements compatibles avec les classificateurs des approches connexionnistes

3 3 1 Codage des augmentations rapides d'énergies importantes

Dans son article [Smith et Fraser, 2004], L Smith propose d'utiliser des neurones à impulsions pour detecter le debut des zones d'energies importantes dans une representation bio-inspiree de la parole Pour effectuer cette tâche, le modele de neurone *IaF* utilise possède un courant de fuite (*Leaky Integrate and Fire* ou *LIaF*), une période refractaire et des synapses dynamiques Puisque le courant de fuite fait que le neurone sans stimulations importantes tend vers un état au repos, il est alors necessaire d'avoir l'arrivee simultanee de plusieurs entrees pour qu'il puisse decharger De plus, le neurone ne peut décharger qu'une seconde fois dans un intervalle de temps couvert par la periode réfractaire et les synapses dynamiques favorisent l'importance des premières valeurs d'entrées, ce qui limite la génération d'impulsions qu'au debut d'un stimulus de courte duree Des neurones *IaF* sont donc connectes à un certain nombre de canaux fréquentiels et s'il y a une énergie importante qui apparaît dans plusieurs canaux à des instants similaires, une impulsion est produite par les neurones concernes

Pour l'approche que l'on propose, un modele tres simple de neurone *LIaF* est implemente pour améliorer la detection des changements importants, de façon similaire à ce qui est presente dans le travail de L Smith Ces neurones sont donc connectés à seize canaux (huit neurones pour l'ensemble des 128 canaux) pour ainsi detecter la coincidence de zones d'energie importantes entre ces canaux Ce nombre est choisit suite à quelques tests

sur des prononciations où l'on a observe que seize canaux couvrait une région suffisamment grande en frequences et huit neurones offraient une diversite adéquate pour caractériser la séquence d'evenements produite

Contrairement au modèle presenté par L Smith, le modele de neurone utilise pour ce travail de recherche ne possede pas de synapses dynamiques, mais le courant de fuite du neurone s'adapte lorsqu'il y a generation d'une impulsion En effet, on n'utilise qu'une couche de neurones, alors que L Smith en utilise deux, et ils sont branches directement a la sortie des filtres Cette modification permet un temps de simulation plus court et comme on indique plus loin dans cette sous-section, elle permet une plus grande souplesse à la periode réfractaire

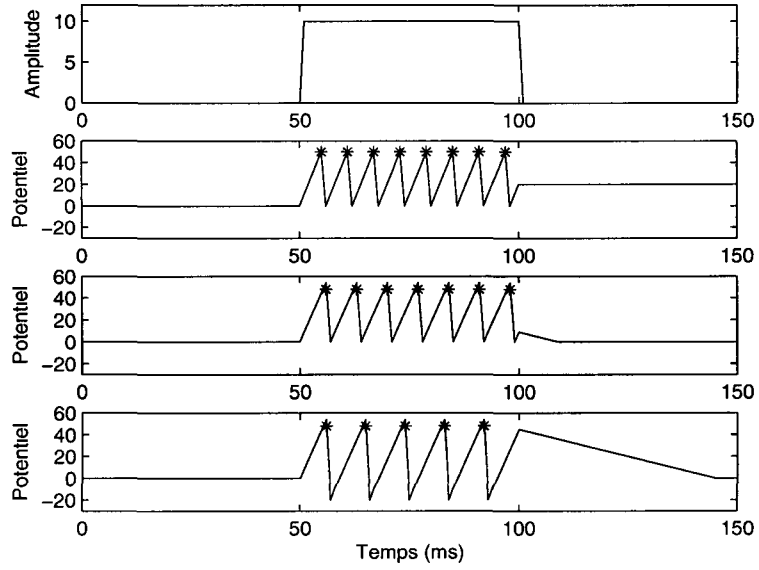

Figure 3 18 Comportement du neurone avec en entrée un stimulus d'amplitude 10 couvrant 50 millisecondes (haut), etat du neurone (deuxième figure), effet du courant de fuite qui fait tendre l'etat du neurone vers l'etat au repos (troisieme figure) et période refractaire qui diminue la sensibilité du neurone suite à une décharge (bas) Les * indiquent les instants où le potentiel interne du neurone est suffisamment eleve pour genérer une impulsion Les paramètres du neurone pour cet exemple sont seuil = 55, courant de fuite = 1 et valeur de la periode refractaire = -20

La figure 3 18 presente le comportement du modele simple de neurone à intégration et décharge (le symbole * marque la position des decharges) lorsqu'un stimulus simple (haut de la figure) lui est presente en entrée La troisième image illustre l'effet du courant de

fuite qui fait tendre l'etat du neurone vers l'etat au repos et la derniere image ajoute une periode refractaire

Le courant de fuite rend plus difficile la generation d'impulsions puisque l'absence de stimulus (ou pour un stimulus trop faible), le potentiel redescend tranquillement à son état initial Sans la presence simultanee de plusieurs entrées dont l'énergie est importante et soutenue sur une courte période de temps, le neurone ne décharge simplement pas

La periode refractaire implementee pour ce modele n'est pas absolue En effet, on permet au neurone de generer une decharge pendant sa periode refractaire, mais pour cet intervalle de temps determine, la sensibilite du neurone est diminuee significativement Le neurone permet alors la detection d'augmentations importantes de l'energie de ses entrees Cette modification est importante puisque pour le traitement proposé, on ne se limite pas qu'à la détection du début du mot Il est donc necessaire au neurone de conserver une sensibilite pour détecter des changements importants tout au long de la durée du signal de parole Pour arriver à ce comportement, on augmente le courant de fuite de façon importante suite a une decharge et il reprend graduellement sa valeur initiale par la suite

En resume, à chaque instant pour un neurone

- on fait la mise à jour de son etat en additionnant ses entrées,
- on calcule l'effet du courant de fuite,
- si le potentiel interne est supérieur au seuil, une impulsion est generée,
- on fait la mise à jour de la periode réfractaire

Comme on peut le voir a la figure 3 19, la detection des changements importants donne une bonne approximation du debut de la prononciation d'un /a/, lorsque le modèle du neurone est utilise avec la représentation obtenue à partir des bases b_{32Hz}

3 3 2 Codage des trajectoires des formants

Une méthode similaire est aussi appliquée aux représentations des trajectoires obtenues precedemment (sous-sections 3 2 1 et 3 2 3) dans le but de détecter les zones ascendantes, descendantes et stables importantes

Pour ce faire, on transforme la trajectoire qui represente F_1 en un vecteur (v_{F_1}) en ignorant la dimension en frequences (canal) et celle qui représente F_2' en un second vecteur $(v_{F_2'})$ pour les trois versions de cette representation (tendances ascendante, stable et descendante) Ces six vecteurs contiennent alors l'énergie de la trajectoire à chaque instant, modifiee

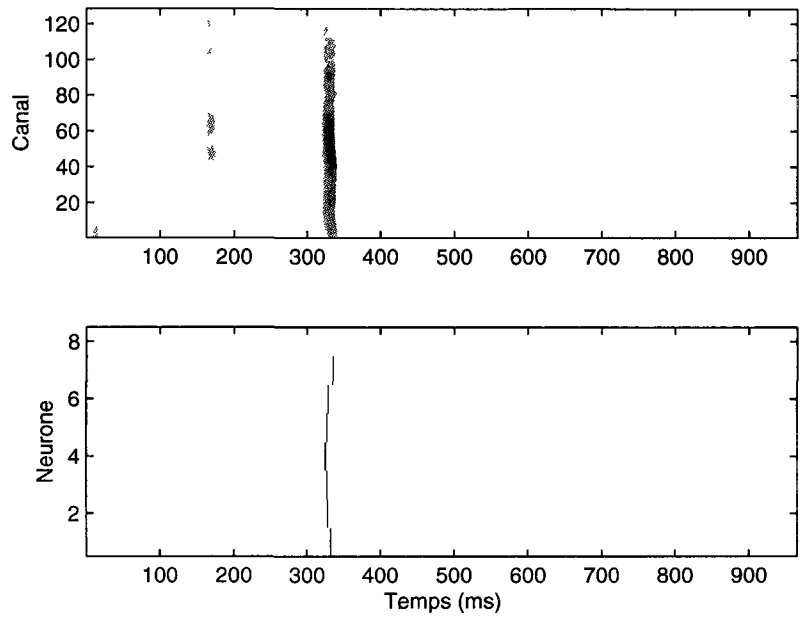

Figure 3 19 Detection de changements importants pour une représentation temps/frequences de la voyelle /a/ prononcee par la locutrice *hgi* (haut) et sorties des neurones *IaF* (bas) Dans l'image du bas, des traits noirs indiquent la génération d'impulsions et chaque neurone couvre 16 canaux

selon son orientation L'information qui represente l'énergie de la trajectoire, ainsi separee de l'information frequentielle, peut alors servir d'entree à un neurone *IaF* La figure 3 20 represente cette etape de façon simple pour la version stable des trajectoires

Il y a donc six vecteurs pour caracteriser les pics d'energie et chacun de ces vecteurs est consideré comme l'entree d'un neurone

- v_{F_1} ascendant,
- v_{F_1} descendant,
- v_{F_1} stable,
- $v_{F_2'}$ ascendant,
- $v_{F_2'}$ descendant,
- $v_{F_2'}$ stable

Dans le cas d'un silence ou d'un bruit faible, le courant de fuite du neurone prévient la génération d'impulsions Lorsque l'energie d'un pic est suffisamment importante, la trajec-

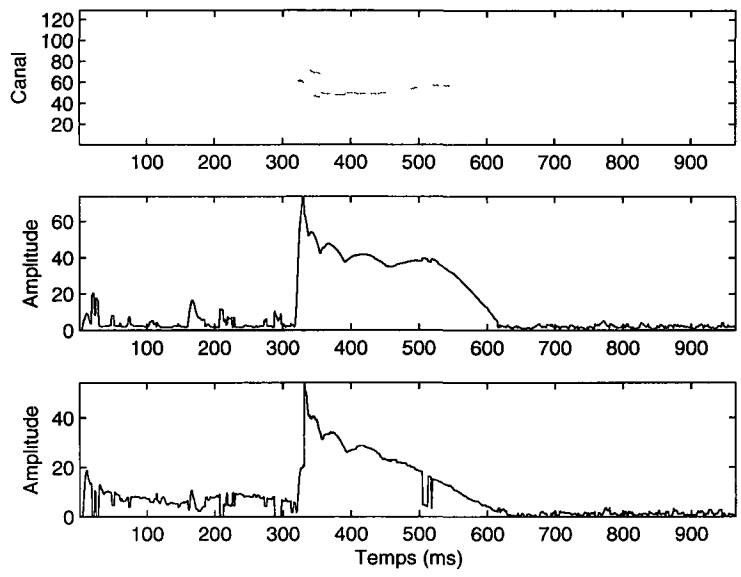

Figure 3 20 Version stable de la detection des pics d'energie pour une representation temps/frequences de la voyelle /a/ prononcee par la locutrice *hgi* (haut), transformee en vecteurs pour être presentées comme entrées à des neurones *IaF* (image du milieu pour le vecteur v_{F_1} et celle du bas pour le vecteur $v_{F_2'}$)

toire qui convient le mieux à la situation (ascendante, descendante ou stable) va generer une impulsion Il est possible que plus d'une trajectoire parviennent à declencher la génération d'impulsion si l'energie est particulièrement importante Par contre, l'ordre dans lequel elles seront produites va debuter par l'orientation qui est favorisee, une caracteristique qui peut d'ailleurs être assimilee à la notion du codage par ordre de rang (*rank order coding*) [Thorpe *et al*, 2001]

Enfin, dans cette section on a montré comment les caractéristiques du signal sont codees en séquence d'impulsions D'abord, huit neurones de type *IaF* signalent la detection de changements d'energie importants et six autres neurones donnent les tendances montantes, stables ou descendantes des zones d'energie importante des trajectoires des formants (F_1 et F_2')

L'utilisation des caracteristiques acoustiques presentees dans ce chapitre pour effectuer la reconnaissance vocale, se retrouve dans la litterature Par contre, l'originalite de ce travail de recherche se situe dans la caracterisation des tendances des trajectoires des formants ainsi que la combinaison de ces detecteurs avec ceux des changements d'energie importants De plus, ces impulsions peuvent ensuite être considérees comme une séquence

d'évenements ımportants pour la realısatıon de la reconnaıssance vocale La performance de notre approche est d'aılleurs presentée dans le chapıtre suıvant

CHAPITRE 4

RECONNAISSANCE DES CHIFFRES ISOLÉS

À travers ce chapitre, les performances de l'approche proposee sont comparees a celles d'un systeme de référence D'abord, on effectue les tests avec un systeme conventionnel sans modification Puis, le traitement presenté au chapitre 3 est adapté à ce systeme dans le but d'améliorer les performances

4 1 Matériel

Les tests sont effectues sur un poste de travail *Intel Core2 Duo* dans un environnement *Windows Vista* Le prototype et le systeme de reference sont developpes dans *Matlab R2006a* en utilisant les outils supplementaires *Voicebox* [Mike Brookes, 2009] et *H2M* [Olivier Cappe, 2001] pour supporter le calcul des coefficients cepstraux (*MFCCs*) et les modeles de Markov (*HMMs*) respectivement

4 1 1 Base de données

Le test utilise pour comparer la performance de l'approche proposee à celle d'un systeme conventionnel consiste à reconnaître les chiffres de zéro à neuf en anglais

Pour ce faire, on utilise la base de donnees *TI 46-Word* qui contient entre autres plusieurs prononciations isolees des chiffres prononces par huit locuteurs et huit locutrices en anglais [Doddington et Schalk, 1981] Ces prononciations sont divisées en deux groupes, *TRAIN* (tableau B 1 en annexe) pour les donnees d'apprentissage et *TEST* (tableau B 2 en annexe) pour la reconnaissance On retrouve dix prononciations de chaque chiffre pour chaque locuteur dans le premier groupe et seize prononciations de chaque chiffre pour chaque locuteur dans le second groupe La parole est enregistree dans un environnement silencieux et le signal est sauvegarde dans le format *NIST SPHERE* à 12 5 KHz avec une quantification de 12 *Bits*

En plus d'offrir des enregistrements de bonne qualité, la base de donnees *TI 46-Word* a été selectionnee parce que chaque prononciation d'un chiffre est enregistree seule dans un fichier, alors que les enregistrements de certaines bases de donnees (comme *TIDIGITS* [R Gary Leonard and George Doddington, 1993]) contiennent des séquences de chiffres

De plus, cette base de donnees est aussi utilisée par d'autres chercheurs comme A Ghani [Ghani *et al* , 2008], M Skowronski [Skowronski et Harris, 2007], D Verstraeten [Verstraeten *et al* , 2005] et A Graves [Graves *et al* , 2004]

Dans le cas de tests en conditions adverses, uniquement les donnees *TEST* sont modifiees par le bruit ou la reverberation

4 2 Modèles pour la reconnaissance des chiffres isolés

Pour modeliser chaque mot, un modele à base de chaîne de Markov (*HMM*) à cinq etats (N), ou les transitions vers les deux prochains voisins suivants sont permises, est utilise et l'algorithme de maximum de vraisemblance (*Expectation-maximisation algorithm* [Wu, 1983, Dempster *et al* , 1977]) permet d'effectuer l'apprentissage Puisque la quantité de données pour l'apprentissage est limitée (dix prononciations d'un chiffre par huit locuteurs) et que les vecteurs d'observations (o_t) contiennent plusieurs coefficients, on force l'utilisation d'une matrice de covariances (Σ) diagonale et la densite de probabilite des observations est une gaussienne simple

En effet, une fonction de densite de probabilite normale de moyenne μ et d'ecart type σ donne la probabilité d'obtenir un coefficient du vecteur d'observations o_t à chaque état j Pour un nombre M de coefficients, il est alors necessaire d'apprendre les paramètres d'un vecteur $\vec{\mu}$ de dimension M et d'une matrice de covariance Σ de dimensions M par M Pour limiter le nombre de ces parametres, on suppose qu'il n'existe pas de correlation entre les différentes dimensions du vecteur d'observations o_t, ce qui nous permet de ne retenir que les valeurs de la diagonale de la matrice de covariances Σ Cette simplification est bien repandue dans le domaine de la reconnaissance vocale et aide aussi a diminuer le temps de calcul necessaire a l'apprentissage [Jurafsky et Martin, 2008]

À la reconnaissance, l'algorithme de Viterbi nous permet de calculer la sequence d'etats la plus probable d'avoir généré les observations pour chaque modèle et le modele qui possède la sequence la plus probable est selectionne comme chiffre prononcé

4 3 Système de référence

Pour le systeme de reférence, des *MFCCs* sont utilises comme observations pour les modeles Douze coefficients sont extraits pour chaque trame de 20 millisecondes et ces trames se recouvrent de moitie Le logarithme de l'energie de la trame est aussi conserve comme treizieme coefficient De plus, on calcule les derivées premières et secondes en temps de

ces treize valeurs pour obtenir les coefficients Delta et Delta-Delta Les vecteurs d'observations ont donc 39 dimensions et contiennent l'information generalement utilisee dans les systèmes de reconnaissance vocale conventionnels [Jurafsky et Martin, 2008] Dans le cas de prononciations isolées des chiffres, les sequences d'observations ainsi générees sont, en moyenne, composee d'une centaine de vecteurs

4 3 1 Reconnaissance de chiffres isolés sur données propres

Dans un premier temps, on effectue la reconnaissance des chiffres avec le systeme de reference directement sur les prononciations de la base de donnees Les modeles sont d'abord appris avec les donnees d'apprentissage et ils servent ensuite à reconnaître les prononciations du groupe *TEST* (tableau 4 1) Une prononciation est bien classee si le modele du chiffre qui est prononcé correspond au modele selectionné

Tableau 4 1 Reconnaissance des chiffres sur donnees propres avec le systeme de réference

99 80%	Fichiers									
Modeles	0	1	2	3	4	5	6	7	8	9
0	249									
1		255								
2			255							
3				254						
4					254					
5						253				
6							251			
7	1							256		
8							3		256	
9						1				254

Dans des conditions optimales, le systeme conventionnel parvient presque a reconnaître parfaitement les chiffres (99 80%) Évidemment, il faut s'attendre à d'excellents resultats dans le cas où les caractéristiques des données d'apprentissage et de reconnaissance sont identiques

4 3 2 Reconnaissance de chiffres isolés avec bruit blanc gaussien

Pour le test suivant, on ajoute un bruit blanc gaussien à differents niveaux d'intensité (SNR 20 dB, 10 dB, 0 dB et -10 dB) aux donnees de reconnaissance pour évaluer la robustesse du système face au bruit La puissance spectrale de ce bruit est pratiquement la même pour toutes les fréquences et l'amplitude du bruit respecte une distribution gaussienne

Le rapport de la puissance du signal sur la puissance du bruit (*RSB* en français ou *SNR* en anglais), donne en decibels (*dB*), permet alors de quantifier le niveau de bruit dans un signal (voir equation 4 1) Plus le rapport est petit, plus le signal est corrompu par le bruit et plus la reconnaissance sera difficile Pour tous les tests de cette section, le niveau de bruit est mesuré sur le signal au complet

$$SNR_{dB} = 10log_{10}(\frac{P_{signal}}{P_{bruit}}) \quad (4\ 1)$$

Tableau 4 2 Reconnaissance des chiffres avec le systeme de référence et un bruit blanc gaussien (*SNR* 20 *dB*)

57 75%	Fichiers									
Modeles	0	1	2	3	4	5	6	7	8	9
0	205	4								1
1		26								
2			25							
3			52	238						1
4		10			160					
5	39	215	25		94	254		120	6	247
6	5		149	16			248	63	15	
7	1		3					73		1
8			1				6		235	
9										4

Tableau 4 3 Reconnaissance des chiffres avec le systeme de réference et un bruit blanc gaussien (*SNR* 10 *dB*)

22 19%	Fichiers									
Modeles	0	1	2	3	4	5	6	7	8	9
0	1									
1										
2										
3				48						
4										
5	63	183			103	254		27		248
6	186	72	255	206	151		254	229	249	6
7										
8									7	
9										

Dans des conditions bruitées (tableaux 4 2, 4 3, 4 4 et 4 5), on remarque une diminution importante de la performance du systeme conventionnel En effet, le taux de reconnaissance chute pratiquement de 40% pour atteindre 57 75% par rapport au test initial (99 80%) pour

Tableau 4 4 Reconnaissance des chiffres avec le systeme de référence et un bruit blanc gaussien (SNR 0 dB)

10 90%	Fichiers									
Modèles	0	1	2	3	4	5	6	7	8	9
0										
1										
2										
3										
4										
5	32	46			9	23				80
6	218	209	255	254	245	231	254	256	256	174
7										
8										
9										

Tableau 4 5 Reconnaissance des chiffres avec le systeme de reference et un bruit blanc gaussien (SNR -10 dB)

10 07%	Fichiers									
Modèles	0	1	2	3	4	5	6	7	8	9
0										
1										
2										
3										
4										
5	15	30	6	6	1	2		1		44
6	235	225	249	248	253	252	254	255	256	210
7										
8										
9										

un SNR de 20 dB et continue de diminuer suivant l'augmentation du bruit (22 19% à 10 dB, 10 90% à 0 dB et 10 07% à -10 dB)

Les confusions observees sont surtout produites par les modeles du *five* et du *six* Puisque ces deux chiffres débutent par une consonne fricative (/f/ et /s/), dont les caracteristiques phonémiques se rapprochent à celles d'un bruit, il n'est pas surprenant de voir l'ensemble des prononciations converger vers ces modèles, suivant l'augmentation du bruit

4 3 3 Reconnaissance de chiffres isolés dans différents environnements bruités

Pour évaluer la performance dans un contexte plus realiste, on poursuit les tests en ajoutant cette fois, des bruits provenant d'enregistrements (echantillonnes à 8 KHz et codes sur 8 $Bits$) de situations réelles (à bord d'un vehicule ou dans un endroit public) La base de donnees *AURORA* [Hirsch et Pearce, 2000] propose huit environnements différents (aeroport, conversation, voiture, exposition, restaurant, rue, metro et train) On ajoute alors au signal propre de la prononciation du chiffre, le bruit désire au niveau souhaite (SNR 20 dB, 10 dB, 0 dB et -10 dB)

Tableau 4 6 Reconnaissance des chiffres avec le système de référence dans différents environnements bruités

Type d'environnement	20 dB	10 dB	0 dB	-10 dB
Aeroport	70 57%	40 13%	17 47%	11 76%
Conversation	66 72%	30 25%	14 12%	10 46%
Voiture	68 88%	28 76%	10 27%	9 99%
Exposition	60 15%	17 82%	15 77%	10 70%
Restaurant	66 01%	31 47%	19 24%	14 79%
Rue	65 82%	21 75%	18 41%	14 72%
Métro	60 23%	39 06%	16 56%	10 86%
Train	72 42%	42 68%	18 84%	13 41%

Comme pour le bruit blanc, la reconnaissance dans ces milieux bruites (tableau 4 6) devient difficile et la performance diminue rapidement lorsque le niveau de bruit augmente, que la source de bruit soit liee à du bavardage, des bruits mecaniques, de la réverberation ou un melange de ces elements

Cependant, les confusions observées convergent cette fois vers les modèles du *nine* et du *zero* Ces prononciations sont habituellement les plus longues et il semble que le systeme de référence ait de la difficulte à isoler le chiffre prononcé dans le signal fortement bruite La duree du signal complet deviendrait alors le critère de classification

4 4 Reconnaissance de chiffres isolés avec le traitement proposé

Pour evaluer la performance de notre traitement en reconnaissance vocale, un prototype a été conçu pour adapter la sequence d'impulsions produite aux *HMMs* En effet, un vecteur d'observations peut être genére à chaque instant où une impulsion est produite

L'information sur le type de composant qui a generé l'impulsion n'est pas conservee pour ce type d'experience Par contre, on sait que chaque vecteur represente la detection de pics d'energie ou un changement important La sequence d'impulsions devient alors une forme de selection des trames

Les impulsions seront donc génerees à partir de la représentation auditive des signaux comme presenté au chapitre précedent (section 3 1) Par contre, puisque la génération des impulsions est sensible au niveau d'energie de la représentation, il est nécessaire de conserver un niveau d'energie similaire à travers l'ensemble des prononciations

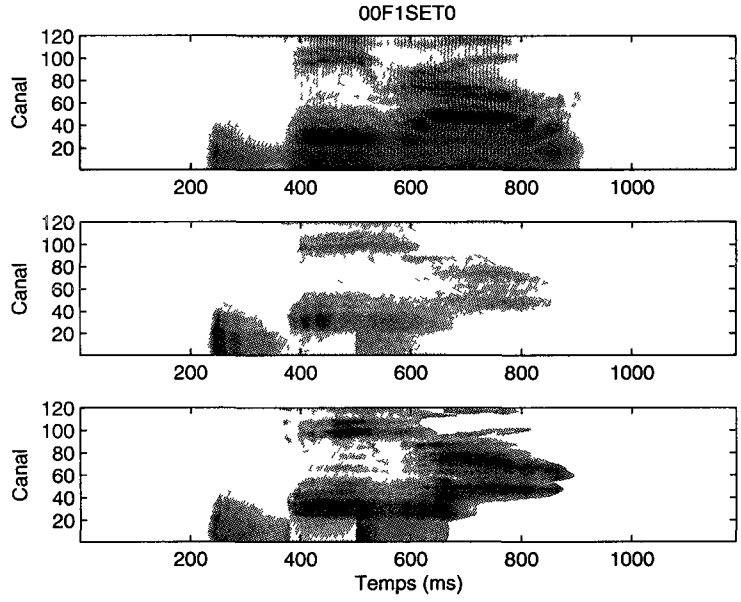

Figure 4 1 Sortie rectifiee et compressee du banc de filtres (haut), maximum des projections sur les bases temporelles b_{f_b} (milieu) et application comme masque sur la representation initiale (bas)

Pour ce faire, la sortie conservee des bases temporelles (*maxsig* de l'équation 3 2) devient un masque (milieu de la figure 4 1) que l'on multiplie point par point à la representation du signal (haut de la figure 4 1) À chaque instant t, les valeurs de ce masque pour chaque canal *ch* sont divisees par la valeur maximale contenue dans ces canaux et à cet instant On obtient alors un masque normalise dont les valeurs se situent entre 0 et 1 La representation finale possede donc le niveau d'energie de la représentation initiale, mais cette énergie ne se concentre que dans les zones importantes de la sortie conservee des bases temporelles b_{f_b} La figure 4 2 presente cette manipulation a un instant t pour un exemple simple

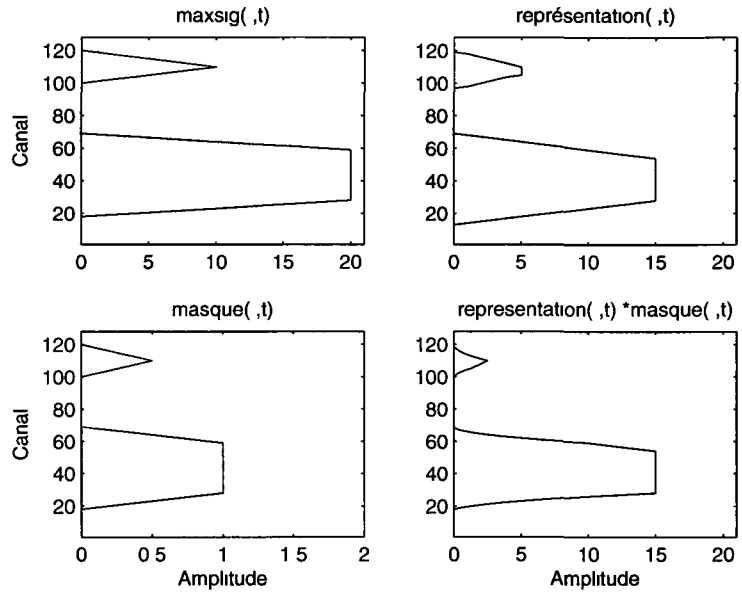

Figure 4 2 Exemple simplifie d'un maximum des projections sur les bases temporelles b_{f_b} a un instant t (en haut a gauche), exemple simplifie de la representation temps/frequences à un instant t (en haut a droite), normalisation du masque (en bas à gauche) et application de ce masque sur la representation initiale a cet instant (en bas à droite)

Par la suite, les impulsions sont generees à partir de la representation ainsi masquée (figure 4 3)

Une autre forme de masquage, basee sur une segmentation temps-fréquences, a aussi éte consideree dans le but de mieux isoler les elements importants du signal Par contre, cette etape necessitait un temps de calcul plus long et n'ameliorait pas les performances Bien qu'elle ne soit pas utilisee dans les tests, cette manipulation est brievement presentee au chapitre suivant (chapitre 5), car elle pourrait offrir un potentiel interessant en analyse et codage du son

Les impulsions génerees à partir de la representation masquee marquent les instants importants (figure 4 4) Un vecteur de coefficients doit être maintenant génere à ces instants pour servir d'observations aux *HMMs*

Dans l'application de la sequence d'impulsions sous forme de selection des trames, on fait l'hypothese que l'ensemble du spectre est interessant, c'est-à-dire qu'il n'y a pas de decoupage spatial On n'effectue qu'une forme de segmentation temporelle pour considerer au complet le spectre du signal lorsqu'une impulsion est generee Des trames de 20 milli-

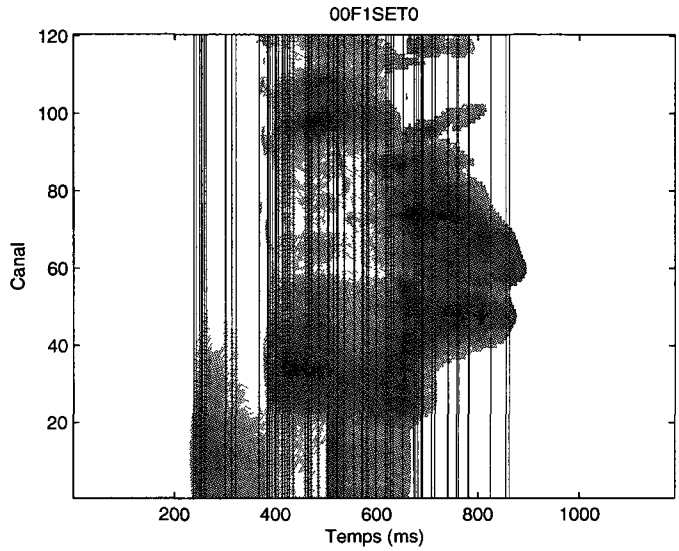

Figure 4 3 Impulsions générées à partir d'une representation masquée (traits verticaux) Ces impulsions se concentrent surtout au debut des zones d'énergies importantes et lorsqu'un transitoire est detecte

secondes sont alors centrees sur le signal aux positions indiquées par le traitement (figure 4 5) Pour chacune de ces trames, on calcule ensuite douze coefficients cepstraux, le logarithme de l'énergie de la trame ainsi que les coefficients Delta et Delta-Delta On tient à souligner que le calcul de ces coefficients est exactement le même que pour le systeme de reférence (section 4 3)

Les modèles des chiffres utilisés possedent aussi les mêmes caractéristiques que ceux utilisés a la section precedente (section 4 2) Chaque chiffre est modélisé par un modèle à base de chaîne de Markov à cinq etats, où les transitions vers les deux prochains voisins suivants sont permises On utilise une matrice de covariances diagonale et la densité de probabilite des observations est une gaussienne simple

Pour les paramètres des neurones (tableau 4 7), les valeurs sont choisies suite à des essais sur quelques prononciations et elles restent constantes pour l'ensemble des tests Les paramètres des neurones responsables de la détection des éléments formantiques et de leurs transitions sont differents de ceux des neurones utilisés pour la detection des *onsets* Ces derniers doivent être plus sensibles à l'augmentation d'énergie pour arriver à détecter une caracteristique plus limitee en temps

Il faut noter que pour la reconnaissance des voyelles (sous-section 3 2 2), les signaux étaient echantillonnés à 16 KHz alors que les prononciations des chiffres sont echantillonnées à

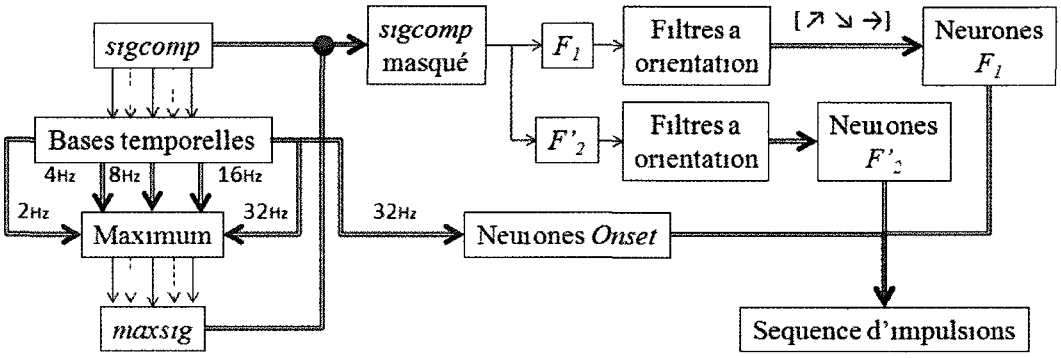

Figure 4 4 Étapes du traitement pour obtenir la sequence d'impulsions

Tableau 4 7 Parametres des neurones

	F_1 et F'_2	*Onset*
Seuil	55	20
Courant de fuite	7	3
Periode refractaire	-150	-10

12 5 KHz Pour respecter cette frequence d'echantillonnage, seulement les 120 premiers filtres sont utilises dans les tests suivants

4 4 1 Reconnaissance de chiffres isolés sur données propres

Dans un premier temps, on evalue la performance de notre approche sur des donnees propres

Tableau 4 8 Reconnaissance des chiffres sur donnees propres avec le prototype

97 80%	Fichiers									
Modèles	0	1	2	3	4	5	6	7	8	9
0	247									2
1		251				1				1
2	3		253				1			
3				254					4	
4		1			254					
5						240				2
6			2				243	2	3	
7		3				8		253		7
8							10		249	
9						5		1		242

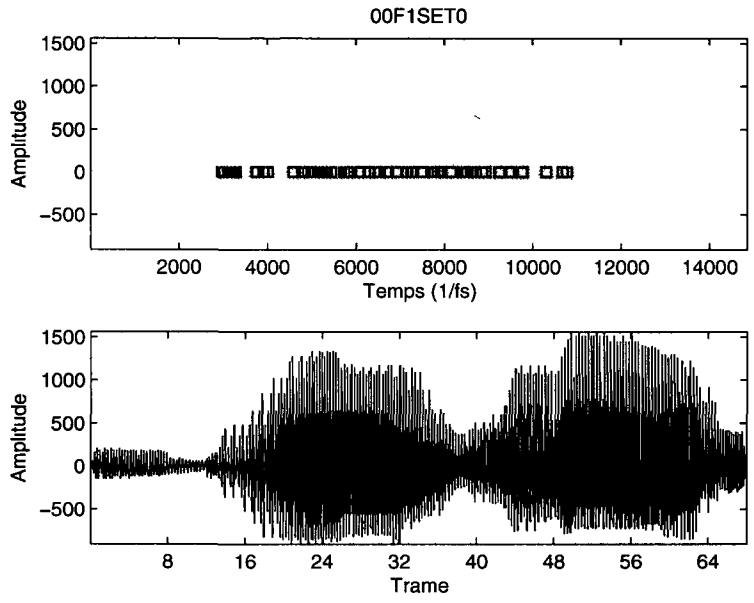

Figure 4 5 Selection des trames (carres) effectuée par le prototype sur une prononciation d'un zero (haut) et le resultat de ces trames mises bout a bout (bas)

Comme on peut le constater, l'approche proposée réussit bien à reconnaître les chiffres dans des conditions optimales à 97 80% (tableau 4 8) De son côte, le systeme de réference (tableau 4 1) offre une performance légerement superieure a 99 80%

La difference entre les deux systèmes est probablement due au fait que notre prototype ne genere des vecteurs d'observations que lorsqu'un evenement interessant a lieu En effet, on note habituellement une diminution du nombre de vecteurs produits par la selection des trames En moyenne, l'approche proposee génere 93 vecteurs pour une prononciation alors que le systeme de reference en genere 107

4 4 2 Reconnaissance de chiffres isolés avec bruit blanc gaussien

Comme pour le systeme de reference, on ajoute un bruit blanc gaussien à différents niveaux d'intensite (SNR 20 dB, 10 dB, 0 dB et -10 dB) aux données de reconnaissance pour evaluer la robustesse face au bruit du systeme propose

Le taux de reconnaissance diminue à 94 34%, 67 19%, 38 99% et 9 28% lorsque le rapport signal à bruit tombe à 20 dB (tableau 4 9), 10 dB (tableau 4 10), 0 dB (tableau 4 11) et -10 dB (tableau 4 12) respectivement

Tableau 4 9 Reconnaissance des chiffres avec le prototype et un bruit blanc gaussien (*SNR* 20 *dB*)

94 34%	Fichiers									
Modèles	0	1	2	3	4	5	6	7	8	9
0	243	3								3
1		215								
2	3		245	24						
3				225					1	
4					251					
5					2	250		3		13
6			6	2			236		1	
7	4	12	4	1		2		252		11
8					1		18		254	
9		25		2		2		1		227

Tableau 4 10 Reconnaissance des chiffres avec le prototype et un bruit blanc gaussien (*SNR* 10 *dB*)

67 19%	Fichiers									
Modeles	0	1	2	3	4	5	6	7	8	9
0	199	70	36	36	21	1			2	22
1		7								
2	16		143	31						
3				141					1	
4		3			129					
5		5			23	251			1	36
6	4		36	21			239	4	7	
7	31	145	35	24	53	1		251	1	92
8		1	5		27		15	1	244	
9		24		1	1	1				104

De son côté, la performance du système de reférence diminue plus rapidement pour les mêmes rapports signal à bruit 57 75% à 20 *dB*, 22 19% à 10 *dB*, 10 90% à 0 *dB* et 10 07% à -10 *dB* (tableaux 4 2, 4 3, 4 4 et 4 5)

Les neurones detectent assez bien les instants où le signal reste dominant, ce qui fait que la selection des trames aide à cerner les elements importants du signal Par contre, plus le rapport signal a bruit diminue, plus il est difficile de separer le signal du bruit par leur énergie et moins de vecteurs sont generes Comparé au systeme de référence, le nombre de vecteurs d'observations utilisé par notre prototype ne cesse de diminuer pour n'en produire en moyenne que la moitie à -10 *dB* (82 vecteurs à 20 *dB*, 71 vecteurs à 10 *dB*, 62 vecteurs à 0 *dB* et 52 vecteurs à -10 *dB*) De plus, des impulsions peuvent être generées par le bruit lorsque son energie devient suffisamment élevee

Tableau 4 11 Reconnaissance des chiffres avec le prototype et un bruit blanc gaussien (SNR 0 dB)

38 99%	Fichiers									
Modeles	0	1	2	3	4	5	6	7	8	9
0	223	112	223	223	116	4		16	28	96
1										
2	5	32	6	3	43	16	1	25	10	41
3										
4										
5						161				4
6	3	3	20	22	2	2	249	51	65	2
7	19	96	6	6	74	35		164		76
8		2			19	3	4		153	
9		10				33				35

Tableau 4 12 Reconnaissance des chiffres avec le prototype et un bruit blanc gaussien (SNR -10 dB)

9 28%	Fichiers									
Modeles	0	1	2	3	4	5	6	7	8	9
0	166	141	198	155	190	77	61	116	82	120
1										
2	82	109	57	98	63	146	189	137	167	132
3										
4										
5										
6	2	4				2	4	1	1	1
7		1			1			2		
8				1		9			6	
9						20				1

Au niveau des confusions, le modele du chiffre *five* qui en causait un nombre important avec le systeme de reference, est rarement choisi à tort avec l'approche proposee Par contre, le modèle du *six* reste une source de confusion et celui du chiffre *seven* est maintenant souvent selectionne de manière incorrecte À -10 dB, les prononciations convergent enfin vers les modeles du *zero* et du *two*

4 4 3 Reconnaissance de chiffres isolés dans différents environnements bruités

Comme pour le système de reference, la reconnaissance des chiffres isolés avec l'approche proposee est aussi testee dans différents environnements bruites

Tableau 4 13 Reconnaissance des chiffres avec le prototype dans differents environnements bruites

Type d'environnement	20 dB	10 dB	0 dB	-10 dB
Aéroport	94 61%	77 70%	35 02%	14 95%
Conversation	94 41%	77 02%	25 96%	14 24%
Voiture	95 24%	74 90%	31 12%	18 76%
Exposition	90 99%	56 61%	23 68%	15 14%
Restaurant	93 55%	61 76%	20 29%	14 16%
Rue	90 83%	79 31%	43 07%	17 11%
Metro	92 33%	59 64%	22 27%	14 39%
Train	89 34%	81 51%	38 12%	14 91%

Encore une fois, l'approche proposee presente des resultats plus robustes aux bruits (tableau 4 13) que ceux obtenus avec le systeme de reference (tableau 4 6) Au niveau des confusions, les modeles responsables restent ceux du *six* et du *seven*

L'approche proposee présente donc un potentiel très interessant pour la reconnaissance vocale dans un milieu bruite en selectionnant de façon automatique les sections où le signal est plus robuste

4.5 Reconnaissance de chiffres isolés avec réverbération

La reverbération est un autre type d'effet nuisible à la reconnaissance qui se definit par la persistance d'un son dans un espace clos ou semi-clos Dans cette situation, on perçoit le signal émis par une source, mais des reflexions de l'onde acoustique vont génerer des echos Les caractéristiques de cette reverberation dépendent entre autres de la taille de la pièce, de la présence et disposition de meubles ainsi que de la propriété d'absorption des matériaux qui constituent les murs, le plancher et le plafond

Le modèle de réverbération applique aux donnees *TEST* provient des travaux de J Allen [Allen et Berkley, 1979] [1] Ce modele simple, mais efficace en temps de calcul simule l'effet d'une reverberation dans une petite pièce rectangulaire comme un local de bureau Ce modele est entre autres utilise par l'équipe de D Wang [Palomaki *et al* , 2004], O Ghitza [Sandhu et Ghitza, 1995] et par D Ward [Ward et Williamson, 2002]

Pour les tests suivants, le modèle de réverberation va genérer les coefficients du filtre pour modifier les donnees a reconnaître selon l'environnement défini Cet environnement est

[1] implemente par Eric A Lehmann

caracterise par les dimensions de la piece (en mètres), l'indice de reflexion des murs, du plancher et du plafond (0 75, 0 75, 0 85, 0 25, 0 3 et 0 9 pour l'ensemble des tests), de la position de la source et du recepteur Quatre pièces sont ainsi definies pour tester la robustesse du systeme propose (tableau 4 14)

Tableau 4 14 Paramètres des reverbérations

	Position de la source (m)	Position du récepteur (m)	Taille de la pièce (m)
X_1	1	4	5
Y_1	1	4	5
Z_1	1 5	1 5	3
X_2	3	13	15
Y_2	3	3	5
Z_2	1 5	1 5	5
X_3	5	26	31
Y_3	6	6	11
Z_3	1 5	1 5	5
X_4	5	35	40
Y_4	7	7	13
Z_4	1 5	1 5	20

Tableau 4 15 Reconnaissance des chiffres avec le systeme de reference et les parametres de réverberation de la petite piece (X_1, Y_1 et Z_1)

94 49%	Fichiers									
Modeles	0	1	2	3	4	5	6	7	8	9
0	250	3	43	1						17
1		247								1
2			170							
3				253					9	
4		4	1		254	1				
5		1				252				6
6			18				248		1	
7			23				1	256		4
8							5		246	
9						1				226

Comme on peut le constater, la reverberation nuit bien à la reconnaissance et la performance des systemes diminue davantage lorsque cette reverbération est produite par une salle aux dimensions importantes Pour les deux systèmes, ce sont surtout les chiffres *two*, *five* et *zero* qui deviennent de plus en plus difficiles à reconnaître

Cependant, on remarque que le systeme de reference est plus affecté par la réverberation En effet, les diminutions de la performance de ce systeme (tableaux 4 15, 4 16, 4 17 et 4 18)

Tableau 4 16 Reconnaissance des chiffres avec le système de reference et les parametres de reverbération de la piece de taille moyenne (X_2, Y_2 et Z_2)

92 57%	Fichiers									
Modèles	0	1	2	3	4	5	6	7	8	9
0	247	13	25			3				6
1		234			1	1				3
2			165							
3			5	252						
4		4			253					
5		2				215		1		1
6	1		37				244		1	
7	2		13			18		255		9
8		1	10	2			10		255	2
9		1				17				233

Tableau 4 17 Reconnaissance des chiffres avec le systeme de reférence et les parametres de reverberation de la grande pièce (X_3, Y_3 et Z_3)

83 36%	Fichiers									
Modeles	0	1	2	3	4	5	6	7	8	9
0	247	28	28		2	8				62
1		211								3
2			77							
3				234						
4		10			249	4				
5						215				1
6	1		93	3	3		247	20		
7	2	4	41	1		24		234		37
8		1	16	16		3	7	2	256	2
9		1								149

sont plus importantes que celles de l'approche proposee (tableaux 4 19, 4 20, 4 21 et 4 22) Le système conventionnel reste légèrement plus efficace dans les deux premiers cas (94 49% et 92 57% contre 93 19% et 91 78%), mais la difference du taux de reconnaissance ne cesse de diminuer avec l'augmentation de la taille de la pièce et la performance de l'approche proposee (85 41% et 82 14%) finit par rattraper et même dépasser celle du système de reference (83 36% et 70 89%) dans le cas des plus grandes pièces

Puisque le traitement se base en partie sur l'énergie, les reflexions plus faibles du signal qui forment les échos affectent peu la sélection des trames De plus, les neurones utilises comme détecteurs perdent de la sensibilité apres la génération d'une impulsion Comme une réflexion sera perçue avec un certain retard, elle a moins de chance de provoquer une impulsion que le signal original

Tableau 4 18 Reconnaissance des chiffres avec le système de reférence et les parametres de reverberation de la plus grande piece (X_4, Y_4 et Z_4)

70 89%	Fichiers									
Modèles	0	1	2	3	4	5	6	7	8	9
0	247	97	32	1	3	3		2		92
1		125								
2			22							
3			6	177						
4		18			247					
5		11			3	251		2		14
6	3		116	32			222	97	1	
7		2						129		13
8		1	79	44	1		32	26	255	8
9		1								127

Tableau 4 19 Reconnaissance des chiffres avec le prototype et les paramètres de reverbération de la petite pièce (X_1, Y_1 et Z_1)

93 19%	Fichiers									
Modèles	0	1	2	3	4	5	6	7	8	9
0	245				1	1				3
1		246				3				
2	2		231		1				1	
3		1		254						4
4		2			250	1				
5						191				2
6	3		23			1	252	16	9	
7		6	1		2	16		238		29
8							2		246	
9						41		2		216

En termes de temps de calcul sur notre poste de travail, les étapes supplementaires nécessaires à la selection des trames ajoutent en moyenne cinq secondes au traitement d'une prononciation Ces etapes comprennent l'application du banc de filtres et des bases temporelles, la recherche des pics d'energie, l'application des filtres à orientation et la generation de la séquence d'impulsions

Enfin, cette section montre bien que le traitement propose, adapte comme selection de trames a un systeme conventionnel, permet d'ameliorer la reconnaissance vocale de chiffres isoles dans des conditions adverses (presence de bruit ou de reverberation) et il offre une performance similaire sur les données propres

Tableau 4 20 Reconnaissance des chiffres avec le prototype et les paramètres de reverberation de la piece de taille moyenne (X_2, Y_2 et Z_2)

91 78%	Fichiers									
Modeles	0	1	2	3	4	5	6	7	8	9
0	245		4	1						1
1		250			26	12				
2	2		244							
3			2	253	1		1		2	
4					222					
5						146				
6	1		4				231		1	
7	2	3	1		5	21	3	249		13
8							19		253	
9		2				75		7		240

Tableau 4 21 Reconnaissance des chiffres avec le prototype et les parametres de reverberation de la grande piece (X_3, Y_3 et Z_3)

85 41%	Fichiers									
Modeles	0	1	2	3	4	5	6	7	8	9
0	244		1	2	4	19	1	3		8
1		248			19	14		1		
2	2		200		1					
3			2	246			3		6	4
4		1			220	1				
5						84				
6	3		45	5			237	9	10	
7		4	7	1	10	39	6	241	1	29
8	1						7		238	
9		2				97		2	1	213

Dans le chapitre suivant, on explore la possibilité d'utiliser une partie du traitement presente pour segmenter en temps et en frequences le signal de parole Une fois realisé, il serait alors possible de trier les segments obtenus pour ne conserver que ceux moins corrompus par le bruit ou selon certains criteres pour orienter le traitement vers une application en codage de la parole ou en rehaussement du signal

Tableau 4 22 Reconnaissance des chiffres avec le prototype et les parametres de reverbération de la plus grande piece (X_4, Y_4 et Z_4)

82 14%	Fichiers									
Modeles	0	1	2	3	4	5	6	7	8	9
0	246		5	10	1	8	16	1		3
1		252			67	45				2
2	1		204	19	4					2
3			18	213		1	6		8	
4					182	1				
5						75				
6	3		21	4			195		3	
7		2	7	4		17	6	245		16
8				4			31	3	245	
9		1				107		7		231

CHAPITRE 5

EXPLORATION DE LA SEGMENTATION EN TEMPS ET EN FRÉQUENCES

Dans le but d'isoler les éléments caracteristiques de la parole d'une representation du signal, comme celle presentee à la figure 3 2, on présente dans cette section une segmentation de la representation en temps et en frequences Cette segmentation permettra de concentrer la reconnaissance sur les elements les plus saillants de la représentation

5 1 Segmentation en temps et en fréquences

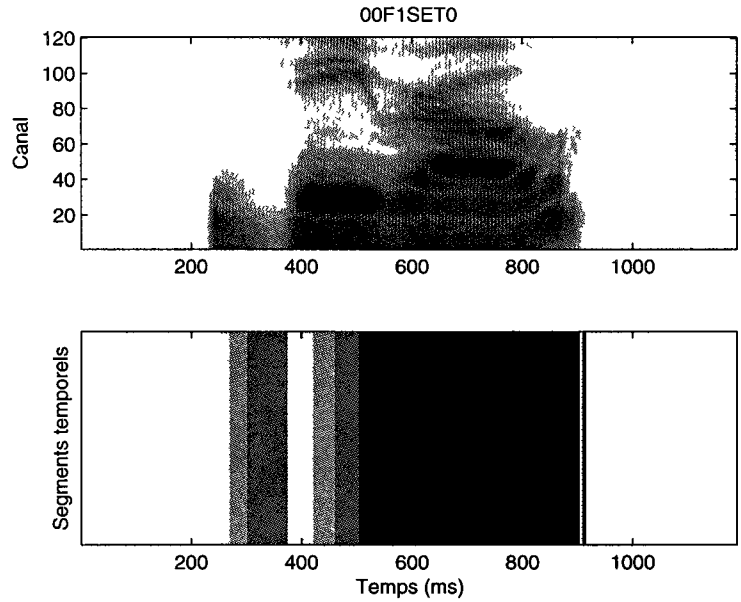

Figure 5 1 Enveloppe compressée de la sortie du banc de filtres, obtenue avec une prononciation d'un *zero* en anglais (haut) et les segments temporels detectes (bas) Les segments sont differencies par les tons de gris Les segments rattaches aux bases courtes sont pâles et ceux rattaches aux bases plus longues sont fonces

Cette manipulation reprend les premieres etapes du traitement présentées à la section 3 1 qui consiste à filtrer et compresser le signal (haut de la figure 5 1) Les bases temporelles b_{f_b} (sous-section 3 1 2) sont ensuite utilisees pour identifier et separer les zones similaires

en temps Les bases plus larges sont plus sensibles aux zones stables alors que les bases plus courtes répondent mieux aux zones de transitions Comme pour l'approche proposee, on filtre la representation par les cinq bases temporelles Par contre, en plus de la valeur maximale des cinq sorties, le systeme utilise aussi une signature de la base temporelle la plus importante

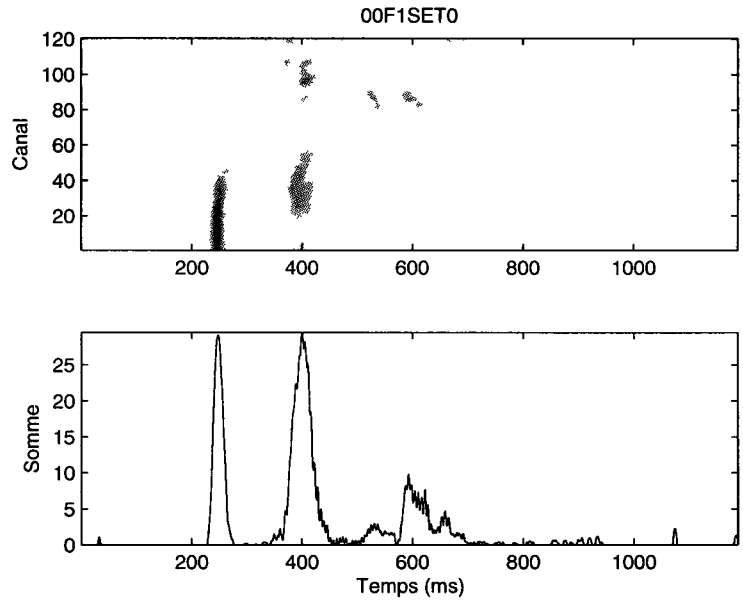

Figure 5 2 Representation du signal filtre par la base b_{32Hz} (haut), somme de la representation filtree (bas)

En effet, à chaque instant, on effectue la somme au niveau de l'energie des canaux de la représentation filtrée pour chaque base temporelle (b_{2Hz}, b_{4Hz}, b_{8Hz}, b_{16Hz} et b_{32Hz}) et on conserve dans un vecteur la valeur de la base qui a produit la plus haute somme à cet instant (la figure 5 2 présente le somme des canaux à chaque instant pour la base b_{32Hz}) Cette étape produit alors un vecteur qui indique quelle base est la plus importante à chaque instant (bas de la figure 5 1) et permet de segmenter la representation en temps (figure 5 3) La longueur du segment est alors determinee par la periode couverte, ou la base qui s'y rattache est la plus importante

Ensuite, des patrons d'excitation (figure 5 4) sont appliques à la representation segmentee en temps pour identifier les zones d'énergies importantes en frequences Cette étape s'inspire du phenomène de masquage que l'on observe au niveau du système auditif [Moore, 1997] Le masquage auditif se manifeste lorsque la perception d'un son est affectée par la presence d'un autre son Dans notre cas, on se limite à un masquage produit par des

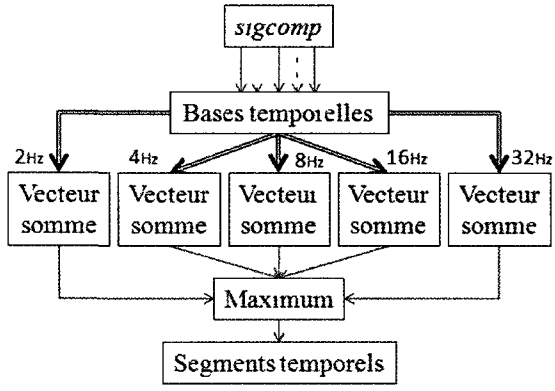

Figure 5 3 Étapes du traitement pour obtenir les segments temporels

canaux de frequences adjacentes Si l'amplitude d'un canal est plus importante que celles des canaux adjacents, ce canal va alors dominer cette zone de fréquences

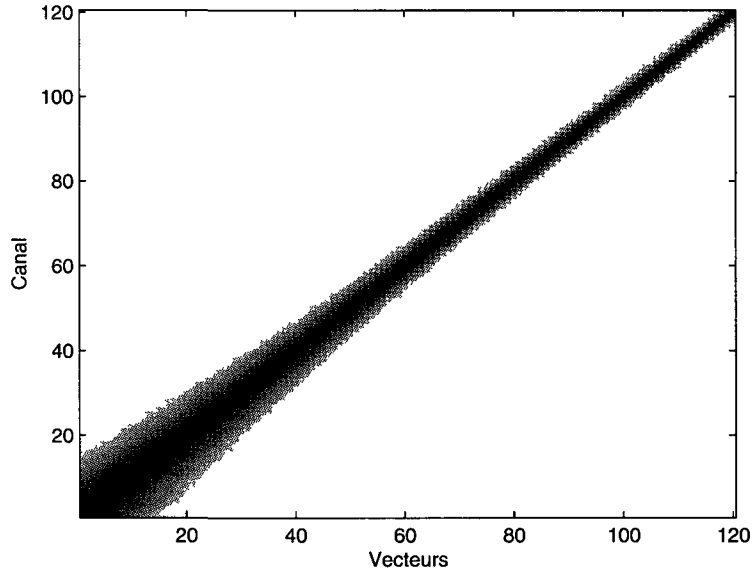

Figure 5 4 Patrons d'excitation géneres avec le banc de filtre en fonction de la frequence centrale du filtre Chaque vecteur presente la réponse des filtres pour une sinusoide de frequence specifique

Pour generer les patrons d'excitation, on filtre des stimuli de frequences pures qui correspondent aux frequences centrales des filtres (120 sinusoides aux frequences presentees dans le tableau A 1) Comme dans l'approche présentee par C Feldbauer [Feldbauer *et al*, 2005] pour le codage de signaux, les réponses maximales des filtres pour un stimulus deviennent le patron du canal qui correspond à la frequence de ce stimulus Par contre, on

ne s'intéresse qu'au composant en frequences de ce patron On obtient alors, pour chaque canal, un vecteur qui contient les reponses maximales des 120 filtres et dont la plus importante valeur est située à la position du canal selectionné Ensuite, on normalise ce vecteur en divisant chacune des reponses des filtres par la plus importante Les eléments d'un patron d'excitation prennent donc des valeurs près de zero lorsqu'ils représentent des canaux eloignes de la frequence du stimulus, mais cette valeur augmente lorsqu'on s'en approche pour finalement atteindre le maximum (1) à cette frequence

Pour chaque segment temporel *segt*, on effectue la somme en temps de l'énergie des canaux pondérée par chaque patron d'excitation pe_{ch} l'un après l'autre Ensuite, pour chaque canal *ch*, on ne conserve que l'identite du patron (Id de l'equation 5 1) qui produit la plus haute des 120 sommes Ces patrons sont alors choisis pour representer les canaux qu'ils couvrent pour la duree du segment temporel (milieu de la figure 5 5) De cette façon, on realise la segmentation frequentielle

$$Id = \max_{ch}(\sum_{t} segt_i(t) \cdot pe_{ch}), ch = 1, 2, 3, \quad , 120 \tag{5 1}$$

Pour chaque segment temporel, on trie les segments en frequences obtenus de celui qui couvre le plus d'énergie de la representation du signal à celui qui en couvre le moins (haut de la figure 5 5) Seuls les segments frequentiels dont l'energie est superieure a la moitie de l'energie du segment le plus important sont enfin conserves ainsi que la portion de la representation temps-fréquences initiale qu'ils representent (bas de la figure 5 5)

5.2 Reconnaissance des chiffres isolés avec la segmentation temps-fréquences

Au chapitre 4, on montre qu'une sequence d'impulsions generées par des detecteurs de caracteristiques permet de realiser la reconnaissance vocale, lorsqu'elle est adaptee a des modeles conventionnels, et offre même une robustesse interessante dans des conditions adverses

Dans cette sous-section, on ajoute la segmentation temps-fréquences au traitement presenté à la section 4 4 pour évaluer son impact Le traitement reste le même, mais la génération des vecteurs d'observations s'effectue sur la représentation segmentée

Dans un premier temps, on evalue la performance de notre approche avec segmentation temps-frequences sur des donnees propres

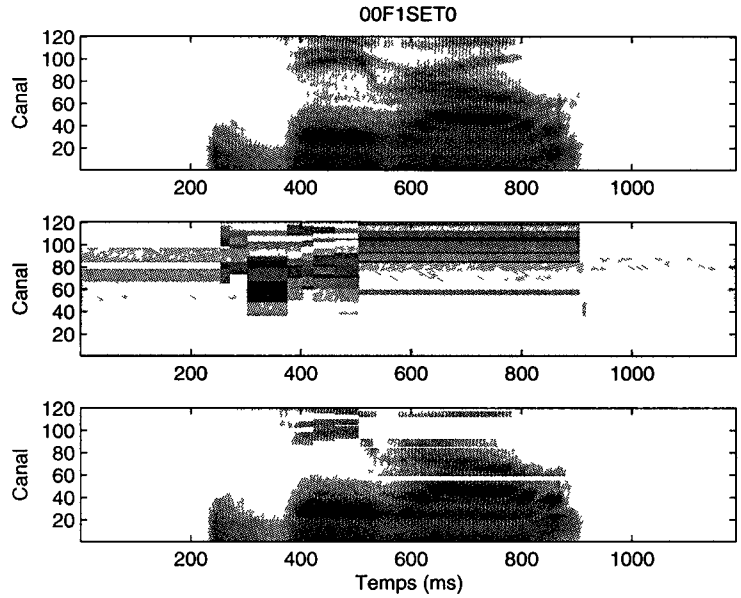

Figure 5 5 Enveloppe compressée de la sortie du banc de filtres, obtenue avec une prononciation d'un *zero* en anglais (haut), segments frequentiels *Id* (milieu), où les differents segments sont identifies par les tons de gris, et les segments conserves (bas)

La segmentation temps-frequences ameliore tres legerement la performance du prototype pour amener le taux de reconnaissance à 97 84% (97 80% pour l'approche originale) Ce taux de reconnaissance ne depasse pas encore celui obtenu avec le systeme conventionnel (99 80% pris du tableau 4 1), mais le potentiel de l'approche se situe surtout au niveau de sa robustesse au bruit On poursuit alors l'evaluation en ajoutant au signal un bruit blanc gaussien à differents niveaux d'intensite (20 dB, 10 dB et 0 dB)

Comme le présente le tableau 5 5, la segmentation temps-fréquences offre des performances similaires à celles du prototype original lorsqu'il y a présence du bruit Cependant, puisque les segments qui possedent peu d'energie sont éliminés de la représentation, les neurones detecteurs, sensibles à l'energie, dechargent moins souvent et les sequences d'impulsions produites à -10 dB sont trop courtes pour permettre l'apprentissage des parametres des modèles

La segmentation temps-frequences permet d'isoler les zones d'energies importantes du signal en ne conservant qu'une partie des segments temps-fréquences généres Cependant, l'integration de cette segmentation dans l'implementation actuelle de l'approche proposee permet difficilement d'evaluer son potentiel Neanmoins, cette manipulation offre une piste

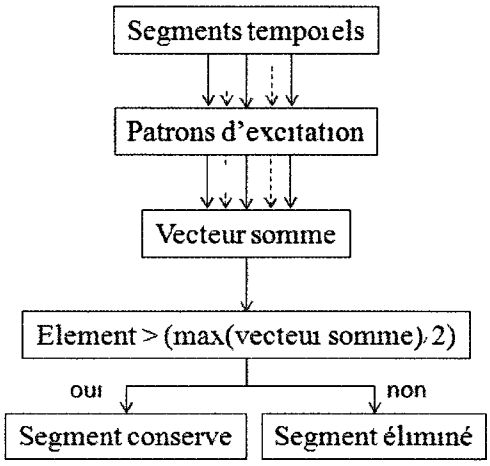

Figure 5 6 Étapes du traitement pour obtenir les segments temps-fréquences

Tableau 5 1 Reconnaissance des chiffres sur données propres avec la segmentation temps-frequences

97 84%	Fichiers									
Modeles	0	1	2	3	4	5	6	7	8	9
0	248									1
1		252			2					1
2	2		255		2		3			
3				253		1			1	
4					249					
5					1	238		1		
6				1			236	1	1	
7		3				12		254		4
8							15		254	
9						3				248

interessante et pourra être adaptee pour des applications en codage de la parole ou en rehaussement du signal, dans des travaux futurs

Une classification basee sur les sequences d'impulsions semble aussi une piste intéressante pour des travaux futurs et permettrait de contourner le problème rencontré De plus, les segments rectangulaires sont présentement trop rigides et ne representent pas bien certaines transitions Une segmentation plus flexible devrait améliorer la performance Enfin, les zones d'energies importantes sont detectées sans reellement en identifier le type (formant, changement important,) On peut imaginer un mécanisme qui pourrait lier l'information du type de detecteur aux segments pour ajouter une information pertinente à la segmentation

Tableau 5 2 Reconnaissance des chiffres avec la segmentation temps-frequences et un bruit blanc gaussien (SNR 20 dB)

93 78%	Fichiers									
Modeles	0	1	2	3	4	5	6	7	8	9
0	246	2	2	1						1
1		192								
2	2		243	8			8			
3				240					2	
4		1			246					
5					3	250	2	1		9
6			3	3	1		222		1	
7	2	13	5	2	3	2		255		7
8			2				22		253	
9		47			1	2				237

Tableau 5 3 Reconnaissance des chiffres avec la segmentation temps-fréquences et un bruit blanc gaussien (SNR 10 dB)

65 07%	Fichiers									
Modeles	0	1	2	3	4	5	6	7	8	9
0	220	44	105	61	30	1	2	5	3	14
1		19								
2	2		90	15			1		1	
3				136			1		2	
4					67					
5		8			49	234	2			23
6	1		17	11			224		8	
7	27	127	37	27	71	8	2	251	1	44
8			6	1	31		22		240	
9		57		3	6	11			1	173

Tableau 5 4 Reconnaissance des chiffres avec la segmentation temps-frequences et un bruit blanc gaussien (SNR 0 dB)

41 58%	Fichiers									
Modeles	0	1	2	3	4	5	6	7	8	9
0	193	65	212	209	37	1	3	6	48	39
1										
2	1								1	
3				9						
4										
5					3	123				12
6	3		16	16			224		43	
7	53	150	25	20	206	28	11	250	11	98
8			2		1		16		153	
9		40			7	102				105

Tableau 5 5 Recapitulatif de la reconnaissance des chiffres pour les systèmes dans differentes conditions

	Propre	*SNR* 20 *dB*	*SNR* 10 *dB*	*SNR* 0 *dB*	*SNR* −10 *dB*
Référence	99 80%	57 75%	22 19%	10 90%	10 07%
Sequences d'impulsions	97 80%	94 34%	67 19%	38 99%	9 28%
Segmentation temps-frequences	97 84%	93 78%	65 07%	41 58%	-

CHAPITRE 6

CONCLUSION

Le systeme presenté aux chapitres precédents permet d'atteindre les objectifs énumérés au début de ce document En effet, cette thèse présente un traitement inspiré des connaissances du fonctionnement du système auditif et des réseaux de neurones pour améliorer la reconnaissance vocale Pour y parvenir, le signal de la parole est filtre par un banc de filtres et compresse pour en produire une representation auditive

L'innovation de l'approche proposee se situe dans l'extraction des elements acoustiques (formants, transitions et *onsets*) à partir de la representation obtenue En effet, une combinaison de detecteurs composes de neurones à decharges permet de révéler la presence de ces elements et genère ainsi une sequence d'evenements pour caractériser le contenu du signal Cette approche n'est donc pas basee sur un traitement par trame tel que l'on retrouve souvent dans la litterature et elle n'est pas limitee par l'utilisation de fenêtres fixes glissantes

Dans le but d'evaluer la performance du traitement presente, la sequence d'evenements est adaptee a un systeme de reconnaissance vocale conventionnel, pour une tâche de reconnaissance de chiffres isoles prononces en anglais Pour ces tests, la sequence d'evenements agit alors comme une sélection de trames automatique pour la génération des observations (*MFCCs*) En comparant les résultats de la reconnaissance du prototype et du systeme de reconnaissance original, on remarque que les deux systèmes reconnaissent tres bien les chiffres prononces dans des conditions optimales et que le système original est légèrement plus performant Par contre, la différence observée au niveau des taux de reconnaissance diminue lorsqu'une reverberation vient affecter les données à reconnaître et les performances de l'approche proposée parviennent à depasser celles du systeme de reference De plus, la selection de trames automatique offre de meilleures performances dans des conditions bruitees

Enfin, l'approche proposee se base sur des caracteristiques dans le temps en fonction de la nature du signal, permet une sélection plus intelligente des données qui se traduit en une parcimonie temporelle, présente un potentiel fort intéressant pour la reconnaissance vocale sous conditions adverses et utilise une detection des caracteristiques qui peut être utilisée comme sequence d'impulsions compatible avec les reseaux de neurones à décharges

6.1 Travaux futurs

Dans un premier temps, evaluer les performances du prototype développe sur une base de donnees plus importante, comme *TIDIGITS* ou *AURORA*, permettrait de mieux analyser ses forces et ses faiblesses Cependant, le traitement propose devra d'abord être adapté pour prendre en compte les zones de silence pour réaliser la reconnaissance vocale en continu avec ces donnees Par la suite, il serait possible de comparer l'approche proposée avec des systemes commerciaux existants

L'utilisation de *HMMs* et de *MFCCs* permet de facilement comparer la performance de l'approche proposee avec celle d'un système conventionnel Par contre, il serait tres interessant d'explorer l'utilisation de la séquence d'impulsions comme entrée pour des classificateurs de type reseaux de neurones à decharges Un classificateur comme la *LSM* [Verstraeten *et al* , 2005] pourrait être un candidat potentiel De plus, les étapes du traitement proposé se prêtent bien à une implementation matérielle

Au niveau du traitement, on a souligne que l'approche proposee etait sensible à l'énergie du signal Puisque les tests se sont effectues sur des enregistrements qui présentaient des caracteristiques similaires, ce probleme n'etait pas contraignant Par contre, il est évident qu'un mecanisme d'adaptation du niveau de l'énergie ou une méthode de detection qui permettrait de diminuer cette sensibilite serait necessaire pour pousser plus loin l'etude du traitement presente Sous cet angle, les travaux de P Gill [Gill *et al* , 2008] attirent notre attention puisqu'il y propose une méthode pour encoder un critère de surprise

Enfin, il existe dans la litterature quelques methodes pour trouver la position des formants et des *onsets* Il serait alors possible de comparer notre approche en explorant ces alternatives

ANNEXE A

FRÉQUENCES CENTRALES DU BANC DE FILTRES

Tableau A 1 Frequences centrales du banc de filtres

Canal	$F_c(Hz)$	Canal	$F_c(Hz)$	Canal	$F_c(Hz)$	Canal	$F_c(Hz)$
1	119	33	515	65	1314	97	3055
2	128	34	532	66	1349	98	3137
3	137	35	550	67	1386	99	3222
4	146	36	568	68	1424	100	3309
5	155	37	586	69	1462	101	3399
6	165	38	605	70	1501	102	3492
7	174	39	624	71	1542	103	3587
8	184	40	644	72	1583	104	3685
9	194	41	664	73	1626	105	3787
10	205	42	685	74	1669	106	3891
11	215	43	706	75	1714	107	3999
12	226	44	727	76	1759	108	4110
13	237	45	749	77	1806	109	4225
14	248	46	772	78	1854	110	4343
15	260	47	795	79	1903	111	4465
16	272	48	818	80	1954	112	4592
17	284	49	842	81	2006	113	4722
18	296	50	867	82	2059	114	4857
19	308	51	892	83	2114	115	4997
20	321	52	918	84	2170	116	5141
21	334	53	944	85	2228	117	5291
22	348	54	971	86	2287	118	5445
23	361	55	999	87	2347	119	5606
24	375	56	1027	88	2410	120	5772
25	389	57	1056	89	2474	121	5944
26	404	58	1086	90	2540	122	6122
27	419	59	1116	91	2607	123	6308
28	434	60	1147	92	2677	124	6500
29	449	61	1179	93	2748	125	6700
30	465	62	1211	94	2822	126	6908
31	481	63	1245	95	2897	127	7124
32	498	64	1279	96	2975	128	7348

ANNEXE B
BASE DE DONNÉES *TI 46-Word*

Tableau B 1 Nombre des prononciations des chiffres pour l'apprentissage

	Chiffres									
Locuteurs	0	1	2	3	4	5	6	7	8	9
F1	10	10	10	10	10	10	10	10	10	10
F2	10	10	10	10	10	10	10	10	10	10
F3	10	10	10	10	10	10	10	10	10	10
F4	10	10	10	10	10	10	10	10	10	10
F5	10	10	10	10	10	10	10	10	10	10
F6	10	10	10	10	10	10	10	10	10	10
F7	10	10	10	10	10	10	10	10	10	10
F8	10	10	10	10	10	10	10	10	10	10
M1	10	10	10	10	10	10	10	10	10	10
M2	10	10	10	10	10	10	10	10	10	10
M3	10	10	10	10	10	10	10	10	10	10
M4	10	9	10	10	10	10	10	10	10	10
M5	10	10	10	10	10	10	10	10	10	10
M6	9	10	10	10	10	10	9	10	10	9
M7	10	10	10	10	10	9	10	10	10	9
M8	10	10	10	10	10	10	10	10	10	10
Total	159	159	160	160	160	159	159	160	160	158

Tableau B 2 Nombre des prononciations des chiffres pour la reconnaissance

	Chiffres									
Locuteurs	0	1	2	3	4	5	6	7	8	9
F1	14	16	16	16	16	16	16	16	16	16
F2	15	16	16	16	16	16	16	16	16	16
F3	15	16	16	16	16	16	16	16	16	16
F4	15	16	16	16	16	16	16	16	16	16
F5	16	16	16	16	16	16	16	16	16	16
F6	16	16	16	16	16	16	16	16	16	16
F7	16	16	16	16	16	16	16	16	16	16
F8	16	16	16	16	16	16	16	16	16	16
M1	16	16	16	16	16	16	16	16	16	16
M2	16	16	16	16	16	16	16	16	16	16
M3	16	16	16	16	16	16	16	16	16	16
M4	15	16	16	16	15	15	16	16	16	15
M5	16	16	15	16	15	16	15	16	16	15
M6	16	16	16	15	16	15	16	16	16	16
M7	16	15	16	15	16	16	15	16	16	16
M8	16	16	16	16	16	16	16	16	16	16
Total	250	255	255	254	254	254	254	256	256	254

LISTE DES RÉFÉRENCES

Ali, A M A , der Spiegel, J V et Mueller, P (May 1998) An acoustic-phonetic feature-based system for the automatic recognition of fricative consonants Dans *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2 IEEE, p 961–964

Ali, A M A , der Spiegel, J V et Mueller, P (November 2001) Acoustic-phonetic features for the automatic classification of stop consonants Dans *IEEE Transactions on Speech and Audio Processing*, volume 9 IEEE, p 833–841

Alkon, D L , Blackwell, K T , Barbour, G S , Rigler, A K et Vogl, T P (March 1990) Pattern-recognition by an artificial network derived from biologic neuronal systems *Biological Cybernetics*, volume 62, n° 5, p 363–376

Allen, J B et Berkley, D A (April 1979) Image method for efficiently simulating small-room acoustics *Journal of the Acoustical Society of America*, volume 65, n° 4, p 943–950

Bandyopadhyay, S et Young, E D (January 2004) Discrimination of voiced stop consonants based on auditory nerve discharges *Journal of Neuroscience*, volume 24, n° 2, p 531–541

Bladon, A et Fant, G (1978) A two-formant model and the cardinal vowels *STL-QPSR*, volume 19, n° 1, p 1–8

Brugge, J F , Dubrovsky, N A , Aitkin, L M et Anderson, D J (1969) Sensitivity of single neurons in auditory cortex of cat to binaural tonal stimulation Effects of varying interaural time and intensity *Journal of Neurophysiology*, volume 32, p 1005–1024

Carlson, R , Granstrom, B et Fant, G (1970) Some studies concerning perception of isolated vowels *STL-QPSR*, volume 11, n° 2–3, p 19–35

Chistovich, L , Sheikin, R et Lublinskaya, V (1978) Centers of gravity and spectral peaks as the determinants vowel quality Frontiers of Speech Communications Research Academic Press, p 145–157

Daugman, J (1980) Two-dimensional spectral analysis of cortical receptive field profiles *Vision Research*, volume 20, n° 10, p 847–856

Dempster, A P , Laird, N M et Rubin, D B (1977) Maximum likelihood from incomplete data via the em algorithm *Journal of the Royal Statistical Society Series B*, volume 39, n° 1, p 1–38

Dimitriadis, D , Maragos, P et Potamianos, A (September 2005) Auditory teager energy cepstrum coefficients for robust speech recognition Dans *9th European Conference on Speech Communication and Technology* p 3013–3016

Doddington, G R et Schalk, T B (September 1981) Speech recognition Turning theory into practice *IEEE Spectrum*, volume 18, n° 9

Edgar, R C (2004) Muscle a multiple sequence alignment method with reduced time and space complexity *BMC Bioinformatics*, volume 5, n° 113

Feldbauer, C , Kubin, G et Kleijn, W B (2005) Anthropomorphic coding of speech and audio A model inversion approach *European Association for Signal Processing Journal on Applied Signal Processing*, volume 1, n° 9, p 1334–1349

Geisler, D C (1998) *From Sound to Synapse Physiology of the Mammalian Ear* Oxford University Press, 381 p

Ghani, A , McGinnity, T M , Maguire, L P et Harkin, J (2008) Neuro-inspired speech recognition with recurrent spiking neurons Dans *International Conference on Artificial Neural Networks* p 513–522

Gill, P , Woolley, S M N , Fremouw, T et Theunissen, F E (2008) What's that sound ? auditory area clm encodes stimulus surprise, not intensity or intensity changes *Journal of Neurophysiology*, volume 99, p 2809–2820

Graves, A , Eck, D , Beringer, N et Schmidhuber, J (2004) Biologically plausible speech recognition with lstm neural nets Dans *Proceedings of Bio-ADIT* p 127–136

Graves, A et Schmidhuber, J (June 2005) Framewise phoneme classification with bidirectional lstm and other neural network architectures Dans *International Joint Conference on Neural Networks*, volume 18 Elsevier Science Ltd, Oxford, UK, p 602–610

Hirsch, H -G et Pearce, D (2000) The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions Dans *ASR-2000* p 181–188

Holmberg, M , Gelbart, D , Ramacher, U et Hemmert, W (2005) Automatic speech recognition with neural spike trains Dans *9th European Conference on Speech Communication and Technology* Lisbon, Portugal

IBM Corporation and Microsoft Corporation (1991) *Multimedia Programming Interface and Data Specifications 1 0* http //www kk ij4u or jp/ kondo/wave/mpidata txt, page consultee le 4 2010

Izhikevich, E M (2006) Polychronization Computation with spikes *Neural Computation*, volume 18, n° 2, p 245–282

Jurafsky, D et Martin, J H (2008) *Speech and Language Processing* Prentice Hall, 1024 p

Kaiser, J F (April 1990) On a simple algorithm to calculate the 'energy' of a signal Dans *International Conference on Acoustics, Speech, and Signal Processing*, volume 1 p 381–384

Kelil, A , Wang, S , Brzezinski, R et Fleury, A (2007) Cluss Clustering of protein sequences based on a new similarity measure *BMC Bioinformatics*, volume 8, n° 286

Lewicki, M S (2002a) Efficient coding of natural sounds *Nature Neuroscience*, volume 5, n° 4, p 356–363

Lewicki, M S (2002b) *Efficient Coding of Time-Varying Signals Using a Spiking Population Code*, chapitre 12 MIT Press, p 223–234

Lewicki, M S et Sejnowski, T J (1999) Coding time-varying signals using sparse, shiftinvariant representations Dans *Advances in Neural Information Processing Systems*, volume 11 MIT Press, p 730–736

Lippman, R P (1997) Speech recognition by machines and humans *Speech Communication*, volume 22, n° 1, p 1–15

Loiselle, S (2004) *Exploration de réseaux de neurones a décharges dans un contexte de reconnaissance de parole* Memoire de maîtrise, Universite du Québec à Chicoutimi, Chicoutimi, Quebec, Canada, 155 p

Loiselle, S , Rouat, J , Pressnitzer, D et Thorpe, S (July 2005) Exploration of rank order coding with spiking neural networks for speech recognition Dans *International Joint Conference on Neural Networks* p 2076–2080

Lyon, R F (1982) A computational model of filtering, detection and compression in the cochlea Dans *International Conference on Acoustics, Speech, and Signal Processing* p 1282–1285

Maass, W (1997) Networks of spiking neurons The third generation of neural network models *Neural Networks*, volume 10, n° 9, p 1659–1671

Maass, W et Bishop, C M (1999) *Pulsed Neural Networks* MIT Press, 377 p

Maass, W , Natschlager, T et Markram, H (2002) Real-time computing without stable states a new framework for neural computation based on perturbations *Neural Computation*, volume 14, n° 11, p 2531–2560

Mariani, J et Lienard, J (May 1977) Acoustic-phonetic recognition of connected speech using transient information Dans *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2 IEEE, p 667–670

Mesgarani, N , David, S V , B, J , Fritz et Shamma, S A (2008) Phoneme representation and classification in primary auditory cortex *Journal of the Acoustical Society of America*, volume 123, n° 2, p 899–909

Messing, D P , Delhorne, L , Bruckert, E , Braida, L D et Ghitza, O (August 2009) A non-linear efferent-inspired model of the auditory system , matching human confusions in stationary noise *Speech Communication*, volume 51, n° 8, p 668–683

Mike Brookes (2009) *VOICEBOX Speech Processing Toolbox for MATLAB* http //www ee ic ac uk/hp/staff/dmb/voicebox/voicebox html, page consultee le 2 fevrier 2010

Moore, B C (1997) *An Introduction to the Psychology of Hearing* Academic Press, 373 p

Oliveri, A , Rizzo, R et Chella, A (August 2007) An application of spike-timing-dependent plasticity to readout circuit for liquid state machine Dans *International Joint Conference on Neural Networks 2007* p 1441–1445

Olivier Cappe (2001) *H2M A set of MATLAB/OCTAVE functions for the EM estimation of mixtures and hidden Markov models* http //perso telecom-paristech fr/ cappe/h2m/, page consultee le 2 fevrier 2010

Olshausen, B A et Field, D J (1997) Sparse coding with an overcomplete basis set A strategy employed by v1 ? *Vision Research*, volume 37, n° 23, p 3311–3325

Palomaki, K J , Brown, G J et Wang, D (2004) A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation *Speech Communication*, volume 43, n° 4, p 361–378

Patnaik, D , Sastry, P et Unnikrishnan, K (2008) Inferring neuronal network connectivity from spike data A temporal data mining approach *Scientific Programming*, volume 16, n° 1, p 49–77

Patterson, R D (1976) Auditory filter shapes derived with noise stimuli *Journal of the Acoustical Society of America*, volume 59, n° 3, p 640–654

Pichevar, R et Rouat, J (2003) Cochleotopic/amtopic (cam) and cochleotopic/spectrotopic (csm) map based sound source separation using relaxation oscillatory neurons Dans *IEEE Neural Networks for Signal Processing Workshop* Toulouse, France

Pickles, J O (1988) *An Introduction to the Physiology of Hearing* Academic Press, 341 p

Picone, J W (September 1993) Signal modeling techniques in speech recognition Dans *Proceedings of the IEEE*, volume 81 Institute of Electrical and Electronics Engineers, New York, United-States, p 1215–1247

Popper, A N et Fay, R R (1992) *The Mammalian Auditory Pathway Neurophysiology* Springer-Verlag, 448 p

R Gary Leonard and George Doddington (1993) *TIDIGITS* http //www ldc upenn edu/Catalog/CatalogEntry jsp?catalogId=LDC93S10, page consultee le 4 mai 2010

Rabiner, L R (February 1989) A tutorial on hidden markov models and selected applications in speech recognition Dans *Proceedings of the IEEE*, volume 77 p 257–285

Ravindran, S , Anderson, D et Slaney, M (May 2004) Low-power audio classification for ubiquitous sensor networks *International Conference on Acoustics, Speech, and Signal Processing*, p 337–340

Rouat, J et Garcia, M (1998) *A Prototype Speech Recognizer based on Associative Learning and Nonlinear Speech Analysis*, chapitre 2 Lawrence Erlbaum Associates, Inc , p 13–26

Sandhu, S et Ghitza, O (May 1995) A comparative study of mel cepstra and eih for phone classification under adverse conditions Dans *International Conference on Acoustics, Speech, and Signal Processing*, volume 1 p 409–412

Schrauwen, B et Verstraeten, D (2007) Compact hardware for real-time speech recognition using a liquid state machine Dans *International Joint Conference on Neural Networks 2007* p 1097–1102

Skowronski, M D et Harris, J G (2007) Noise-robust automatic speech recognition using a predictive echo state network Dans *IEEE Transactions on Audio, Speech and Language Processing*, volume 15 p 1724–1730

Smit, W et Barnard, E (April 2009) Continuous speech recognition with sparse coding *Computer Speech and Language*, volume 23, n° 2, p 200–219

Smith, E et Lewicki, M S (January 2005) Efficient coding of time-relative structure using spikes *Neural Computation*, volume 17, n° 1, p 19–45

Smith, L S et Fraser, D S (May 2004) Sound feature detection using leaky integrate-and-fire neurons Dans *International Conference on Acoustics, Speech, and Signal Processing*, volume 1 p 617–620

Sroka, J J et Braida, L D (April 2005) Human and machine consonant recognition *Speech Communication*, volume 45, n° 4, p 401–423

Stevens, S S , Volkman, J et Newman, E (1937) A scale for the measurement of the psychological magnitude of pitch *Journal of the Acoustical Society of America*, volume 8, n° 3, p 185–190

Suga, N (1990) Cortical computational maps for auditory imaging *Neural Networks*, p 3–21

Thorpe, S , Delorme, A et Rullen, R V (2001) Spike-based strategies for rapid processing *Neural Networks*, volume 14, n° 6–7, p 715–725

Unnikrishnan, K , Hopfield, J et Tank, D (1992) Speaker-independent digit recognition using a neural network with time-delayed connections *Neural Computation*, volume 4, n° 1, p 108–119

Verstraeten, D , Schrauwen, B et Stroobandt, D (April 2005) Isolated word recognition using a liquid state machine Dans *Proceedings of the 13th European Symposium on Artificial Neural Networks* p 435–440

Ward, D B et Williamson, R C (May 2002) Particle filter beamforming for acoustic source localization in a reverberant environment Dans *International Conference on Acoustics, Speech, and Signal Processing*, volume 11 p 1777–1780

Wu, C F J (1983) On the convergence properties of the em algorithm *The Annals of Statistics*, volume 11, n° 1, p 95–103

Zotkin, D N , Chi, T , Shamma, S A et Duraiswami, R (June 2005) Neuromimetic sound representation for percept detection and manipulation *European Association for Signal Processing Journal on Applied Signal Processing*, volume 1, n° 9, p 1350–1364

Zwicker, E (1961) Subdivision of the audible frequency range into critical bands *Journal of the Acoustical Society of America*, volume 23, n° 2, p 228–248